Poročilo 2.3: Dobre prakse uvajanja UI sistemov

Pripravili: Maja Škrjanc (IJS), Maruša Škrjanc (IJS)

Datum: 30.12.2025

Dopolnjeno 27.2.2026

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Kazalo

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

1 Namen in obseg

Ta dokument predstavlja poglobljene opise štirih izbranih kandidatov dobrih praks, ki so posebej uporabni za pripravo poročila dobrih praks v okviru projekta NOSI-AI. Poudarek je na primerih z visoko kredibilnimi viri (regulatorji oziroma uradni javni organi) ter z dovolj tehnično-operativnimi podrobnostmi, da jih je mogoče prevesti v konkretne zahteve, smernice in kontrolne sezname (npr. za uporabo LLM/RAG ter "klasičnih" ML modelov).

Opisana je tudi preliminarna navezava na Akt o UI(tveganostne kategorije), pri čemer je treba poudariti, da je dokončna kvalifikacija tveganosti vedno odvisna od *konkretne namere uporabe* (intended purpose) in vloge subjekta v dobavni verigi. Oznake "visoko- tvegan" kandidat" in "LIMITED-RISK kandidat" so zato razumljene kot pragmatične oznake za strukturiranje poročila, ne kot pravno zavezujoča odločitev.

2 Merila izbora in metodologija preverjanja

Izbor primerov v tem poročilu ni bil zasnovan kot splošen pregled "zanimivih" projektov umetne inteligence, temveč kot ciljno usmerjena identifikacija tistih primerov, ki so za javni sektor in regulirano okolje hkrati verodostojni, operativno uporabni in dovolj podrobno dokumentirani, da iz njih lahko izpeljemo prenosljive dobre prakse. Osnovno merilo je bila zato kredibilnost vira: prednost so imeli primeri, za katere obstajajo primarni uradni viri, zlasti dokumenti regulatorjev ali javnih organov, ki so sami nosilci, spremljevalci ali objavitelji konkretnega sistema. V obravnavanem naboru to pomeni predvsem dokumente in uradne strani CNIL, Datatilsynet, DINUM/Albert API in City of Helsinki AI Register.

Drugo merilo je bila operativna uporabnost primera. V poročilo so bili vključeni predvsem tisti sistemi, pri katerih uradna dokumentacija ne ostane na ravni splošnega opisa namena, temveč vsebuje tudi elemente, ki so za dejansko implementacijo odločilni: opis arhitekturnega pristopa (npr. RAG, federativno učenje, skupna platformna infrastruktura), osnovne informacije o upravljanju (*governance*), varnostnih kontrolah, obdelavi podatkov, človeškem nadzoru in mehanizmih izboljševanja delovanja. Takšna raven opisa je bila ključna, ker omogoča, da se iz posameznega primera ne izlušči le "kaj sistem dela", temveč tudi "kako je njegovo delovanje upravljano".

Tretje merilo je bila vsebinska in tveganostna raznolikost nabora. Primeri so bili izbrani tako, da skupaj pokrivajo vsaj dva različna tipa uporabe: na eni strani sisteme, ki se približujejo področjem povečanega regulatornega pomena v javnih storitvah ali financah, na drugi strani pa primere, kjer je težišče predvsem na transparentnosti, nadzoru uporabe in kakovosti storitve. Namen takšne razporeditve ni bil podati formalne pravne kvalifikacije v tem poglavju, temveč zagotoviti, da poročilo ne ostane omejeno le na eno vrsto AI rešitev, ampak zajame tako bolj "kritične" kot tudi bolj "vsakodnevne" javnosektorske scenarije.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Četrto merilo je bila možnost prenosa v operativne zahteve. Izbrani so bili tisti primeri, pri katerih je bilo mogoče identificirati elemente, ki jih je mogoče neposredno prevesti v organizacijske, procesne in tehnične zahteve. Sem sodijo na primer vprašanja presoje vplivov in internih politik uporabe, usposabljanja uporabnikov, infrastrukturnih odločitev (npr. lokalna oziroma suverena oblačna namestitev), upravljanja logov in hrambe, nadzora dostopa, spremljanja kakovosti ter drugih kontrol, ki so v praksi nujne za odgovorno uvajanje sistemov UI.

Metodologija preverjanja je bila pri vseh primerih enotna. Ključne trditve so bile preverjene neposredno v primarnih dokumentih in uradnih spletnih virih, zlasti v PDF poročilih, odločbah, pravilih uporabe, uradnih predstavitvah produktov in registrskih vnosih. V besedilu so zato referencirane predvsem tiste navedbe, ki so neposredno podprte z objavljenimi viri; kjer dokumentacija ne daje dovolj natančnih tehničnih ali organizacijskih informacij, je to v opisu izrecno omejeno ali pojasnjeno. Uporabljeni viri so v izvirnih jezikih – predvsem v francoščini, norveščini in angleščini – zato so prevodi in povzetki v poročilu nujno interpretativni, vendar so pripravljeni z namenom čim bolj operativne in terminološko natančne uporabe.

3. Izbrani primeri

3.1 CNIL »regulativni peskovnik« – France Travail: projekt »Personalizirani nasveti« (LLM + RAG v javni storitvi)

France Travail je francoski javni zavod in osrednji nosilec javne službe za zaposlovanje (fr. service public de l'emploi). Njegova temeljna naloga je podpora posameznikom pri iskanju zaposlitve in ponovni vključitvi na trg dela, pri čemer v okviru svojih pristojnosti sodeluje tudi pri postopkih in storitvah, povezanih z brezposelnostjo ter aktivnimi politikami zaposlovanja.

V operativni praksi to pomeni, da svetovalci France Travail dnevno delajo z velikim številom uporabnikov in jim pomagajo oblikovati izvedljive, individualizirane korake do zaposlitve, pri čemer ima pomembno vlogo tudi usposabljanje (preusposabljanje, nadgradnja znanj, pridobivanje specifičnih kompetenc). Eden osrednjih izzivov, ki je motiviral projekt, je bila kompleksnost izbire ustreznih usposabljanj: ponudba je obsežna, v uporabi je več različnih informacijskih virov in orodij, podatki o usposabljanjih prihajajo iz različnih kanalov, svetovalec pa mora kljub temu v kratkem času oblikovati priporočilo, ki je smiselno, realistično in prilagojeno posamezniku.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Projekt »Conseils Personnalisés« je bil zasnovan kot odgovor na to operativno težavo: namenjen je bil temu, da svetovalcem olajša orientacijo v obsežni ponudbi in jim omogoči, da hitreje in bolj dosledno prepoznajo možnosti, ki so dejansko primerne za konkreten profil iskalca zaposlitve.

Kdo so uporabniki in komu je rešitev namenjena

Primarni uporabniki sistema so svetovalci France Travail (t. i. strokovni uporabniki), ki orodje uporabljajo v procesu svetovanja kot podporo pri pripravi predlogov usposabljanj.

Končni naslovniki priporočil so iskalci zaposlitve, vendar je za pravilno razumevanje vloge orodja bistveno, da je rešitev zasnovana kot orodje za podporo strokovnemu odločanju (decision support), ne pa kot storitev, ki bi sama neposredno odločala o pravicah posameznika, dostopu do storitev ali o dodelitvi koristi. Vloga človeka – svetovalca – ostaja osrednja, orodje pa je omejeno na priporočanje oziroma pripravo predlogov, ki jih svetovalec presodi in po potrebi prilagodi.

Cilj rešitve

Orodje je zasnovano kot priporočilni mehanizem: na podlagi (i) profila iskalca zaposlitve ter (ii) dodatnih navodil oziroma usmeritev svetovalca predlaga usposabljanja, ki bi bila za posameznika najprimernejša.

CNIL (francoski nadzorni organ za varstvo osebnih podatkov, fr. Commission nationale de l'informatique et des libertés) izrecno poudari, da je namen projekta olajšati izbor relevantnih storitev v situaciji, ko je ponudba velika in ko je v praksi hkrati v uporabi več orodij in virov. Rešitev lahko predlaga usposabljanja, vključno z usposabljanji za mehke veščine (npr. komunikacija, javno nastopanje), ter tudi lokalne dogodke znotraj mreže France Travail.

Za razumevanje izhodiščnega problema je pomembno tudi, da katalog obsega več kot 20.000 usposabljanj, ki izvirajo iz različnih virov (France Travail in partnerji). To pomeni dodatne izzive glede usklajevanja formatov, konsistentnosti podatkov in kakovosti opisov – pri čemer je prav kakovost vhodnih podatkov ključna za smiselnost priporočil.

Tehnična zasnova sistema

Rešitev uporablja veliki jezikovni model (LLM), zgrajen na podlagi temeljnega modela (*foundation model*), in ga dopolnjuje z mehanizmom RAG (*retrieval augmented generation*). V praksi to pomeni, da sistem pred pripravo odgovora najprej poišče relevantne informacije v vnaprej določenih virih (»retrieval«), nato pa na tej osnovi generira besedilo (»generation«). CNIL pri tem izrecno pojasni, da RAG vključuje mehanizem iskanja informacij v vektorizirani

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

bazi (tj. bazi z *embeddingi*), kar omogoča pripravo odgovorov, ki so praviloma bolj specifični in hkrati lažje posodobljivi kot sam model (ker se lahko posodablja znanje v bazi, ne da bi bilo treba ponovno učiti model).

RAG v tem primeru vključuje zlasti:

- katalog usposabljanj France Travail,
- profil France Travail iskalca zaposlitve,
- dodatne specifikacije, ki jih svetovalec poda prek besedilnih navodil (t. i. *promptov*), da se zahteva natančneje usmeri (npr. časovne omejitve, lokalna dostopnost, prednostne kompetence).

Veliki jezikovni model Mixtral

CNIL navaja, da je bil v projektu uporabljen model Mixtral, in sicer lokalno nameščen ("on-premise"), torej nameščen in upravljan znotraj okolja France Travail. Mixtral je družina jezikovnih modelov podjetja Mistral AI; v CNIL opombi je kot referenca naveden prav Mistralov uradni opis *Mixtral of Experts*, ki se nanaša na Mixtral 8×7B.

Ključna posebnost Mixtral 8×7B je arhitektura SMoE (Sparse Mixture of Experts) – "redka mešanica ekspertov". To pomeni, da ima model na posamezni plasti več "ekspertnih" podsistemov (npr. 8), vendar pri vsakem tokenu praviloma aktivira samo dva eksperta, ki ju izbere usmerjevalna mreža (*router*). Posledično ima model dostop do zelo velikega števila parametrov, a pri inferenci aktivno uporabi le del, kar izboljša razmerje zmogljivost/strošek.

V primarnih virih je navedeno, da ima Mixtral 8×7B približno 46,7B skupnih parametrov, pri čemer se na token aktivira približno 12,9B parametrov; poleg tega je model treniran za kontekst 32k tokenov.

V kontekstu javne storitve, kot je France Travail, je Mixtral relevanten predvsem zato, ker je:

- odprto-utežen in licenciran pod Apache 2.0, kar omogoča izvedbo v lastnem okolju, ter
- zasnovan z dobrim kompromisom med kakovostjo in stroški izvajanja, kar je pomembno pri operativnih sistemih z velikim številom uporab.

Lokalna postavitev

CNIL navaja, da je France Travail izbral Mixtral po primerjalni študiji različnih jezikovnih modelov. Čeprav CNIL v tem delu ne našteje vseh kriterijev izbire, je iz regulatornega vidika jasno, da je lokalni (on-premise) pristop posebej pomemben zaradi zaupnosti in nadzora nad osebnimi podatki, ker sistem uporablja (a) profile iskalcev zaposlitve in (b) podatke iz kataloga usposabljanj, dodatno pa lahko osebni podatki vstopajo tudi prek promptov.

Ključne prednosti izbranega pristopa so zato zlasti:

- Zaupnost in suverenost podatkov: CNIL izrecno opozori, da v primeru, ko se. storitve izvajajo neposredno v omrežnem okolju ponudnika (off-premise) lahko pomenijo tveganje izgube zaupnosti osebnih podatkov. Lokalna namestitev to tveganje praviloma zmanjša, ker obdelava ostane v nadzorovanem okolju upravljavca.

- Boljša ujemljivost z regulatornimi priporočili za generativno UI: CNIL v splošnih usmeritvah za uvajanje generativne UI priporoča izbiro robustnega sistema in varnega načina namestitve, pri čemer kot primer navaja lokalne, varne in specializirane sisteme; če se uporablja zunanji ponudnik, je treba posebej oceniti možnost, da bi ponudnik ponovno uporabil podatke, in uporabo prilagoditi temu tveganju.

- Operativna učinkovitost (zlasti v kombinaciji z RAG): RAG omogoča, da se sistem pri priporočilih opira na ažurne in domensko specifične podatke (katalog, profil, navodila svetovalca), kar je praktično bistveno pri velikem številu in dinamičnosti ponudb.

- Primerna tehnična lastnost "daljšega konteksta": Mixtral 8×7B je zasnovan za obdelavo daljšega konteksta (32k), kar lahko v praksi olajša vključevanje relevantnih delov profila in opisov usposabljanj v enoten odgovor (seveda ob ustrezni minimizaciji).

Slabosti in omejitve velikega jezikovnega modela Mixtral

Tudi pri lokalni namestitvi in uporabi RAG pristopa ostanejo ključne omejitve generativnih modelov, ki so v javnem sektorju še posebej občutljive:

- Probabilistična narava, "halucinacije" in navidezna prepričljivost: generativni sistemi lahko proizvedejo netočne ali zavajajoče rezultate, ki so kljub temu videti verodostojni; to je eden razlogov, zakaj CNIL v primeru France Travail tako močno poudari smiselno človeško presojo in pogoje dela, ki omogočajo preverjanje.

- Kompleksnost upravljanja in obratovanja (MLOps/LLMOps): lokalna namestitev pomeni tudi, da organizacija sama nosi breme varnega obratovanja (posodobitve, nadzor dostopov, beleženje, odzivanje na incidente, upravljanje zmogljivosti). To je prednost z vidika nadzora, vendar zahteva višjo organizacijsko zrelost. (To je logična posledica odločitve za lokalno nameščeno storitev; CNIL jo posredno naslavlja skozi poudarek na varnem načinu namestitve in upravljanja)

- Varnostno-vedenjska omejitev (model sledi navodilom): Mistral AI izrecno opozori, da brez ustreznih "varoval" v promptih ali dodatnega prilagajanja (npr. preference tuning) model sledi navodilom, kar je relevantno pri tveganjih zlonamernih ali neustreznih navodil (npr. prompt injection).

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

- Pristranskosti: Mistral opisuje, da merijo halucinacije in pristranskosti ter primerjalno poročajo o rezultatih; kljub temu pristranskost ni "odpravljena" lastnost, temveč kontinuirano tveganje, ki ga je treba upravljati na ravni podatkov, promptov, testiranja in nadzora.

Namestitev

Priporočila CNIL za ta primer so podana v kontekstu arhitekturne izbire, da se veliki jezikovni model Mixtral izvaja v načinu "on-premise", tj. kot lokalno nameščen in upravljan sistem v infrastrukturnem okolju France Travail. CNIL ob tem izrecno navede, da je France Travail takšno rešitev izbral po primerjalni študiji več različnih jezikovnih modelov, kar kaže na zavestno presojo kriterijev (zmogljivost, stroški, zrelost izvedbe, predvsem pa profil tveganja).

V tej terminologiji on-premise pomeni, da se obdelava izvaja znotraj nadzorovanega omrežja in upravljavskih pristojnosti organizacije (npr. v lastnem podatkovnem centru ali zasebnem oblaku pod njenim upravljanjem), tako da organizacija ohrani neposreden nadzor nad: (i) dostopi, (ii) beleženjem (revizijskimi sledmi), (iii) hrambo, (iv) segmentacijo omrežja, (v) varnostnimi politikami in (vi) pogodbeno-upravljavskimi razmerji glede podatkov. Nasprotno off-premise pomeni uporabo modela kot storitve, ki se izvaja v omrežnem okolju ponudnika licence (tj. pri ponudniku storitve), pri čemer CNIL posebej opozori, da takšna izvedba lahko poveča tveganje izgube zaupnosti osebnih podatkov, ki se v sistemu obdelujejo.

Pomembno je poudariti, da on-premise namestitev tveganj ne odpravi, jih pa praviloma drugače razporedi: namesto izpostavljenosti tretjemu ponudniku se težišče premakne na notranjo varnostno zasnovo, upravljanje dostopov, zagotavljanje dokazljivosti ukrepov ter operativno disciplino. Z vidika regulatorja je zato infrastruktura "ključni nosilni element" skladnosti: čeprav je model lahko tehnično zmogljiv, je za javno storitev odločilno, ali je celotni sistem (model, RAG, baze znanja, dostopi, revizijske sledi in organizacijski postopki) zasnovan tako, da je zaupnost in integriteta podatkov realno zavarovana.

Infrastruktura

Pri arhitekturi LLM + RAG je treba infrastrukturo razumeti kot sestav več komponent, ne le kot "strežnik z modelom". Tipično vključuje:

- Strežniški sloj za izvajanje LLM (inferenčni strežnik), ki izpostavi nadzorovan API za klice modela.
- Sloj za RAG:
    - vektorsko bazo (npr. baza z "embeddingi") in iskalne mehanizme,
    - cevovod za pripravo/posodabljanje dokumentov (ETL: čiščenje, vektorizacija, verzioniranje),
    - pogosto tudi ločen (manjši) model za embeddinge.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

- Integracijski in varnostni sloj (API prehodi, avtentikacija, avtorizacija, omejitve dostopa, upravljanje ključev).
- Opazovanje in dokazljivost (beleženje klicev in pozivov, sledljivost dokumentov, metrike kakovosti, varnostni dogodki, obravnava incidentov).
- Upravljanje sprememb (posodobitve modela, posodobitve baze znanja, testiranje, "rollback", kontrola konfiguracij).

Zavedanje o takšni arhitekturi in postavitvi je pomembna za določitev tveganj. Namreč, pri RAG pristopu se pomemben del tveganj premakne na bazo znanja: kdo ima dostop, kako se izvaja hramba, ali je mogoče dokazati, na katere vire se je sistem opiral, in kako se prepreči, da bi v bazo "ušli" neustrezni ali preobčutljivi podatki (npr. nepotrebni deli profila iskalca zaposlitve).

Okvirna strojna oprema (HW)

Ker CNIL v priporočilih ne podaja tehničnih specifikacij strežnikov, je pri načrtovanju strojne opreme priporočeno ostati pri preverljivih in konservativnih okvirjih. Za model Mixtral 8×7B (ki je po primarnih virih model z arhitekturo *Mixture-of-Experts* in je objavljen pod licenco Apache 2.0) velja, da je izvedba v polni natančnosti pomnilniško zelo zahtevna, zato se v praksi pogosto uporablja več strežniških GPU-jev ali pa kvantizacija.

Iz javno dostopnih tehničnih navedb (vključno z razpravami na uradnem repozitoriju modela) izhaja približen red velikosti potrebnega grafičnega pomnilnika (VRAM):

- polna natančnost (FP16): približno ~90–100 GB VRAM,
- 8-bitna kvantizacija: približno ~45 GB VRAM,
- 4-bitna kvantizacija: približno ~22–23 GB VRAM, pri čemer se izrecno opozarja na degradacijo kakovosti pri 4-bitnih nastavitvah.

To pomeni, da izbira za lokalno postavitev praviloma ni enostavna: za stabilno in odzivno storitev je treba zagotoviti:

- ustrezno GPU kapaciteto (ali kvantizacijo in posledično testiranje kakovosti),
- dovolj sistemskega RAM-a in hitrega diskovnega podsistema (SSD/NVMe) za RAG podatkovne tokove,
- zanesljivo omrežje in segmentacijo, ter
- redundanco, če je sistem operativno kritičen.

Pri tem velja praktično pravilo, da je dejanska potrebna kapaciteta odvisna od ciljne prepustnosti (števila sočasnih uporabnikov), dolžine konteksta in velikosti RAG zadetkov, zato je treba infrastrukturo dimenzionirati na podlagi pričakovanih scenarijev uporabe, ne zgolj na podlagi "ali model teče oziroma se odziva".

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Relevanca in tveganja

V kontekstu France Travail je infrastrukturna odločitev povezana s tveganji na dveh ravneh:

- Prvič, sistem obdeluje podatke, ki lahko vsebujejo osebne informacije o iskalcih zaposlitve; CNIL zato izpostavi, da *off-premise* izvedba (storitev v okolju ponudnika) poveča tveganje izgube zaupnosti. Lokalna (on-premise) izvedba to tveganje načeloma znižuje, ker zmanjšuje izpostavljenost podatkov zunanjim subjektom in omogoča strožje notranje kontrole.
- Drugič, lokalna izvedba odpira drugačen razred tveganj: organizacija mora sama zagotoviti varno obratovanje (dostopi, revizijske sledi, segmentacija, nadzor nad izpadi, upravljanje incidentov).

V javni storitvi je to še posebej pomembno, ker lahko pod-dimenzionirana ali nestabilna infrastruktura povzroči operativne pritiske (npr. večja latenca, slabša razložljivost, "copy-paste" uporaba), kar posredno oslabi tudi smiselnost človeškega nadzora in poveča tveganje, da se priporočilom sledi nekritično. Z drugimi besedami: infrastruktura ni le "IT" vprašanje, temveč pomemben dejavnik, ki vpliva na to, ali so pravno-organizacijske varovalke v praksi dejansko učinkovite.

Čeprav je arhitektura (LLM + RAG, lokalna namestitev) zasnovana tako, da zmanjšuje določene razrede tveganj, se pri takšni rešitvi tveganja ne pojavijo le na ravni modela, temveč tudi na ravni vnosov (promptov), podatkovnih virov, načina gostovanja ter dejanske rabe v vsakodnevnem delu svetovalcev. CNIL v okviru spremljanja zato tveganja ne obravnava abstraktno, temveč jih poveže s konkretnimi kontrolnimi točkami, ki jih je mogoče organizacijsko in tehnično uvesti.

(i) Prvi kritični sklop tveganj je povezan z **razkritjem osebnih podatkov prek besedilnih navodil (promptov)**. Pri generativnih sistemih je vnos "po privzetem" odprt: svetovalec lahko iz želje po večji natančnosti v prompt vključi preveč podrobnosti o iskalcu zaposlitve. CNIL izrecno opozori, da se lahko na ta način v sistem nehote vnesejo podatki, ki niso nujni za namen priporočanja usposabljanj, in zato priporoča omejevanje vnosa ter navaja primere podatkov, ki se jim je treba izogibati (npr. družinske okoliščine, zdravstveno stanje ipd.). To tveganje je v kontekstu France Travail posebej relevantno zato, ker gre za javno storitev in za obdelavo podatkov posameznikov, ki so lahko v ranljivejšem položaju; že majhna odstopanja pri praksi vnosa lahko v takem okolju pomenijo nesorazmeren učinek na zasebnost.

(ii) Drugi sklop tveganj zadeva **zaupnost pri izvedbi zunaj nadzorovanega okolja (off-premise)**. CNIL posebej poudari, da lahko uporaba generativnega modela kot zunanje storitve pomeni povečano tveganje izgube zaupnosti osebnih podatkov. Ravno zato je dejstvo, da je bil v tem primeru model nameščen lokalno (on-premise), regulatorno pomemben element arhitekture: obdelava ostane v okolju upravljavca, kar načeloma zmanjša izpostavljenost podatkov tretjim

subjektom. Hkrati CNIL (tudi v svojih splošnih priporočilih) poudarja, da mora organizacija pri zunanji storitvi vedno presoditi možnost, da bi ponudnik posredovane podatke nadalje obdeloval ali ponovno uporabljal, in temu prilagoditi dovoljene primere uporabe (npr. prepoved posredovanja osebnih podatkov ali prepoved uporabe za odločanje, kjer so učinki za posameznika pomembni). V praksi to pomeni, da izbira načina namestitve ni le tehnična odločitev, temveč odločitev o "profilu tveganja" celotnega sistema.

(iii) Tretji sklop tveganj se nahaja v **RAG sloju**, torej v "bazi znanja" in mehanizmih iskanja. Ker RAG temelji na vektorizirani bazi (embeddingih), se del tveganj prestavi iz samega modela na **upravljanje baze**: kdo ima dostop do nje, ali obstaja revizijska sled poizvedb, kakšna je politika hrambe, kako se posodabljajo viri in kako se zagotavlja njihova kakovost. Če v RAG vstopajo tudi podatki iz profila iskalca zaposlitve, gre za osebne podatke, kar nujno zahteva stroge kontrole dostopa, minimizacijo in jasno opredelitev namenov obdelave. CNIL v opisu mehanizma RAG zato izrecno poudari vlogo vektorizirane baze kot ključne komponente sistema — s tem nakazuje, da se skladnost in varnost ne moreta presojati samo na ravni "ali je model dober", temveč tudi na ravni "ali je dobro urejen ekosistem znanja in dostopov".

(iv) Četrti sklop tveganj je tveganje **napačnih ali pristranskih priporočil**, ki bi lahko povzročila nepravične učinke. CNIL kot primere možnih virov pristranskosti omenja tudi geografske podatke in posredno sklepanje o osebnih lastnostih (npr. inferiranje spola iz imen). To kaže, da je pri priporočilih na podlagi profila realno tveganje diskriminatornih učinkov, četudi sistem formalno "samo priporoča". V takih sistemih sta zato minimizacija podatkov in kontrola uporabe promptov regulatorno centralni temi: manj kot sistem nepotrebno "ve", manj je priložnosti za neželene korelacije, hkrati pa se zmanjšajo možnosti, da se v model pretihotapijo občutljive informacije, ki niso potrebne za namen priporočanja.

Opisani sklopi tveganj (prompti, RAG in infrastrukturna umestitev) neposredno pojasnijo, zakaj CNIL v tem primeru toliko pozornosti namenja učinkovitemu človeškemu nadzoru: tudi če sistem formalno deluje kot priporočilno orodje, lahko v praksi zaradi avtoritete izpisa, časovnih pritiskov ali neustreznih organizacijskih pogojev postane "de facto" odločilni dejavnik. Prav zato je ključno, da je človeška intervencija smiselna in dejanska (ne zgolj nominalna) ter podprta z razumljivostjo izhodov, možnostjo utemeljenega odstopa, usposabljanjem in nadzorovanimi delovnimi pogoji. Na tej podlagi se v naslednjem poglavju utemelji razmerje do 22. člena GDPR, tj. preprečevanje situacij, ko bi priporočilni sistem po učinku prešel v režim izključno avtomatiziranega odločanja s pravnimi ali podobno pomembnimi učinki za posameznika.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Upravljanje odločanja in »smiselna človeška intervencija« (GDPR, 22. člen)

CNIL je pri obravnavi projekta *Conseils Personnalisés* osrednjo pozornost usmerila v vprašanje, kako preprečiti, da bi priporočilni sistem v praksi deloval kot prikrita (de facto) avtomatizirana individualna odločitev. Tudi kadar orodje formalno ne sprejema dokončne odločitve, lahko zaradi organizacijskih pogojev dela, avtoritete izpisa ali časovne stiske svetovalec začne izhodom slediti rutinsko. V takih okoliščinah bi se lahko ustvaril učinek, ki je regulatorno problematičen: posameznik sicer nominalno prejme "priporočilo", vendar se to priporočilo v praksi uporablja kot odločitev, ki usmerja obravnavo in lahko pomembno vpliva na njegove možnosti.

Zato CNIL koncept »človeške intervencije« razume kot kombinacijo (1) tehničnih lastnosti sistema in (2) organizacijskih pogojev, v katerih se sistem uporablja. V okviru presoje glede člena 22 GDPR (prepoved oziroma stroga omejitev izključno avtomatiziranega odločanja s pravnimi ali podobno pomembnimi učinki) CNIL izpostavi dve operativni zahtevi, ki sta v tem primeru "nosilni":

1. Svetovalec mora imeti na voljo razumljiv in obvladljiv prikaz predlogov – ne le v smislu jezikovne berljivosti, temveč v smislu uporabniške ergonomije: predlogi morajo biti predstavljeni tako, da jih je mogoče kritično presojati, primerjati in po potrebi zavrniti ali prilagoditi.
2. Svetovalcu morajo biti na voljo informacije, ki omogočajo strokovno presojo predloga, tj. takšna raven pojasnil ali konteksta, da svetovalec ne deluje zgolj kot "posrednik" izpisa, ampak kot strokovni odločevalec, ki razume, na čem predlog temelji in kje so njegove meje.

CNIL s tem dejansko nakaže standard, ki presega zgolj formalno prisotnost človeka "v zanki" (*human-in-the-loop*). Poudarek ni na tem, da človek obstaja kot procesna kljukica, temveč na tem, da ima človek resnično možnost in praktične pogoje, da uveljavi presojo. Človeška intervencija je torej merljiva predvsem po tem, ali je svetovalec sposoben:

(a) razumeti, kaj sistem predlaga,
(b) razumeti, zakaj je predlog sploh nastal oziroma katera dejstva so zanj relevantna, ter
(c) sprejeti drugačno odločitev brez sankcij, implicitnega pritiska ali organizacijske penalizacije.

Usposabljanje, organizacijski ukrepi in "operacionalizacija" nadzora

France Travail je za krepitev avtonomije svetovalcev in za zmanjšanje tveganja nekritičnega sledenja izhodom vzpostavil kombinacijo kompetenčnih in organizacijskih ukrepov.

Na kompetenčni ravni je uvedel programe ozaveščanja in usposabljanja, ki vključujejo: stalno izobraževanje o umetni inteligenci (da se zmanjšata "mistifikacija" modela in pretirano zaupanje) ter specifično usposabljanje za uporabo orodja (da svetovalci razumejo namen, omejitve in pravilno uporabo).

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Na organizacijski ravni je vzpostavil tudi mrežo t. i. »AI korespondentov«, ki delujejo kot komunikacijski most med posameznimi agencijami in centralno ravnijo: omogočajo zbiranje povratnih informacij iz prakse, kanaliziranje težav, posodabljanje navodil ter vzdrževanje skupnih standardov uporabe.

Ta sklop ukrepov je regulatorno pomemben, ker naslavlja ključno "mehko" tveganje generativnih sistemov v javnih storitvah: tveganje avtomatizacije prek navade. Tehnično lahko sistem ostaja priporočilen, a če organizacija ne investira v usposabljanje, podporo, komunikacijske kanale in delovne pogoje, se lahko odločanje vseeno postopno "preseli" na sistem. Mreža korespondentov in usposabljanja zato nista zgolj podporna aktivnosti, temveč del varovalne arhitekture: ustvarjata mehanizem, ki ohranja človeka kot dejanskega nosilca odgovornosti, hkrati pa omogoča učenje organizacije iz konkretnih primerov uporabe.

Zaključna presoja CNIL in pomen za uvrstitev v okvir 22. člena GDPR

CNIL na tej podlagi ocenjuje, da ob navedenih garancijah uporaba orodja verjetno ne predstavlja prepovedane oblike izključno avtomatiziranega odločanja po 22. členu GDPR. Pri tem je pomemben tudi element namenske zasnove: orodje svetovalcu ponuja izbor možnosti, ne pa enega samega rezultata, ki bi determiniral ravnanje; poleg tega sistem ne odloča o dostopu do storitve kot take (storitev svetovanja ostaja na voljo), temveč pomaga pri izbiri usposabljanj znotraj že obstoječega nabora.

Za namene dobrih praks je iz te presoje mogoče izpeljati jasen izvedbeni nauk: kadar se v javni storitvi uporablja LLM-podprto priporočanje na podlagi profila posameznika, je treba "meaningful human intervention" načrtovati kot sistem ukrepov, ki vključujejo:

1. oblikovanje uporabniškega vmesnika in prikaza rezultatov,
2. zadostno raven kontekstnih informacij za presojo,
3. procesne pravice in realno možnost odstopa, ter
4. stalno usposabljanje in organizacijsko podporo.

Šele kombinacija teh elementov omogoča, da človek ne ostane formalni člen v postopku, temveč je dejanski nosilec odločanja, kot to zahteva standard člena 22 GDPR.

Minimizacija podatkov in upravljanje promptov (privacy-by-design)

To podpoglavje obravnava dve tesno povezani temi, ker sta v primeru CNIL-ovega regulatorno vodenega peskovnika in projekta France Travail – »Conseils Personnalisés« (»Personalizirani nasveti«) dejansko del istega vprašanja: kako zagotoviti, da generativna rešitev deluje z najmanjšim potrebnim obsegom osebnih podatkov in da se ta obseg ne povečuje "po poti najmanjšega odpora" prek prostega besedilnega vnosa.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Pri klasičnih informacijskih sistemih je obseg osebnih podatkov pogosto vnaprej določen s strukturo obrazcev in podatkovnih shem; pri generativni UI pa se pomemben del interakcije seli v prosto besedilo, zato morajo biti ukrepi minimizacije kombinacija:

- tehnične zasnove,
- pravil uporabe in
- organizacijskih praks.

CNIL zato minimizacijo obravnava kot tipičen primer privacy-by-design: zahteve se ne izpolnijo zgolj z dokumentom, temveč s konkretnimi kontrolami v podatkovnih tokovih in uporabniški praksi.

Kategorizacija osebnih podatkov (sistematična minimizacija vhodov)

CNIL izrecno poudari, da minimizacija pri generativni UI zahteva sistematično identifikacijo kategorij osebnih podatkov, ki lahko napajajo orodje, ter njihovih virov. Pri projektu »Conseils Personnalisés« se to ne nanaša le na en "podatkovni vir", temveč na več vhodnih tokov, ki se v RAG-arhitekturi združujejo v proces generiranja priporočil. CNIL zato obravnava minimizacijo kot kartiranje: *kaj lahko sploh pride v sistem, od kod to prihaja, in zakaj je to za namen potrebno*.

V konkretnem primeru so bile pregledane zlasti:

(a) podatkovne zbirke, povezane s katalogom izobraževanj, in
(b) elementi osebnega dosjeja iskalca zaposlitve.

Ključno je, da CNIL ne govori o minimizaciji na splošni ravni, temveč o granularni analizi uporabnosti posameznih polj: za vsako informacijo se presoja, ali je dejansko potrebna za namen priporočanja usposabljanj. CNIL navaja tudi, da so bile identificirane informacije, ki za ta namen niso nujne, hkrati pa lahko povečajo tveganje pristranskosti ali diskriminacije. To je pomembna regulatorna logika: minimizacija ni samo "manj podatkov = manj tveganja", temveč tudi "manj nerelevantnih podatkov = manj prostora za nepravične korelacije" pri priporočilih.

Upravljanje promptov kot "vsebinska minimizacija" (prompt-higiena)

Posebnost generativnih sistemov v javnih storitvah je, da se dodatni podatki pogosto ne vključijo le prek strukturiranih baz, temveč prek besedilnih navodil ali pozivov (promptov), ki jih vnaša svetovalec. CNIL zato prompt obravnava kot samostojen "podatkovni kanal": ker je vnos po naravi prost, obstaja tveganje, da svetovalec iz želje po natančnosti v prompt nehote vnese prekomerne ali neprimerne osebne podatke, vključno s podatki, ki so občutljivi ali za namen nerelevantni (npr. družinske razmere, zdravstveno stanje ali druge okoliščine, povezane s socialno reintegracijo). CNIL zato upravljanje promptov uvrsti v jedro minimizacije: če se prompti ne upravljajo, lahko tudi sicer dobro zasnovana minimizacija na ravni podatkovnih polj "razpade" v praksi.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

CNIL predlaga kumulativne ukrepe, kar pomeni, da ne zadošča en sam "mehak" ukrep (npr. navodilo v dokumentaciji), ampak več med seboj dopolnjujočih se varovalk. Med predlaganimi ukrepi so:

- standardizirani predlogi promptov (vnaprej pripravljene strukture vprašanj, ki svetovalca vodijo k relevantnim informacijam in ga odvračajo od nepotrebnih osebnih podrobnosti),
- filtri oziroma »črne liste« za blokado posebej občutljivih pojmov ali kategorij,
- opozorila v uporabniškem vmesniku, ki svetovalca sproti opominjajo, kaj se ne sme vnašati,
- obvezna navodila dobrih praks in usposabljanje za pravilno oblikovanje promptov,
- ter dodatno predlog uvedbe listine o uporabi, ki formalizira pravila uporabe in jih "prizemlji" v konkretne primere dobre in slabe prakse.

Takšen nabor ukrepov je pomemben tudi zato, ker prompt-higiena ni zgolj zasebnostno vprašanje, temveč vpliva na kakovost priporočil: prekomerni ali neustrezni osebni podatki lahko povečajo šum, povzročijo neželene pristranskosti in znižajo predvidljivost izhodov. S tem postane upravljanje promptov hkrati ukrep privacy-by-design in ukrep kakovosti Sistema, ki ju zahtevata Akt o UI in GDPR.

Hranjenje promptov za testiranje in revizijo (dokazljivost, roki, namen)

CNIL v analizi prepozna, da obstaja legitimna potreba po hranjenju vhodnih promptov (in po potrebi izhodov) za namene testiranja, izboljšav, analize napak ter revizije – vključno z revizijo pristranskosti. V generativnih sistemih so prompti del "dejanskega delovanja" sistema; brez njih je težko reproducirati, zakaj je sistem predlagal določeno priporočilo ali zakaj je prišlo do napake. Zato je hranjenje promptov lahko pomembno za dokazljivost in za sistematično izboljševanje.

Hkrati pa CNIL jasno opozori, da takšno hranjenje ni "samoumevno": prompti lahko vsebujejo osebne podatke in zato njihovo shranjevanje zahteva:

(i) ustrezno pravno podlago,
(ii) presojo združljivosti namena (zlasti kadar se prompti, ki so nastali pri svetovanju, kasneje uporabljajo za analitiko ali razvoj; CNIL se tu sklicuje na potrebo po presoji skladno z GDPR art. 6(4)),
(iii) določitev primernih rokov hrambe in pravil dostopa. CNIL s tem nakaže tipično dilemo dobrih praks: po eni strani želimo dovolj podatkov za revizijo in izboljšave, po drugi strani pa mora veljati načelo, da se podatki hranijo le toliko časa in v takem obsegu, kot je nujno potrebno za določen, jasno opredeljen namen.

Povzetek dobrih praks

Iz obravnave CNIL izhaja jasna, prenosljiva ugotovitev: pri generativni UI minimizacija ni enkratna odločitev o "katerih poljih ne uporabljamo", ampak kontinuiran sklop varovalk na treh ravneh:

1. na ravni strukturiranih podatkovnih virov,
2. na ravni promptov kot nestrukturiranega kanala ter
3. na ravni evidence (logov) za dokazljivost in izboljšave.

V projektu »Conseils Personnalisés« CNIL ravno s to kombinacijo pokaže, kako se privacy-by-design v generativnih sistemih operacionalizira: z granularnim kategorizacijo in mapiranjem podatkov, z disciplineranjem promptov in z jasno urejeno politiko hranjenja podatkov za revizijo.

Obravnava pristranskosti in (ne)diskriminacije

CNIL v okviru spremljanja projekta *France Travail – »Conseils Personnalisés«* vprašanje pristranskosti ne obravnava kot abstraktno "etično" temo, temveč kot konkretno operativno tveganje, ki se lahko materializira že pri navidezno benignem priporočilnem sistemu. Ker orodje predlaga usposabljanja na podlagi profila iskalca zaposlitve, obstaja realna možnost, da sistem – bodisi zaradi podatkov, bodisi zaradi načina sklepanja modela – ustvarja diferencirane predloge, ki v praksi pomenijo neenako obravnavo. To tveganje je v javni storitvi še posebej občutljivo, ker lahko vpliva na priložnosti posameznika (npr. na usmerjanje v določene vrste usposabljanj, na dostop do podpornih programov ali na dinamiko svetovalnega procesa), četudi orodje formalno ne odloča o pravicah.

CNIL posebej izpostavi možnost, da priporočila postanejo "profilno pogojena" na način, ki ni legitimen z vidika enake obravnave. Kot primer tveganja navaja situacije, v katerih bi sistem na podlagi elementov profila (npr. spol ali družinske razmere) predlagal različne izobraževalne poti, pri čemer bi se lahko reproducirali stereotipi ali strukturne neenakosti (npr. "primerni" poklici, predpostavke o razpoložljivosti časa, implicitne ocene motivacije ipd.). Ključno je, da se pri takih sistemih diskriminatorni učinki pogosto ne pojavijo kot eksplicitno pravilo ("če X, potem Y"), temveč kot rezultat kombinacije več posrednih signalov, ki se v modelu sestavijo v vzorec.

Zato CNIL kot potencialne vire pristranskosti ne omeji le na očitne kategorije, ampak opozori tudi na bolj subtilne poti, po katerih sistem lahko "rekonstruira" občutljive lastnosti ali socialne okoliščine. Med možnimi viri pristranskosti omenja tudi geografske podatke, ki lahko delujejo kot posredni približek socialno-ekonomskega položaja ali etničnega ozadja, ter možnost inferiranja spola iz imen ali drugih identifikatorjev. Poleg tega izpostavi tudi občutljive informacije, ki jih sistem lahko izpelje iz življenjske poti posameznika, tj. iz kombinacije podatkov o preteklih zaposlitvah, prekinitvah, statusih ali drugih elementih profila. To je posebej pomembno v generativnih sistemih, kjer model ne dela samo "klasične"

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

statistične napovedi, temveč generira besedilne predloge in utemeljitve, v katerih se lahko implicitne pristranskosti izrazijo kot navidezno "razumna" priporočila.

V tem okviru CNIL opozori na praktično pravno dilemo: če želimo pristranskost preverjati sistematično, se pogosto pojavi potreba po t. i. testiranju po zaščitenih kategorijah (npr. primerjava rezultatov med spoloma ali drugimi relevantnimi skupinami). Vendar pa GDPR glede posebnih vrst osebnih podatkov (t. i. "občutljivi podatki") postavlja stroge omejitve, in CNIL izrecno poudari, da GDPR praviloma ne omogoča enostavnega "obvoda" teh omejitev zgolj zato, ker bi organizacija želela izvesti revizijo ali "fairness audit" UI sistema. Z drugimi besedami: dobra namera (testiranje nediskriminacije) sama po sebi še ne pomeni, da je pravna podlaga za obdelavo občutljivih podatkov avtomatično zagotovljena.

Hkrati pa CNIL nakaže, da se regulativni okvir na ravni EU razvija in da bi lahko določene zahteve prihodnjega režima (npr. EU Akt o umetni inteligenci) v določenih kontekstih okrepile pravno-operativni prostor za sistematično testiranje pristranskosti – predvsem skozi zahteve glede upravljanja podatkov in kakovosti podatkovnih zbirk pri sistemih visokega tveganja. Pri tem je treba takšno možnost razumeti previdno: ne gre za to, da bi AI Akt "razveljavil" GDPR, temveč za to, da lahko pri določenih sistemih in ob izpolnjenih pogojih nastane bolj strukturiran nabor obveznosti in pričakovanj glede testiranja, dokumentiranja in dokazovanja, kar organizacijam olajša utemeljevanje nujnosti določenih testov in vzpostavljanje dokazljivega okvira skladnosti.

Implikacija za dobre prakse v tem primeru je zato dvojna:

1. Prvič, obravnava pristranskosti mora biti zasnovana kot del celotne arhitekture *privacy-by-design*: minimizacija podatkov, nadzor promptov in organizacijski ukrepi zmanjšujejo verjetnost, da sistem sploh dobi signale, ki bi omogočali neželene diskriminatorne vzorce.
2. Drugič, kadar je testiranje pristranskosti vseeno potrebno, mora organizacija že vnaprej načrtovati pravni in procesni okvir testiranja (kateri podatki, za kateri namen, za koliko časa, kdo dostopa, kako se rezultati dokumentirajo), saj ugotavljanje nepristranskosti ("fairness audit") v javnih storitvah ni zgolj tehnično vprašanje, temveč tudi vprašanje zakonitosti in dokazljivosti.

Preliminarna navezava na Akt o UI (tveganost)

Pri tem primeru je koristno jasno ločiti dve pravni "optiki", ki se deloma prekrivata, vendar ne odgovarjata na isto vprašanje. CNIL je v svoji analizi primarno presojala skladnost z GDPR, zlasti ali lahko uporaba orodja *»Conseils Personnalisés«* v danem scenariju pomeni izključno avtomatizirano individualno odločanje s pravnimi ali podobno pomembnimi učinki (GDPR, člen 22). Takšna presoja je vezana na vprašanje *učinka na posameznika* in na to, ali je odločanje *izključno avtomatizirano*.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

V Aktu o UI se tveganost UI sistema presoja na način, da je razvrstitev »visoko tveganih« (high-risk) sistemov v 6. členu vezana na dve poti, in sicer

1. na UI sisteme, ki so varnostna komponenta ali del proizvoda, ki spada pod harmonizacijsko zakonodajo EU iz Priloge I in je zanj praviloma predvidena presoja skladnosti s strani tretje osebe, ter
2. na samostojne UI sisteme, katerih namen uporabe spada med primere iz Priloge III. V v tem smislu je Priloga III ključna za "use-case" pot iz člena 6(2), ne pa edini vir opredelitve visoko-tveganih sistemov.

To pomeni, da je povsem možno, da sistem po GDPR ne zapade v 22. člen (ker človek dejansko odloča), hkrati pa bi bil po Aktu o UI še vedno lahko obravnavan kot visoko tvegan – če bi bil njegov namen uporabe umeščen v relevantno področje Priloge III.

Za primer *France Travail* sta z vidika Akta o UI posebej relevantni dve področji iz Priloge III:

1. Zaposlovanje, upravljanje delovnih razmerij in dostop do samozaposlitve (Priloga III, točka 4): sem sodijo sistemi za rekrutiranje/izbor kandidatov ter sistemi, ki vplivajo na odločitve v delovnih razmerjih ali spremljanje in vrednotenje vedenja/uspešnosti v delovnem razmerju.

   V trenutnem opisu *»Conseils Personnalisés«* (priporočanje usposabljanj za iskalce zaposlitve) orodje ni neposredno sistem za rekrutiranje ali upravljanje delovnega razmerja, kot je tipično razumljeno v točki 4. Vendar pa je področje zaposlovanja kot takšno za France Travail naravno "bližnje okolje" in je pomembno, ker se podobna tehnologija lahko relativno hitro razširi npr. na rangiranje zaposlitvenih možnosti, ujemanje kandidatov, ali druge funkcionalnosti, ki bi že bolj neposredno posegale v sfero zaposlitvenih odločitev. Ta opomba je v poročilu relevantna zato, ker kaže, da se tveganost po Aktu o UI pogosto ne spremeni zaradi tehnologije same, temveč zaradi razširitve namena uporabe.

2. Dostop do in uživanje bistvenih zasebnih storitev ter bistvenih javnih storitev in koristi (Priloga III, točka 5): sem med drugim sodijo sistemi, ki jih javni organi (ali v njihovem imenu) uporabljajo za ocenjevanje upravičenosti fizičnih oseb do bistvenih javnih pomoči/storitev in za dodelitev, zmanjšanje, odvzem ipd. takih koristi in storitev.
   Ta kategorija je za primer France Travail konceptualno posebej pomembna, ker se primer odvija v kontekstu javne službe, kjer lahko priporočila in usmeritve vplivajo na pot obravnave posameznika. Če bi orodje *»Conseils Personnalisés«* (ali njegova nadgradnja) postalo del procesa, ki bi dejansko vplival na upravičenost, prioritizacijo, dodelitev ali omejitev določenih javnih storitev ali koristi, bi se njegova umestitev v točko 5 lahko okrepila.

Iz tega sledi tudi ključna metodološka posledica za naš dokument dobrih praks: ta primer je operativno zelo uporaben kot"visoko-tvegan" kandidat (v smislu "najstrožjega" pristopa k

dobri praksi), ker leži v neposredni bližini področij Priloge III, kjer Akt o UI pričakuje najvišjo raven upravljanja tveganj, človeškega nadzora, podatkovnega upravljanja in dokazljivosti. V praksi to pomeni, da je smiselno, da dobre prakse v tem primeru oblikujemo tako, kot da gre za visokotvegani scenarij (zlasti glede: upravljanja tveganj, smiselnega človeškega nadzora, minimizacije podatkov in upravljanja promptov), ker to predstavlja robusten standard, ki ostane uporaben tudi v primeru kasnejše razširitve namena uporabe.

Hkrati pa je mora biti poročilo terminološko natančeno: končna kvalifikacija po Aktu o UI ni stvar "občutka", temveč zahteva natančno opredelitev namena uporabe in vloge izhoda v postopku. V tem konkretnem primeru bi to pomenilo jasno odgovoriti vsaj na naslednje:

- ali sistem zgolj predlaga usposabljanja kot strokovna podpora svetovalcu,
- ali pa (dejansko ali formalno) začne vplivati na upravičenost, dodelitev ali odločanje glede storitev/koristi oziroma na odločitve v sferi zaposlovanja.

Ta razmejitev je odločilna, ker je vezanost na "visoko-tveganje" po Aktu o UI utemeljuje prav preko namena uporabe v kategorijah Priloge III.

Ključne prenosljive dobre prakse

To podpoglavje povzema dobre prakse, ki izhajajo neposredno iz konkretnega primera regulatorno vodenega "peskovnika" za umetno inteligenco v javnih storitvah. Namen je dvojni:

- bralcu omogočiti jasno povezavo med opisanim projektom in izvedenimi varovalkami, ter
- pokazati, kako so bile ključne zahteve (zasebnost, zaupnost, nepristranskost in človeški nadzor) operacionalizirane v praksi.

Zato so spodaj navedene prakse omejene na tiste, ki jih CNIL v tem primeru izrecno obravnava ali ki so bile v okviru projekta dejansko uvedene (npr. usposabljanje, mreža »AI korespondentov«, prompt-disciplineriranje, granularna minimizacija polj profila, ureditev hrambe promptov za revizijo).

1) "Človeška intervencija" je bila obravnavana kot izvedbeni pogoj, ne kot formalna klavzula

CNIL je v tem primeru najprej jasno problematizirala tveganje, da lahko priporočilni sistem – četudi formalno ne sprejema odločitev – v praksi postane de facto odločilni dejavnik. Zato je kot jedrno dobro prakso izpostavila zahtevo, da mora svetovalec prejeti predloge v razumljivem in obvladljivem formatu, ter da mora imeti na voljo informacije in pogoje, ki mu

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

omogočajo strokovno presojo predloga, vključno z realno možnostjo, da predlogu ne sledi. Ključna izvedbena lekcija primera je torej, da "human oversight" pri generativnih sistemih ni zgolj prisotnost človeka v procesu, temveč kombinacija vmesnika, podpore in delovnih pogojev, ki omogočajo dejansko presojo.

2) France Travail je človeški nadzor podprl z organizacijskimi ukrepi: usposabljanje + mreža »AI korespondentov«

Kot konkretno dobro prakso CNIL navaja, da je France Travail vzpostavil stalno ozaveščanje in usposabljanje (splošno izobraževanje o UI ter specifično usposabljanje za uporabo orodja) ter mrežo »AI korespondentov« kot strukturiran komunikacijski kanal med lokalnimi agencijami in centralno ravnjo. To je v CNIL-ovi logiki ključna varovalka proti "avtomatizaciji po navadi": sistem lahko ostane priporočilen, vendar brez organizacijske podpore svetovalci hitro začnejo izhodom slediti rutinsko. V tem primeru je bila mreža korespondentov uporabljena kot mehanizem za povratne informacije, prenos navodil in standardizacijo uporabe.

3) Minimizacija podatkov je bila izvedena kot "granularna" presoja polj profila in virov podatkov

CNIL v tem primeru minimizacije ne obravnava splošno, temveč jo veže na konkretno projektno prakso: pregledani so bili

- podatki iz kataloga izobraževanj in
- elementi osebnega dosjeja iskalca zaposlitve, pri čemer je bila izvedena granularna analiza uporabnosti posameznih polj.

Posebno pomembna dobra praksa je, da so bile identificirane tudi informacije, ki za namen priporočanja niso potrebne in lahko hkrati povečajo tveganje pristranskosti ali diskriminacije. Iz francoskega primera zato izhaja jasna izvedbena lekcija: minimizacija pri LLM + RAG pomeni sistematično mapiranje *kaj, od kod, zakaj* in hkrati presojanje *ali ta podatek odpira neželene korelacije*.

4) Prompti so bili obravnavani kot samostojen "podatkovni kanal" in predmet posebne higiene

CNIL je izrecno opozorila, da so prompti privzeto vsebinsko prosti, zato obstaja tveganje, da svetovalci vnašajo prekomerne ali neprimerne osebne podatke (navedeni so primeri, kot so družinske razmere in zdravstveno stanje). V tem primeru je dobra praksa prav v tem, da je bil prompt obravnavan kot del zasnove "privacy-by-design": CNIL predlaga kumulativne ukrepe, kot so standardizirani predlogi promptov, filtri/»blacklist«, opozorila v uporabniškem vmesniku ter najmanj obvezna navodila dobrih praks in usposabljanje za pisanje promptov; dodatno omenja tudi listino (charte) o uporabi. Ključno sporočilo primera je, da se minimizacija podatkov pri generativni UI pogosto "zlomi" ravno na promptih, zato mora biti prompt-higiena formalizirana in podprta z več nivoji varovalk.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

5) Hranjenje promptov je bilo priznano kot legitimno za revizijo – vendar le pod strogimi pogoji

CNIL v tem primeru prepoznal, da obstaja legitimna potreba po hranjenju promptov za testiranje, izboljšav in revizijo (vključno s testiranjem pristranskosti). Hkrati pa je opozoril, da takšno hranjenje zahteva jasen pravni okvir: ustrezno pravno podlago, presojo združljivosti namena (GDPR člen 6(4)) ter določitev primernih rokov hrambe in pravil dostopa. Dobra praksa, ki jo iz primera lahko neposredno izluščimo, je torej dvojna: (i) hranjenje promptov je del dokazljivosti in kvalitete, (ii) vendar mora biti strogo omejeno, namensko utemeljeno in časovno zamejeno, ker prompti lahko vsebujejo osebne podatke.

6) Infrastrukturna izbira (lokalna namestitev) je bila obravnavana kot regulatorno relevantna varovalka zaupnosti

CNIL izrecno navaja uporabo modela Mixtral v lokalni namestitvi (on-premise) in opozori, da lahko off-premise uporaba poveča tveganje izgube zaupnosti osebnih podatkov. V tem primeru je dobra praksa v tem, da je bil izbor infrastrukture obravnavan kot del upravljanja tveganj, ne kot nevtralna IT odločitev.

Izvedbena lekcija je tudi, da lokalni pristop sicer zmanjša določena tveganja (izpostavljenost podatkov tretjemu ponudniku), vendar zahteva visoko zrelost notranjega upravljanja (dostopi, revizijske sledi, incidenti), kar CNIL implicitno naslavlja skozi širšo logiko "varne namestitve" in organizacijskih varovalk.

Viri (primer 3.1)

1. CNIL (2025). *Bac à sable « IA et services publics » – Les recommandations de la CNIL aux lauréats* (PDF). https://www.cnil.fr/sites/cnil/files/2025-04/bac_a_sable_recommandations.pdf
2. CNIL (2023). *« Bac à sable » intelligence artificielle et services publics : la CNIL accompagne 8 projets innovants* (spletna stran). https://www.cnil.fr/fr/bac-sable-intelligence-artificielle-et-services-publics-la-cnil-accompagne-8-projets-innovants
3. CNIL (2025). *Intelligence artificielle et services publics : la CNIL publie le bilan de son « bac à sable »* (spletna stran). https://www.cnil.fr/fr/bilan-bac-a-sable-IA-services-publics
4. CNIL (2023). *Comment déployer une IA générative ? La CNIL apporte de premières précisions* (Q&A). https://www.cnil.fr/fr/comment-deployer-une-ia-generative-la-cnil-apporte-de-premieres-precisions
5. CNIL (2024). *Revoir le webinaire : Comment déployer un système d'IA générative : les premières préconisations* (stran + posnetek).

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

https://www.cnil.fr/fr/revoir-le-webinaire-comment-deployer-un-systeme-dia-generative-les-premieres-preconisations
(posnetek) https://video.cnil.fr/w/wXUgWCbt1fze2rBiFYazRj
6. CNIL. *Les fiches pratiques IA* (zbirka praktičnih listov). https://www.cnil.fr/fr/les-fiches-pratiques-ia

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

3.2 Datatilsynet sandbox – Finterai: federated learning za AML (ML brez deljenja podatkov)

Norveški nadzorni organ Datatilsynet je v okviru regulativnega sandbox programa preučil, ali lahko federativno učenje (federated learning) predstavlja izvedljiv pristop v primerih, ko je deljenje podatkov med organizacijami težko ali pravno omejeno. Študija se osredotoča na primer uporabe v kontekstu preprečevanja pranja denarja (AML), kjer banke potrebujejo učinkovitejše metode zaznavanja sumljivih transakcij, vendar hkrati obdelujejo občutljive osebne podatke in so zavezane načelom GDPR.

Datatilsynet je norveški nadzorni organ za varstvo osebnih podatkov. V okviru svojega regulativnega peskovnika (*regulatory privacy sandbox*) skupaj z izbranimi udeleženci preizkuša in pravno-tehnično presoja nove digitalne rešitve z vidika varstva osebnih podatkov, pri čemer Datatilsynet nudi svetovanje in usmeritve, ugotovitve pa ne pomenijo zavezujoče odločitve ali predhodne odobritve.

V tem okviru je Datatilsynet obravnaval primer podjetja Finterai, norveškega zagonskega podjetja, ki želi federativno učenje uporabiti v boju proti pranju denarja in financiranju terorizma (AMLTF). Izhodiščni problem v AML praksi je opisan zelo operativno: posamezna banka ima pogosto "premalo" potrjeno kriminalnih transakcij, da bi lahko z visoko zanesljivostjo razločila sumljive vzorce od običajnih transakcij. Posledica so sistemi nadzora transakcij, ki označijo preveč transakcij kot sumljivih (veliko lažnih pozitivnih), kar sproži časovno in stroškovno zahtevne ročne preiskave. Datatilsynet poudari, da bi se ta problem lahko omilil, če bi se modeli gradili na več podatkih, vendar banke potrebnih podatkov praviloma ne morejo deliti, ker transakcije vsebujejo osebne podatke.

Projekt Finterai je bil zasnovan kot odgovor na to strukturno dilemo: kako omogočiti učenje iz "širšega" vzorca transakcij, ne da bi banke med seboj dejansko delile podatke o strankah. Finterai predlaga uporabo federativnega učenja (*federated learning*), ki ga Datatilsynet opiše kot decentraliziran pristop v strojnem učenju, kjer posamezni udeleženci (npr. banke) učijo modele lokalno, med seboj pa izmenjujejo učno informacijo na način, ki ne zahteva prenosa surovih osebnih podatkov. Ključna obljuba tega pristopa v obravnavanem primeru je, da banke lahko "učijo druga od druge" brez dejanskega deljenja podatkov o svojih strankah.

Kdo so uporabniki in komu je rešitev namenjena

Primarni uporabniki rešitve so banke oziroma njihovi oddelki in strokovni profili, ki izvajajo transakcijski nadzor in obravnavajo zaznane AML alarme. To so praviloma strokovni uporabniki, ki delujejo v okviru obveznosti po AML pravilih in izvajajo ročne preglede ter nadaljnjo obravnavo transakcij, ki jih sistemi označijo kot sumljive.

Vloga Finterai v tem scenariju ni v tem, da bi nadomestil AML presojo bank, temveč da ponudi tehnološki pristop in (potencialno) infrastrukturno-organizacijski okvir, ki bankam omogoča

sodelovanje pri učenju modelov ob ohranjanju decentralizacije podatkov. Datatilsynet posebej izpostavi, da je eden osrednjih sklopov presoje v peskovniku ravno vprašanje vlog in odgovornosti pri obdelavi v takem modelu sodelovanja.

Cilj rešitve

Cilj obravnavanega primera je torej omogočiti, da banke izboljšajo zaznavanje sumljivih transakcij z učenjem iz širšega spektra vzorcev, ne da bi morale med seboj deliti transakcijske podatke, ki vsebujejo osebne podatke. Datatilsynet povzame, da je bil v peskovniku pri tem pristopu analiziran nabor ključnih vprašanj, ki jih federativno učenje odpira v razmerju do pravil varstva osebnih podatkov, in sicer zlasti v zvezi z (i) obdelovalnimi vlogami (kdo je upravljavec/obdelovalec), (ii) načelom minimizacije podatkov in (iii) varnostnimi izzivi pri takšni decentralizirani zasnovi.

Tehnična zasnova: federativno učenje + standardizacija podatkov + decentralizirana obdelava

Arhitekturno izhodišče: federativno učenje kot "učenje brez deljenja podatkov"

Federativno učenje je zasnovano kot distribuiran postopek učenja modela, pri katerem se surovi podatki ne zbirajo centralno, temveč ostanejo pri posameznih udeležencih (npr. bankah). Udeleženci prejmejo enak začetni model, ga lokalno trenirajo na svojem podatkovnem naboru, nato pa navzven posredujejo zgolj učne posodobitve (npr. parametre ali posodobitve parametrov), ki so lahko dodatno zaščitene s kriptografskimi mehanizmi. Centralna strežniška komponenta izvede varno agregacijo (*secure aggregation*), s čimer združi prejete posodobitve v novo različico modela, in jo ponovno razpošlje udeležencem. Postopek se ponavlja v več krogih, dokler model ne doseže želenega stanja oziroma konvergence.

Ključna značilnost takšne arhitekture je, da se med udeleženci in centralno komponento izmenjujejo modelne informacije (»učenje«), ne pa izvorni transakcijski ali drugi osebni podatki. Namen tega pristopa je omogočiti izboljševanje modela na podlagi širšega spektra vzorcev, hkrati pa omejiti potrebo po prenosu in združevanju podatkov med organizacijami, kjer bi bilo to pravno, varnostno ali operativno težko izvedljivo.

Datatilsynet v zaključnem poročilu najprej pojasni standardni (referenčni) potek federativnega učenja: udeleženci prejmejo isti model, ga **lokalno trenirajo** na svojem podatkovnem naboru, nato pa posredujejo **šifrirane učne posodobitve** (paket posodobitev modela) ki naj **ne bi vsebovale osebnih podatkov**; centralna strežniška komponenta izvede **varno agregacijo (secure aggregation)** in z agregiranimi posodobitvami posodobi centralni model; postopek se ponavlja do konvergence, nato udeleženci prejmejo izboljšan model. Datatilsynet izrecno poudari, da je temeljni namen pristopa prav v tem, da se lokalni podatki

(ki pogosto vključujejo osebne podatke) **v teoriji ne prenašajo** med udeleženci niti na centralni strežnik, temveč se izmenjuje "učenje" (modelni parametri).

Kaj se v resnici izmenjuje: modelni parametri in hiperparametri in ter zakaj to ni "ničelno tveganje"

Pri federativnem učenju surovi podatki (npr. transakcije, identifikatorji strank, SWIFT sporočila) po zasnovi ostanejo v okolju vsake banke. Kljub temu pa se med udeleženci in koordinatorjem v procesu učenja **ne izmenjuje "nič"**: izmenjuje se **model** in/ali **modelne posodobitve**, ki predstavljajo "znanje", pridobljeno iz lokalnih podatkov.

V praksi to pomeni, da se navzven prenašajo predvsem:

- **modelni parametri (uteži)** oziroma njihove posodobitve (*model updates*): to so numerične vrednosti, ki se pri učenju spreminjajo in določajo, kako model obravnava vhodne podatke; v nekaterih implementacijah se prenašajo posodobitve (npr. Δuteži ali gradienti), v drugih pa – kot je v tem primeru opisano – lahko kroži tudi **celoten model**;
- **hiperparametri** (vsaj v določenih fazah): to so nastavitve učenja, ki ne predstavljajo "znanja" o posameznih primerih, temveč določajo, kako poteka učenje (npr. hitrost učenja, število iteracij, izbrana arhitektura, regularizacija). V opisu Finterai se izpostavi, da banke same določijo model in hiperparametre, kar pomeni, da je ta del konfiguracije v praksi del dogovorjenega protokola učenja.

Takšna izmenjava p< vendarle **ni brez tveganja**, čeprav se surovi podatki ne prenašajo. Modelni parametri so rezultat učenja na podatkih; z drugimi besedami, uteži so "odtis" statističnih zakonitosti, ki jih je model zaznal v lokalnih transakcijah. V večini primerov iz uteži **ni mogoče neposredno "prebrati"** posameznega transakcijskega zapisa ali identitete stranke. Vendar pa obstajajo znani razredi napadov in zlorab, pri katerih se lahko iz modela ali modelnih posodobitev poskuša izpeljati informacija o učnih podatkih – bodisi o tem, ali je bil določen posameznik v učnem naboru, bodisi o značilnostih podatkovnih vzorcev, ki so bili uporabljeni pri učenju.

Zaradi tega je bilo v peskovniku postavljeno osrednje tehnično-varnostno vprašanje: ali bi lahko model, ki kroži med udeleženci, v določenih okoliščinah posredno razkril osebne podatke, zlasti če bi imel ranljivosti ali bi bil izpostavljen zlonamernemu ravnanju (npr. napadom tipa *model inversion*).

Zato je v tem use-caseu federativno učenje najbolj pravilno razumeti kot pristop, ki zmanjša potrebo po čez-organizacijskem prenosu in centraliziranju surovih transakcijskih podatkov, s čimer praviloma zniža določena klasična tveganja (npr. velika "koncentracija" osebnih podatkov v eni točki in z njo povezana izpostavljenost).

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Vendar pa tveganje ni ničelno, ker proces kljub temu vključuje **izmenjavo modela oziroma modelnih posodobitev**, ki so rezultat učenja na osebnih podatkih. Posledično je treba varnostno presojo razširiti: poleg varovanja podatkovnih zbirk je treba obravnavati tudi varnost modelov in protokola izmenjave (npr. integriteto in zaupnost prenosov, možnost zlonamernih udeležencev, ter tveganja posrednega razkritja informacij iz modela). Ta premik težišča – od "deljenja podatkov" k "deljenju učenja" – je tudi jedro razlogov, zakaj federativno učenje v AML okolju deluje kot obetavna, vendar ne samodejno varna rešitev.

Specifika Finterai pristopa: serijsko (ne paralelno) učenje in prenos celotnega modela

Datatilsynet nato opiše, da želi Finterai federativno učenje uporabiti drugače kot referenčni (Googlov) pristop. Ključna razlika nastopi že na začetku: Finterai predvideva, da stranke (banke), ne Finterai, določijo, *kakšen model* se bo treniral, in same definirajo hiperparametre. Posledično je Finterai zasnoval federativno učenje kot serijski postopek: model se najprej uči pri eni banki, nato se pošlje naslednji banki itd. Datatilsynet izrecno izpostavi še eno pomembno razliko: pri referenčnem pristopu udeleženci običajno vračajo zgolj "model update" (npr. gradiente), medtem ko Finterai predvideva, da se v njihovem procesu vrača celoten model; to je sicer težje z vidika prenosa (več prometa), vendar Datatilsynet navaja, da se Finterai za to odloča zaradi kombinacije tehničnih, varnostnih in poslovnih razlogov ter da tudi »secure aggregation« implementira drugače, delno zaradi drugačnega use-casea.

Finterai primer: potek učenja

Datatilsynet poda poenostavljen, a operativno zelo uporaben opis poteka Finterai federativnega učenja:

1. Sodelujoča banka pošlje Finterai zahtevo za vzpostavitev modela in posreduje svoje hiperparametre ter navodila za trening.
2. Finterai na tej podlagi zgradi začetni model.
3. Model in hiperparametre pošlje prvi banki, kjer se model **lokalno trenira**.
4. Lokalni trening poteka na **standardiziranih transakcijskih podatkih** in drugih podatkih (Datatilsynet navaja tudi KYC in tretje-stranske podatke).
5. Po končanem treningu banka model vrne Finterai; Finterai ga shrani v svojo bazo.
6. Finterai izvede **kakovostno-varnostne kontrole** (med drugim preverjanje morebitnih podatkovnih uhajanj in pristranskosti).
7. Finterai nato posreduje posodobljeni model naslednji banki; postopek se ponavlja do konvergence.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

8. Končni model je shranjen na strežniku in je dostopen sodelujočim bankam, ki ga lahko prenesejo ter uporabijo na svojih lokalnih naborih podatkov za prepoznavanje sumljivih transakcij.

Za arhitekturo je posebej pomembna Datatilsynetova eksplicitna opomba: v tej zasnovi naj bi vsa hramba podatkov v zvezi s procesi (vključno z nadzorom transakcij) potekala pri bankah, Finterai pa ne bi imel dostopa do lokalnih transakcijskih podatkov bank pri razvoju ali obratovanju storitve.

Standardizacija transakcijskih podatkov kot predpogoj za sodelovanje med bankami

Datatilsynet poudari, da federativno učenje (vsaj v "horizontalnem" smislu) predpostavlja, da imajo udeleženci podatke v primerljivih točkah in formatu, zato je standardizacija transakcijskih podatkov ena ključnih tehničnih pred-faz. V poročilu Datatilsynet navaja, da Finterai bankam ponuja programsko opremo za standardizacijo, ki uporablja AI-podprto obdelavo naravnega jezika (natural language processing) z namenom standardizacije podatkov – vključno s SWIFT sporočili – tako, da so podatki v enotnem formatu med sodelujočimi bankami.

Za varnostno-zasebnostni profil je ključno še, kako je ta komponenta umeščena: Datatilsynet navede, da se programje namesti in izvaja v IT okolju posamezne banke, brez dostopa Finterai ali drugih bank do podatkov, ki se standardizirajo. Datatilsynet prav tako zapiše, da naj bi se trening algoritmov v standardizacijski programski opremi poteka interno pri posamezni banki, pri čemer se naučene standardizacijske preslikave/parametri (tj. rezultat treninga standardizacijskega modula) ne delijo niti s Finterai niti z drugimi bankami; izmenjava med bankami je predvidena šele na ravni federativnega učenja AML modela."

Podatkovne kategorije

V obravnavanem use-caseu je izhodiščni podatkovni prostor tipičen za AML okolje: jedro obdelave predstavljajo **transakcijski podatki**, ki nastajajo pri izvajanju plačil in prenosov. Ti zapisi pogosto vključujejo tako strukturirane elemente (npr. znesek, valuta, datum, računi, banke posrednice) kot tudi opisna polja, kjer se prenašajo dodatne informacije o namenu transakcije ali identifikaciji strank. V tem okviru poročilo izrecno izpostavi tudi SWIFT sporočila, ki lahko vsebujejo osebne podatke, kadar je pošiljatelj ali prejemnik fizična oseba, oziroma kadar opisna polja vsebujejo identifikacijske podatke ali druge osebne okoliščine transakcije.

Poleg transakcijskih podatkov so pomemben vir tudi **KYC podatki** (*know-your-customer*), tj. podatki, ki jih banke zberejo in vzdržujejo za identifikacijo stranke, razumevanje profila tveganja in izpolnjevanje zakonskih obveznosti skrbnega preverjanja. V praksi gre za kombinacijo identifikacijskih in atributnih podatkov (npr. status stranke, poslovni profil, lastništvo pri pravnih osebah, geografske povezave), ki se uporabljajo bodisi neposredno bodisi kot vhodne značilnice pri zaznavanju sumljivih vzorcev.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Tretji sklop predstavljajo t.i. **zunanji podatki**, ki jih banke uporabljajo kot dodatne indikatorje tveganja. Poročilo kot primere navaja podatke o PEP (politično izpostavljene osebe), podatke iz sankcijskih seznamov ter podatke iz virov, povezanih z negativnimi medijskimi objavami (t. i. adverse media). Ti viri imajo pomembno vlogo pri oceni tveganja, vendar pogosto prihajajo iz različnih ponudnikov, imajo različne podatkovne strukture in lahko vsebujejo napake ali nepopolnosti, zato jih je treba uporabljati previdno in z jasnimi pravili o ažurnosti ter kakovosti.

- Takšna razčlenitev podatkovnih kategorij je za tehnično zasnovo ključna iz dveh razlogov. Prvič, pojasni, zakaj je **standardizacija** v tem use-caseu zahtevna: transakcijski podatki (zlasti SWIFT sporočila) in z njimi povezani opisi so heterogeni, pogosto polstrukturirani in se med bankami razlikujejo; brez standardizacije ni mogoče zagotoviti, da bi federativno učenje nad različnimi bankami delovalo na primerljivih vhodih.
- Drugič, razlaga, zakaj je **minimizacija podatkov** v AML kontekstu specifična: banke imajo regulatorno obveznost izvajati nadzor in zaznavati sumljive transakcije, vendar to ne pomeni, da je dopustno vključiti »vse podatke, ki obstajajo«.

Ravno zaradi občutljivosti transakcijskih in KYC podatkov je treba pri zasnovi modela in standardizacijskih cevovodov sistematično utemeljiti, kateri podatki so dejansko potrebni za namen zaznavanja vzorcev pranja denarja, ter kako se ob tem zmanjša nepotrebna izpostavljenost osebnih podatkov (npr. s selekcijo polj, pseudonimizacijo, omejitvijo dostopov in jasnimi roki hrambe).

Infrastruktura: centralne komponente in vloga oblačnih storitev

Čeprav Datatilsynet ne podaja podrobne implementacijske specifikacije infrastrukture, iz opisa izhaja, da ima Finterai v procesu vlogo centralnega koordinatorja (gradnja začetnega modela, hranjenje modelov, preverjanje uhajanj in pristranskosti, distribucija modela med bankami ter objava končnega modela). Datatilsynet hkrati opozori, da rešitve na področju ML pogosto uporabljajo specializirano programsko in strojno opremo, ki se v praksi pogosto realizira z uporabo oblačnih storitev, ter da takšna raba zahteva ustrezne varnostne kompetence; vendar Datatilsynet tudi izpostavi, da federativno učenje omogoča zmanjšanje klasičnih tveganj, povezanih s centraliziranim nalaganjem in agregiranjem velikih količin podatkov, ker omogoča lokalizacijo obdelave.

Operativni cilj tehnične zasnove

Operativni cilj tehnične zasnove v tem primeru je omogočiti bankam, da pri zaznavanju sumljivih transakcij izkoristijo »**kolektivno učenje**« – tj. učenje iz širšega spektra transakcijskih vzorcev, kot jih ima na voljo posamezna banka – brez klasičnega pristopa, pri

katerem bi bilo treba transakcijske podatke in druge osebne podatke prenašati in združevati v skupnem (centraliziranem) okolju.

V AML praksi je to operativno relevantno predvsem zato, ker je veliko sumljivih vzorcev redkih, heterogenih in se razkrije šele ob večjem naboru primerov; če so podatki razpršeni po bankah, ima vsaka banka le del "slike", zato lahko modeli, trenirani zgolj na lokalnih podatkih, ostajajo manj občutljivi ali pa povzročajo več lažnih alarmov.

Federativno učenje v tem kontekstu deluje kot tehnični kompromis: podatki ostanejo pri bankah, izmenjuje pa se **učenje** – tj. model ali modelne posodobitve (uteži, posodobitve uteži oziroma drugi modelni artefakti, ki nastanejo kot rezultat učenja na lokalnih podatkih). Takšna zasnova praviloma zmanjša potrebo po pravno in varnostno zahtevnem čez-organizacijskem deljenju surovih podatkov, vendar ne odpravi vseh tveganj.

Ravno zato je pomemben drugi del operativne logike: federativna rešitev mora biti zasnovana in upravljana kot **sistem z integriranimi zahtevami informacijske varnosti in varstva osebnih podatkov** (tj. kot "security- in privacy-by-design" rešitev), ne zgolj kot podatkovno-znanstveni eksperiment, kjer je edini kriterij uspešnosti natančnost modela.

V praksi to pomeni, da je treba poleg učinkovitosti zaznavanja (npr. zmanjšanje lažnih pozitivnih in povečanje zaznavne moči) sistematično obravnavati tudi tveganja, ki izvirajo iz samega protokola izmenjave modelov: modelne posodobitve lahko postanejo nova napadalna površina, zlasti če bi zlonamerni udeleženec ali napadalec poskušal iz modela oziroma posodobitev posredno izpeljati informacije o učnih podatkih. Eden tipičnih primerov takega napada so napadi tipa model inversion, kjer napadalec poskuša rekonstruirati določene lastnosti učnih podatkov ali sklepati o njih na podlagi dostopa do modela.

Zato operativni cilj tehnične zasnove v tem primeru ni zgolj "naučiti boljši model", temveč vzpostaviti takšen režim obratovanja, kjer so:

(i) prenosi modelov zaščiteni z vidika zaupnosti in integritete,
(ii) (ii) vloge in dostopi strogo kontrolirani,
(iii) (iii) uvedene kontrole proti zlorabam in napadom, ter
(iv) (iv) zagotovljena dokazljivost ukrepov in odzivnost na incidente.

V tem smislu "kolektivno učenje" ni samo tehnična inovacija, temveč je tudi **organizacijsko-varnostni projekt**: uspeh se ne meri zgolj po tem, ali model statistično izboljša zaznavanje sumljivih transakcij, temveč po tem, ali je mogoče v več-akterjskem okolju (več bank + koordinator) vzpostaviti in vzdrževati **operativni režim varnega obratovanja**, ki ohranja obvladljivo raven tveganj za zasebnost in varnost.

Ključna razlika je prav v tem, da se del tveganj iz klasičnega "deljenja podatkov" premakne v področje **"deljenja učenja"** – tj. v področje varnosti modelov, protokolov izmenjave in obratovalnih kontrol – kar zahteva, da se poleg tehničnih kontrol uvedejo tudi jasni organizacijski dogovori.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Organizacijska dimenzija se v praksi kaže predvsem v tem, da morajo sodelujoči subjekti vnaprej dogovoriti **vloge in odgovornosti** v celotnem življenjskem ciklu (kdo je odgovoren za konfiguracijo modela, varnost prenosov, hranjenje modelov, zaznavo anomalij, odziv na incidente ter odločitev, kdaj se model umakne ali zamenja).

Podobno ključna je ureditev **dostopov in pooblastil** (kdo lahko sproži učenje, kdo dostopa do modelnih paketov, kdo izvaja analize modela, kdo ga lahko uvede v produkcijo in kdo ima pooblastilo za izredni izklop). Ker se federativni protokol praviloma izvaja iterativno, je nujen tudi formaliziran **proces upravljanja sprememb**: sprememba hiperparametrov, standardizacijskih pravil ali učnega protokola lahko vpliva na kakovost rezultatov in hkrati spremeni varnostna tveganja, zato mora biti sprememba kontrolirana, testirana in odobrena.

Nadalje mora obstajati usklajen **incidentni režim in koordinacija**, saj v več-organizacijskem okolju ni dovolj, da incident rešuje vsak udeleženec "po svoje": že vnaprej morajo biti dogovorjeni postopki za ravnanje ob sumu zlorabe (npr. kompromitiran modelni paket, sum model inversion ali sum zastrupljanja posodobitev), načini obveščanja drugih udeležencev, izolacija problema in forenzična obravnava.

Nenazadnje pa je treba zagotoviti tudi **dokazljivost in revizijsko sled**, ki je v AML kontekstu posebej pomembna: razvidno mora biti, kateri model je bil poslan komu, kdaj in s katero konfiguracijo, kako je bil validiran, ter kako je potekala odločitev o njegovi uporabi. To ni zgolj tehnično beleženje (logiranje), temveč organizacijski dogovor o tem, kaj se beleži, kdo ima dostop do teh evidenc in kako dolgo se hranijo.

Vse navedeno je "varnostno" zato, ker ne zadeva le zaščite infrastrukture, temveč zaščito celotnega protokola federativnega učenja, kjer modeli in protokoli izmenjave postanejo nova napadalna površina; in je hkrati "organizacijsko" zato, ker te varovalke ne nastanejo samodejno, temveč šele z jasno določenimi pravili, odgovornostmi in postopki med več sodelujočimi institucijami.

Pravno-organizacijska dimenzija: vloge in odgovornosti

Federativno učenje v AML okolju ni zgolj tehnični protokol, temveč model sodelovanja, v katerem se zaradi razpršenega učenja in centralne koordinacije hitro odpre vprašanje: **kdo je odgovoren za katero fazo obdelave in na kateri pravni podlagi**.

V tem use-caseu se obdelujejo transakcijski podatki, KYC podatki in drugi podatki, ki pogosto vključujejo osebne podatke, zato je vprašanje vlog in odgovornosti ("kdo je upravljavec, kdo je obdelovalec, kdo je zgolj ponudnik tehnologije") nosilno za celotno skladnost. Posebnost federativnega scenarija je, da se klasični občutek "upravljavca" (tisti, ki ima podatke) in "obdelovalca" (tisti, ki podatke obdeluje) v praksi razsloji: podatki ostanejo lokalno pri bankah, vendar se sistem kot celota koordinira prek skupnega modela in protokola izmenjave.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

V poročilu je zato jasno izhodišče, da banke v tem kontekstu nastopajo kot **upravljavci osebnih podatkov (controllers)**. Banke namreč določajo namen obdelave (AML nadzor, zaznavanje sumljivih transakcij) in ključna sredstva obdelave v svojem okolju (kateri podatki se uporabijo, kako se podatki standardizirajo in kako se rezultati uporabijo v AML procesu).

To izhodišče je pomembno zato, ker banke tudi v federativnem scenariju ne "odložijo" svoje odgovornosti: dejstvo, da podatki ne zapustijo njihove infrastrukture, ne pomeni, da se odgovornost razprši ali prenese, temveč predvsem pomeni, da ostaja odgovornost bank za zakonitost in varnost lokalne obdelave v celoti aktivna.

Vloga Finterai je v takem ekosistemu bolj subtilna in jo je treba opredeliti funkcionalno.

Poročilo izpostavi, da Finterai **praviloma ni upravljavec** osebnih podatkov bank, lahko pa glede na konkretne funkcije in pogodbeno ureditev nastopa kot **obdelovalec (processor)**. Razlog za to razlikovanje je preprost: v federativni arhitekturi

Finterai tipično ne prejema surovih transakcijskih podatkov, vendar lahko izvaja določene aktivnosti, ki so del obdelave v širšem smislu – npr. koordinacijo učenja, distribucijo modela, vzdrževanje infrastrukture za modele ali izvajanje testov nad modeli. Če so te aktivnosti izvedene **"v imenu bank"** in v skladu z njihovimi navodili, gre praviloma za procesorsko vlogo.

Če pa bi Finterai samostojno določal namene ali bistvena sredstva obdelave (npr. neodvisno odločal, katere podatkovne kategorije so potrebne, ali model uporabljal za lastne namene), bi se razmerje lahko približalo upravljavski vlogi — in ravno zato poročilo poudari, da mora biti razmejitev vlog izrecna in dokazljiva.

Iz tega sledi ključna dobra praksa na ravni organizacije: v federativnem scenariju ni dovolj, da se v pogodbi navede "podatki se ne delijo", temveč je treba natančno opredeliti:

1. katere funkcije izvaja Finterai,
2. ali te funkcije pomenijo obdelavo osebnih podatkov (tudi posredno prek modelnih paketov),
3. katere varnostne ukrepe mora zagotoviti posamezen akter, in
4. kako se zagotovi dokazljivost.

Ker se v takih sistemih del tveganj premakne v področje varnosti modelov in protokolov izmenjave, mora pogodbeno-organizacijski okvir praviloma zajeti vsaj: politike dostopa do modelnih paketov, varnost prenosa in hrambe modelov, odgovornost za varnostne incidente, procese posodabljanja in sprememb (change management) ter režim revizije. V praksi to pomeni, da se "pravna razmejitev" in "tehnična razmejitev" ne smeta razhajati: če Finterai npr. hrani modele, izvaja preverjanja ali distribuira model med bankami, mora biti jasno, ali to počne kot procesor in pod kakšnimi navodili, ter kakšne varnostne in organizacijske obveznosti iz tega izhajajo.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Nazadnje je treba poudariti še eno značilnost federativnega ekosistema: več bank je lahko vključenih v skupni proces, vendar to samo po sebi še ne pomeni, da so banke nujno skupni upravljavci v smislu skupnega določanja namenov in sredstev obdelave.

Vloga vsake banke kot upravljavca izhaja iz njenega lastnega AML procesa in lastnega odločanja o uporabi rezultatov. Vprašanje, ali sodelujoče banke skupaj določajo bistvene elemente protokola in namena (in bi zato postale skupni upravljavci), je stvar konkretne ureditve in governance modela federacije. Ravno zato je v poročilu poudarek na tem, da se vloge ne domnevajo, temveč se izrecno opredelijo – in da se skladnost v takem ekosistemu gradi predvsem z jasnim dogovorom o odgovornostih, ukrepih in mejah delovanja vsakega akterja.

Minimizacija podatkov: usklajevanje AML zahtev in GDPR načela minimizacije

V obravnavanem primeru poročilo izpostavi **praktično kolizijo zahtev**, ki je za federativno učenje v AML okolju skoraj neizogibna: AML obveznosti bank pogosto zahtevajo zbiranje in obdelavo širšega nabora podatkov, pri čemer se obseg in vrsta podatkov lahko razlikujeta glede na tveganostni profil stranke (npr. stopnja skrbnega preverjanja). To oteži standardizacijo podatkovnih kategorij med bankami, saj vsi udeleženci ne razpolagajo nujno z enakimi podatkovnimi polji ali pa jih ne zbirajo v enakem obsegu in ob istem času. Posledično standardizacija za potrebe federativnega učenja ni zgolj tehnična naloga, temveč mora upoštevati tudi razlike, ki izhajajo iz regulativnih obveznosti in tveganostnih ocen posameznih bank.

V takem okviru poročilo usmerja k zasnovi, ki omogoča **časovno minimizacijo** oziroma načelo "zbiraj, ko je potrebno": sistem naj bo oblikovan tako, da banke lahko z zbiranjem ali vključitvijo določenih podatkov počakajo do trenutka, ko je to dejansko potrebno za konkretni namen obdelave, namesto da bi se podatki zajemali ali vključili "za vsak slučaj" že vnaprej. Takšen pristop je pomemben zato, ker zmanjšuje nepotrebno izpostavljenost osebnih podatkov v fazah, ko ti podatki še niso upravičeno potrebni, hkrati pa bankam omogoča izpolnjevanje AML zahtev, ko se tveganje oziroma potreba po dodatni skrbnosti dejansko materializira.

Poročilo hkrati poudari klasično vsebino načela minimizacije: uporabljeni podatki morajo biti **ustrezni, relevantni in omejeni na nujno potrebno**, dodatno pa je treba zagotoviti tudi, da se podatki **izbrišejo ali anonimizirajo**, ko niso več potrebni za namen obdelave. V kontekstu federativnega učenja to pomeni, da minimizacija ni le vprašanje, katere podatke model potrebuje, temveč tudi vprašanje, kako so podatki obdelani v standardizacijskih cevovodih, kako dolgo se hranijo, kdo ima dostop in kako se zagotovi, da se odvečni podatki ne ohranjajo v sistemu zaradi tehničnih ali organizacijskih navad.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Čeprav ponudnik tehnologije (npr. Finterai) v takem scenariju praviloma ni upravljavec osebnih podatkov, poročilo jasno nakaže še eno pomembno implikacijo: tehnologija mora biti zasnovana tako, da upravljavcem (bankam) omogoča **izvajanje načela minimizacije v praksi**. To vključuje možnost konfiguracije, selekcije podatkovnih polj, omejitve vključevanja podatkov v učne procese ter podporo pravilom hrambe in brisanja. Če sistem takšnih mehanizmov ne omogoča, banke svojih obveznosti minimizacije ne morejo učinkovito izpolnjevati, kar lahko pomeni, da uporaba rešitve kot take ne bi bila zakonita oziroma skladna z varstvom osebnih podatkov.

Varnostni vidiki: napadi na modele (npr. model inversion) in robustnost

Federativno učenje v tem use-caseu zmanjšuje potrebo po deljenju surovih podatkov med bankami, vendar s tem ne odpravi varnostnih tveganj; nasprotno, del tveganj se premakne iz področja "varovanja podatkovnih zbirk" v področje **varovanja modelov, modelnih posodobitev in protokolov izmenjave**. Poročilo zato izrecno izpostavi, da je federativno učenje lahko ranljivo za napade, ki ogrožajo **robustnost modela** ali **zasebnost posameznikov**, katerih podatki se uporabljajo pri lokalnem učenju v bankah.

Ključna implikacija je, da je treba federativno arhitekturo obravnavati kot varnostno kritično tudi v primerih, ko surovi podatki formalno ne zapuščajo bank.

Poročilo poudari, da se napadi lahko pojavijo v **dveh fazah**.

Prvič, v fazi treniranja, ko se modeli ali "učni paketi" izmenjujejo med udeleženci (ali prek centralne komponente): v tej fazi lahko napadalec poskuša vplivati na učenje, zastrupiti posodobitve ali pridobiti vpogled v modelne artefakte.

Drugič, v fazi uporabe (inference), ko banke uporabljajo natreniran model za zaznavanje sumljivih transakcij: tudi tu je model lahko tarča napadov, saj dostop do modela in njegovih izhodov lahko omogoča sklepanje o podatkih, na katerih je bil model treniran, ali manipulacijo obnašanja modela.

Takšna dvostopenjska ranljivost je pomembna za zasnovo kontrol, ker pomeni, da varnostne zahteve ne smejo veljati le v fazi "razvoja", temveč morajo pokrivati celoten življenjski cikel obratovanja.

Kot konkreten primer poročilo izpostavi **napad tipa »model inversion«**, pri katerem napadalec z analizo modela poskuša rekonstruirati oziroma izluščiti informacije o podatkih, na katerih je bil model treniran. Pri tem je posebej pomembna naučena lekcija, ki jo je smiselno neposredno prevzeti v dobre prakse: tudi če se predpostavlja, da model *ne vsebuje* osebnih podatkov v dobesednem smislu (ker so v modelu "samo uteži"), je pogosto varneje, da se z modelom ravna, **kot da vsebuje osebne podatke**, saj obstaja možnost, da napadalec z dovolj znanja, dostopa in tehnikami sklepanja iz modela izpelje informacije o učnih podatkih. Takšna predpostavka je operativno pomembna, ker vpliva na režim zaščite

modelov: kdo sme dostopati do modelov, kako se modeli prenašajo, kako se hranijo, kako se izvaja revizijska sled, in kako se obravnavajo incidenti.

Poročilo nadalje opozori, da federativno učenje v praksi pogosto pomeni obsežno uporabo **oblačnih storitev** in z njimi povezane tehnične odvisnosti. To ima dve neposredni posledici.

1. Prvič, izvajanje rešitve zahteva ustrezne **kompetence informacijske varnosti** (npr. upravljanje identitet in dostopov, šifriranje, segmentacija, nadzor nad prenosom in hrambo modelov, varnostno spremljanje).
2. Drugič, večja je potreba po sistematičnem obvladovanju **dobavne verige** in tehničnih odvisnosti: ko del rešitve temelji na zunanji infrastrukturi, knjižnicah ali storitvah, mora organizacija upravljati tveganja, ki izhajajo iz ranljivosti, posodobitev, konfiguracijskih napak in pogodbenih razmerij.

V federativnem scenariju to ni obrobno vprašanje, ker varnostni incident v eni komponenti (npr. v delu za izmenjavo modelov) lahko vpliva na zaupanje in varnost celotnega sodelovanja med bankami.

Skupni zaključek je zato jasen: federativno učenje je lahko pomemben korak k zmanjšanju potrebe po deljenju podatkov, vendar varnostne zahteve ne postanejo manjše – postanejo drugačne. Namesto da je glavno tveganje "uhajanje surovih podatkov", postanejo osrednja tveganja povezana z **varnostjo modelov**, z **integriteto protokola učenja**, z možnostjo napadov na zasebnost prek modela ter z zrelostjo upravljanja infrastrukturnih odvisnosti.

Preliminarna navezava na Akt o UI (tveganost)

Obravnavani primer je umeščen v kontekst preprečevanja pranja denarja (AML), pri čemer takšna uporaba kot taka ni izrecno navedena med primeri uporabe v Prilogi III (Annex III) Akta o UI. Kljub temu je primer operativno zelo relevanten za razumevanje tveganostnih profilov v finančnem sektorju, saj Akt o UI med "visoko-tvegane" primere v Prilogi III zajema vsaj nekatere odločitvene sisteme v financah (npr. sisteme za ocenjevanje kreditne sposobnosti oziroma dostop do bistvenih zasebnih storitev), pri čemer so določene izjeme opredeljene v samem besedilu Priloge III.

Finterai je zato v tem poročilu uporaben predvsem kot **referenčna arhitektura**: prikazuje, kako se pri občutljivih finančnih podatkih lahko zasnuje učenje modelov brez centralnega deljenja podatkov, hkrati pa zelo jasno razkrije, da se tveganja premaknejo v smer **varnosti modelov, protokola izmenjave in dobavne verige**. Posledično je primer posebej primeren kot izhodišče za poglavja o (i) varovanju podatkov in tehničnih odvisnosti v dobavni verigi, (ii) minimizaciji osebnih podatkov ob regulatorno zahtevni obdelavi ter (iii) modelno-varnostnih tveganjih (npr. napadi na modele), ki so prenosljiva tudi na scenarije, ki bi bili v finančnih institucijah po Aktu o UI lahko kvalificirani kot "Visoko-tvegani".

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Ključne prenosljive dobre prakse (povzetek)

To podpoglavje povzema dobre prakse, ki izhajajo neposredno iz regulatorno vodenega peskovnika Datatilsynet in konkretnega primera Finterai (federativno učenje za AML). Namen je dvojni:

- prvič, bralcu omogočiti jasno povezavo med opisano tehnično zasnovo (federativno učenje + standardizacija podatkov + decentralizirana obdelava) in dejanskimi varovalkami, ki jih primer odpira;
- drugič, pokazati, kako se v federativnem okolju zahteve varstva osebnih podatkov in informacijske varnosti operacionalizirajo v praksi, pri čemer se tveganja premaknejo iz »deljenja podatkov« v področje »deljenja učenja« (modelov in posodobitev).

Zato so spodaj navedene prakse omejene na tiste, ki jih poročilo Datatilsynet v tem primeru izrecno obravnava:

- federativno učenje kot alternativni arhitekturni vzorec,
- razmejitev vlog in odgovornosti,
- načelo minimizacije podatkov v AML kontekstu,
- modelno-varnostna tveganja (npr. model inversion) in razlika med treningom ter uporabo, ter
- upravljanje dobavne verige in oblačnih odvisnosti.

1. **Federativno učenje kot arhitekturna alternativa centralizaciji podatkov**
   Jedrna dobra praksa primera je, da se federativno učenje uporabi kot arhitekturni vzorec za situacije, kjer bi bilo deljenje transakcijskih podatkov med organizacijami pravno ali varnostno težko izvedljivo. V tem modelu podatki po zasnovi ostanejo v bankah, izmenjuje pa se "učenje" (model ali modelne posodobitve). Praksa je prenosljiva zato, ker pokaže izvedbeno pot, kako lahko organizacije z visokimi zahtevami glede zaupnosti vseeno sodelujejo pri izboljševanju modelov – vendar ob jasnem razumevanju, da to ni "brez tveganja", temveč drugačna porazdelitev tveganj.

2. **Razmejitev vlog (upravljavec/obdelovalec) kot nosilni pogoj skladnosti**
   Poročilo kot ključno izpostavi, da banke v takem scenariju nastopajo kot upravljavci osebnih podatkov, medtem ko je vloga ponudnika tehnologije (Finterai) odvisna od konkretnih funkcij in pogodbenega okvira: praviloma ni upravljavec, lahko pa je obdelovalec. Iz tega izhaja prenosljiva dobra praksa: v federativnih ekosistemih je treba vloge opredeliti funkcionalno in dokazljivo (kdo določa namene, kdo določa sredstva obdelave, kdo izvaja katere faze protokola) ter to pretvoriti v pogodbeno-

tehnične obveznosti, ki upravljavcem omogočajo dejansko izvajanje skladnosti (dostopi, hramba, revizijske sledi, incidenti).

3. **Minimizacija osebnih podatkov kot zasnova** (ne le kot načelo)
   Primer izpostavi, da AML obveznosti bank pogosto zahtevajo širši nabor podatkov, pri čemer se obseg razlikuje glede na tveganostni profil stranke; to lahko oteži enotno standardizacijo podatkov med bankami. Kot dobra praksa se v tem okviru izpostavi zasnova, ki omogoča **časovno minimizacijo**: z zbiranjem ali vključevanjem določenih podatkov se čaka do trenutka, ko je to dejansko potrebno za konkreten namen obdelave, namesto da bi se podatki vključili "za vsak slučaj". Pri tem je pomembna tudi druga izvedbena lekcija: standardizacija kategorij se izvaja le tam in v takem obsegu, kjer je to upravičeno z namenom, sicer lahko standardizacija sama postane vektor za nepotrebno širitev obdelave.

4. **Modelna varnost: tveganja napadov na modele in ločena presoja treninga ter uporabe**
   Poročilo izrecno obravnava, da federativno učenje odpira specifična modelno-varnostna tveganja: napadi se lahko pojavijo tako v fazi učenja (trening) kot v fazi uporabe (inference). Kot konkreten primer je izpostavljen napad tipa model inversion, kjer napadalec z analizo modela poskuša izluščiti informacije o učnih podatkih. Prenosljiva dobra praksa je zato dvojna:

   (i) z modelnimi artefakti in posodobitvami je smiselno ravnati, kot da predstavljajo občutljiv objekt (tudi če ne vsebujejo surovih podatkov), ter

   (ii) varnostne kontrole je treba načrtovati ločeno za fazo učenja in fazo produkcijske uporabe, saj se napadalna površina in vektorji razlikujejo.

5. **Dobavna veriga in oblačne odvisnosti: varnost kot pogoj izvedljivosti**
   Poročilo opozori, da takšni ML sistemi v praksi pogosto vključujejo oblačne storitve, specializirano infrastrukturo in več tehničnih odvisnosti. Zato je dobra praksa, da se varnost ne obravnava kot dodatna plast, temveč kot pogoj izvedljivosti: organizacije morajo imeti ustrezne kompetence informacijske varnosti, vzpostavljen nadzor nad konfiguracijami, upravljanje identitet in dostopov, ter sistematično upravljanje dobavne verige (knjižnice, orodja, platforme, storitve). V federativnem okolju je to posebej pomembno, ker incident v eni ključni komponenti (npr. v izmenjavi modelov) lahko vpliva na zaupanje in varnost celotnega sodelovanja.

Viri (primer 3.2)

1. Datatilsynet (2022). Maskinlæring uten datadeling: Bruk av føderert læring for antihvitvasking – sluttrapport fra sandkasseprosjektet med Finterai (PDF).

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

https://www.datatilsynet.no/contentassets/d837496134c8421590a3d2b69b1e4440/sluttrapport-til-publisering-og-oversettelse2.pdf
2. Datatilsynet (n.d.). Finterai: Machine learning without data sharing (spletna stran, uradni povzetek projekta). https://www.datatilsynet.no/en/regulations-and-tools/reports-on-specific-subjects/reports/finterai-machine-learning-without-data-sharing/
3. Datatilsynet (n.d.). Finterai: Maskinlæring uten datadeling (spletna stran, norveška različica povzetka projekta). https://www.datatilsynet.no/regelverk-og-verktoy/rapporter-og-utredninger/rapporter-fra-sandkassa/finterai-maskinlaring-uten-datadeling/
4. Datatilsynet (n.d.). Reports from the sandbox projects (zbirna stran poročil iz regulativnega peskovnika). https://www.datatilsynet.no/en/regulations-and-tools/reports-on-specific-subjects/reports/
5. Datatilsynet (n.d.). Rapporter og funn fra sandkassa (zbirna stran poročil iz regulativnega peskovnika; norveška različica). https://www.datatilsynet.no/regelverk-og-verktoy/rapporter-og-utredninger/rapporter-fra-sandkassa/
6. Datatilsynet (n.d.). Personvernfremmende teknologi (PET) – Erfaringer fra sandkassen (spletna stran; kontekstualna predstavitev PET pristopov, vključno z izkušnjami iz projekta federativnega učenja za AML). https://www.datatilsynet.no/personvern-pa-ulike-omrader/internett-og-apper/personvernfremmende-teknologi/erfaringer-fra-sandkassen/
7. Datatilsynet (2022). Webinar: Kunsten å lære av data du ikke har (spletna stran + gradivo o webinarju). https://www.datatilsynet.no/aktuelt/aktuelle-nyheter-2022/webinar-om-foderert-laring/
8. Datatilsynet (n.d.). Personvernpodden – SandKasten #6: "Hvitsnippkrim vs. vidunderbarn" (podcast epizoda; predstavitev projekta in ključnih vprašanj). https://www.datatilsynet.no/regelverk-og-verktoy/personvernpodden/sandkasten/6-sandkasten-hvitsnippkrim-vs.-vidunderbarn/
9. Finterai (n.d.). Optimize AML systems (spletna stran; opis podjetja in njihovega pristopa – ne nadomešča uradnih virov Datatilsynet). https://www.finterai.com/

----------do tu2----------

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

3.3 Francoska javna uprava (DINUM/Etalab) – Albert API: platforma za GenAI v javnem sektorju

V francoskem javnem sektorju se je pri uvajanju generativne umetne inteligence hitro pokazal tipičen sistemski izziv: posamezne administracije potrebujejo dostop do naprednih modelov in z njimi povezanih storitev, vendar je ad hoc uporaba zunanjih ponudnikov ali razdrobljena vzpostavitev lastnih rešitev povezana z večjimi tveganji (zaupnost podatkov, neenotni varnostni standardi, podvajanje stroškov in kompetenčnih naporov) ter otežuje prehod od pilotov k stabilni produkcijski rabi. Na tej podlagi je bil kot cilj postavljen razvoj skupne državne platforme, ki bi generativno UI omogočila v upravljanem, standardiziranem in suverenem okolju, s čimer bi administracijam olajšala razvoj konkretnih aplikacij in hkrati poenotila osnovne varnostne ter organizacijske kontrole.

V tem kontekstu je **DINUM** (*Direction interministérielle du numérique*) – medresorski organ francoske države za usmerjanje in podporo digitalni preobrazbi javne uprave – skupaj z notranjimi deležniki (vključno z **Etalab**, ki je v javno dostopnih predstavitvah naveden kot izvajalska enota pri razvoju) vzpostavil storitev **Albert API**. Albert API je umeščen v programski okvir **ALLiaNCE**, v okviru katerega se razvijajo in upravljajo medresorske (skupne) komponente umetne inteligence za uporabo v javnem sektorju. Rezultat je konkretna platforma/API, ki centralizira dostop do generativnih modelov in spremljevalnih funkcij ter omogoča, da se posamezne administracije osredotočijo na vsebinske use-case aplikacije, medtem ko je infrastruktura in osnovni režim upravljanja zagotovljen na ravni skupnega okolja.

V operativni praksi Albert API naslavlja zelo konkreten problem javnega sektorja: administracije, ki uporabljajo zunanje komercialne API-je za generativno UI, prevzemajo tveganja razkritja občutljivih informacij in tehnološke odvisnosti; hkrati pa vsaka ekipa, ki poskuša vzpostaviti lasten "AI backend" (izbor in gostovanje modelov, GPU infrastruktura, skaliranje, nadzor dostopov), praviloma podvaja stroške in razvoj. Albert API je zato zasnovan kot skupna infrastruktura oziroma "odobreno okolje" na ravni države: centralizira in varuje uporabo generativnih modelov ter omogoča, da se posamezne administracije osredotočijo na razvoj konkretnih aplikacij, medtem ko so ključne tehnične in organizacijske kontrole vzpostavljene na ravni platforme. V javnih opisih je platforma predstavljena kot infrastruktura, ki podpira celoten cikel od eksperimentiranja do produkcijskega obratovanja.

Kdo so uporabniki in komu je rešitev namenjena

Primarni uporabniki Albert API so tehnične in produktne ekipe v javni upravi, ki gradijo ali vzdržujejo digitalne storitve države. Platforma je v tem smislu standardizirana vstopna točka (backend): administracije jo integrirajo v svoje aplikacije prek API-ja, namesto da bi vsaka institucija samostojno vzpostavljala infrastrukturo za poganjanje modelov. Albert API je

obenem zasnovan tako, da znižuje prag vstopa: javno je navedeno, da lahko agenti državne javne uprave zaprosijo za dostopni ključ in s tem izvajajo eksperimente ter gradijo rešitve v lastni kodi.

Cilj rešitve

Cilj rešitve je zagotoviti varno, suvereno in skupno okolje za uporabo generativne UI v javnem sektorju, ki je dovolj enostavno za integracijo in hkrati dovolj robustno za produkcijsko rabo. V javnih predstavitvah je izrecno navedeno, da Albert API ponuja standardiziran vmesnik, združljiv s konvencijami OpenAI, in omogoča dostop do odprtokodnih modelov (npr. Llama, Mistral) ter do naprednih storitev, kot so RAG, OCR, klasifikacija in vektorizacija dokumentov.

To administracijam omogoča gradnjo aplikacij, ki se opirajo na interne dokumente in baze znanja, brez potrebe, da bi vsaka ekipa zase postavljala celoten podporni ekosistem.

Kot indikator zrelosti rešitve se v javno dostopni predstavitvi navaja tudi, da je Albert API že uporabljen v več kot 70 javnih projektih in obdeluje več kot 100.000 zahtevkov tedensko, kar kaže na to, da gre za platformo, zasnovano za dejansko obratovanje in širšo rabo.

Varnostna arhitektura in podatkovna suverenost

Albert API je zasnovan kot skupna (medresorska) platforma, pri kateri se varnost in podatkovna suverenost uveljavita že na ravni infrastrukture in načina obratovanja. Jedrni varnostni režim je opisan z dvema operativnima praviloma, ki sta posebej pomembni za javni sektor:

(i) platforma ne hrani nobene sledi vsebine pogovorov, poslanih modelom, in
(ii) platforma ne pošilja nobenih uporabniških podatkov na internet.

Hkrati je platforma umeščena v suvereno oblačno okolje s certifikacijo SecNumCloud pri gostitelju Outscale, kar pomeni, da je varnostni profil zasnovan tudi na infrastrukturni ravni (lokacija obdelave, režim gostovanja in varnostne zahteve okolja).

Z vidika arhitekture je Albert API predstavljen kot enotna vstopna točka (API Gateway) za dostop do različnih modelov in povezanih storitev: administracije svoje aplikacije (digitalne produkte) povežejo na Albert API prek standardiziranega vmesnika, združljivega s konvencijami OpenAI, kar omogoča hitrejšo integracijo in lažjo zamenjavo zunanjih API-jev s suvereno rešitvijo. V javnem opisu ALLiaNCE je dodatno navedeno, da je Albert API instanca odprtokodnega orodja OpenGateLLM, razvitega v okviru francoske države, kar potrjuje platformni vzorec “gateway + katalog modelov + upravljanje uporabe”.

Na ravni funkcionalne arhitekture platforma ne ponuja le “golega” klepeta z LLM, temveč nabor storitev, ki so tipične za produkcijske javne use-case scenarije. Med ključnimi funkcionalnostmi so izrecno navedeni:

(i) dostop do generativnih modelov,
(ii) RAG na zahtevo (retrieval augmented generation) in
(iii) upravljanje projektov ter spremljanje uporabe (governance/operativna kontrola porabe).

Poleg tega so na isti strani opisane še spremljevalne komponente, ki tvorijo tipičen "RAG ekosistem": embeddings (vektorizacija), rerank (ponovno rangiranje zadetkov), OCR ter storitev "RAG on demand", ki povezuje generativni model z administrativnimi dokumentarnimi bazami. To je za varnost in kakovost pomembno, ker se pri takšni zasnovi znaten del domenskega "znanja" seli v nadzorovane podatkovne vire (baze dokumentov), medtem ko model ostaja splošnejši, uporaba pa je organizacijsko in tehnično omejena na platformni režim.

Podatkovna suverenost je dodatno podprta z načinom, kako Albert API obravnava modele: platforma javno navaja, da modeli izhajajo iz odprtokodnega ekosistema (npr. modeli, objavljeni s strani tretjih na Hugging Face), vendar so gostovani na strežnikih Albert API, pri čemer "nobeni podatki" uporabnikov niso posredovani ponudnikom modelov. Hkrati je izrecno navedeno, da se portfelj modelov redno posodablja in da se nabor modelov spreminja, zato je pri prenosu dobrih praks ključno, da se organizacije opirajo na kontrole platforme (način gostovanja, režim obdelave, standardiziran dostop, upravljanje projektov in uporabe), ne pa na statični seznam modelov.

Platforma nudi dostop do 'open-weights' modelov, gostovanih na SecNumCloud infrastrukturi, ter da se seznam modelov in licenc lahko spreminja. To je pomembno, ker pomeni, da se morajo organizacije pri prenosu dobrih praks osredotočiti na *kontrole* (okolje, governance), ne na statični seznam modelov.

Albert API je zasnovan kot skupna (medresorska) platforma, pri kateri se varnost in podatkovna suverenost uveljavita že na ravni infrastrukture in načina obratovanja. Jedrni varnostni režim je opisan z dvema operativnima praviloma, ki sta posebej pomembni za javni sektor: platforma **ne hrani nobene sledi vsebine pogovorov**, poslanih modelom, in platforma **ne pošilja nobenih uporabniških podatkov na internet**. S tem je tveganje nenamernega "odtoka" vsebin iz uporabe generativne UI zasnovno omejeno, kar je pri javnih storitvah ključno zaradi pogoste obdelave notranjih (tudi občutljivih) vsebin. [ALB-1]

Podatkovna suverenost je dodatno podprta z izbiro gostovanja: modeli in API delujejo v **suverenem oblačnem okolju**, pri katerem je izrecno navedena **certifikacija SecNumCloud** gostitelja **Outscale**. Takšna umestitev ima operativni pomen, ker platforma ne gradi zgolj na

“aplikacijskih” pravilih (npr. ne-shranjevanje pogovorov), temveč tudi na infrastrukturnem režimu, ki je predmet strožjih varnostnih zahtev in nadzora. [ALB-1]

Pomemben element suverenosti je tudi način, kako Albert API obravnava modele. Platforma navaja, da omogoča dostop do širokega nabora **odprtokodnih temeljnih (foundation) generativnih modelov**, ki izvirajo od tretjih ponudnikov (npr. Mistral, Meta) in so javno objavljeni na platformah, kot je Hugging Face, vendar so **gostovani na strežnikih Albert API**; ob tem je izrecno zapisano, da se **nobeni uporabniški podatki ne pošiljajo tem ponudnikom modelov**. To je za javno upravo pomembno zato, ker omogoča uporabo modelov iz odprtokodnega ekosistema brez klasičnega modela “podatki gredo k ponudniku API-ja”. [ALB-2]

Hkrati dokumentacija opozori, da je portfelj modelov **dinamičen**: modeli se redno posodabljajo in nabor razpoložljivih modelov se lahko spreminja. V kontekstu dobrih praks to pomeni, da se je smiselno opreti predvsem na **kontrole platforme** (suvereno gostovanje, pravila glede hrambe in iznosa podatkov, standardiziran vstopni vmesnik, upravljanje uporabe), ne pa na statični seznam modelov, ki je po naravi podvržen spremembam. [ALB-2]

Osnovni arhitekturni vzorec platforme (kar je razvidno iz javne dokumentacije)

Albert API je v javnih opisih predstavljen kot **enotna vstopna točka (platforma/API)**, namenjena integraciji v digitalne produkte administracij. Platforma izrecno navaja, da ponuja več ključnih funkcionalnosti: **dostop do generativnih modelov**, **RAG “na zahtevo”** ter **upravljanje projektov in spremljanje uporabe** (governance/operativna kontrola porabe). [ALB-3]

Za integracijo je posebej pomembno, da Albert API uporablja **standard OpenAI** (»conventions OpenAI«) in s tem omogoča lažjo uporabo z obstoječimi orodji v ekosistemu; v dokumentaciji sta kot primera eksplicitno navedena tudi OpenWebUI in LangChain. To zmanjšuje strošek integracije in omogoča, da se administracije osredotočijo na vsebinsko aplikacijo, ne na prilagajanje infrastrukturnega vmesnika. [ALB-3]

Platforma poleg “chat” vmesnika ponuja tudi podporne gradnike, ki so tipični za produkcijske dokumentno-usmerjene primere rabe v javnem sektorju: **embeddings** (vektorizacija besedil), **OCR**, ter **rerank** (ponovno razvrščanje rezultatov za povečanje relevantnosti pri dokumentnem iskanju in s tem večjo zanesljivost RAG odgovorov). Ti gradniki so predstavljeni kot del “orodjarne” generativne UI, ki olajša gradnjo rešitev, vezanih na interne dokumentne korpuse (npr. zakoni, okrožnice, poročila, baze znanja). [ALB-3]

Pomembna omejitev: javna dokumentacija na teh straneh jasno navede, da obstaja storitev »RAG à la demande« in da platforma ponuja embeddings/ocr/rerank, vendar **ne opiše do te mere**, da bi bilo mogoče z gotovostjo povzeti notranjo topologijo (npr. ali platforma upravlja tudi celotno vektorsko bazo za uporabnika ali zgolj zagotovi gradnike in vmesnike). Zato je pri prenosu dobrih praks priporočljivo ostati pri tem, kar je izrecno navedeno: platforma nudi RAG kot funkcionalnost in nabor podpornih storitev, podrobnosti izvedbe pa je treba

preveriti v dodatni tehnični dokumentaciji (API reference) ali v internih dokumentih posamezne administracije.

Homologacija (odobritev) in varnostni nadzor države

V francoskem sistemu državne informacijske varnosti "homologacija" pomeni formalno odločitev pristojne avtoritete, da je informacijski sistem glede na identificirana tveganja in uvedene ukrepe primeren za obratovanje, pri čemer avtoriteta sprejme preostala (rezidualna) tveganja in določi pogoje ter trajanje veljavnosti odobritve. Takšna homologacija je običajno časovno omejena in je povezana z načrtom nadaljnjih varnostnih izboljšav (npr. akcijskim načrtom).

Za Albert API je bila izdana formalna "Décision d'homologation" (odločba o homologaciji) z datumom 14. 9. 2025, v kateri je kot predmet odločbe navedena homologacija informacijskega sistema Albert API. Odločba med razlogi izrecno navaja, da so bile odpravljene vse ranljivosti, identificirane prek programa Bug Bounty, ter da rezultati izvedenih varnostnih auditov niso pokazali večjih ali kritičnih ranljivosti, ki bi omogočile kompromitacijo sistema ali podatkov, ki jih sistem obdeluje.

Homologacija je bila izdana po pregledu elementov, obravnavanih na komisiji 5. 6. 2025, in velja za obdobje enega leta, pri čemer je izrecno določena tudi "rezerva" – izvedba dodatnih ukrepov, predvidenih za leto 2025 v okviru plana ukrepov informacijske varnosti (plan d'action SSI). To je pomembno sporočilo za razumevanje "odobrenega okolja": homologacija ni enkratna izjava o varnosti, temveč okvir, ki vključuje tudi časovno zamejene obveznosti izboljšav in nadzor nad njihovim izvajanjem.

Odločba hkrati vsebuje ključno opozorilo glede meja odobritve. Če bi posamezna administracija želela prek Albert API obdelovati podatke posebej visoke občutljivosti (npr. podatke, ki bi jih v praksi šteli za "sensitive" v smislu GDPR), pa tak scenarij ni predviden v okviru obstoječe homologacije, mora pristojna avtoriteta znotraj uporabniške administracije izvesti lastno ponovno oceno varnostne ravni, pri čemer se opre na homologacijski dosje, ki ga DINUM da na voljo. To pomeni, da homologacija platforme sicer vzpostavi osnovni državni varnostni okvir, vendar ne "prenese" avtomatično varnostne presoje za vsako možno vrsto podatkov ali vsak specifičen namen uporabe.

Nazadnje odločba določa tudi režim sprememb: vsaka večja sprememba arhitekture, vsaka pomembna evolucija uporabe ali dodatek funkcionalnosti, ki bi lahko dodal ali povečal tveganja, mora biti predmet novega pregleda v okviru ponovno sklicane komisije. Ta element je z vidika dobrih praks še posebej relevanten, ker pokaže, da "odobreno okolje" v javni upravi ni statična tehnična postavitev, temveč sistem upravljanja sprememb, kjer je varnostno dovoljenje vezano na dejanski obseg uporabe in arhitekturne lastnosti v danem trenutku.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Funkcionalnosti z vidika governance (API, računi, telemetrija, RAG)

Dostop, namen uporabe in omejitve (kvote)

Albert API je opisan kot storitev, ki jo DINUM zagotavlja administracijam znotraj državnega informacijskega ekosistema za uporabo s strani njihovih zaposlenih (agents) ter tudi izvajalcev (service providers / prestataires), ki jih administracije povabijo v okviru izvajanja nalog. Storitev je dana v uporabo brez plačila in je v osnovi namenjena eksperimentiranju (experimentation), pri čemer je uporaba vezana na vnaprej določene omejitve (limits / limitations), ki morajo ustrezati predvidenemu načinu uporabe.

DINUM si hkrati pridržuje možnost, da omejitve določi in jih po presoji tudi spremeni, vključno z možnostjo povišanja, če to utemeljuje dejanska raba in potrebe administracije. Poleg standardnega režima je predviden tudi privilegiran dostop (privileged access / accès privilégié), ki se lahko uredi na podlagi posebne konvencije (convention) med administracijo in DINUM.

Identiteta uporabnika, računi in API ključi (sledljivost dostopa)

Upravljanje dostopa temelji na osebnem uporabniškem profilu: uporabnik je fizična oseba, ki deluje za administracijo (kot zaposleni ali izvajalec), registracija pa je individualna. Kot priporočena možnost avtentikacije je navedena uporaba ProConnect, alternativno pa je dovoljena registracija z lokalnim geslom (local password / mot de passe), pri čemer so podane zahteve glede uporabe močnega, neponovljenega gesla, namenjenega izključno tej storitvi.

Za integracije in uporabo v aplikacijah platforma omogoča generiranje API ključev (API keys / clés d'API) v uporabniškem prostoru. S tem je vzpostavljen standardni mehanizem sledljivosti in upravljanja: profil uporabnika → API ključ → klici API, kar je temelj za nadzor nad rabo, forenzično obravnavo incidentov in obvladovanje tveganj (npr. kompromitacija ključa).

Metrike uporabe in operativni nadzor (telemetrija)

V uporabniškem prostoru so na voljo merilniki uporabe, ki omogočajo operativno upravljanje storitve. Med izrecno navedenimi elementi so spremljanje klicev API, skupno število zahtevkov, poraba tokenov, prikaz ogljičnega odtisa/porabe (carbon consumption / consommation carbone) ter prikaz veljavnosti oziroma podatkov o API ključu. Poleg tega je omogočen tudi vpogled v omejitve po posameznem modelu (per-model limits / limites par modèle).

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Takšen nabor metrik je za javni sektor praktično pomemben, ker omogoča več nivojev nadzora:

(i) hitro zaznavanje nenavadnih vzorcev rabe (indikator zlorab ali napak v integraciji),
(ii) upravljanje porabe in obremenitev po modelih, ter
(iii) osnovo za dokazljivost in notranje evidence o tem, kako in v kolikšnem obsegu se storitev uporablja.

Dokumentacija ne opisuje podrobnosti notranje implementacije teh mehanizmov, vendar jasno določa, da je nadzor uporabe sestavni del platforme.

RAG kot vgrajena funkcionalnost (baza besedil in iskalni mehanizem)

Platforma vključuje funkcionalnost za izvajanje RAG (Retrieval-Augmented Generation): omogoča bazo za shranjevanje besedilnih podatkov (text database / base de données de textes) in pripadajoč iskalni mehanizem (search mechanism / moteur de recherche), ki sta namenjena temu, da se generativni model pri odgovorih lahko opira na dokumente oziroma besedila, ki jih zagotovi uporabnik.

Pri tej funkcionalnosti je iz governance vidika ključna jasna razmejitev zasebnosti med uporabniki in hrambe v platformi. Vneseni podatki niso deljeni z drugimi uporabniki, vendar so lahko v okviru funkcionalnosti za nalaganje vsebin in pripravo baze shranjeni znotraj sistema Albert API. Posledično je za administracije nujno, da že pred uporabo RAG funkcionalnosti določijo interne smernice za:

(i) dovoljene vrste in razrede dokumentov,
(ii) (klasifikacijo in označevanje vsebin,
(iii) roke hrambe ter
(iv) odgovornosti za zagotavljanje, da se v bazo ne nalagajo neustrezne ali pretirano občutljive vsebine.

Pravila o (ne)uporabi občutljivih podatkov in odgovornost uporabnika

Albert API je zasnovan kot skupna državna platforma, ki uporabnikom ponuja standardizirano in varno okolje za uporabo generativne UI, vendar dokumentacija hkrati jasno postavi mejo glede vrst podatkov, ki naj bi se prek storitve obdelovali. V pravilih uporabe je izrecno navedeno, da platforma ni namenjena prejemanju ali obdelavi podatkov, ki bi jih šteli za **občutljive osebne podatke (sensitive personal data / données sensibles)**, pri čemer so kot referenca navedeni predvsem podatki iz **GDPR člena 9** (posebne vrste osebnih podatkov) in **GDPR člena 10** (podatki v zvezi s kazenskimi obsodbami in prekrški).

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Takšna formulacija ima v praksi dve pomembni implikaciji:

- Prvič, administracijam in njihovim ekipam daje jasno izhodiščno pravilo: generativna uporaba prek Albert API je predvidena za primere, kjer se obdelujejo neobčutljive vsebine oziroma vsebine, ki so skladno s klasifikacijo in internimi pravili primerne za tak režim obdelave.
- Drugič, pravilo služi kot governance mehanizem, ki zmanjšuje tveganje "neopaznega drsenja" uporabe v scenarije z višjo občutljivostjo, kjer bi bile potrebne dodatne varnostne presoje, strožji tehnični ukrepi in pogosto tudi drugačna pravna utemeljitev.

Dokumentacija hkrati nedvoumno določa, da je uporabnik (tj. administracija in konkretna oseba, ki storitev uporablja) odgovoren za podatke, ki jih v storitev vnese ali prek nje obdeluje. To velja tako za podatke, poslane v pozivih (prompts), kot tudi za podatke, ki se v sistem naložijo prek funkcionalnosti vnosa datotek (file ingestion / ingestion de fichiers), na primer pri uporabi RAG.

V praksi to pomeni, da platforma ne prevzema vloge "filtra" ali "odobritve" vsebine, temveč zagotavlja infrastrukturno okolje in osnovne kontrole, medtem ko je odgovornost za ustreznost, zakonitost in klasifikacijo podatkov (kaj je dopustno uporabiti in pod kakšnimi pogoji) na strani uporabnika.

Za prenos dobrih praks je ta razporeditev odgovornosti posebej pomembna, ker pokaže tipičen model upravljanja v javnem sektorju: centralna platforma zagotovi varnostno osnovo in standardiziran režim uporabe, vendar skladnost uporabe v konkretnem administrativnem kontekstu ostane naloga posamezne administracije.

To vključuje predvsem:

(i) opredelitev dovoljenih podatkovnih kategorij in primerov rabe,
(ii) notranje usmeritve za klasifikacijo in ravnanje z dokumenti (še posebej pri RAG), ter
(iii) izvedbo lastnih ocen tveganj in skladnosti za konkretne aplikacije (npr. v okviru DPIA (Data Protection Impact Assessment / AIPD), kjer je to potrebno).

Preliminarna navezava na Akt o UI(tveganost)

Albert API je primarno v funkciji infrastrukture, ki omogoča storitve (platforma za dostop do modelov in povezanih storitev), zato se tveganost po Aktu o UI v praksi praviloma presoja na ravni konkretne aplikacije in njenega namena, ne na ravni same platforme kot take. Ključno vprašanje je, ali je posamezni AI sistem, ki je na platformi zgrajen ali prek nje uporabljen, visoko tvegan po pravilih iz člena 6 (tj. po poti Priloge I ali Priloge III).

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Ker Albert API sam po sebi ne predstavlja enega specifičnega, "namenskega" primera uporabe iz Priloge III, je pri številnih tipičnih upravnih uporabah generativne UI (npr. pomoč pri pisanju, povzemanje, priprava osnutkov) bolj realno, da bo relevantnejši predvsem režim transparentnosti in obveščanja za določene sisteme generativne UI, ne pa režim, ki velja za visoko tvegane sisteme.

Hkrati pa ista infrastruktura lahko podpira tudi aplikacije, ki bi se po svojem namenu lahko približale področjem iz Priloge III (npr. sistemi v zaposlovanju ali pri dostopu do bistvenih javnih storitev/koristi). Zato je dobra praksa, da se na ravni platforme standardizirajo vsaj osnovne varnostne in procesne kontrole (upravljanje dostopov, sledljivost, upravljanje sprememb), saj to olajša skladno uporabo tudi v zahtevnejših domenah, če in ko se na platformi razvijejo takšni primeri.

Ključne prenosljive dobre prakse (povzetek)

To podpoglavje povzema dobre prakse, ki izhajajo neposredno iz primera Albert API kot državne, medresorske platforme za generativno UI. Namen je dvojni: (i) bralcu omogočiti jasno povezavo med opisano platformo in kontrolami, ki jih platforma dejansko uvaja na ravni infrastrukture in uporabe, ter (ii) pokazati, kako se "approved environment" (environnement approuvé) v javni upravi operacionalizira prek kombinacije infrastrukturnih varnostnih garancij, formalne odobritve (homologation) in upravljanja uporabe (governance). Zato so spodaj navedene prakse omejene na tiste, ki so v javni dokumentaciji Albert API izrecno opisane (varnostna stran, odločba o homologaciji, pravila uporabe):

1. **Paradigma "approved environment"** (environnement approuvé)**: suvereno gostovanje + formalna homologacija + varnostni nadzor**
   Osrednja prenosljiva praksa je, da se generativna UI v javnem sektorju umesti v predhodno upravljano in formalno odobreno okolje, namesto da bi vsaka administracija vzpostavljala lasten "AI backend". Albert API pri tem kombinira:

   (i) suvereno gostovanje v okolju, ki je vezano na SecNumCloud certifikacijo gostitelja, ter

   (ii) (formalno odločbo o homologaciji (décision d'homologation), ki časovno omeji odobritev, opredeli pogoje in zahteva nadaljnje varnostne ukrepe (plan d'action SSI).

   V odločbi so kot elementi varnostnega nadzora izrecno navedeni tudi odpravljeni izsledki programa nagrajevanega prijavljanja ranljivosti (bug bounty) ter rezultati izvedenih varnostnih pregledov (auditov), pri katerih niso bile ugotovljene večje ali

kritične ranljivosti, ki bi omogočile kompromitacijo sistema ali obdelovanih podatkov.

2. **Jasno opredeljen varnostni "baseline": brez hrambe vsebin pogovorov in brez iznosa podatkov na internet**
   Albert API na ravni platforme postavi dve zelo konkretni in za javni sektor uporabni garanciji: platforma ne hrani vsebine pogovorov in ne pošilja uporabniških podatkov na internet. Takšen baseline (baseline guarantees) je prenosljiv zato, ker ga je mogoče neposredno prevesti v minimalne zahteve za "approved environment" v javnih institucijah: kjer je to mogoče, naj se obdelava izvaja v nadzorovanem okolju, brez "odtoka" podatkov v javno omrežje in brez nepotrebne hrambe vsebine interakcij.

3. **Upravljanje dostopa kot del governance: identiteta, API ključi in merilniki porabe**
   Platforma obravnava generativno UI kot storitev, ki zahteva sledljivost rabe: uvedena je identitetna plast (ProConnect (ProConnect) ali lokalno geslo (mot de passe)), možnost generiranja API ključev (clés d'API) ter uporabniški vpogled v metrike porabe (npr. število klicev, poraba tokenov, omejitve po modelih, prikaz "carbon consumption" (consommation carbone)). Ta kombinacija je prenosljiva dobra praksa, ker ustvarja operativne pogoje za nadzor uporabe: od zaznave anomalij in zlorab do stroškovnega in kapacitetnega upravljanja, brez katerega generativna UI v javnih okoljih hitro postane neobvladljiva.

4. **RAG kot vgrajena funkcionalnost – ob izolaciji podatkov med uporabniki in jasni odgovornosti za vsebino**

   Albert API omogoča RAG (RAG / retrieval augmented generation) prek baze za shranjevanje besedil in iskalnega mehanizma, pri čemer dokumentacija izrecno določa, da vneseni podatki niso deljeni z drugimi uporabniki, vendar so lahko (v okviru vnosa v sistem) shranjeni znotraj platforme. To je prenosljiva dobra praksa, ker združuje dve zahtevi javnega sektorja:

   (i) omogočiti dokumentno utemeljene odgovore na podlagi internih virov in
   (ii) hkrati ohraniti izolacijo med organizacijami/projekti.

   Neposredna posledica za dobre prakse je, da mora biti upravljanje baz znanja ("knowledge base governance")del uvedbe: klasifikacija dokumentov, dovoljen obseg vsebin in pravila hrambe ne morejo biti prepuščeni ad hoc uporabi.

5. **Eksplicitna omejitev glede občutljivih podatkov in obveznost ponovne varnostne presoje v posebnih primerih**
   Pravila uporabe izrecno določajo, da platforma **ni namenjena** prejemanju občutljivih osebnih podatkov (sensitive personal data / données sensibles), zlasti v smislu GDPR člena 9 in 10, ter da je uporabnik odgovoren za podatke, ki jih uporablja (vključno z vnosom datotek).

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Odločba o homologaciji dodatno poudari, da morajo administracije pri načrtovani obdelavi podatkov "posebej visoke občutljivosti" izvesti lastno ponovno oceno varnostne ravni na podlagi homologacijskega dosjeja; večje spremembe arhitekture ali uporabe pa zahtevajo nov pregled (nova komisija). Ta kombinacija jasno pokaže prenosljiv governance model: platforma zagotovi varnostno osnovo in odobreno okolje, odgovornost za konkretni scenarij uporabe (in za "dvig" varnostne ravni, kadar se uporaba spremeni) pa ostane pri administraciji.

Viri (primer 3.3)

1. Albert API (n.d.). Infrastructure sécurisée (uradna varnostna stran; "ne conserve aucune trace des conversations"; "n'envoie aucune de vos données sur internet"; SecNumCloud/Outscale). https://albert.sites.beta.gouv.fr/solutions/security/
2. DINUM (n.d.). Modalités d'utilisation – Albert API (PDF; dostop in omejitve/kvote; ProConnect; računi in API ključi; metrike porabe; RAG (ingestion + search); odgovornost uporabnika; omejitev glede občutljivih podatkov). https://albert.sites.beta.gouv.fr/documents/2/MU_albert_api.pdf
3. DINUM (2025). Décision d'homologation du système d'information Albert API (PDF; datum 14. 9. 2025; odpravljene ranljivosti iz programa nagrajevanega prijavljanja ranljivosti (bug bounty); rezultati auditov brez večjih/kritičnih ranljivosti; veljavnost 1 leto; plan d'action SSI; obveznost ponovne ocene pri obdelavah posebej občutljivih podatkov; nova komisija ob večjih spremembah). https://albert.sites.beta.gouv.fr/documents/3/20250828_DINUM_INT_AlbertAPI_Decision-Homologation_vDEF-1.pdf
4. ALLiaNCE / DINUM (n.d.). Albert API (uradna predstavitev produkta; OpenAI-kompatibilen API Gateway; instanca OpenGateLLM; umestitev med medresorske AI produkte). https://alliance.numerique.gouv.fr/produit/produits-ia-interministeriels/albert-api/
5. Albert API (n.d.). Fonctionnalités (uradna stran; dostop do modelov; RAG à la demande; upravljanje projektov in spremljanje uporabe; embeddings, OCR, rerank; OpenAI standard). https://albert.sites.beta.gouv.fr/solutions/fonctionnality/
6. Albert API (n.d.). Modèles (uradna stran; modeli iz odprtokodnega ekosistema; gostovanje na strežnikih Albert API; podatki se ne pošiljajo ponudnikom modelov; portfelj modelov se redno posodablja). https://albert.sites.beta.gouv.fr/solutions/models/

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

3.4 Mesto Helsinki – AI Register: "Summer Job Voucher Chatbot" (transparentna evidenca + praktični governance)

Mesto Helsinki je vzpostavilo javno dostopno evidenco AI sistemov (AI Register) kot "okno" v algoritme in AI rešitve, ki jih uporablja v mestnih storitvah. Evidenca je namenjena temu, da imajo prebivalci in strokovna javnost dostop do razumljivih in ažurnih informacij o tem, kje in kako AI vpliva na mestne storitve, hkrati pa omogoča tudi oddajo povratnih informacij.

V praksi je AI Register zasnovan tako, da posamezni vnosi niso zgolj seznam, temveč vsebujejo tudi podrobnejše informacije o uporabljenih podatkih, logiki obdelave in upravljanju (governance) – torej elemente, ki so ključni za transparentnost in notranji nadzor uporabe AI v javnem sektorju.

Opis sistema in namen uporabe

»Summer Job Voucher Chatbot« je v evidenci opisan kot nov digitalni kontaktni kanal mestne enote za kulturo in prosti čas (Culture and leisure), ki odgovarja na vprašanja o prijavi in uporabi poletnega zaposlitvenega vavčerja (summer job voucher). Storitev je zasnovana za izboljšanje dostopnosti podpore uporabnikom tudi izven običajnih ur službe: chatbot odgovarja 24/7, uporabniku pa omogoča tudi prehod na živi klepet s predstavnikom službe za pomoč uporabnikom v rednih urah.

Kot dodatni cilj je navedeno, da kanal podpira samopostrežno uporabo, zagotavlja ažurne informacije o tem, kako vavčer poiskati in uporabiti, ter vodi mlade in delodajalce skozi ključne korake zaposlovanja in prijavnih postopkov.

Kdo so uporabniki in komu je rešitev namenjena

Primarni uporabniki so zunanji uporabniki mestne storitve – predvsem mladi, ki želijo uporabiti poletni vavčer, ter delodajalci, ki iščejo informacije o postopku in pogojih. Vnos v AI Register izrecno poudari, da kanal izboljšuje uporabniško izkušnjo in pospeši iskanje relevantnih informacij v primerjavi z ročnim brskanjem po spletnih straneh.

Cilj rešitve

Cilj rešitve je torej povečati dostopnost in odzivnost podpore uporabnikom (tudi izven uradnih ur), hkrati pa izboljšati kakovost samopostrežnega kanala z rednim učenjem iz pogostih vprašanj in rabe storitve. AI Register navaja, da je storitev trenirana na podlagi gradiv, zbranih iz interakcij s službo za pomoč uporabnikom, da lahko odgovarja na najpogostejša vprašanja o vavčerju.

Podatkovni viri (datasets) in hrambe

Pri tem primeru je izhodiščni vir informacij posamezni registrski vnos v Helsinki AI Register, tj. v javno dostopni evidenci sistemov umetne inteligence, ki jih uporablja mesto Helsinki. Register je zasnovan kot pregledno "okno" za uporabo umetne inteligence v mestnih storitvah: za vsak sistem ponuja kratek pregled in podrobnejši strukturiran opis, ki vključuje tudi podatke o namenu, uporabljenih virih in upravljanju sistema.

Vnos za »Summer Job Voucher Chatbot« podatkovne vire predstavi jasno in jih razdeli na dve ločeni skupini:

1. Prvo skupino predstavlja **učni oziroma konfiguracijski material** (*AI training material*), ki služi pripravi in testiranju delovanja sistema.
2. Drugo skupino predstavljajo **pogovorni dnevniki** (*conversation logs*), torej vprašanja in odgovori, ki nastajajo pri dejanski uporabi storitve.

Takšna razmejitev je z vidika upravljanja pomembna zato, ker omogoča ločeno obravnavo podatkov, ki so namenjeni zasnovi sistema, in podatkov, ki nastajajo v operativni rabi ter zato zahtevajo jasna pravila glede analize, hrambe in zaščite.

1) AI training material: izvor, vsebina in (ne)prisotnost osebnih podatkov

"AI training material" je opisan kot nabor vzorčnih vprašanj s področja storitve. Vprašanja so bila zbrana v času razvojnega projekta pri strokovnjakih mesta Helsinki in naj bi odražala vprašanja, ki jih prebivalci tipično zastavljajo v situacijah podpore uporabnikom. Hkrati register izrecno navede tri ključne lastnosti tega materiala:

(i) material je v lasti mesta Helsinki,
(ii) (ii) ni licenciran, ter
(iii) (iii) ne vključuje osebnih podatkov.

Register pojasni tudi, kako se material uporablja: vzorčna vprašanja se naložijo v sistem, na njihovi podlagi sistem "zgradi model", del vzorčnih vprašanj pa se uporablja tudi za testiranje delovanja modela.

(Opomba: register tu ne opisuje tehničnih podrobnosti učenja, temveč funkcionalno razlaga, da se na osnovi naloženega materiala sistem konfigurira in preverja.)

2) Pogovorni dnevniki (conversation logs): kaj se beleži, zakaj in kako dolgo

Pri uporabi storitve sistem beleži vprašanja in odgovore v pogovorni dnevnik. Namen beleženja je izboljševanje storitve: razvojna ekipa dnevnike redno analizira, da ugotovi, kako je sistem odgovarjal na vprašanja uporabnikov, katere vsebinske vrzeli so se pokazale in kje je treba dopolniti obstoječe pogovorne poti. Na podlagi te analize se lahko posodobi učni material, doda nova vzorčna vprašanja ali oblikuje nove poti pogovora; pri tem se upošteva tudi povratne informacije uporabnikov.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Register hkrati določa tudi režim hrambe. Pogovorni dnevniki se po opravljeni analizi odstranijo iz sistema po šestih mesecih, medtem ko se za dolgoročno spremljanje razvoja storitve hranijo agregirana poročila o uporabi, kot so npr. stopnje uporabe, odzivnosti in drugi parametri. Tudi za pogovorne dnevnike je navedeno, da so v lasti mesta Helsinki, vendar niso licencirani. To pomeni, da mesto z njimi upravlja kot z lastnim internim gradivom za razvoj in spremljanje storitve, ne pa kot z odprto ali posebej licencirano zbirko za zunanjo ponovno uporabo

Celovita slika zasebnosti: sistem ne zahteva osebnih podatkov, vendar obvladuje tveganje neželenega vnosa

Helsinki AI Register pri tem chatbotu zasebnost obravnava na zelo praktični ravni. Vnos izrecno navaja, da storitev nikoli ne zahteva osebno določljivih podatkov (*personally identifiable information*), kar pomeni, da zbiranje takih podatkov ni del njenega predvidenega namena uporabe. Kljub temu pa register realistično izpostavi, da lahko uporabnik v pogovor kljub temu vnese nepotrebne osebne podatke, zato je bilo treba ob sami storitvi predvideti tudi varovalke za obvladovanje tega tveganja.

Kot prvi ukrep je v registru navedena avtomatizirana funkcija, ki iz pogovornih dnevnikov takoj odstrani nekatere očitne osebne identifikatorje, kot so številke socialnega zavarovanja in elektronski naslovi. Takšna rešitev je pomembna zato, ker tveganje naslavlja že na ravni operativne uporabe: sistem ne čaka šele na kasnejši ročni pregled, temveč poskuša najbolj tipične občutljive vnose odstraniti takoj po njihovem pojavu v dnevnikih.

Drugi ukrep predstavlja redni strokovni pregled pogovornih dnevnikov. Register izrecno navaja, da dnevnike pregledujejo strokovnjaki, ki odstranijo tiste osebne podatke, ki jih avtomatizirana funkcija morda ni zaznala. S tem je vzpostavljen dvostopenjski mehanizem zaščite: avtomatizirano filtriranje za najpogostejše identifikatorje ter dodatni človeški pregled za preostale primere.

Za bralca je pri tem primeru pomembno predvsem to, da register ne predstavlja zasebnosti kot abstraktnega načela, temveč kot **kombinacijo zasnove storitve in konkretnih operativnih ukrepov.**

- Prvi element je omejitev namena uporabe — storitev ni zasnovana za pridobivanje osebnih podatkov.
- Drugi element je tehnična varovalka — samodejno odstranjevanje najbolj očitnih osebnih identifikatorjev.
- Tretji element pa je organizacijska varovalka — redni strokovni pregled dnevnikov.

Skupaj ti ukrepi kažejo razmeroma zrelo ureditev za javni chatbot, kjer je glavno tveganje prav to, da uporabnik v pogovor vnese več, kot je za uporabo storitve dejansko potrebno.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Obdelava podatkov in arhitektura modela

AI Register pri tem sistemu vsebuje tudi ločen sklop informacij o obdelavi podatkov (*data processing*) in arhitekturi modela (*model architecture*), kar je za javni register posebej koristno, saj bralcu omogoča osnovno razumevanje tehnološke zasnove storitve. V dokumentaciji navajajo, da chatbot temelji na programskem produktu IBM Watson Assistant, ki je ponujen kot programska storitev v oblaku (*software as a service, SaaS*) v okviru IBM Cloud, pri čemer je kot lokacija podatkovnega centra naveden Frankfurt v Evropski uniji. Register hkrati pojasni, da storitev uporablja modele umetne inteligence in strojnega učenja za obdelavo vprašanj v naravnem jeziku.

Z vidika uporabniškega toka je delovanje opisano razmeroma jasno. Uporabnik zastavi vprašanje prek digitalnega vmesnika, ki deluje na različnih napravah. Če umetna inteligenca na vprašanje ne zna ustrezno odgovoriti ali če je to iz drugega razloga primerneje, se lahko uporabnik poveže s predstavnikom službe za pomoč uporabnikom (*customer service representative*), ki nadaljuje pogovor. Register dodatno navaja, da lahko uporabnik poda tudi povratno informacijo o storitvi, ki se nato uporablja pri njenem nadaljnjem razvoju.

Takšna arhitektura je pomembna iz dveh razlogov.

- Prvič, jasno kaže, da sistem ni zamišljen kot popolnoma zaprt avtomatiziran kanal, temveč kot digitalni samopostrežni sloj, ki se po potrebi poveže z živim človekom.
- Drugič, register s tem posredno pokaže tudi osnovno logiko upravljanja kakovosti: vprašanja, na katera sistem ne zna odgovoriti, in povratne informacije uporabnikov postanejo vir za nadaljnji razvoj vsebine in pogovornih poti.

To dodatno potrjujejo tudi deli registra o človeškem nadzoru, kjer je navedeno, da sistem samodejno prepoznava teme, na katere ne zna odgovoriti, te teme pa se nato posredujejo strokovnjakom mesta Helsinki za nadaljnje izboljšave.

Register vsebuje tudi povezavo do logične arhitekture sistema (*system architecture description*), vendar besedilni del javnega vnosa ne opisuje natančnejše notranje topologije ali tehničnih komponent tega diagrama. Zato je mogoče z gotovostjo povzeti predvsem to, kar je izrecno zapisano: storitev temelji na IBM Watson Assistant kot SaaS rešitvi v podatkovnem centru v Frankfurtu, uporablja AI in ML za obdelavo vprašanj v naravnem jeziku ter je zasnovana tako, da omogoča prehod na človeško podporo in uporabo povratnih informacij za nadaljnji razvoj.

Dobaviteljski model in zunanje tehnološke odvisnosti

Z vidika upravljanja dobavne verige je pri tem primeru pomembno, da je v registrskem vnosu izrecno naveden zunanji dobavitelj (external supplier / fournisseur externe), in sicer IBM. Hkrati je navedeno, da chatbot temelji na produktu IBM Watson Assistant, ki je zagotovljen kot programska storitev v oblaku (software as a service, SaaS) v okviru IBM Cloud, pri čemer

je kot lokacija podatkovnega centra naveden Frankfurt v Evropski uniji. To pomeni, da mesto Helsinki storitev uporablja prek zunanje tehnološke platforme, ne pa kot lastno v celoti interno razvito in gostovano rešitev.

Takšna zasnova je za javni sektor pomembna predvsem zato, ker jasno pokaže razmerje med upravljavcem storitve in tehnološkim dobaviteljem. Mesto Helsinki ostaja nosilec javne storitve in skrbnik njene vsebine, medtem ko je temeljna pogovorna infrastruktura zagotovljena prek zunanjega SaaS produkta. To potrjujejo tudi drugi deli vnosa v register: učni material in pogovorni dnevniki so opredeljeni kot gradivo v lasti mesta Helsinki, čeprav tehnološko okolje zagotavlja zunanji ponudnik. S tem je nakazana tipična ureditev v javnem sektorju, kjer javni organ ne razvija nujno vse tehnologije sam, vendar mora kljub temu ohraniti zadosten nadzor nad podatki, namenom uporabe in izboljševanjem storitve.

Z vidika arhitekture dobavne verige je pomembno tudi, da javni register ne opisuje le dobavitelja, temveč tudi meje njegove vloge, kolikor so razvidne iz objavljenih podatkov. Vnos potrjuje, da storitev uporablja modele umetne inteligence in strojnega učenja za obdelavo vprašanj v naravnem jeziku, vendar hkrati pokaže, da je končna storitev organizirana tako, da omogoča prehod na človeško podporo, kadar sistem ne zna ustrezno odgovoriti ali kadar je to primerneje. To pomeni, da zunanji tehnološki produkt ni postavljen kot popolnoma avtonomen kanal, temveč kot del širše storitvene ureditve, v kateri mesto ohranja operativni nadzor nad uporabniško izkušnjo.

Javno objavljena dokumentacija ne gre do te mere v podrobnosti, da bi bilo mogoče natančno rekonstruirati celotno tehnično dobavno verigo (npr. podizvajalce, notranje komponente SaaS okolja ali podrobno topologijo integracij). Register sicer vsebuje povezavo do opisa logične arhitekture sistema (system architecture description), vendar besedilni del vnosa teh elementov ne razčleni podrobneje. Zato je iz javnega registra mogoče z gotovostjo povzeti predvsem naslednje: storitev uporablja zunanji produkt IBM Watson Assistant kot SaaS iz podatkovnega centra v Frankfurtu, IBM je naveden kot zunanji dobavitelj, mesto Helsinki pa ohranja vlogo nosilca storitve ter lastništvo nad ključnimi vsebinskimi podatki, uporabljenimi za razvoj in spremljanje chatbot storitve.

Kakovost, človeški nadzor in operativno upravljanje

Helsinki AI Register pri tem chatbotu kakovosti ne predstavlja kot enkratne lastnosti sistema, temveč kot nekaj, kar se spremlja in sproti izboljšuje. Vnos izrecno navaja, da se razvoj uporabe in kakovosti storitve stalno vrednoti, pri čemer se redno pripravljajo poročila o uporabi in kakovosti. Med kakovostnimi kazalniki so navedeni zlasti delež pravilnih odgovorov glede na vsa zastavljena vprašanja ter neposredne povratne informacije uporabnikov, zbrane ob uporabi storitve. To pomeni, da je spremljanje kakovosti zasnovano kot kombinacija kvantitativnih podatkov o delovanju sistema in kvalitativne ocene uporabniške izkušnje.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Z vidika upravljanja je pomembno, da se kakovost storitve ne meri le zato, da bi mesto Helsinki vedelo, "koliko se chatbot uporablja", ampak predvsem zato, da se lahko ugotovi, ali sistem dejansko odgovarja pravilno in uporabno. Register s tem pokaže razmeroma praktičen pristop: uspešnost klepetalnika se ne ocenjuje zgolj z obsegom uporabe, temveč z vprašanjem, ali so odgovori vsebinsko ustrezni in ali jih uporabniki prepoznavajo kot koristne.

Tudi pri človeškem nadzoru register podaja konkretno, operativno razlago. Navedeno je, da storitev samodejno prepoznava in zbira teme, na katere ne zna odgovoriti, te teme pa se nato posredujejo strokovnjakom mesta Helsinki, ki na tej podlagi razvijajo nove pogovorne poti. To pomeni, da človek ni vključen šele ob morebitnem "incidentu", temveč je del rednega cikla izboljševanja: sistem sam zazna svoje vsebinske vrzeli, strokovnjaki pa jih nato pretvorijo v nove rešitve znotraj storitve.

Register dodatno navaja, da se nadzor izvaja z rednim pregledom povratnih informacij, pregledom pogovornih dnevnikov in analizo statistike uporabe. Na tej podlagi strokovnjaki prepoznajo teme, ki uporabnike posebej zanimajo, ter izboljšujejo pogovorne poti in tudi iskljivost priporočil z uporabo novih ključnih besed. Tak opis je za širšo strokovno javnost pomemben zato, ker pokaže, da je človeški nadzor pri tem primeru organiziran kot stalni operativni proces, ne le kot formalna možnost ročnega posega.

Če to podpoglavje beremo skupaj z drugimi deli registrskega vnosa, se pokaže razmeroma koherentna slika upravljanja:

- chatbot je zasnovan kot samopostrežni kanal, vendar njegovo delovanje ves čas spremljajo ljudje, ki analizirajo uporabo, obravnavajo povratne informacije in na tej podlagi posodabljajo vsebino ter logiko pogovora.

Javno objavljene informacije sicer ne opisujejo podrobnejših notranjih postopkov odločanja ali formalnih odgovornosti posameznih ekip, vendar dovolj jasno pokažejo, da mesto Helsinki kakovost in človeški nadzor razume kot del rednega vzdrževanja in razvoja storitve

Upravljanje tveganj: omejitev namena uporabe in nadzor nad neželenim vnosom osebnih podatkov

Pri tem primeru je pomembno, da register tveganja za zasebnost ne obravnava abstraktno, temveč jih poveže z dejanskim načinom uporabe chatbot storitve. Izhodiščno pravilo je jasno: sistem ni zasnovan za pridobivanje osebno določljivih podatkov in jih od uporabnika tudi ne zahteva. Vendar pa je v registru hkrati izrecno priznano, da lahko uporabniki v pogovor vnesejo več informacij, kot je za uporabo storitve dejansko potrebno.

Upravljanje tega tveganja temelji na kombinaciji treh elementov. Prvi je že sama omejitev namena uporabe: storitev je zasnovana kot informacijski in podporni kanal, ne kot sistem, ki bi za svoje delovanje potreboval osebne podatke uporabnika. Drugi element predstavlja

tehnična varovalka v obliki samodejnega odstranjevanja nekaterih najbolj tipičnih osebnih identifikatorjev iz pogovornih dnevnikov. Tretji element pa je organizacijski nadzor, saj strokovnjaki mesta redno pregledujejo dnevnike in odstranijo osebne podatke, ki jih avtomatizirana funkcija morebiti ne bi zaznala.

Za širšo strokovno javnost je pri tem primeru bistveno predvsem to, da register pokaže razmeroma zrelo upravljavsko logiko: tveganje ni obravnavano zgolj z navodilom, naj uporabnik osebnih podatkov ne vnaša, temveč s kombinacijo zasnove storitve, avtomatiziranih varovalk in rednega človeškega pregleda. S tem mesto Helsinki pokaže, da je pri javnih samopostrežnih kanalih treba upravljanje tveganj graditi okoli najbolj verjetnega praktičnega scenarija – ne da bo sistem osebne podatke zahteval, temveč da jih bo uporabnik vnesel po nepotrebnem.

Nediskriminacija in dostopnost

Helsinki AI Register pri tem chatbotu vprašanja nediskriminacije in dostopnosti obravnava razmeroma zadržano, vendar dovolj konkretno, da je iz dokumentacije mogoče razbrati dejanske omejitve storitve. Register izrecno navaja, da je storitev trenutno na voljo le v finščini, obenem pa tudi, da ne popravlja samodejno tipkarskih napak. Ti dve okoliščini nista predstavljeni kot abstraktno tveganje, temveč kot zelo praktična omejitev uporabe: sistem je manj dostopen uporabnikom, ki finščine ne obvladajo kot maternega jezika, in tudi tistim, ki v vprašanjih delajo pravopisne ali tipkarske napake.

Za javni sektor je takšna transparentnost pomembna že sama po sebi. Register namreč ne ustvarja vtisa, da je storitev univerzalno dostopna ali enako primerna za vse uporabnike, temveč odkrito pokaže, kje so njene trenutne meje. To je z vidika upravljanja dober pristop, ker omogoča bolj realistično presojo dejanskih učinkov storitve na različne skupine uporabnikov in hkrati postavlja jasno izhodišče za nadaljnje izboljšave.

Pomembno je tudi, da register teh omejitev ne skriva znotraj tehničnega opisa sistema, temveč jih obravnava v posebnem sklopu nediskriminacije. S tem mesto Helsinki posredno pokaže, da vprašanje enake dostopnosti ni razumljeno zgolj kot tehnična pomanjkljivost sistema, ampak kot relevanten upravljavski vidik javne storitve. Javno objavljeni vnos sicer ne navaja podrobnejših korektivnih ukrepov ali formalne analize vplivov na posamezne skupine uporabnikov, vendar že sama izrecna navedba jezikovne omejitve in neodzivnosti na tipkarske napake pomeni koristno prakso transparentnosti, ki je pri javnih chatbotih pogosto odsotna.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Preliminarna navezava na Akt o UI AI (LIMITED-RISK)

Pri tem primeru je treba najprej poudariti, da Akt o UI ne uporablja formalne kategorije "limited-risk" kot enotnega pravnega režima, temveč za nekatere sisteme, ki niso "visoko-tvegani", določa predvsem posebne obveznosti transparentnosti. Za sisteme umetne inteligence, ki neposredno komunicirajo z naravnimi osebami, Akt o UI t določa, da morajo biti posamezniki obveščeni, da komunicirajo z AI sistemom, razen kadar je to iz okoliščin že očitno.

Na podlagi javno dostopnega opisa je »Summer Job Voucher Chatbot« predvsem informacijski in podporni kanal za odgovarjanje na vprašanja o mestni storitvi, z možnostjo prehoda na človeško podporo. V takšni zasnovi je regulatorno težišče zato predvsem na jasni transparentnosti do uporabnika, ne pa na obravnavi sistema kot očitno "visoko tveganega" po pravilih iz člena 6 in Priloge III. Pri Aktu o UI je namreč za presojo "visoke-tveganosti" ključen namen uporabe konkretnega sistema in njegova umestitev v poti iz 6. člena, zlasti v primerih iz Priloge III.

V tem smislu je primer mesta Helsinki uporaben predvsem kot zgled, kako se lahko zahteve transparentnosti in odgovorne uporabe operacionalizirajo že na ravni javno objavljenega registrskega vnosa: uporabnik lahko razume, da komunicira z AI sistemom, kakšen je namen sistema, kateri podatki se uporabljajo, kako dolgo se hranijo pogovorni dnevniki, kako je organiziran človeški nadzor in kateri ukrepi so uvedeni za preprečevanje nenamernega vnosa osebnih podatkov. Čeprav tak register sam po sebi ni obveznost iz 50. člena Akta o UI, je v praksi zelo dobra podpora izpolnjevanju temeljne zahteve po pregledni in pošteni uporabi chatbot storitve v javnem sektorju.

Hkrati pa je treba ostati terminološko previden: ista tehnična zasnova klepetalnika bi lahko v drugačnem kontekstu ali z drugačnim namenom uporabe prešla v zahtevnejši regulativni okvir. Zato je pri tem primeru najprimerneje reči, da gre po javno opisanem namenu uporabe predvsem za primer, kjer je v ospredju transparentnost in upravljanje uporabe, ne pa za primer, ki bi bil že po samem opisu očitno zajet med "visoko-tvegane" sisteme iz Priloge III.

Ključne prenosljive dobre prakse (povzetek)

To podpoglavje povzema dobre prakse, ki izhajajo neposredno iz primera Summer Job Voucher Chatbot in iz načina, kako je ta sistem opisan v Helsinki AI Register. Namen ni splošna razprava o transparentnosti v javnem sektorju, temveč prikaz, kako je mesto Helsinki pri konkretnem chatbotu javno objavilo ključne informacije o namenu sistema, uporabljenih podatkih, upravljanju tveganj, človeškem nadzoru in omejitvah storitve. Prav v tem je glavna vrednost primera: ne gre le za opis posameznega klepetalnika, temveč za prikaz, kako je mogoče AI sistem v javni storitvi predstaviti na strukturiran, preverljiv in za uporabnika razumljiv način.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

1. **Javno dostopna in strukturirana evidenca AI sistemov kot orodje transparentnosti**
   Prva in najbolj očitna dobra praksa je že sama oblika registra. Helsinki AI Register ni zgolj seznam uporabljenih AI rešitev, temveč javna evidenca, ki za vsak sistem objavlja strukturirane informacije po vnaprej določenih sklopih, med drugim o podatkovnih virih (datasets), obdelavi podatkov (data processing), nediskriminaciji (non-discrimination), človeškem nadzoru (human oversight) in upravljanju tveganj (risk management). Takšna struktura je za javni sektor posebej uporabna, ker omogoča, da so bistvene informacije o sistemu objavljene na enoten in primerljiv način, hkrati pa olajša tudi notranji governance in poznejše presoje skladnosti.

2. **Jasna ločitev med učnim gradivom, pogovornimi dnevniki in agregirano analitiko**
   Vnos za ta chatbot jasno razlikuje med učnim oziroma konfiguracijskim materialom in pogovornimi dnevniki, ki nastajajo pri dejanski uporabi storitve. Dodatno pa določa še ločeno obravnavo agregiranih poročil o uporabi, ki se hranijo za spremljanje dolgoročnega razvoja. To je pomembna dobra praksa, ker že na ravni javnega opisa pokaže, da niso vsi podatki v sistemu obravnavani enako: eni služijo pripravi in testiranju storitve, drugi izboljševanju operativne rabe, tretji pa dolgoročnemu spremljanju uspešnosti. Takšna razmejitev je praktično uporabna tudi kot model za interne evidence in politike upravljanja podatkov.

3. **Časovno omejena hramba vsebinskih logov ob hkratni ohranitvi agregatov**
   Register izrecno navaja, da se pogovorni dnevnik izbriše šest mesecev po analizi, medtem ko se poročila o uporabi (npr. stopnje uporabe, odzivnosti in drugi parametri) hranijo za spremljanje dolgoročnega napredka. To je zelo prenosljiva dobra praksa, ker vzpostavlja razumno ravnotežje med dvema ciljema: na eni strani omogoča dovolj dolg dostop do logov za izboljševanje sistema, na drugi strani pa omejuje obdobje hrambe podrobnih interakcijskih zapisov, ki lahko nosijo večje zasebnostno tveganje. Hkrati mesto ohrani agregirane kazalnike, ki so potrebni za spremljanje kakovosti storitve, ne da bi bilo treba dolgoročno hraniti vsebinsko bogate zapise posameznih pogovorov.

4. **Dobaviteljska transparentnost: izrecna navedba ponudnika in lokacije obdelave**
   V registrskem vnosu je jasno navedeno, da je zunanji dobavitelj IBM, in da storitev temelji na IBM Watson Assistant kot programski storitvi v oblaku (SaaS), ponujeni iz podatkovnega centra v Frankfurtu v Evropski uniji. To je pomembna dobra praksa, ker javni register ne ostane na abstraktni ravni opisa “chatbota”, temveč razkrije tudi osnovni dobaviteljski in infrastrukturni kontekst sistema. Za javni sektor je takšna informacija posebej pomembna z vidika upravljanja dobavne verige, presoje zunanjih tehnoloških odvisnosti ter osnovnega razumevanja, kje in prek katerega ponudnika se storitev dejansko izvaja.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

5. **Večplastno obravnavanje tveganja neželenega vnosa osebnih podatkov** Register izrecno navaja, da storitev nikoli ne zahteva osebno določljivih podatkov, vendar hkrati priznava, da lahko uporabnik takšne podatke vseeno vnese. Za obvladovanje tega tveganja sta navedena dva konkretna ukrepa: avtomatizirano odstranjevanje določenih osebnih identifikatorjev (npr. številk socialnega zavarovanja in elektronskih naslovov) iz pogovornih dnevnikov ter redni strokovni pregled dnevnikov, pri katerem strokovnjaki odstranijo podatke, ki jih avtomatika ne zazna. Ta kombinacija tehnične in organizacijske varovalke je posebej dobra praksa, ker pokaže, da upravljanje tveganj ne temelji samo na pravilih uporabe, temveč na dejanskih operativnih ukrepih.

6. **Transparentno navajanje omejitev sistema kot del odgovorne uporabe** Register odkrito navede tudi omejitve sistema: chatbot je trenutno na voljo le v finščini in ne popravlja samodejno tipkarskih napak, kar lahko omeji uporabo pri osebah, ki finščine ne govorijo kot maternega jezika. Takšna transparentnost je sama po sebi pomembna dobra praksa, ker uporabniku in strokovni javnosti omogoča realno presojo dejanske dostopnosti sistema. Register s tem ne prikazuje storitve kot univerzalno primerne za vse, temveč jasno pove, kje so njene trenutne meje in kje so potencialna področja nadaljnjih izboljšav.

7. **Stalno spremljanje kakovosti in človeški nadzor kot del rednega obratovanja** Register navaja, da se razvoj uporabe in kakovosti storitve stalno vrednoti, pri čemer se vodijo poročila o uporabi in kakovosti, med kazalniki pa so navedeni zlasti delež pravilnih odgovorov in neposredne povratne informacije uporabnikov. Poleg tega sistem samodejno prepoznava teme, na katere ne zna odgovoriti, te teme pa se posredujejo strokovnjakom mesta, ki na tej podlagi razvijajo nove pogovorne poti in izboljšujejo iskljivost priporočil. To je pomembna prenosljiva praksa zato, ker kaže, da človeški nadzor ni omejen na izredne posege, temveč je vgrajen v vsakodnevni cikel spremljanja, analize in izboljševanja storitve.

Viri (primer 3.4)

1. City of Helsinki AI Register (2026). Summer Job Voucher Chatbot – opis podatkovnih virov, hrambe logov (6 mesecev), IBM Watson Assistant (Frankfurt), human oversight, risk management z avtomatskim brisanjem SSN/email. https://ai.hel.fi/en/summer-job-voucher-chatbot/
2. City of Helsinki AI Register (2026). AI Register – predstavitev registra kot javno dostopne evidence AI sistemov mesta Helsinki ter njegove vloge pri transparentnosti in oddaji povratnih informacij. https://ai.hel.fi/en/ai-register/
3. City of Helsinki AI Register (2026). Get to know AI Register – dodatna pojasnila o namenu registra, strukturi informacij in vlogi registra kot “window into the artificial

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

intelligence systems used by the City of Helsinki”.
https://ai.hel.fi/en/get-to-know-ai-register/

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

4 Vsebinska sinteza in priporočila

Na podlagi poglobljenih analiz štirih izbranih primerov je mogoče izluščiti več skupnih vzorcev dobrih praks, ki presegajo posebnosti posameznih use-caseov in so zato relevantni za nadaljnjo pripravo poročila NOSI-AI. Namen tega poglavja ni ponavljanje ugotovitev iz opisov primerov, temveč njihova sinteza v obliko, ki je neposredno uporabna za pripravo priporočil, zahtev in operativnih kontrolnih seznamov.
Obravnavani primeri so med seboj tehnološko in organizacijsko različni: segajo od LLM + RAG sistema v javni službi zaposlovanja, preko federativnega učenja v AML okolju, do skupne državne platforme za generativno UI in javno dostopnega registra chatbot storitve. Prav ta raznolikost pa omogoča, da se pokažejo tisti elementi dobrih praks, ki se pojavljajo ne glede na razliko v arhitekturi, namenu uporabe ali regulatornem okolju. Skupni imenovalec vseh primerov je, da dobra praksa ne nastane šele na ravni "končnega modela", temveč se vzpostavlja skozi celoten življenjski cikel sistema: od opredelitve namena uporabe, izbire infrastrukture in podatkovnih virov, do upravljanja vhodov, človeškega nadzora, spremljanja kakovosti in upravljanja sprememb.

4.1 Skupni vzorci dobrih praks, ugotovljeni čez primere

Prvi ponavljajoči se vzorec je jasno razlikovanje med vsebino interakcij in operativnimi evidencami uporabe.
V vseh primerih se, čeprav v različnih oblikah, pojavlja potreba po razmejitvi med vsebinskimi vnosi oziroma izhodi sistema in tistimi podatki, ki služijo spremljanju uporabe, kakovosti ali varnosti. Pri France Travail je to vprašanje posebej izrazito pri promptih in pogojih njihove hrambe za revizijo, testiranje in izboljšave. Helsinki pri chatbotu jasno loči med pogovornimi dnevniki, ki se po analizi izbrišejo po šestih mesecih, in agregiranimi poročili o uporabi, ki se hranijo za dolgoročno spremljanje kakovosti. Albert API pa kot del platformne zasnove izrecno poudari ne-hrambo vsebine pogovorov, ob hkratni prisotnosti telemetrije uporabe in merilnikov porabe. Iz teh primerov izhaja jasna izvedbena lekcija: že v zasnovi sistema je treba razlikovati med vsebinsko občutljivimi podatki in tistimi evidencami, ki so potrebne za governance, kapacitetno upravljanje, spremljanje kakovosti ali varnostni nadzor.

Drugi ponavljajoči se vzorec je, da mora biti človeški nadzor zasnovan kot dejanski operativni mehanizem, ne kot formalna klavzula.
V primeru France Travail je to najbolj očitno: CNIL pogojuje dopustnost uporabe 62ystem62 z dejansko možnostjo svetovalca, da rezultat razume, presodi in mu po potrebi ne sledi. Pri Helsinkiju je človeški nadzor operacionaliziran drugače, vendar enako konkretno: 62ystem samodejno zaznava teme, na katere ne zna odgovoriti, te teme pa se posredujejo strokovnjakom, ki vsebino in pogovorne poti sproti dopolnjujejo. Pri Albert API pa se težišče premakne na odgovornost uporabniških administracij za pravilno uporabo platforme, klasifikacijo podatkov ter ponovno presojo v primerih občutljivejših obdelav. Skupna

ugotovitev je, da učinkovitega človeškega nadzora ni mogoče izpeljati zgolj iz dejstva, da "je človek nekje v procesu", temveč iz kombinacije razumljivosti rezultatov, določenih odgovornosti, podpore uporabnikom, procesov eskalacije in ustreznih delovnih pogojev.

Tretji ponavljajoči se vzorec je upravljanje vhodov kot samostojno področje nadzora.
V sodobnih UI sistemih pomemben del tveganj ne nastaja šele pri izhodu modela, temveč že na vhodni strani. France Travail to pokaže z obravnavo promptov kot posebnega podatkovnega kanala, ki zahteva minimizacijo, standardizacijo, filtre in usposabljanje uporabnikov. Helsinki pri chatbotu podobno naslovi tveganje neželenega vnosa osebnih podatkov z avtomatiziranim odstranjevanjem identifikatorjev in rednim strokovnim pregledom dnevnikov. Albert API pa vhodno upravljanje obravnava že na ravni pravil uporabe, saj prepoveduje uporabo občutljivih podatkov in odgovornost za vsebino ohranja pri administraciji, ki sistem uporablja. Vsi ti primeri potrjujejo, da je pri generativnih in drugih sodobnih UI sistemih upravljanje vhodov eden osrednjih gradnikov skladnosti, kakovosti in zmanjševanja tveganj.

Četrti ponavljajoči se vzorec je obravnava infrastrukture in dobavne verige kot vprašanja prvega reda.
Infrastrukturna odločitev se v vseh primerih kaže kot vsebinsko pomemben element governance, ne kot tehnična podrobnost. Pri France Travail je lokalna (on-premise) postavitev izrecno obravnavana kot varovalka zaupnosti in nadzora nad osebnimi podatki. Albert API odgovarja na isto vprašanje z drugačnim, a funkcionalno sorodnim modelom: s suverenim oblačnim okoljem, formalno homologacijo in standardiziranim državnim okvirom uporabe. Helsinki pa v javnem registru pregledno navede zunanji produkt, SaaS model in lokacijo podatkovnega centra, s čimer pokaže, da je tudi pri manjšem javnem chatbotu dobaviteljska transparentnost pomemben del odgovorne uporabe. Iz primerov sledi, da je treba infrastrukturo, lokacijo obdelave, vlogo dobaviteljev, certifikacije in pogodbeno-organizacijska razmerja obravnavati kot integralni del upravljanja sistema.

Peti ponavljajoči se vzorec je, da modelna varnost zahteva ločeno obravnavo in je ni mogoče zreducirati na klasično informacijsko varnost.
Ta vidik je najbolj razviden v primeru Finterai. Datatilsynet izrecno pokaže, da federativno učenje sicer zmanjšuje potrebo po centralnem deljenju surovih podatkov, vendar hkrati odpira nova tveganja, vezana na modele, modelne posodobitve in protokole izmenjave. Pri tem posebej izpostavi napade tipa *model inversion* in potrebo po ločeni obravnavi tveganj v fazi učenja ter v fazi produkcijske uporabe. Ta nauk je širše prenosljiv tudi na druge sisteme: varovanje modelov, parametrov, posodobitev in API vmesnikov je danes samostojna varnostna tema, ki je ni mogoče zadovoljivo obravnavati le s klasičnimi ukrepi za varovanje podatkovnih zbirk ali omrežja.

Šesti ponavljajoči se vzorec je transparentno navajanje omejitev in stalno spremljanje kakovosti.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

Helsinki v registru odkrito navede, da je chatbot trenutno na voljo le v finščini in da ne popravlja samodejno tipkarskih napak. Poleg tega kakovost spremlja z deležem pravilnih odgovorov in povratnimi informacijami uporabnikov. Tudi v drugih primerih se kaže podobna logika: sistemi niso predstavljeni kot "samoumevno ustrezni", temveč kot rešitve, katerih uporabnost in dopustnost sta odvisni od stalnega spremljanja, usposabljanja, merjenja kakovosti, analiziranja napak in posodabljanja vsebine ali modela. To pomeni, da je ena ključnih dobrih praks že sama sposobnost organizacije, da omejitve sistema jasno opredeli, jih ustrezno komunicira in jih obravnava kot predmet stalnega izboljševanja.

4.2 Priporočila za prenos v operativni kontrolni seznam

Na podlagi navedenih primerov je priporočljivo, da se nadaljnje poglavje poročila NOSI-AI oblikuje kot **operativni kontrolni seznam po življenjskem ciklu sistema**, in ne le kot nabor pravnih ali tehničnih zahtev po posameznih temah. Takšna zasnova je primernejša za dejansko uporabo v praksi, saj omogoča, da se v vsaki fazi razvoja, uvajanja in uporabe preveri, katere odločitve je treba sprejeti, katere kontrole vzpostaviti in katere dokaze o tem zbrati.
Hkrati je smiselno v kontrolnem seznamu ločiti med:

- (a) sistemi tipa LLM/RAG ter
- (b) sistemi tipa strolnega učenja (Machine Learning) za analitiko, detekcijo ali priporočanje,

saj so nekatera vprašanja skupna, druga pa izrazito arhitekturno specifična.

V fazi zasnove naj bo osrednja kontrolna točka jasna opredelitev **namena uporabe** (*intended purpose*) in vloge sistema v konkretnem postopku. To vključuje vprašanja, komu je sistem namenjen, ali gre za podporo odločanju ali za neposredno avtomatizirano obdelavo, katere odločitve sistem ne sme sprejeti, ali sistem neposredno komunicira z naravnimi osebami ter ali se use-case približuje področjem iz Priloge III Akta o UI. V isti fazi je treba opredeliti tudi osnovni podatkovni režim: katere kategorije podatkov so dopustne, katere so prepovedane ali omejene, in ali predvidena uporaba zahteva dodatno presojo vplivov oziroma posebno notranjo odobritev.

V fazi razvoja in priprave podatkov naj kontrolni seznam zahteva sistematično upravljanje vhodov in minimizacijo podatkov. Pri LLM/RAG sistemih to pomeni pravila za proste besedilne vnose, dokumentne baze, nalaganje datotek, klasifikacijo vsebin, filtriranje in roke hrambe. Pri ML sistemih pa pomeni selekcijo podatkovnih polj, standardizacijo formatov, časovno minimizacijo ter jasno razmejitev med lokalno obdelavo in izmenjavo modelnih artefaktov. Operativno to zahteva vsaj kartiranje virov, klasifikacijo vhodov, pravila za dopustne vsebine ter dokumentirane odločitve o tem, kateri podatki lahko vstopijo v sistem in zakaj.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

V fazi testiranja in predprodukcijske presoje naj se kontrolni seznam osredotoči na tri sklope: kakovost, pristranskost in varnost modela.
Pri kakovosti je treba določiti merila uspešnosti, testne scenarije in način dokumentiranja omejitev sistema. Pri pristranskosti je treba opredeliti potencialne vire neželenih korelacij ter, kjer je to pravno in tehnično izvedljivo, predvideti način preverjanja in dokumentiranja rezultatov. Pri varnosti modela pa je treba posebej nasloviti tiste grožnje, ki jih klasična IT-varnost ne zajame dovolj: pri LLM sistemih npr. prompt injection, zlorabe vhodov ali nenamerni vnos občutljivih vsebin, pri federativnem učenju pa varnost modelov, parametrov, modelnih posodobitev in protokolov izmenjave.

V fazi uvedbe naj kontrolni seznam preverja infrastrukturne in organizacijske predpogoje za varno in skladno uporabo. To vključuje izbiro okolja gostovanja, razmejitev vlog med upravljavcem, uporabnikom in dobaviteljem, način avtentikacije in avtorizacije, upravljanje dostopov, sledljivost uporabe, evidenco ključev, režim eskalacije ter načrt odzivanja na incidente. Pri platformnih rešitvah, kot je Albert API, je posebej pomembno preveriti, ali so platformne kontrole skladne z dejanskim use-caseom administracije. Pri SaaS rešitvah, kot v primeru Helsinkija, pa je ključno, da so dobaviteljski model, lokacija obdelave in meje vloge zunanjega ponudnika dovolj jasno razumljeni.

V fazi obratovanja naj bo kontrolni seznam usmerjen v tri stalne obveznosti: spremljanje uporabe, izvajanje človeškega nadzora in upravljanje sprememb. Spremljanje uporabe naj vključuje ločitev med vsebino interakcij in telemetrijo, roke hrambe, postopke brisanja ali anonimizacije ter pregled agregiranih kazalnikov. Človeški nadzor mora biti organiziran tako, da nepravilni odgovori, vsebinske vrzeli ali neustrezna priporočila dejansko pridejo do odgovornih strokovnjakov. Upravljanje sprememb pa mora zajemati tako tehnične spremembe kot spremembe namena uporabe, saj lahko prav slednje pomenijo prehod v drugačen regulatorni ali tveganostni profil.
Zato priporočljivo, da se končni operativni kontrolni seznam strukturira v:

- **skupni del**, ki velja za vse UI sisteme, ter
- **dodatna modula**, ločena za LLM/RAG in za ML/detekcijske sisteme.

Skupni del naj zajema najmanj: namen uporabe, klasifikacijo podatkov, infrastrukturo, razmejitev odgovornosti, logging, hrambo, kakovost, človeški nadzor in upravljanje sprememb. Modul za **LLM/RAG** naj dodatno naslavlja prompts, dokumentne baze, halucinacije, vsebinske omejitve in vhodne filtre. Modul za **ML/detekcijske sisteme** pa naj poudari standardizacijo podatkov, modelne artefakte, varnost protokola učenja, lažno pozitivne rezultate ter tveganja, povezana z modelno in ne le podatkovno varnostjo.

Takšna zasnova bi po eni strani zvesto odražala ugotovitve iz analiziranih primerov, po drugi strani pa bi bila dovolj operativna, da jo je mogoče uporabiti kot osnovo za naslednjo iteracijo dokumenta in za nadaljnjo pripravo konkretnih zahtev, smernic in kontrolnih točk v okviru projekta NOSI-AI.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

5 Zaključek

To poročilo je bilo pripravljeno z namenom, da za potrebe projekta NOSI-AI identificira in analizira take primere uporabe umetne inteligence iz evropskega javnega sektorja in regulatornih peskovnikov, ki so hkrati dovolj verodostojni, vsebinsko bogati in operativno uporabni, da iz njih lahko izpeljemo prenosljive dobre prakse za slovenski javni sektor. V tem smislu dokument ni zasnovan kot splošen pregled aktualnih trendov na področju umetne inteligence, temveč kot strokovno utemeljen nabor referenčnih primerov, iz katerih je mogoče izluščiti konkretne zahteve, opozorila in izvedbene usmeritve za nadaljnje delo.

Analiza štirih izbranih primerov je pokazala, da se vprašanje odgovorne in skladne uporabe umetne inteligence v javnem sektorju ne rešuje na eni sami ravni. Noben obravnavani primer namreč ne potrjuje poenostavljene predstave, da bi bilo mogoče ustreznost sistema zagotoviti zgolj z izbiro "boljšega modela", z dodatnim pravnim aktom ali z enkratnim varnostnim pregledom. Nasprotno: v vseh primerih se kot odločilna pokaže potreba po večplastni zasnovi, ki povezuje namen uporabe, podatkovne tokove, infrastrukturo, človeški nadzor, upravljanje vhodov, spremljanje kakovosti in upravljanje sprememb. Prav v tem je tudi glavna vrednost analiziranih primerov za projekt NOSI-AI: pokažejo, da se dobra praksa pri umetni inteligenci ne začne šele pri modelu, temveč pri celotnem režimu njegove uporabe. Posebej pomembna ugotovitev dokumenta je, da se ključna tveganja v sodobnih sistemih umetne inteligence pogosto ne pojavljajo tam, kjer bi jih pričakovali po logiki klasičnih informacijskih sistemov. Pri generativnih sistemih se težišče pomembno premika na upravljanje vhodov, promptov, dokumentnih baz in vsebinskih omejitev. Pri federativnem učenju se del tveganj premakne iz področja deljenja podatkov v področje deljenja modelov, modelnih posodobitev in varnosti protokolov izmenjave. Pri platformnih rešitvah, kakršna je Albert API, pa se kot ključno vprašanje pokaže, kako zgraditi "odobreno okolje", v katerem administracije ne prevzemajo vsaka zase celotnega bremena infrastrukture, vendar hkrati ohranijo odgovornost za konkretne načine uporabe. Iz tega izhaja širši sklep: tehnološka arhitektura ni nevtralna izvedbena podrobnost, temveč eden nosilnih dejavnikov skladnosti, varnosti in obvladljivosti sistema.

Dokument obenem potrjuje, da je v javnem sektorju posebej pomembno razlikovati med sistemi, pri katerih je težišče predvsem na transparentnosti in kakovosti storitve, ter sistemi, ki po svoji funkciji, kontekstu ali možnih učinkih posegajo v občutljivejša področja odločanja. Vendar analiza hkrati pokaže, da je tudi pri navidezno manj tveganih sistemih – denimo javnih chatbotih ali skupnih platformah za generativno UI – nujno že vnaprej vzpostaviti temeljne mehanizme upravljanja, ker se lahko prav prek širjenja namena uporabe ali postopne operativne rutinizacije tveganost sistema bistveno spremeni. Zato so v poročilu preliminarne navezave naAkt o UI obravnavane previdno in metodološko zadržano: ne kot

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)

dokončne pravne kvalifikacije, temveč kot pomoč pri strukturiranju razmisleka o tem, kdaj je treba posamezen sistem obravnavati strožje in bolj previdno.

Na tej podlagi je mogoče kot osrednji rezultat poročila razumeti dvoje. Prvič, dokument prinaša vsebinsko sintezo dobrih praks, ki se ponavljajo čez različne primere: jasno razlikovanje med vsebino in telemetrijo, dejanska operacionalizacija človeškega nadzora, upravljanje vhodov kot samostojno področje, infrastrukturna in dobaviteljska transparentnost ter ločena obravnava modelne varnosti. Drugič, te ugotovitve so bile prevedene v predlog operativnih kontrolnih seznamov, ki predstavljajo pomemben most med analitičnim delom poročila in njegovo praktično uporabnostjo v nadaljnjih fazah projekta.
Prav zato je smiselno razumeti to poročilo kot vmesni, vendar vsebinsko ključni korak v širšem procesu priprave smernic in priporočil projekta NOSI-AI. Njegova vrednost ni le v opisu štirih dobrih primerov, temveč v tem, da iz njih oblikuje strukturiran okvir za nadaljnje delo: za pripravo operativnih checklist, za razvoj notranjih metod presojanja use-caseov, za oblikovanje zahtev glede infrastrukture in vhodov, ter za bolj enotno razumevanje, kaj v praksi pomeni odgovorna uporaba umetne inteligence v javnem sektorju.

Za nadaljnje delo je zato priporočljivo, da se v naslednji iteraciji projekta operativni kontrolni seznami iz poglavja 5 dodatno razvijejo v uporabna delovna orodja, prilagojena potrebam javnih organov in nosilcev konkretnih primerov uporabe. S tem bi se analiza iz tega dokumenta neposredno prevedla v instrumente, ki ne bodo služili zgolj kot referenčno gradivo, temveč kot praktična podlaga za načrtovanje, preverjanje in uvajanje sistemov umetne inteligence v slovenskem javnem sektorju.

CRP NOSI-AI (št. V2-24065), ki ga (so)financirata Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost RS (ARIS) in Ministrstvo za digitalno preobrazbo (MDP)