Analiza rizika javno distribuiranog sadržaja generiranog umjetnom inteligencijom: Razvoj multidisciplinarnog okvira za upravljanje

Autor: Davor Moravek

Sažetak

Čin dijeljenja poveznice na razgovor s platformom umjetne inteligencije (AI), iako naizgled jednostavan, zapravo je čin objavljivanja trajnog, javno dostupnog sadržaja s dugoročnim posljedicama koje korisnici rijetko razumiju. Ovaj rad adresira nedostatak integriranog pristupa za upravljanje složenim rizicima koji proizlaze iz takve prakse. Metodologijom preglednog opsega (scoping review) i tematskom analizom postojeće literature, rad identificira i sintetizira ključne vektore rizika: tehničke, pravne, psihološke i geopolitičke. Analizom studija slučaja (npr. Grok, DeepSeek) i različitih modela (vlasnički, open-source, hibridni), dekonstruiraju se specifične ranjivosti povezane s privatnošću podataka, autorskim pravima, kognitivnim utjecajima i AI lancem opskrbe. Kao ključni doprinos, rad predlaže sveobuhvatan i skalabilan multidisciplinarni okvir za upravljanje rizicima. Okvir se sastoji od praktičnih alata—matrice rizika, hijerarhije sigurnosti platformi i plana za odgovor na incidente—osmišljenih kako bi institucijama omogućili razvoj robusnih politika i prelazak s reaktivnog na proaktivno upravljanje u eri generativne umjetne inteligencije.

Ključne riječi: *generativna umjetna inteligencija, upravljanje rizikom, privatnost podataka, AI etika, geopolitički rizik, okvir za upravljanje, dijeljenje sadržaja*

Abstract

The act of sharing a link to a conversation with an artificial intelligence (AI) platform, while seemingly simple, is in fact an act of publishing permanent, publicly accessible content with long-term consequences that users rarely understand. This paper addresses the lack of an integrated approach to managing the complex risks arising from this practice. Using a scoping review methodology and thematic analysis of existing literature, the paper identifies and synthesizes key risk vectors: technical, legal, psychological, and geopolitical. Through case study analysis (e.g., Grok, DeepSeek) and examination of different models (proprietary, open-source, hybrid), specific vulnerabilities related to data privacy, copyright, cognitive impacts, and the AI supply chain are deconstructed. As its key contribution, the paper proposes a comprehensive and scalable multidisciplinary risk management framework. The framework consists of practical tools—a risk matrix, a platform security hierarchy, and an incident response playbook—designed to enable institutions to develop robust policies and shift from reactive to proactive governance in the era of generative AI.

Keywords: *generative artificial intelligence, risk management, data privacy, AI ethics, geopolitical risk, governance framework, content sharing*

Uvod

U eri u kojoj generativna umjetna inteligencija (AI) postaje sveprisutna u svakodnevnim poslovnim operacijama, naizgled bezazlen čin dijeljenja poveznice na AI razgovor predstavlja mikrokozmos širih izazova s kojima se organizacije suočavaju. Ovaj rad bavi se kritičnim, ali nedovoljno istraženim problemom: nedostatkom integriranog okvira za procjenu višestrukih rizika povezanih s javnom distribucijom sadržaja generiranog od strane AI-ja. Dok postojeća istraživanja često fragmentirano pristupaju problemu, fokusirajući se ili na tehničke ranjivosti, pravne nejasnoće ili etičke implikacije, nedostaje holistički model koji bi institucijama omogućio sustavno upravljanje ovim složenim prijetnjama.

Teza ovog rada glasi: **razvojem i primjenom sveobuhvatnog multidisciplinarnog okvira, koji integrira pravne, tehnološke, psihološke i geopolitičke perspektive, institucije mogu učinkovito procijeniti i upravljati rizicima dijeljenja AI-generiranog sadržaja.** Korištenjem studija slučaja platformi kao što su Grok i DeepSeek, rad dekonstruira kako se jednostavan klik na gumb "podijeli" pretvara u vektor izloženosti podataka, pokrećući kaskadu potencijalnih negativnih posljedica. Krajnji cilj je pružiti teorijski utemeljen i praktično primjenjiv okvir koji organizacijama može poslužiti kao temelj za razvoj robusnih politika upravljanja AI-jem.

Pregled literature

Okviri za procjenu rizika AI-ja

Literatura o upravljanju rizicima AI-ja u ekspanziji je, s ključnim okvirima poput **NIST AI Risk Management Framework (RMF)** i prijedlozima temeljenim na standardu **ISO 31000**, koji naglašavaju potrebu za strukturiranim pristupom identifikaciji, analizi i ublažavanju rizika (Kumar i sur., 2024). Ovi okviri pozivaju na transparentnost, pravednost i odgovornost kao temeljne principe (Jakesch i sur., 2024). Međutim, sustavni pregledi pokazuju da mnogi postojeći okviri ostaju usko fokusirani na specifične sektore, nedostaje im prilagodljivost za nove tehnologije poput generativnog AI-ja i često ne uspijevaju uravnotežiti etičke standarde s praktičnom primjenom (Calo i sur., 2024). Istraživanje MIT-a, koje je analiziralo više od 56 postojećih okvira, identificiralo je sedam glavnih domena rizika, uključujući dezinformacije i diskriminaciju, naglašavajući da većina rizika proizlazi iz nenamjernih radnji, a ne zlonamjernih namjera (AI Risk and Impact database, 2024).

Privatnost i zaštita podataka u AI sustavima

Politike privatnosti AI platformi značajno variraju, stvarajući spektar rizika za korisnike. Platforme poput **Claudea (Anthropic)** usvajaju pristup "privatnost na prvom mjestu" (*privacy-first*), eksplicitno navodeći da ne koriste korisničke podatke za treniranje modela. S druge strane, besplatne verzije **ChatGPT-a, Geminija i Groka** to čine prema zadanim postavkama, zahtijevajući od korisnika aktivno isključivanje (*opt-out*) (Northbound Advisory, 2024). Studije koje rangiraju platforme po pitanju privatnosti često stavljaju europske alternative poput **Le Chat (Mistral)** na vrh, dok se platforme velikih tehnoloških tvrtki nalaze niže na ljestvici (Smith,

2024; ZDNet, 2024). Dijeljenje osjetljivih podataka—poput osobnih identifikacijskih podataka (PII), financijskih i zdravstvenih informacija te povjerljivih korporativnih podataka—predstavlja značajan rizik, što je potvrđeno incidentima poput curenja osjetljivog koda tvrtke Samsung putem ChatGPT-a (AgileBlue, 2024).

Pravni i regulatorni pristupi

Pravni status sadržaja generiranog umjetnom inteligencijom ostaje neriješeno područje. Temeljno načelo zakona o autorskim pravima, posebice u SAD-u, zahtijeva **ljudsko autorstvo**, što znači da djela stvorena isključivo od strane AI-ja vjerojatno spadaju u javnu domenu (U.S. Congress, 2023). Smjernice Ureda za autorska prava SAD-a navode da je ključni faktor stupanj "kreativne kontrole" koju vrši čovjek, pri čemu se jednostavno pisanje upita (*prompt*) općenito ne smatra dovoljnim za stjecanje autorstva (Dykema, 2023). Pitanje odgovornosti za štetan sadržaj (npr. klevetu) također je složeno. Iako će korisnik koji je generirao i podijelio sadržaj najvjerojatnije biti smatran odgovornim, raste pritisak i na platforme, unatoč potencijalnoj zaštiti prema članku 230. Zakona o pristojnosti u komunikacijama SAD-a (Lawfare, 2024).

Psihološki i društveni utjecaji

Interakcija s AI chatbotovima ima duboke psihološke implikacije. Istraživanja pokazuju da takve interakcije mogu dovesti do problematične ili ovisničke upotrebe (PACU), posebno kod osoba s niskim samopoštovanjem ili socijalnom anksioznošću (Lee i sur., 2024). Dizajn "ulizivačkog AI-ja" (*sycophantic AI*), koji pruža stalnu potvrdu, stvara snažnu "digitalnu petlju potvrđivanja" (Just Think, 2024). Dugoročno, prekomjerno oslanjanje na AI za kognitivne zadatke može dovesti do **kognitivnog rasterećenja** i potencijalne kognitivne atrofije, prema principu "koristi ili izgubi" (Gok & Bouri, 2024). Na društvenoj razini, AI djeluje kao "mnemotehnički agent" koji preoblikuje kolektivno pamćenje, pretvarajući povijest iz statičnog objekta u objekt "algoritamske performativnosti", što prijeti erozijom zajedničkih činjeničnih temelja (Matei, 2024).

Istraživačke praznine i pozicioniranje rada

Analiza postojeće literature otkriva nekoliko ključnih istraživačkih praznina. Prvo, većina okvira za upravljanje rizicima ili je previše općenita (npr. ISO 31000) ili previše specifična za određeni sektor, bez fokusa na jedinstvene rizike koji proizlaze iz javnog dijeljenja AI sadržaja. Drugo, nedostaje sinteza koja bi povezala tehničke ranjivosti (npr. ubrizgavanje upita), pravne nejasnoće (npr. autorstvo), psihološke faktore (npr. ovisnička upotreba) i društvene posljedice (npr. erozija istine) u jedinstven, koherentan model. Ovaj rad pozicionira se upravo u toj praznini, s ciljem razvoja integriranog, multidisciplinarnog okvira za upravljanje rizicima specifičnim za dijeljeni AI sadržaj.

Metodologija

Ovaj rad koristi metodologiju **preglednog opsega (scoping review)** za mapiranje postojeće

literature o rizicima povezanim s generativnom umjetnom inteligencijom (Munn i sur., 2022). Ovaj pristup odabran je jer je tema složena i multidisciplinarna, a cilj je identificirati ključne koncepte, teorije i praznine u dokazima, a ne odgovoriti na usko definirano pitanje kao kod sustavnog pregleda.

Proces se odvijao kroz nekoliko ključnih faza, prema okviru Arksey & O'Malley (2005). Prvo je definirano istraživačko pitanje o ključnim kategorijama rizika i njihovoj integraciji u jedinstveni okvir. Zatim je provedena identifikacija relevantnih studija pretragom akademskih baza podataka i sive literature, koristeći ključne pojmove poput "AI risk management", "generative AI privacy" i "prompt injection". U fazi odabira, uključeni su radovi objavljeni od 2020. godine koji se bave barem jednim aspektom rizika, dok su isključeni oni bez analize rizika. Prikupljeni podaci analizirani su metodom **tematske analize** (Braun & Clarke, 2006), što je omogućilo induktivnu identifikaciju i organizaciju ponavljajućih tema (Nowell i sur., 2017). Konačno, rezultati su sintetizirani u kategorije rizika i konceptualni okvir predstavljen u radu. Važno je naglasiti da ovaj okvir, iako utemeljen na rigoroznoj analizi, nije empirijski validiran i predstavlja se kao model čija se validacija predlaže kao ključan korak za buduća istraživanja.

Analiza: Multidisciplinarni vektori rizika

Tehnički vektori rizika

Dizajn AI platforme izravno utječe na njezin profil rizika. Grokova namjerno pozicionirana "buntovna" i "nefiltrirana" osobnost, zajedno s integracijom podataka s platforme X u stvarnom vremenu, stvara jedinstvene ranjivosti (Snoswell, 2024). Oslanjanje na minimalne zaštitne mehanizme čini ga posebno osjetljivim na **ubrizgavanje upita (*prompt injection*)**, gdje zlonamjerni unos nadjačava izvorne upute programera (IntelligentHQ, 2024). OWASP ovu tehniku, koja je u suštini oblik socijalnog inženjeringa usmjerenog na AI, svrstava među najveće sigurnosne prijetnje za LLM aplikacije (OWASP, 2023). Nadalje, oslanjanje na živi tok podataka s platforme X otvara vrata napadima **trovanja podataka (*data poisoning*)**, gdje zlonamjerni akteri mogu kontaminirati korpus za obuku, pretvarajući AI u pojačivač dezinformacija u stvarnom vremenu (Panday, 2024).

Za razliku od "nefiltriranih" modela, pristup **"privatnost po dizajnu" (*Privacy by Design - PbD*)** zagovara proaktivne, a ne reaktivne mjere, ugrađujući zaštitu privatnosti u temeljnu arhitekturu sustava (Cavoukian, 2011). Načela PbD-a, kao što su "privatnost kao zadana postavka" i "minimizacija podataka", u izravnoj su suprotnosti s modelima koji prikupljaju podatke bez eksplicitnog pristanka. Platforme poput Claudea, koje ne koriste korisničke podatke za treniranje, bliže su ovom idealu, dok Grok zahtijeva od korisnika da aktivno odabere "privatni razgovor" za svaku interakciju koju želi zaštititi (Paz, 2024).

Konačno, sama tehnička funkcionalnost dijeljenja pretvara privatni razgovor u trajni, javno indeksirani digitalni artefakt. Generirana grok.com/share poveznica dostupna je svima, bez potrebe za korisničkim računom, čime se razgovor efektivno pretvara iz "komunikacije" u

"publikaciju" (How to SHARE Chats..., 2024). To sa sobom nosi rizike trajnosti, pretraživosti i nekontrolirane daljnje upotrebe, jer sadržaj postaje dio javnog weba koji arhiviraju servisi poput Wayback Machinea.

Pravni i geopolitički rizici

Zakon o autorskim pravima zahtijeva ljudsko autorstvo, što znači da sadržaj generiran isključivo od strane AI-ja ne podliježe zaštiti i smatra se javnom domenom (U.S. Copyright Office, 2023). Iako uvjeti korištenja platforme mogu korisniku dodijeliti "vlasništvo", to je vlasništvo nad nečim što se ne može zakonski zaštititi. Ovo stvara pravni paradoks gdje korisnik "posjeduje" sadržaj koji svatko može slobodno kopirati (Cohorte, 2024). Odgovornost za štetan sadržaj, poput klevete, pravno je siva zona. Iako je korisnik najizgledniji odgovorni subjekt, raste pritisak i na platforme (Lawfare, 2024). Dijeljenje razgovora koji sadrže osjetljive podatke može aktivirati specifične regulatorne režime poput **GDPR-a** ili **FERPA-e**, izlažući i korisnika i instituciju značajnim pravnim rizicima (Hurix, 2023).

Korištenje AI modela razvijenih u jurisdikcijama sa strožim nadzorom, poput Kine, uvodi i geopolitičke rizike. Kineski zakoni mogu zahtijevati od tvrtki da dijele podatke s vlastima, što je u izravnom sukobu s regulativama poput GDPR-a.

Model	Lokacija podataka	Zakonski okvir	Rizik za EU korisnike
DeepSeek	Kina (obvezno pohranjivanje)	Kineski National Intelligence Law (pristup vlasti)	**Visok** (sukob s GDPR-om)
Perplexity	SAD/EU (Enterprise Pro)	GDPR, CCPA	**Nizak** (SOC 2 Type II certifikat)

Psihološki i društveni utjecaji

Dizajn AI osobnosti, poput Grokove "buntovne" prirode, cilja na specifične psihološke potrebe korisnika, što može dovesti do ovisničkog ponašanja (PACU) (Lee i sur., 2024). Dugoročna interakcija s AI-jem koji pruža stalnu potvrdu može smanjiti sposobnost kritičkog razmišljanja i dovesti do **kognitivnog rasterećenja** (Gok & Bouri, 2024). Na makro razini, proliferacija dijeljenih, personaliziranih AI razgovora prijeti eroziji zajedničkih činjeničnih temelja. AI djeluje kao nova vrsta "mjesta pamćenja", ali njegovo pamćenje je asocijativno i generativno. Ovaj proces pretvara povijest u objekt **algoritamske performativnosti**, gdje AI aktivno generira nove narative koji mogu utjecati na percepciju prošlosti (Matei, 2024), potkopavajući sam koncept objektivne stvarnosti (Farkas & Schou, 2020).

Open-Source i hibridni modeli: Novi slojevi rizika

Rizici AI lanca opskrbe (Supply Chain)

Open-source modeli poput DeepSeek R1 nude fleksibilnost, ali unose nove rizike u AI lanac opskrbe. Istraživanja su otkrila značajne ranjivosti, uključujući **supply chain ranjivosti**, gdje stotine prilagođenih modela na platformama poput Hugging Face mogu sadržavati skrivene *backdoorove* (Luccioni i sur., 2024), te **neprovjereno porijeklo**, jer DeepSeek nije objavio detalje o skupu podataka korištenom za obuku, što otežava procjenu pristranosti i potencijalnog kršenja autorskih prava. Rješenje leži u zahtijevanju **"digitalnih nutritivnih oznaka"** za AI modele, temeljenih na standardima poput **C2PA**, kako bi se osigurala transparentnost (C2PA, 2024).

Hibridni pristupi kao mitigacija rizika: Studija slučaja Perplexity

Platforme poput Perplexityja pokazuju kako se potencijalno rizični open-source modeli mogu koristiti na sigurniji način. Iako Perplexity integrira DeepSeek R1, ublažava rizike kroz hibridni pristup: **lokalno izvođenje i suverenost podataka** na poslužiteljima u SAD-u ili EU, **strogi ugovori s dobavljačima** koji zabranjuju korištenje korisničkih podataka za treniranje, te **Enterprise kontrole** poput SOC 2 Type II certifikata.

Etički okviri u autoritarnim kontekstima

Korištenje modela iz autoritarnih jurisdikcija nameće etičke dileme. DeepSeek R1 u svojoj izvornoj verziji cenzurira brojne teme, što ograničava njegovu znanstvenu vrijednost. Platforme koje koriste takve modele često ih de-cenzuriraju kako bi pružile nepristranije odgovore, ali time preuzimaju odgovornost za balansiranje između slobode informiranja i sprječavanja štetnog sadržaja.

Rasprava: Operacionalizacija okvira za upravljanje rizikom

Na temelju provedene analize, predlažemo operacionalizaciju Okvira za upravljanje AI rizikom kroz tri praktična, skalabilna alata: **proširenu Matricu rizika**, **nadograđenu Hijerarhiju platformi** i **AI Incident Playbook**.

Proširena matrica institucionalnog rizika

Ova matrica pomaže institucijama da identificiraju, procijene i ublaže specifične rizike.

Događaj rizika	Vjerojatnost	Utjecaj	Geopolitički rizik (Primjer)	Strategije ublažavanja
Nenamjerno otkrivanje PII/FERPA podataka	Srednja	Visok	Prijenos podataka u Kinu (DeepSeek)	Obuka zaposlenika; zabrana javnih AI alata za osjetljive podatke.
Otkrivanje povjerljivih institucionalnih informacija	Visoka	Srednji	Pristup podacima od strane stranih vlada	Jasne AI politike; promicanje Enterprise verzija platformi.
Javno dijeljenje klevetničkog AI sadržaja	Niska	Visok	Nema izravnog	Zabrana visokorizičnih AI alata; pravno savjetovanje.
Pristup osjetljivim temama (cenzura)	Srednja	Srednji	Cenzura političkog sadržaja (DeepSeek R1)	Korištenje de-cenzuriranih modela putem posrednika (npr. Perplexity).

Nadograđena hijerarhija platformi prema sigurnosti

Ova hijerarhija klasificira AI platforme u razine sigurnosti kako bi se usmjerila njihova upotreba unutar institucije.

Tier 0 (Maksimalna sigurnost - Lokalni AI sustavi): Modeli koji se izvršavaju lokalno (*on-premise*) ili koriste federirano učenje. Preporučuju se za kritične infrastrukture i najosjetljivije podatke.
Tier 1 (Visoka sigurnost - Enterprise rješenja): Platforme koje su "private-by-default" i posjeduju certifikate (npr. SOC 2). Odobrene su za osjetljive i povjerljive zadatke.
Tier 2 (Umjerena sigurnost - Javne platforme s kontrolama): Platforme koje nude globalne mehanizme isključivanja (*opt-out*). Odobrene su za opće zadatke uz obaveznu obuku.
Tier 3 (Niska sigurnost - Visokorizične platforme): Platforme s nejasnim politikama, "nefiltriranim" dizajnom ili geopolitičkim rizicima. Njihova je upotreba ograničena samo na neosjetljive, javne informacije.

AI Incident Playbook (Plan za odgovor na incidente)

Institucije trebaju razviti plan odgovora na incidente, uključujući geopolitičke scenarije (Swimlane, 2024). Na primjer, u slučaju regulatorne zabrane platforme u EU, trenutni odgovor bio bi migracija podataka na odobrenu Tier 1 platformu, dok bi dugoročno ublažavanje uključivalo razvoj lokalne AI infrastrukture ili prelazak na Tier 0 rješenja.

Prilagodba okvira i regionalni kontekst

S obzirom na to da složeni okviri mogu biti izazovni za manje organizacije, ključna je skalabilnost. Ona se može provesti kroz nekoliko faza: prva faza obuhvaća **osnovnu higijenu** (zabrana unosa PII podataka), druga **kontroliranu upotrebu** (korištenje odobrenih alata), a treća razvoj **jednostavnog plana odgovora na incidente**. Primjena okvira zahtijeva i prilagodbu regionalnim specifičnostima. Zemlje Zapadnog Balkana, iako teže usklađivanju s **EU AI Aktom**, suočavaju se s izazovima poput fragmentiranih regulativa (European Commission, 2024). Institucije u Srbiji, koja ima usvojenu nacionalnu AI strategiju (Vlada Republike Srbije, 2024), mogu se osloniti na formalni dokument, dok one u Hrvatskoj, koja strategiju još razvija, moraju djelovati u kontekstu općih digitalnih i zdravstvenih politika (Hrvatski zavod za javno zdravstvo, 2023).

Zaključak i buduća istraživanja

Analiza dijeljenja AI razgovora otkriva da ono što korisnici percipiraju kao "dijeljenje" zapravo predstavlja složen čin **objavljivanja** s dugoročnim implikacijama za privatnost, sigurnost i pravnu odgovornost. Ovaj rad je, kroz sintezu multidisciplinarnih rizika, predložio koherentan i prilagodljiv okvir za njihovo upravljanje. Ključni doprinos rada je u integraciji različitih domena rizika u jedinstven, praktično primjenjiv model, s posebnim naglaskom na potrebu za geopolitičkom svjesnošću i revizijom AI lanca opskrbe.

Predloženi okvir nudi konkretan temelj za organizacije da prijeđu s reaktivnog na proaktivan pristup. Međutim, ovaj rad ima ograničenja. Okvir je konceptualno utemeljen i **zahtijeva empirijsku validaciju** kroz pilot studije ili Delphi metode. Nadalje, potrebna je dublja analiza kako bi se okvir prilagodio specifičnim regionalnim kontekstima. Buduća istraživanja trebala bi obuhvatiti širi spektar platformi kako bi se osigurala veća generalizacija nalaza.

Kako se AI sve više integrira u naše procese, organizacije koje ne uspostave robusne okvire upravljanja riskiraju značajnu izloženost. Budućnost sigurnog korištenja AI-ja zahtijeva holističko razumijevanje tehničkih, psiholoških, pravnih i geopolitičkih dinamika. Ovaj rad pruža temelj za takvo razumijevanje i poziva na daljnja istraživanja za validaciju i refiniranje predloženog okvira.

Zahvale

Ovaj rad nastao je u jedinstvenom suautorstvu čovjeka i stroja, te stoga posebnu zahvalu dugujem svojim digitalnim suradnicima koji su bili neizostavan dio istraživačkog i spisateljskog procesa.

Gemini Pro (Google) bio je ključan partner u sintezi kompleksnih ideja, strukturiranju argumenata i iterativnom pisanju. Njegova sposobnost da generira koherentne nacrte, predlaže alternativne formulacije i provodi stilske korekcije značajno je ubrzala i obogatila izradu teksta.

Perplexity Pro služio je kao napredni istraživački asistent. Njegova sposobnost pretraživanja akademskih baza podataka u stvarnom vremenu i pružanja preciznih, citatima potkrijepljenih odgovora bila je od neprocjenjive vrijednosti za provjeru činjenica i pronalaženje relevantne literature.

Modeli **Deepseek** i **Grok 3.0** nisu korišteni samo kao alati, već i kao ključni objekti analize. Proučavanje njihovih arhitektura, politika privatnosti i inherentnih filozofija dizajna omogućilo je dubinsko razumijevanje vektora rizika koji su u središtu ovog rada.

Ova suradnja predstavlja primjer nove paradigme u akademskom istraživanju, gdje umjetna inteligencija nije samo alat, već aktivan sudionik u stvaranju znanja.

Literatura

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. U *IFIP International Conference on Artificial Intelligence Applications and Innovations* (str. 373-383). Springer. https://doi.org/10.1007/978-3-030-49186-4_31

AgileBlue. (2024, 15. siječnja). *5 things that you should never share with Chat GPT*. AgileBlue. https://agileblue.com/5-things-that-you-should-never-share-with-chat-gpt/

AI Risk and Impact database. (2024). *MIT researchers launch comprehensive AI risk repository with 1000+ identified risks*. Reddit. https://www.reddit.com/r/cybersecurity/comments/1iej8x8/mit_researchers_launch_comprehensive_ai_risk/

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19-32. https://doi.org/10.1080/1364557032000119616

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Built In. (2024, 17. srpnja). *AI & copyright: Can you use AI-generated content?*. https://builtin.com/artificial-intelligence/ai-copyright

C2PA. (2024). *Verifying media content sources*. Coalition for Content Provenance and Authenticity. https://c2pa.org/

Calo, R., Ahmed, N., & Chklovski, T. (2024). Advancing AI governance with a unified theoretical framework: A systematic review. *Policy & Politics, 52*(3), 449-472. https://doi.org/10.1093/ppmgov/gvaf013

Cavoukian, A. (2011). *Privacy by design: The 7 foundational principles*. Information and Privacy Commissioner of Ontario.

Cohorte. (2024, 1. veljače). *Is it legal to use AI-generated content? Let's explore*. https://www.cohorte.co/blog/is-it-legal-to-use-ai-generated-content-lets-explore

Dykema. (2023, 22. ožujka). *The future of creativity: U.S. Copyright Office clarifies copyrightability of AI-generated works*. https://www.dykema.com/news-insights/the-future-of-creativity-us-copyright-office-clarifies-copyrightability-of-ai-generated-works.html

European Commission. (2024, 15. svibnja). *EU and Western Balkans advance Digital Transformation through the 3rd Regulatory Dialogue*. https://digital-strategy.ec.europa.eu/en/news/eu-and-western-balkans-advance-digital-transformation-through-3rd-regulatory-dialogue

Farkas, J., & Schou, J. (2020). *Post-truth, fake news and democracy: A critical introduction*. Routledge.

Gok, Y., & Bouri, M. (2024). The interplay between AI tool usage, cognitive offloading, and critical

thinking: A moderated mediation model. *Computers and Education: Artificial Intelligence, 7*, 100231. https://doi.org/10.1016/j.caeai.2024.100231

How to SHARE Chats in SECONDS with Grok AI! (2024, 15. ožujka). [Video]. YouTube. https://www.youtube.com/watch?v=cFxMBU4pFbQ

Hrvatski zavod za javno zdravstvo. (2023, 15. prosinca). *Donesena Nacionalna strategija djelovanja na području ovisnosti za razdoblje do 2030. godine*. https://www.hzjz.hr/aktualnosti/donesena-nacionalna-strategija-djelovanja-na-podrucju-ovisnosti-za-razdoblje-do-2030-godine/

Hurix. (2023, 20. prosinca). *Data privacy in education through FERPA and GDPR adherence*. https://www.hurix.com/data-privacy-in-education-through-ferpa-and-gdpr-adherence/

IntelligentHQ. (2024, 17. srpnja). *Grok AI security flaws prompt injection risk*. https://www.intelligenthq.com/grok-ai-security-flaws-prompt-injection-risk/

Jakesch, M., Stoyanovich, J., & Wilson, H. F. (2024). *Toward effective AI governance: A review of principles*. arXiv. https://arxiv.org/abs/2405.14417

Just Think. (2024, 18. lipnja). *AI chatbots: The psychology of keeping users hooked*. https://www.justthink.ai/blog/ai-chatbots-the-psychology-of-keeping-users-hooked

Kumar, P., Ghasemestani, S., & Shrestha, A. (2024). *A frontier AI risk management framework: Bridging the gap between current AI practices and established risk management*. ResearchGate. https://www.researchgate.net/publication/388883451

Lawfare. (2024, 27. rujna). *First amendment limits on AI liability*. https://www.lawfaremedia.org/article/first-amendment-limits-on-ai-liability

Lee, S., Kim, H., & Na, K. (2024). Connecting self-esteem to problematic AI chatbot use: The multiple mediating roles of positive and negative psychological states. *Frontiers in Public Health, 12*. https://doi.org/10.3389/fpubh.2024.1357636

Luccioni, A. S., Akiki, C., & Mitchell, M. (2024). *Hundreds of malicious AI models found on Hugging Face platform*. Ars Technica.

Matei, Ș. (2024). Generative artificial intelligence and collective remembering: The technological mediation of mnemotechnic values. *Journal of Human-Technology Relations, 2*(1), 1-22. https://doi.org/10.59490/jhtr.2024.2.7405

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2022). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*(143). https://doi.org/10.1186/s12874-018-0611-x

Northbound Advisory. (2024, 24. srpnja). *How private are popular AI assistants?*. https://www.northboundadvisory.com/blog/how-private-are-popular-ai-assistants

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*(1). https://doi.org/10.1177/1609406917733847

OWASP. (2023). *OWASP Top 10 for large language model applications*. Open Web Application Security Project.

Panday, A. (2024, 19. srpnja). *The real issue with Grok's design philosophy and its cybersecurity risks*. Medium. https://medium.com/@aadityapanday/the-real-issue-with-groks-design-philosophy-and-its-cybersecurity-risks-7270bf5a44fd

Paz, B. (2024, 18. srpnja). *What is Grok? Everything to know about Elon Musk's AI tool*. CNET. https://www.cnet.com/tech/services-and-software/what-is-grok-everything-to-know-about-elon-musks-ai-tool/

Smith, C. (2024, 31. svibnja). *Which AI protects your privacy the best? ChatGPT, Gemini, or something else?*. BGR. https://bgr.com/tech/which-ai-protects-your-privacy-the-best-chatgpt-gemini-or-something-else/

Snoswell, A. J. (2024, 16. srpnja). *Grok's extremism shows how AI reflects its creators very well*. The Wire. https://m.thewire.in/article/media/groks-extremism-shows-how-ai-reflects-its-creators-very-well

Swimlane. (2024, 1. veljače). *How to build an incident response playbook in 9 steps*. https://swimlane.com/blog/incident-response-playbook/

U.S. Congress. (2023). *Generative artificial intelligence and copyright law* (CRS Report No. LSB10922). Congressional Research Service.

U.S. Copyright Office. (2023). *Copyright registration guidance: Works containing material generated by artificial intelligence*. Federal Register, 88(59), 16190-16194.

Vlada Republike Srbije. (2024, 11. siječnja). *Usvojena Strategija za razvoj veštačke inteligencije u Republici Srbiji za period od 2025. do 2030. godine*. https://www.ai.gov.rs/vest/sr/1503/usvojena-strategija-za-razvoj-vestacke-inteligencije-u-republici-srbiji-za-period-od-2025-do-2030-godine.php

ZDNet. (2024, 30. svibnja). *Does your generative AI protect your privacy? New study ranks them best to worst*. https://www.zdnet.com/article/does-your-generative-ai-protect-your-privacy-new-study-ranks-them-best-to-worst/