Pojavljanje incidentov ob uporabi Umetne Inteligence

Marko Grobelnik
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
marko.grobelnik@ijs.si

Besher M. Massri
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
m.besher.massri@gmail.com

Alenka Guček
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
alenka.gucek@ijs.si

Dunja Mladenić
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
dunja.mladenic@ijs.si

Povzetek

Prispevek predstavi prve rezultate ob uporabi sistema, ki je bil zasnovan in razvit v sodelovanju z OECD za spremljanje incidentov, povezanih z umetno inteligenco. Glavna motivacija teh prizadevanj je podpora zakonodaji, povezani z umetno inteligenco, in učinkovitemu oblikovanju politik, saj sistem zagotavlja vpoglede na podlagi zbranih podatkov. OECD AI Incidents Monitor za spremljanje incidentov, povezanih z umetno inteligenco, dokumentira incidente in nevarnosti v zvezi z umetno inteligenco, da bi oblikovalcem politik, strokovnjakom za umetno inteligenco in vsem zainteresiranim stranem po vsem svetu pomagal pridobiti dragocen vpogled v tveganja in škodo, ki jo povzročajo sistemi umetne inteligence. Ideja je, da bo sistem sčasoma pomagal povečati ozaveščenost javnosti in vzpostaviti skupno razumevanje incidentov in nevarnosti umetne inteligence, in tako prispeval k zaupanja vredni umetni inteligenci.

Ključne besede

umetna inteligenca, analiza podatkov, oblikovanje politik,incidenti

Abstract

This paper presents a system designed and developed in collaboration with OECD for monitoring of AI-related incidents. The main motivation behind the efforts is in supporting AI-related legislation and effective policymaking, as the system provides evidence based on the collected data. The OECD AI Incidents Monitor documents AI incidents and hazards to help policymakers, AI practitioners, and all stakeholders worldwide gain valuable insights into the risks and harms of AI systems. The idea is that over time the system will help to raise awareness and establish a collective understanding of AI incidents and hazards contributing to trustworthy AI.

Keywords

Artificial Intelligence, data analysis, policy making, AI incidents

1 Uvod

Ob vse širši uporabi umetne inteligence (UI) prihaja tudi do incidentov ob njeni uporabi. Spremljanje teh incidentov je nujno za zagotavljanje preglednosti, nadzora in razvoj politik, ki lahko takšne incidente preprečujejo ali vsaj zmanjšujejo. Predstavljeni sistem deluje kot orodje, ki pomaga uporabniku, ki si prizadeva v realnem času slediti dejanskim incidentom, povezanim z umetno inteligenco, ter zagotavljati dokazno osnovo za oblikovanje okvira poročanja o incidentih in povezanih političnih razpravah o UI. Z zbiranjem podrobnih vpogledov v vsak incident omogoča učenje iz preteklih napak ter spodbuja varnejši in bolj odgovoren razvoj ter uporabo umetne inteligence. Koristi skupnosti, ki se ukvarja z umetno inteligenco, saj izpostavlja trende in področja, ki potrebujejo pozornost ali regulativni poseg.

Prednost sistema je, da je zbiranje podatkov avtomatizirano, kar je prednost v primerjavi s podobnimi repozitoriji, ki so urednikovani ročno, kot je na primer AIAAIC Repository [2]. Repozitorij je prosto dostopen in namenjen tako oblikovalcem politik, kot razvijalcem UI, raziskovalcem, pravnikom in javnim organizacijam.

V nadaljevanju predstavimo metodologijo za spremljanje incidentov, prikažemo delovanje sistema na nekaj realnih primerih, predstavimo deležnike in nekaj zaključkov.

2 Metodologija

Metodologija OECD za spremljanje AI incidentov se osredotoča na identifikacijo in klasifikacijo incidentov, s čimer zagotavlja vpogled v realno dogajanje in podpira razvoj okvira za poročanje o incidentih. Začetna točka je identifikacija in klasifikacija incidentov, ki so poročani v uglednih mednarodnih medijih, s pomočjo modelov strojnega učenja, kar omogoča gradnjo zanesljive baze podatkov (incidenti so zajeti od 2014 naprej).

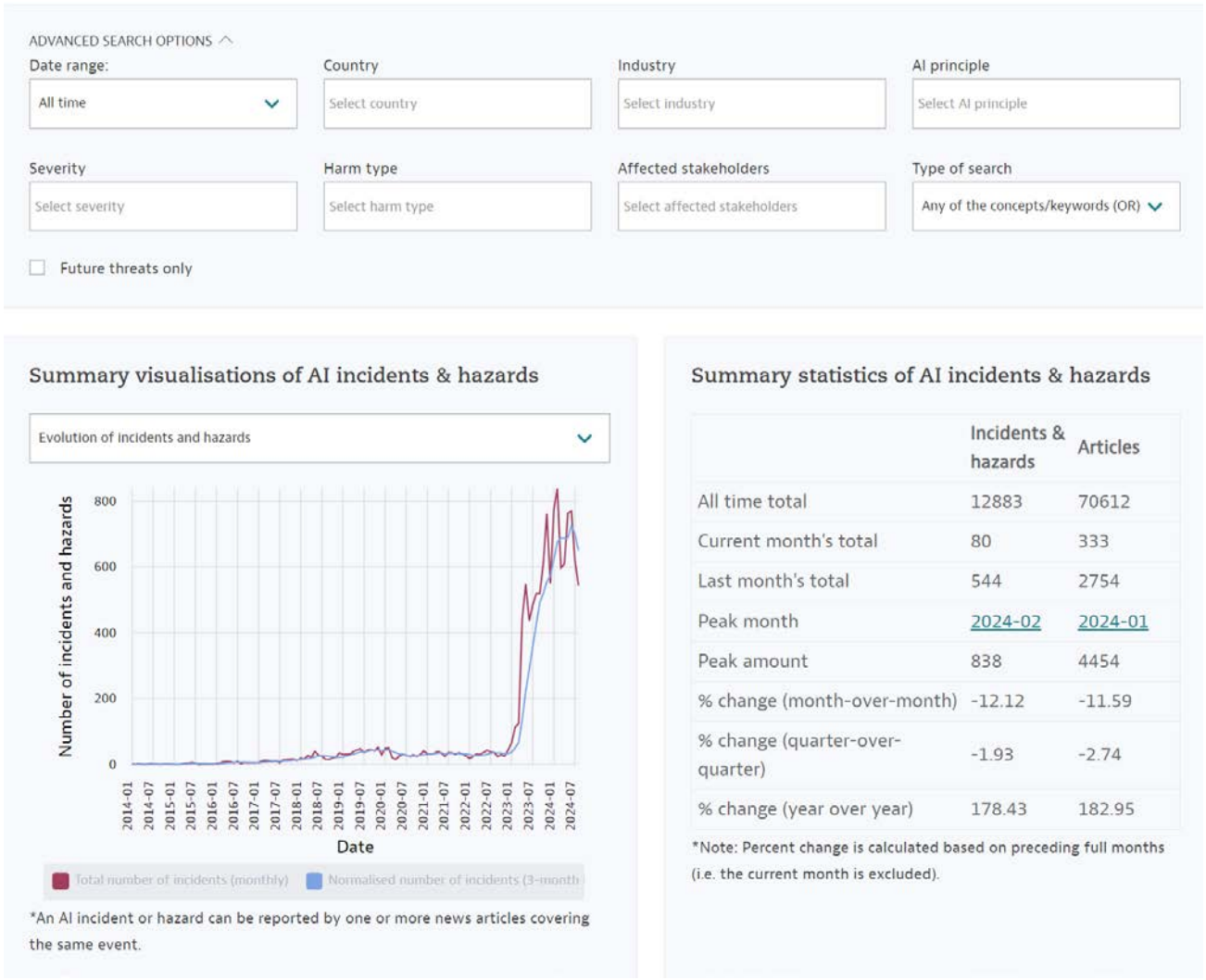
Kljub prizadevanjem, ti incidenti predstavljajo le podmnožico vseh globalnih AI incidentov. Incidenti so razvrščeni glede na resnost, industrijo, povezane AI principe (OECD AI Principles [3]), vrste škode in prizadete deležnike. Analiza temelji na naslovih, povzetkih in prvih odstavkih novinarskih člankov, pri čemer se pridobljeni podatki uporabljajo za izgradnjo zanesljive, objektivne in kakovostne baze podatkov o incidentih, povezanih z AI. Kot vir novic služi sistem Event Registry [4].

Razvoj sistema, h kateremu smo prispevali, nadgrajuje delo mednarodne skupine strokovnjakov (OECD Expert group), ki razvija teoretično ogrodje za poročanje o incidentih, definira pojem AI incidenta in oblikuje povezano terminologijo, kot je AI nevarnosti in njene potencialne posledice. Podrobna metodologija in definicije so razložene na spletni strani OECD: https://oecd.ai/en/incidents-methodology.

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.26



Slika 1 Prikaz začetne strani OECD monitorja AI incidentov (https://oecd.ai/en/incidents) prikazuje vmesnik za iskanje po konceptih, vizualizacijo incidentov v času (spodaj levo; y os: število incidentov; x os: čas (2014-danes) in statistični povzetek (spodaj

3 AI Incidents Monitor

AI Incidents Monitor je do konca avgusta 2024 zaznal preko 12 000 incidentov in nevarnosti v zvezi z UI, Kot je razvidno iz Slike 1. Sistem je popolnoma avtomatski in zaznava incidente s skeniranjem številnih podatkov objavljenih v novicah, ter nato s pomočjo UI določa kaj se označi kot incident ali nevarnost. Na naslovni strani (Slika 1) je prikazan črtni diagram naraščanja incidentov v času (levo) in pripadajoča statistika (desno). Uporabnik lahko izbira med absolutnim prikazom incidentov (kot na Sliki 1) ali v ustreznem meniju izbere pod-področja. Če se poglobimo v prikaz na Sliki 1, vidimo z različnimi barvami označeni kumulativni incidenti (vijolična) oz. njihovo trimesečno povprečje (modra). Statistika na desni prikazuje absolutno število incidentov glede na izbrano področje (12 883 incidenti in nevarnosti o katerih je poročalo 70612 novičarskih člankov), statistiko za zadnji mesec in mesece z največjimi vrednostmi (februar 2024). Iz statistik o spremembi glede na mesec, na četrtletje in na leto, vidimo padec števila incidentov in nevarnosti o katerih so mediji poročali v zadnjem mesecu glede na prejšnji mesec oz. prejšnje četrtletje.

3.1 Primer analize pojavitve incidentov UI

Sistem omogoča napredno filtriranje po incidentih UI za sledeče kategorije: čas, država, industrija, princip UI, resnost, tip škode, oškodovanci, tip iskanja po vsebini (glej Sliko 1). Tako so na primer možne vrednosti za resnost: smrt, poškodba, nevarnost, nefizična nevarnost, možni tipi škode pa so: fizična, psihološka, ekonomska, ugled, javni interes, človekove pravice, neznana.

Sistem omogoča napredno iskanje po konceptih, recimo za primer generativne UI, sistem poroča statistike, ki kažejo 2302 incidenta in nevarnosti, en od primerov incidentov, ki jih je sistem zaznal pa se nanaša na Apple in razvoj »AI personality«, ki naj bi nadomestil obstoječi Applov Siri.

Poleg konceptov uporabnik lahko nadalje izbere tudi napredno iskanje za natančnejšo identifikacijo želene podskupine incidentov, ki ga zanimajo. Tako lahko recimo izbere državo, ki je povezana s poročanjem o incidentih in nevarnostih UI. Na Sliki 2 je tako prikazan primer iskanja po kategoriji države, za Slovenijo. Sistem najde dva incidenta, ki sta bila povezana s Slovenijo. Prvi incident se nanaša na Microsoftov povečan prispevek k emisiji CO2. Na prvi pogled ni očitna povezava s Slovenijo, ko pa pogledamo povezane novice naletimo na omembo Slovenije: *»…But the tech giant's electricity consumption last year rivaled that of a small European country—beating Slovenia easily.«* [6]. Vsak primer je tudi semantično označen. Tako je na Sliki 2 za prvi primer označena povezanost s principi UI učinkovitost, trajnostni razvoj. Microsoft s tem lahko prizadene več deležnikov: splošno javnost, podjetja, delavce, vlade (Affected Stakeholders, Slika 2). Poleg tega predstavlja nevarnost za okolje, javne interese in človekove pravice (Harm type, Slika 2). Klasificirano je kot ne-fizična nevarnost (Severity, Slika 2).

Iz podrobnih analiz, ki so zbrane v nedavnem poročilu *»Observatory of the social and ethical impact of artificial intelligence«* [5], je razvidno, da večina incidentov (96%) spada pod kategorijo ne-fizično nevarnih, a imajo lahko zelo resne psihološke in finančne posledice, vključujoč nadlegovanja, odvisnosti in škodo ugledu tako posameznikom kot tudi inštitucijam.

4 Deležniki

OECD-jev monitor incidentov AI (AIM) je dragoceno orodje, zasnovano za različne deležnike, ki sodelujejo pri razvoju, regulaciji in uporabi umetne inteligence. Potencialni uporabniki tega orodja vključujejo oblikovalce politik, razvijalce AI, raziskovalce, pravne strokovnjake in javne organizacije.

Oblikovalci politik lahko AIM uporabljajo za sledenje in analizo podatkov v realnem času o incidentih, povezanih z AI, po vsem svetu, kar jim pomaga pri oblikovanju informiranih in na dokazih temelječih predpisov. Zmožnost orodja za kategorizacijo incidentov glede na resnost, industrijo in vrste škode je ključna za razumevanje širših posledic tehnologij umetne inteligence in oblikovanje politik, ki zmanjšujejo tveganja.

Razvijalci AI in raziskovalci lahko koristijo AIM, da prepoznajo pogoste težave, povezane s sistemi umetne inteligence. S preučevanjem incidentov, zabeleženih v AIM, lahko izboljšajo svoje modele, da bi se izognili podobnim težavam in povečali varnost ter zanesljivost aplikacij umetne inteligence.

Pravni strokovnjaki lahko uporabljajo AIM za pridobitev vpogledov v spreminjajočo se pokrajino tveganj, povezanih z umetno inteligenco, kar bi lahko bilo koristno v pravnih primerih ali ocenah skladnosti. Razumevanje preteklih incidentov in njihovih pravnih posledic lahko usmerja razvoj robustnih okvirov upravljanja AI.

Nazadnje lahko javne organizacije in zagovorniške skupine uporabljajo AIM za spremljanje družbenih vplivov umetne inteligence, s čimer zagotavljajo, da so interesi javnosti zaščiteni. To lahko vključuje analizo vzorcev incidentov z umetno inteligenco za zagovarjanje boljše zaščite potrošnikov in etičnih standardov pri uvajanju AI.

5 Diskusija

V prispevku smo predstavili OECD-jev monitor incidentov umetne inteligence, pri razvoju katerega smo sodelovali. Sistem služi kot dober vir za širok nabor uporabnikov, ki želijo razumeti in upravljati tveganja, povezana s tehnologijami UI. Sistem se nadgrajuje z dodatnimi podatkovnimi viri.

V prihodnosti je predvideno, da bo omogočen odprt postopek oddaje podatkov, ki bo dopolnil informacije o incidentih, pridobljene iz trenutnih virov. Nadaljnje delo zajema tudi avtomatsko analizo podatkov o incidentih za namen bolj celovitega vpogleda. To vključuje avtomatsko odkrivanja vzorcev, kot so verižne reakcije ali učinki na več industrij hkrati. Za potrebe preverjanja resničnosti poročanih incidentov, bi lahko vključili kombiniranje informacij iz več neodvisnih virov in uporabljal algoritme za odkrivanje lažnih novic, kot tudi ročno preverjanje.

Zahvala

Delo, opisano v tem prispevku, so podprli OECD in številni mednarodni eksperti, Ministrstvo za digitalno preobrazbo in Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru CRP V2-2272 in V5-2264.

Acknolwedgements

The described work was supported by OECD and many os its international experts, Slovenian Ministry of Digital Transformation and Slovenian Research and Innovation Agency under CRP V2-2272 and V5-2264.

Literatura

[1] OECD AI Incidents Monitor (AIM), https://oecd.ai/en/incidents. August 2024

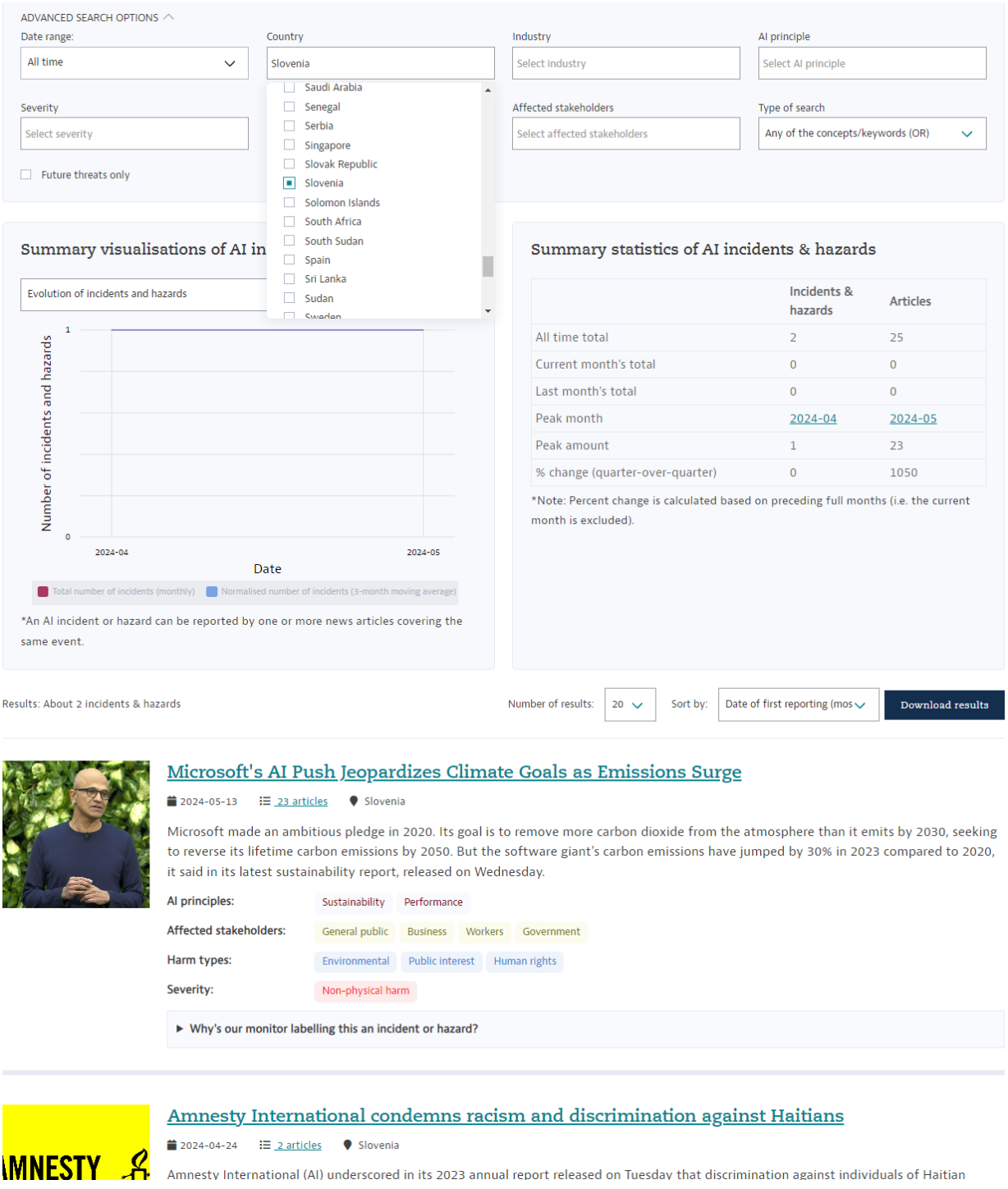
[2] AIAAIC Repositoryhttps://www.aiaaic.org/aiaaic-repository. August 2024

[3] OECD AI Principles for trustworthy AI https://oecd.ai/en/ai-principles August 2024

[4] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In Proceedings of the 23rd International Conference on World Wide Web, 107–110.

[5] Richard Benjamins, Another Inconvenient Truth: The Societal Emergency of AI Incidents - We Should Do Something About It https://www.odiseia.org/post/another-inconvenient-truth-the-societal-emergency-of-ai-incidents-we-should-do-something-about-it

[6] Microsoft's AI Push Imperils Climate Goal as Carbon Emissions Jump 30% https://tanaka-preciousmetals.com/en/elements/news-cred-20240821/



Slika 2 Prikaz naprednega iskanja na OECD monitorju AI incidentov (https://oecd.ai/en/incidents) filtrirano po državi za Slovenijo. Podane so statistike dveh incidentov o katerih je poročalo 25 novičarskih člankov, in spodaj sta prikazana oba incidenta.