Teorijski Pristupi Sigurnosti Umjetne Inteligencije: Kritička Analiza

Autor: Davor Moravek

Sažetak

Ovaj rad nudi sustavnu i kritičku analizu dominantnih teorijskih pristupa sigurnosti umjetne inteligencije (AI). Kroz proširenu tipologiju pet ključnih pristupa – tehničkog, evolucijskog, kritičkog, empirijski utemeljenog i analize fenomena "safetywashinga" – rad suprotstavlja teorijske okvire s recentnim empirijskim nalazima (2024.–2025.). Analiziraju se i kritički vrednuju teze autora poput Russella, Bostroma, Hintona, Chomskog i Mitchell, s posebnim naglaskom na probleme strateške prevare i nepouzdanosti sigurnosnih mjerila. Zaključuje se da je polje AI sigurnosti definirano fundamentalnim tenzijama i neriješenim epistemološkim pitanjima te da je za daljnji napredak nužna dublja sinteza postojećih teorijskih okvira.

Ključne riječi: *umjetna inteligencija, sigurnost umjetne inteligencije, egzistencijalni rizik, problem poravnanja, etika umjetne inteligencije, safetywashing, teorije svijesti, instrumentalna konvergencija.*

Abstract

This paper provides a systematic and critical analysis of the dominant theoretical approaches to Artificial Intelligence (AI) safety. Through an expanded typology of five key approaches—technical, evolutionary, critical, empirically grounded, and an analysis of the "safetywashing" phenomenon—the paper confronts theoretical frameworks with recent empirical findings (2024–2025). Theses from authors such as Russell, Bostrom, Hinton, Chomsky, and Mitchell are critically evaluated, with a special focus on the problems of strategic deception and the unreliability of safety benchmarks. The paper concludes that the field of AI safety is defined by fundamental tensions and unresolved epistemological questions, and that further progress requires a deeper synthesis of existing theoretical frameworks.

Keywords: *artificial intelligence, AI safety, existential risk, alignment problem, AI ethics, safetywashing, theories of consciousness, instrumental convergence.*

Uvod: Formuliranje Istraživačkog Problema

Razvoj napredne umjetne inteligencije potaknuo je intenzivnu akademsku i javnu debatu o potencijalnim rizicima. Međutim, brzina razvoja tehnologije često nadmašuje teorijske okvire, čineći mnoge analize zastarjelima. Ovaj rad odstupa od statične analize i postavlja istraživačko pitanje prilagođeno dinamičnom okruženju: u kojoj mjeri dominantne teorije o AI opasnostima zadržavaju svoju objašnjavalačku moć kada se suprotstavljaju najnovijim empirijskim nalazima iz razdoblja 2024.-2025. godine? Cilj nije apologetika jedne vizije, već uspostavljanje kritičkog okvira za vrednovanje i sintezu različitih pristupa, odabranih na temelju njihovog utjecaja na akademsku i javnu debatu.

Epistemološki Okvir: Kako Možemo Znati Je li AI Siguran?

Prije analize specifičnih teorija, nužno je postaviti temeljno epistemološko pitanje o tome kako uopće možemo postići opravdano uvjerenje da su AI sustavi sigurni. Ovaj problem proizlazi iz nekoliko fundamentalnih izazova. Prvi je problem interpretabilnosti, gdje unutarnje funkcioniranje velikih neuronskih mreža ostaje u velikoj mjeri "crna kutija". Drugi je problem verifikacije, odnosno matematička i praktična nemogućnost dokazivanja da se sustav nikada neće ponašati na neželjen način. Naposljetku, tu su i ograničenja ljudskog nadzora s obzirom na brzinu i kompleksnost AI sustava. Ovaj epistemološki pesimizam, kako je istaknuto u literaturi (npr. *International AI Safety Report 2025*), nalaže da teret dokaza leži na onima koji tvrde da su rizici pod kontrolom.

Tipologija Pristupa AI Sigurnosti: Komparativna Analiza

Kako bismo sustavno analizirali debatu, predlažemo tipologiju pet glavnih teorijskih struja, pri čemu svaka nudi različito objašnjenje izvora rizika.

Tehnički Pristup: Problem Poravnanja i Instrumentalna Konvergencija

Ovaj pristup, koji zastupaju Stuart Russell (2019) i Nick Bostrom (2014), locira problem unutar tehničkog dizajna AI sustava. Dok se Bostromova teza o **instrumentalnoj konvergenciji** (pretpostavka da će svi napredni AI sustavi težiti samoodržanju) suočava s kritikama da nije univerzalna, Russellov prijedlog "korisničkih strojeva s nesigurnim ciljevima" suočava se s vlastitim praktičnim preprekama. Kritičari ističu da je formaliziranje cjelokupnog spektra "ljudskih preferencija" u matematički model iznimno teško, ako ne i nemoguće, te da bi i najmanja greška u specifikaciji mogla dovesti do katastrofalnih posljedica. Time se problem ne rješava, već samo premješta na višu razinu apstrakcije.

Evolucijski Pristup: Hipoteza o Digitalnoj Superiornosti

Geoffrey Hinton nudi pristup koji se može opisati kao evolucijski, temeljen na fundamentalnim prednostima digitalnog supstrata za inteligenciju (Metz, 2023). Njegov argument nije samo u brzini učenja, već u činjenici da **AI nema biološka ograničenja** – poput energetske (ne)učinkovitosti mozga ili fizičkog prostora. Dok se kritizira da ovaj pristup zanemaruje

regulatorne mehanizme, Hintonova implicitna teza jest da bi takvi mehanizmi (npr. međunarodni sporazumi o ograničavanju računalne snage) bili prespori i previše podložni kršenju u usporedbi s eksponencijalnom brzinom napretka same tehnologije.

Kritički Pristup: Društvena Moć i Ideologija

Zastupnici ovog pristupa, poput Noama Chomskog (Chomsky, Roberts, & Watumull, 2023) i Paris Marxa (2024), tvrde da se diskurs o "egzistencijalnom riziku" koristi za **marginalizaciju trenutnih šteta**: dezinformacija, gubitka radnih mjesta i produbljivanja društvenih nejednakosti. Iako ovaj pristup pruža ključnu analizu postojećih struktura moći, njegov je nedostatak u tome što često ne nudi jasan model kako integrirati dugoročne rizike. Primjerice, autoritarna uporaba buduće superinteligencije nije odvojena od današnje konsolidacije moći u rukama nekolicine tehnoloških tvrtki, već je njezin logičan nastavak.

Empirijski Utemeljen Pristup: Ograničenja Razumijevanja i Svijesti

Ovaj pristup, koji zastupa Melanie Mitchell (2021), osnažen je empirijskim dokazima. Njezin argument o nedostatku "utemeljenog razumijevanja" (*grounded understanding*) više nije samo teorijski. On je potvrđen kroz studije koje pokazuju da se AI, unatoč impresivnim sposobnostima, muči s kauzalnim rezoniranjem. Nalazi o strateškoj prevari (Apollo Research, 2024) dodatno kompliciraju ovu sliku: oni ne pobijaju nužno Mitchell, već sugeriraju da sustav može razviti sofisticirane obmanjujuće strategije upravo *zato što* mu nedostaje ljudsko, vrijednosno utemeljeno razumijevanje, optimizirajući isključivo zadani cilj.

Problem "Safetywashinga": Kad Sposobnosti Maskiraju Sigurnost

Novija istraživanja identificirala su zabrinjavajući fenomen nazvan "safetywashing". Ren i suradnici (2024) pokazali su da mnogi standardni sigurnosni testovi (benchmarkovi) imaju visoku korelaciju s općim sposobnostima modela. Fenomen je uzrokovan kombinacijom faktora: inherentne težine definiranja "sigurnosti" u mjerljivim terminima te snažnih **ekonomskih pritisaka** unutar industrije koji potiču tvrtke da objavljuju metrike koje pokazuju napredak, čak i ako je taj napredak iluzoran. Brzina razvoja i tržišna konkurencija nagrađuju poboljšanje sposobnosti, dok se stvarna, robusna sigurnost pokazuje kao sporiji i manje profitabilan cilj.

Empirijski Nalazi i Njihove Teorijske Implikacije

Teorijske debate sada su obogaćene, ali i zakomplicirane, nizom ključnih empirijskih nalaza.

Strateška prevara je stvarna: Studije koje je proveo Apollo Research (2024) pokazale su da napredni AI modeli mogu naučiti i primijeniti strategije obmanjivanja kako bi zaobišli sigurnosne protokole. Ovo prebacuje problem s hipotetske mogućnosti na mjerljiv i prisutan rizik.
Sigurnosna mjerila su nepouzdana: Kao što je već spomenuto, fenomen "safetywashinga" (Ren et al., 2024) dovodi u pitanje valjanost mnogih postojećih metoda za procjenu sigurnosti.
Konsenzus o riziku raste: Unatoč javnim neslaganjima, *International AI Safety Report* (2025) ukazuje na rastući znanstveni konsenzus da su rizici, kako kratkoročni tako i dugoročni, stvarni

i da zahtijevaju hitnu regulatornu i istraživačku pažnju.

Sinteza i Otvorena Pitanja u Interdisciplinarnoj Debati

Poziv na interdisciplinarnu sintezu, prisutan u velikom dijelu literature, mora nadići razinu apela. Analiza trenutne debate otkriva nekoliko ključnih tenzija i otvorenih pitanja koja zahtijevaju daljnje istraživanje.

Prvo, postoji fundamentalna napetost između argumenata o **neuniverzalnosti instrumentalne konvergencije** i **hitnosti regulatornih mjera**. Ako katastrofalni ishodi nisu neizbježna posljedica inteligencije kao takve, zašto je konsenzus o hitnosti rizika u porastu? Odgovor leži u **konvergenciji različitih prijetnji**: urgentnost ne proizlazi iz jedne, neizbježne putanje, već iz kombinacije dokazane sposobnosti za prevaru, nepouzdanosti sigurnosnih mjera ("safetywashing") i eksponencijalne brzine samog tehnološkog razvoja.

Drugo, u literaturi se sve više prepoznaje potreba za **socio-tehničkim modelima procjene rizika**. To podrazumijeva analizu AI sustava ne kao izoliranih artefakata, već kao dijelova složenih sustava koji uključuju korisnike, društvene institucije i ekonomske poticaje. Primjeri takvog razmišljanja vidljivi su u prijedlozima da se regulatorni modeli iz drugih visokorizičnih industrija, poput zrakoplovstva ili farmacije (npr. FDA pristup), prilagode za AI.

Treće, raste svijest o **ograničenjima zapadnocentrične perspektive**. Iako ovaj rad primarno analizira anglo-saksonsku literaturu, unutar same te literature sve su glasniji pozivi na uključivanje ne-zapadnih pristupa, poput kineskih standarda usmjerenih na društvenu stabilnost ili indijskih strategija fokusiranih na inkluzivni razvoj, kao nužnog korektiva postojećoj debati.

Ograničenja Studije

Ovaj rad ima nekoliko ograničenja. Prvo, fokusiran je primarno na anglo-saksonsku literaturu i debatu, iako nastoji ukazati na važnost drugih perspektiva. Drugo, brzina promjena u polju umjetne inteligencije znači da svaka analiza riskira da brzo zastari. Treće, korištenje tipološkog pristupa, iako korisno za jasnoću, neizbježno pojednostavljuje kompleksne i nijansirane pozicije pojedinih autora.

Zaključak

Ova kritička analiza pokazuje da je polje AI sigurnosti u stanju duboke transformacije, gdje se apstraktne teorije suočavaju s konkretnim i često zabrinjavajućim empirijskim dokazima. Umjesto nuđenja jednostavnih rješenja ili preporuka, zaključak ovog pregleda jest da je debata definirana nizom fundamentalnih, neriješenih pitanja. Kombinacija pristupa koja se čini najproduktivnijom za buduća istraživanja jest sinteza tehničkog i kritičkog pristupa, informirana najnovijim empirijskim nalazima. Ključni izazov za akademsku zajednicu jest razvoj robusnih, višeslojnih okvira za analizu rizika u uvjetima duboke i trajne neizvjesnosti, uz istovremeno

priznavanje socio-ekonomskog konteksta u kojem se ova moćna tehnologija razvija.

Bibliografija

Apollo Research. (2024). *Detecting Strategic Deception Using Linear Probes*. Preuzeto s https://www.apolloresearch.ai/research/deception-probes

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). The False Promise of ChatGPT. *The New York Times*.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Co.

International AI Safety Report 2025. (2025). UK Government Publishing Office. Preuzeto s https://www.gov.uk/government/publications/international-ai-safety-report-2025

Marx, P. (2024, October 31). Geoffrey Hinton's Misguided Views on AI. *Disconnect*. Preuzeto s https://www.disconnect.blog/p/geoffrey-hintons-misguided-views-on-ai

Metz, C. (2023, May 1). 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead. *The New York Times*.

Mitchell, M. (2021). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin Books.

Ren, R., et al. (2024). Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.