Ali nas uporaba velikih jezikovnih modelov v znanstvenem raziskovanju približuje časovni točki, ko bo stroj nadvladal človeka?

Does the use of large language models in scientific research bring us closer to the point in time when machines will surpass humans?

Franc Mali
Faculty of Social Sciences
University of Ljubljana
Ljubljana, Slovenia
franc.mali@fdv.uni-lj.si

Povzetek

Prispevek se ukvarja z vprašanjem, ali veliki jezikovni modeli v okviru generativne umetne inteligence že danes odpirajo vrata v fazo splošne umetne inteligence in morda – kot naslednji korak – v fazo umetne superinteligence. S tem bi bili dani predpogoji za prevlado strojev nad ljudmi. Pozornost je namenjena zlasti uporabi velikih jezikovnih modelov v procesu znanstvenega raziskovanja. Raziskovalna dejavnost predstavlja eno najbolj ustvarjalnih človekovih intelektualnih dejavnosti. Logično vprašanje je, ali je ravno znanstvena dejavnost, predvsem zaradi svoje kreativne narave, najbližja prečkanju te meje, ki predstavlja pomembno eksistenčno tveganje za celotno človeštvo. Osrednji del razprave je namenjen vprašanju, v katerih fazah današnjega znanstvenega raziskovanja je vloga velikih jezikovnih modelov že postala nepogrešljiva.

Ključne besede

generativna umetna inteligenca, veliki jezikovni modeli, znanstvena kreativnost, eksistenčno tveganje, okrepljeno učenje

Abstract

The article addresses the question of whether large language models within the framework of generative artificial intelligence are already opening the door to the phase of artificial general intelligence and, perhaps, as the next step, to the phase of artificial superintelligence. This would set the conditions for machines to dominate humans. Particular attention is given to the use of large language models in the process of scientific research. Research activity represents one of the most creative human intellectual endeavors. The logical question arises whether scientific activity, especially due to its creative nature, is the closest to crossing this boundary, which poses a significant existential risk to all of humanity. The central part of the discussion focuses on the question of which phases of today's scientific research the role of large language models has already become indispensable.

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

https://doi.org/10.70314/is.2024.cog.9

Keywords

generative artificial intelligence, large language model, scientific creativity, existential risk, reinforcement learning

1 Uvod

V okviru pričujoče obravnave izhajam iz predpostavke, da se je skozi celoten zgodovinski razvoj umetne inteligence implicitno zastavljalo vprašanje, ali lahko ta doseže oziroma celo preseže človeško inteligenco. Že od začetkov razvoja umetne inteligence so bila tovrstna razmišljanja spodbujena z različnimi testi, ki naj bi med drugim nakazovali, ali se strojna "inteligenca" približuje človeški inteligenci. Pomembni premik v teh razmišljanjih se je zgodil, ko je tehnologija umetne inteligence prešla od klasičnih načel strojnega učenja k načelom delovanja globokih nevronskih mrež. V moji razpravi me v prvi vrsti zanima, ali najnovejši razvoj generativne umetne inteligence že kaže znake prehoda v fazo umetne splošne inteligence in morda – kot naslednji korak – umetne super inteligence. Posebej me zanima, ali najbolj kreativna področja človekovega intelektualnga delovanja, kot to predstavlja znanstveno raziskovanje, že odpirajo vrata nastopu umetne splošne inteligence. To namreč pomeni, da se počasi trasira pot nadvladi strojev nad človekom, kar je sicer predmet precej distopičnih razmislekov filozofov in družboslovcev, tako pri nas kot drugje v svetu. Moja obravnava ostaja na ravni nekoliko bolj splošne družboslovne refleksije o tej kompleksni tematiki in se ne ukvarja z ožjimi tehničnimi vidiki delovanja umetne inteligence, zato se bom v primeru sklicevanj na algoritme delovanja umetne inteligence oprl na nekoliko bolj poljudne definicije, kot so na primer tiste, ki jih je predstavil Partha Ray [1]. Po Rayu generativna umetna inteligenca (GUI) spada v skupino modelov umetne inteligence, ki lahko ustvarjajo nove podatke (informacije) na podlagi vzorcev in struktur, naučenih iz obstoječih podatkov (informacij). Ti modeli lahko

generirajo vsebine na najrazličnejših področjih, naj si bo besedil, slik ali glasbe. Pri analizi, razumevanju in ustvarjanju teh vsebin, ki vedno bolj spominjajo na človeške stvaritve, se opirajo na tehnike globokega učenja in nevronske mreže. Veliki jezikovni modeli (VJM), ki se razvijajo pod okriljem GUI, pa so zasnovani za generiranje naravnega jezika, kot so stavki, odstavki ali celotni dokumenti. Njihova ključna lastnost je zmožnost predhodnega učenja na velikih količinah besedilnih podatkov ter nato prilagajanje za specifične naloge uporabnikov. V prispevku v štirih krajših poglavjih razpravljam (1) o umetni splošni oziroma super inteligenci kot dejavniku tveganja o človeku, (2) o dilemah, ki so povezane z nadvlado stroja nad človekom, (3) o danes vedno bolj nepogrešljivi vlogi VJM v posameznih fazah znanstvenega raziskovanja, (4) o specifičnih problemih uporabe VJM na področju družboslovnega raziskovanja. Na koncu prispevka je podanih še nekaj zaključnih misli.

2 Umetna super inteligenca kot dejavnik eksistenčnega tveganja za človeka

Potem ko je Nick Bostrom pred desetimi leti postavil in utemeljil tezo, da obstaja verjetnost, da bo nadaljnji razvoj umetne inteligence pripeljal do nastopa umetne superinteligence, ki naj bi bila neprimerno bolj kognitivno zmožna kot človek, kar bi lahko predstavljalo eksistenčno tveganje za celotni človeško vrsto, ta tema, zlasti po nastopu GPT 4 in drugih vrst VJM (Bard, Claude, Llama, Gemini, itd.) vzbuja vedno večjo pozornost med strokovnjaki, tako med naravoslovci in tehniki kot tudi med družboslovci in humanisti [2]. V svojem prispevku se bom izognil (spekulativnim) ocenam, ki se vrtijo okrog problema časovnih mejnikov, ko (če) naj bi pametni stroji nadvladali ljudi. Ena skupina ekspertov namreč trdi, da se to ne bo zgodilo niti v sto letih [3], druga skupina ekspertov spet trdi, da gre zgolj za vprašanje dveh ali treh desetletij [4]. Bolj kot to, me zanima, ali uporaba GUI v takšni kreativni človekovi dejavnosti kot je znanost resnično na široko odpira vrata nastopu umetne splošne oziroma umetne super inteligence, ki naj bi se sicer zgodila v bližnji prihodnosti. To vprašanje je treba povezati s konkretno prisotnimi strahovi pred katastrofičnimi in celo eksistenčnimi tipi tveganj GUI, ki bi lahko imeli negativne družbene posledice. Če katastrofično tveganje ocenjujemo po kriteriju maksimalne razširjenosti (število ljudi, ki bi bili prizadeti), intenziteti (trpljenju, ki ga povzroča) in trajanja škodljivih družbenih posledic nekontroliranega razvoja posamezne tehnologije, potem pri eksistenčnem tveganju, ki naj bi bil povzean z umetno inteligenco, odločilno vlogo igra samo en kriterij: nevarnost iztrebljanja človeške vrste zaradi prevlade stroja nad človekom.

Neredko se srečujemo z ocenami, da pomembni predpogoj za varni prihodnji razvoj GUI, v okviru katerega se lahko izognemo eksistenčnim tipom tveganj, predstavlja »algoritem okrepljenega učenja« (v ang.: »reinforcement learning algorithms«) [5, 6]). Pri »okrepljenem učenju« gre za to, da se v procesu sprejemanja odločitev nagradi to, kar vodi v dobrobit ljudi. Vendar v konkretnih situacijah težavo predstavlja praktično usklajevanje funkcij umetne inteligence z sprejetimi družbenimi vrednotami. Čeprav se ta problem na prvi pogled zdi trivialen, temu ni tako. Družbene vrednote so raznolike, amorfne in jih je težko zapopasti v kvantitativnih kategorijah. Problem, kako »okrepljeno učenje« uskladiti z sprejemljivimi družbenimi vrednotami, zato ni nekaj, kar se da na zelo enostaven in samoumevni način razrešiti. Njegova razrešitev je odvisna od več dejavnikov. Eden izmed teh je možnost, da se modeli GUI razvijajo kot odprtokodni modeli, kar je seveda v nasprotju z sedanjo strategijo multinacionalk, da preko lastniškega nadzora novih naprednih tehnologij javnosti prikrivajo ključne informacije.

Negativna posledica lastniškega odnosa do VJM je, da znanje o notranjih mehanizmih delovanja VJM, ki predstavljajo vrh razvoja umetne inteligence danes, še vedno predstavlja izziv za večino uporabnikov, (deloma) pa tudi za strokovnjake s področja računalništva. Težko je namreč analizirati in priti na tej osnovi do razumevanja VJM, ki delujejo v okviru kompleksnih notranjih struktur z milijoni parametrov. Četudi lahko v vlogi uporabnikov ali celo računalniških razvijalcev vidimo končni rezultat delovanja VJM, pa je pojasnitev oziroma interpretacija njihovih notranjih struktur izjemno zahtevna. Skratka, veliki jezikovni modeli še vedno nastopajo kot »črne skrinjice« (»black boxes«). Thomas Arnold je za opis te nevzdržne situacije uporabil naslednjo posrečeno analogijo: »To je tako kot da bi se prizadevali za razlago delovanja kompleksne kemijske reakcije, ne da bi poznali natančno strukturo in interakcijo molekul.« [7] V strokovni literaturi se sicer omenja tudi nekaj izjem. Za modele kot so BLOOM, Cerebras-GPT ali Llama, naj bi podjetja, ki se ukvarjajo z umetno inteligenco, dopuščala večji javni vpogled [8]. Spet za druge so informacije za javnost odprli, potem pa ponovno zaprli. Četudi vrhunski znanstveniki, ki se ukvarjajo z UI in prihajajo iz akademske sfere znanosti, v vedno večjem številu opozarjajo, da je prosti dostop do vseh informacij na tem področju eden ključnih dejavnikov, ki lahko zagotovi verodostojno in zanesljivo raziskovanje, saj le tako lahko dostopamo do informacij o celotni »arhitekturi« VJM (t.j. od uporabljenih podatkovnih baz do algoritmov), v zvezi s tem še vedno ni bilo storjenih veliko sprememb.

3 Ali lahko ustvarjalno dimenzijo znanstvenega dela dokončno prevzame umetna inteligenca?

Na prihodnje izzive, ki so povezani z nastopom umetne splošne oziroma umetne super inteligence, je treba gledati tudi v luči današnjih dogajanj. Že danes si lahko zastavljamo vprašanje, ali bo ustvarjalno znanstveno delo dokončno prevzela GUI: ali je res upravičeno trditi, da kar je nekoč kalkulator pomenil za številke, in kar internet za globalni značaj komunikacijo, to danes pomeni za znanstveno kreativnost razvoj GUI? Znanstveno kreativnost lahko subsumiramo pod bolj splošni pojem inteligence. Ta naj bi načeloma izkazovala celo paleto zmožnosti, od kreativnih do racionalnih oblik (znanstvenega, umetniškega, itd.) mišljenja, od načrtovanja do učenja na temelju izkušenj, itd. Četudi danes spekuliramo, da bo splošna umetna inteligenca dosegla ali presegla inteligenčne zmožnosti ljudi, pa bomo v strokovni literaturi težko naleteli na neke soglasne kriterije, ki naj bi povedali, kaj predstavlja »inteligenca« pri strojih in kaj predstavlja inteligenca pri ljudeh. Formalne definicije, ki vztrajajo ne nekem skupnem imenovalcu, nam niso vedno v pomoč. Nobena izmed teh formalnih definicij ne ponuja nekega dokončnega kriterija, ki bi nam omogočal primerjavo »inteligentnosti« različnih entitet. Če se za hip ustavimo ob najnovejšem delu Yuval Noaha Harareja, ki nosi naslov »Nexus. A Brief History of Information Networks from the Stone Stage to AI« [9], bomo pri njemu hitro prepoznali besednjak, ki naj bi

nedvoumno nakazoval, da GUI poseduje moment intencionalnosti, t.j. sposobnost GUI slediti delovanju, ki izhaja iz njih samih. (Avtor knjige govori o tem, da se pametni stroji, ki jih vodi GUI, sami odločajo, izbirajo, delujejo, itd.). Ob prebiranju najnovejšega Hararejevega dela se lahko vprašamo, zakaj vsiljuje intencionalnost kot ključni kriterij za izenačevanje »inteligentnosti« človeka in stroja. Lahko bi uporabil širšo definicijo inteligence in bi le to pripisal že entitetam, ki so pasivne, torej ne vključujejo momenta intencionalnosti, vendar vseeno reagirajo na okolje in lahko opravljajo kompleksne naloge. To je na primer storil Sebastien Bubeck, ki je skupaj z soavtorji preučeval, ali so v jezikovnem modelu GTP-4 že dani zametki umetne splošne inteligence. Postavil je namreč tezo, da si neko inteligentno entiteto lahko predstavljamo tudi kot »orakelj«, ki nima notranjih vzgibov ali želja za delovanje, vendar lahko natančno in koristno zagotavlja informacije o kateri koli temi ali domeni vedenja [10]. Definicijo inteligence, ki izhaja zgolj iz kriterija intencionalnosti, imamo lahko za restriktivno še iz enega razloga. Če namreč pri tej definiciji izhajamo iz notranjih motivov za doseganja ciljev našega delovanja v kar se da širokem okolju, kjer se soočamo z nikoli zaključenim spektrom situacij, potem v primeru rabe takšne definicije implicitno predpostavljamo, da je pojem inteligence neizogibno vezan na univerzalnost in optimalnost. To pomeni, da spet operiramo z apriorno definiranim in ne aposteriorno preverjenim konceptom inteligence. Dejansko oziroma realno inteligenco človeka namreč nikakor ne moremo opredeliti kot absolutno univerzalno in optimalno.

S podobnimi dilemami se soočamo, če naš pogled usmerimo na kreativnost kot eno izmed dimenzij človekove inteligence. Tudi v tem primeru odgovor na vprašanje, ali umetna inteligenca enostavno privzema kreativne moči znanosti, ni enoznačen. Ne gre samo za to, da se že pri vprašanju kreativnosti človeka srečujemo z ogromnim številom definicij (znanstveniki uporabljajo danes več kot 50 definicij [11]), zadeve postanejo še bolj kompleksne, ko iščemo skupni imenovalec med definicijo človeške kreativnosti in kreativnostjo, ki jo pripisujemo umetni inteligenci. Na eni strani imamo avtorje, kot so Marc Ruco [12] ali Stephen Rice [13], ki pravijo, da kolikor k standardnim definicijam človekove ustvarjalnosti – ta vključuje dimenzijo originalnosti in učinkovitosti - dodamo tudi dimenziji avtentičnosti, potem GUI ne more tekmovati z ljudmi.

Na drugi strani imamo avtorje, kot na primer Hubert Kent, za katere je GPT- 4 že dosegel izredno visoko stopnjo znanstvene kreativnosti, vsaj kar zadeva t.i. odprti tip mišljenja, saj naj bi empirične analize pokazale, da GPT-4 že zmore doseči rezultate, ki so enaki rezultatom, ki jih doseže zgolj 1% najbolj inteligentnih ljudi [14]. Rezultati dodatnih študij naj bi ravno tako dokazovali, da model GPT-4 izkazuje veliko stopnjo fleksibilnosti zunaj ustaljenih okvirov mišljenja in naj bi imel na področju odprtega tipa mišljenja celo višji kreativni potencial od ljudi. Pričakovati torej je, pravijo avtorji, ki so opravili te in podobne študije, da bo model GPT-4, kolikor bo dosežen napredek glede povečanih zmožnosti učenja na velikih bazah podatkov in bolj napredni arhitekturi nevronskih mrež, kmalo storil pomembni korak v smeri umetne splošne inteligence.[15]. V tem primeru Turingovi testi že zvenijo zastarelo. V okviru rabe Turingovega testa gre namreč za to, da se kot kriterij izenačitve dveh inteligenc vzame situacijo, ko nek uporabnik, ki komuncira z klepetalnikom UI, ne zna več ločiti, ali je na drugi strani človek ali stroj [8].

4 Zakaj umetna inteligenca postaja vedno bolj nepogrešljiv pomočnik v vseh fazah znanstvenega raziskovanja?

V okviru moje razprave me ob bolj načelnem epistemološkem vprašanju, t.j. ali nova tehnologija umetne inteligence postopoma zavzema prostor znanstvene kreativnosti, zanima tudi bolj konkretno vprašanje: ali ta nova napredna tehnologija dobiva status nepogrešljivosti v vseh drugih fazah znanstvenega raziskovanja?

Sodobno znanstveno raziskovanje je multidimenzionalni proces, ki vključuje različne faze, ki od začetka raziskovanja do končne objave znanstvenih rezultatov segajo od najbolj rutinskih pa do najbolj ustvarjalnih aktivnosti. VJM v tem primeru prevzema vlogo koristnega in vedno bolj nepogrešljivega »asistenta« v vseh fazah znanstvenega raziskovanja. Bo ta »asistent« v bolj ali manj oddaljeni prihodnosti postal »profesor«, ki bo dokončno nadomestil človeka – znanstvenika?

1. Najprej je treba izpostaviti, da GUI zaradi svoje učinkovitosti vedno bolj nadomešča znanstvenike v postopkih pridobivanja podatkov. Podobno je z učinkovitostjo GUI v vsebinskem pregledovanju, povzemanju in sumiranju množice informacijskih virov, ki so kot »state of the art« relevantne v vsakem začetnem procesu znanstvenega raziskovanja. GUI je sposoben obdelave in analize velike količine podatkov. Vloga GUI postaja neprecenljiva pri pregledu in sintezi vsebin iz znanstvene literature, povzemanju podatkov in sinteze kompleksnih podatkovnih baz, samodejnem prepoznavanju vzorcev in trendov, ki jih je mogoče izpeljati iz podatkov, modeliranju in napovedovanju na temelju zbranih podatkov, itd.
2. Vedno bolj se povečuje vloga GUI pri ustvarjanju novih idej. Glede na današnjo eksponentno rast znanstvenih informacij je prenos te raziskovalne funkcije iz človeka na GUI hkrati povezana z zmožnostjo GUI, da učinkovito in predvsem avtonomno ustvarja nova raziskovalna vprašanja in hipoteze. Eden največjih izzivov najbolj naprednih področij znanosti je skorajda neskončno število hipotez, ki se nanašajo na raziskovalne probleme, zaradi česar se včasih zdi natančno sistematično raziskovanje, ki bi omogočalo sprejetje hevristično najbolj obetavne hipoteze, brez sodelovanja UI skorajda nemogoče. Primer: v biokemiji naj bi obstajalo približno 10^{60} molekul, to pa je praktično enako številu zdravil, ki jih je treba na temelju ogromnega števila molekul šele odkriti [16]. Pri tem imajo ravno najnovejši modeli GUI potreben potencial, da revolucionarno posežejo v to fazo znanstvenega raziskovanja, ko gre za biokemijo. Podobne primere bi lahko navedli za področje genomike, astronomije, kvantne fizike, itd. Ne moremo mimo omembe še ene funkcije GUI. Ta funkcija GUI je vezana na njeno zmožnost usmerjanja k bolj interdisciplinarno zasnovanim revolucionarnim znanstvenim odkritjem, saj so njeni potenciali pri obdelavi in sintezi informacij iz različnih disciplin skorajda neomejeni.

3. Vloga GUI se povečuje tudi v procesih evalvacije končnih rezultatov znanstvenega raziskovanja. Če izhajam iz bolj splošnih epistemoloških predpostavk in se na tem mestu izognemo razpravi o prednostih in tudi tveganjih uporabe GUI v konkretnih recenzentskih postopkih, potem naj na kratko omenimo zgolj eno izredno pomembno vlogo te nove napredne tehnologije, t.j. preverjanje rezultatov eksperimentalnih in drugih empiričnih raziskav. V preteklosti je v glavnem veljalo, da ni problematična ponovljivost dobljenih znanstvenih rezultatov, bodisi na temelju javno dostopnih znanstvenih objav ali ustreznih eksperimentalnih protokov. Sodobna znanost se nahaja v vedno večji krizi, kar zadeva zmožnost replikacije, saj je tako z vidika časa kot tudi stroškov v številnih, če ne kar vseh vseh znanstvenih disciplinah težko izvesti potrebne eksperimentalne in druge znanstvene ponovitve. O tveganjih za povečanje goljufij in prevar v moderni znanosti, ki izhajajo iz teh kompleksnih situacij raziskovalnega dela, sem več pisal na drugih mestih [17]. GUI lahko odigra zelo relevatno funkcijo v današnjem času enormne produkcije znanstvenih rezultatov, ko je vedno težje izvajati ponovitve eksperimentov z namenom izvajanja kontrole znanstvenih rezultatov. Njen predikativni pristop namreč lahko zagotovi učinkovito, hitro, sistematično in natančno napoved ponovljivosti posameznih znanstvenih odkritij ali pa celo vseh spoznanj na posameznem področju znanosti.

4. Pozitivna vloga GUI se danes povečuje tudi v okviru širših družbenih in kognitivnih predpostavk, ki so relevantne za delovanje moderne znanosti. V zvezi s to širšo funkcijo bi izpostavil vlogo GUI pri spodbujanju komunikacij znotraj znanstvene skupnosti, pa tudi komunikacije znanstvenikov navzven. To zadnje naj bi se dogajalo predvsem s pomočjo modela ChatGPT, ki generira takšne tipe pojasnitev, ki vodijo k premagovanju komunikacijskih prepadov med eksperti in laiki. Vendar je to funkcijo, kot smo že opozorili v enem izmed predhodnih poglavij, mogoče izvajati le, če bo prišlo do uveljavitve nove paradigme odprtokodne znanosti. V zadnjem času strokovnjaki, ki delujejo na področju GUI, vedno bolj poudarjajo, da je treba razviti modele, ki bodo čim bolj korespondirali z fizično realnostjo. Menijo, da je treba največ naporov usmeriti v nadaljnji razvoj multimodalnih sistemov GUI. Demis Hassabis, izvršni direktor firme DeepMind, je v intervjuju za angleški dnevnik Guardian konec prejšnjega leta dejal, da je bil storjen na tem področju največji korak z modelom Gemini, ki ga razvija njegovo podjetje [18].

5 Ali nova tehnologija generativne umetne inteligence v okviru družboslovnega raziskovanja nujno in vedno zagotavlja znanstveno objektivnost?

Kot družboslovca me seveda zanima tudi vprašanje vedno večje rabe VJM na področju mojega področja znanstvenega raziskovanja. Kar takoj je treba reči, da na področju družbenih ved VJM izkazujejo velik (hevristični) potencial v razvijanju novih pristopov k anketnim raziskavam in ponovljivosti eksperimentov na področju vedenjske ekonomije [19], diskurzivnih analizah tekstov, ki jih je mogoče izvajati na avtomatizirani način [20] in končno tudi na področju razvijanja modelov, ki simulirajo stvarno obnašanje ljudi. V tem zadnjem primeru gre predvsem za t.i. »agent-based« modele, ki preučujejo, kako delovanje oziroma vedenje na mikro ravni (npr.: odločitve individualnih agentov) vodi do posledic na makro (družbeni) ravni (npr.: oblikovanje družbenih vzorcev delovanja oziroma obnašanja). V okviru teh modelov se seveda lahko preučuje tudi obratni vpliv: kako makro-nivo vpliva na obnašanje na mikro ravni [8]. V okviru sociologije se s temi »agent-based« modeli preučuje socialna omrežja, oblikovanje sosedskih skupnosti, itd.

Se pa v zvezi z družboslovnim raziskovanjem pojavlja določen paradoks, na katerega želim opozoriti v tem sklepnem delu moje razprave. Ta paradoks predstavlja dejstvo, da postopki »okrepljenega učenja« (ang. »reinforcement learning«), ki naj bi odpravili »halucinacije« in raznovrstne pristranosti, predstavljajo oviro za doseganje objektivno veljavnih znanstvenih rezultatov. Če pride skozi delovanje t.i. »reinforcement self-learning by human feed-back« (RLHF) do idealiziranja sveta, t..j. sveta, kakršen naj bi bil, ne pa sveta, kakršen dejansko je, takšna prizadevanja za zmanjšanja pristranskosti algoritmov, katerih cilj je promovirati liberalne vrednote, lahko ogrozijo veljavnost raziskav v družboslovju, ki jih podpira umetna inteligenca. »Požarni zid«, ki se ga želi danes pospešeno graditi preko RLHF, odpravlja tveganja GUI, kar zadeva njeno široko uporabo (in preprečuje tveganja, ki so se, kot pravi Yuval Harare, že zažrla v civilizacijski kod sodobnih družb), po drugi strani pa predstavlja epistemološko tveganje za objektivni značaj današnjih družboslovnih raziskav. Tudi to predstavlja dilemo današnjega in prihodnjega razvoja umetne inteligence, ki zahteva naš celovit interdisciplinarni razmislek, saj se je le na tej osnovi mogoče izogniti negativnim družbenim in tudi epistemološkim implikacijam njenega razvoja.

6 Zaključek

V zadnjem času je tako v znanstvenih krogih kot tudi zunaj znanstvenih krogov veliko govora o možnih tveganjih današnjega in prihodnjega razvoja umetne inteligence. Znanstveniki iz Massachusetts Institute of Technology, ene najbolj uglednih akademskih institucij v ZDA, so v letošnjem letu pripravili javno dostopni repozitorij z umetno inteligenco povezanih primerov tveganj. V omenjenem repozitoriju se trenutno nahaja kar 777 opisov takšnih tveganj. To je še en dokaz, kako veliko zanimanje obstaja danes za ta vprašanja. V mojem kratkem prispevku sem se dotaknil zgolj enega izmed teh številnih problemov, ki je vezan bolj na epistemologijo znanstvenega raziskovanja, ne pa toliko na družbene posledice razvoja umetne inteligence. V tem kontekstu me je predvsem zanimalo, ali pospešena raba VJM v okviru različnih znanstveno-raziskovalnih aktivnosti predstavlja eno izmed domen, kjer se na stežaj odpirajo vrata nastopu umetne splošne oziroma umetne super inteligence. Še posebej me je zanimalo vprašanje, zakaj GUI postaja že danes nepogrešljivo »orodje« v seh fazah znanstvenega raziskovanja. V sklepnem delu sem se na kratko ustavil ob nekaterih specifičnih dilemah uporabe GUI na področju družboslovnega raziskovanja.

Literatura

[1] Ray Parta, 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 3: 21–154; https://doi.org/10.1016/j.iotcps.2023.04.003.

[2] Nick Bostrom, 2014. Superintelligence: Paths, dangers, strategies. Oxford University Press.

[3] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, & Owain Evans, 2022. When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62(July), 29-754. https://doi.org/10.1613/jair.1.11222.

[4] Max Roser, 2023. AI timelines: What do experts in artificial intelligence expect for the future? Our World in Data. https://ourworldindata.org/ai-timelines.

[5] Rishi Bommasani et al., 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258v3 [cs.LG] 12 Jul 2022.

[6] Yogesh Dwivedi et al., 2023. Opinion Paper - So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71 (2023) 102642; https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[7] Thomas Arnold, 2024. Herausforderungen in der Forschung: Mangelnde Reproduzierbarkeit und Erklärbarkeit. V: L. Ohly in G. Schreiber (Hrsg.) KI:Text.Diskurse über KI-Textgeneratoren, str. 67-83. Berlin/Boston: De Gruyter Verlag.

[8] Christopher Bail, 2024. Can Generative AI improve social science?. PNAS, May 9, 2024 1211) e2314021121; https://doi.org/10.1073/pnas.2314021121.

[9] Yuval Noah Harari, 2024. Nexus. A Brief History of Information Networks from the Stone Stage to AI. New York: Penguin Random House.

[10] Bubeck S. et al., 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4 (2023), arXiv:2108.07258v3 [cs.LG] .

[11] L.B. Soros, Alyssa Adams, Stefano Kalonaris, Olaf Witkowski, Christian Guckelsberger, 2024. On Creativity and Open-Endedness. arXiv:2405.18016v4 [cs.AI] 23 Jun 2024.

[12] Marc Runco, 2023. AI can only produce artificial creativity. Journal of Creativity 33 (2023) 100063. https://doi.org/10.1016/j.yjoc.2023.100063.

[13] Stephen Rice, Winter Scott, Rice Connor, 2024. The advantages and limitations of using ChatGPT to enhance technological research. Technology in Society 76 (2024) 102426. https://dx.doi.org/10.2139/ssrn.4416080.

[14] Hubert Kent, Kim Awa, Darya Zabelina, 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. Scientific Reports 14(1):3440; https://doi.org/10.1038/s41598-024-53303-w.

[15] Mohamed Salah et al., 2023. Chatting with ChatGPT: decoding the mind of Chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. Current Psychology, 43, 7843–7858 (2024). https://doi.org/10.1007/s12144-023-04989-0.

[16] Hanchen Wang et al., 2023. Scientific discovery in the age of artificial intelligence. Nature, 620: Vol 620 | 3 August 2023. DOI: 10.1038/s41586-023-06221-2.

[17] Franc Mali, 2011. Razvoj moderne znanosti. Socialni mehanizmi. Ljubljana: Založba FDV.

[18] Demis Hassabis, 2023. Google releases new AI model with claim it can outperform ChatGPT in most tests. The Guardian, 7. December, 2023.

[19] John Horton, 2023. Large language models as simulated economic agents: What can we learn from homo silicus? arXiv [Preprint] (2023). 10.48550/arXiv.2301.07543.

[20] Igor Grossmann, Cassandra Parker, Mathew Feinberg, Nicholas Christakis, 2023. AI and the transformation of social science research. Science, 380(6650):1108-1109. DOI:10.1126/science.adi1778.Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.