Pregledni znanstveni članek/Article (1.02)
Bogoslovni vestnik/Theological Quarterly 83 (2023) 4, 839—852
Besedilo prejeto/Received:09/2023; sprejeto/Accepted:11/2023
UDK/UDC: 14:2:004.89
81:004.89:2
DOI: 10.34291/BV2023/04/Strahovnik

Vojko Strahovnik

Etični in teološki izzivi velikih jezikovnih modelov
Ethical and Theological Challenges of Large Language Models

Povzetek: V prispevku obravnavamo etične in teološke izzive, ki so povezani z umetno inteligenco, posebej s področjem velikih jezikovnih modelov. Po uvodu v drugem razdelku na kratko predstavimo, kaj so veliki jezikovni modeli in kakšen njihov razvoj. V tretjem razdelku obravnavamo etične izzive teh modelov. Dotikamo se tudi obstoječih etičnih smernic in izpostavljamo, v kolikšni meri se omenjenih izzivov sploh lotevajo. V četrtem razdelku izpostavljamo teološke izzive, ki jih obravnavani modeli odpirajo – ti so tesno povezani z etičnimi vidiki. V zaključku navajamo nekaj premislekov o jezikovnih tehnologijah v našem imaginariju in o nadaljnjem razvoju velikih jezikovnih modelov ter spremembah, ki jih bi takšen razvoj lahko prinesel.

Ključne besede: umetna inteligenca, veliki jezikovni modeli, tveganja, etične smernice, teološki izzivi umetne inteligence

Abstract: In this article we discuss the ethical and theological challenges related to artificial intelligence, especially in the area of large-scale language models. In the second section, we briefly introduce what large language models are and their development. In the third section, we discuss the ethical challenges of these models. I also touch on existing ethical guidelines and highlight to what extent they address these challenges at all. In the fourth section, we highlight the theological challenges raised by these models. These are closely linked to ethical considerations. In conclusion, we give some reflections on language-related technologies in our imagination and further development of large-scale linguistic models, including the changes that such a development might bring about.

Keywords: artificial intelligence, large language models, risks, ethical guidelines, theological challenges of artificial intelligence

1. Uvod

Veliki jezikovni modeli (gre za modele, kot so npr. ChatGPT-4, BERT, LaMDA, PaLM, LLaMA) so področje umetne inteligence, ki v zadnjih letih pospešeno napreduje, hkrati pa vzbuja zelo raznolike odzive (Floridi 2023).[1] Ti po eni strani segajo do apokaliptičnih napovedi o koncu sveta, kot ga poznamo, pogosto pa jih spremljajo pozivi k ustavitvi razvoja omenjenih modelov vse dotlej, dokler ne bomo znali predvideti vplivov njihovega razvoja (The Future of Life Institute 2023). Na drugi strani so presoje, da ta tehnologija ne prinaša nič novega in da gre za povsem ‚neinteligentno' ponavljanje, prerazporejanje oz. sestavljanje nizov besed, ki je podobno ‚govoru' papige (Bender in Koller 2020; Bender idr. 2021). Prvi razdelek podaja jedrnat pregled, kaj so veliki jezikovni modeli, njihov razvoj, poseben podrazdelek pa se ukvarja z že ugotovljenimi tveganji, povezanimi z njihovo uporabo. Drugi razdelek raziskuje etične izzive, povezane s temi modeli; vključuje tudi pregled obstoječih etičnih smernic in njihovo učinkovitost pri obravnavi teh izzivov. Tretji razdelek izpostavlja teološke posledice, ki jih ti modeli sprožajo in so tesno prepletene z etičnimi vprašanji. Prispevek zaključujemo z razmislekom o prihodnjem razvoju obsežnih jezikovnih modelov in možnih spremembah, ki jih tak razvoj bo ali bi jih lahko prinesel.

2. Veliki jezikovni modeli

Veliki jezikovni modeli so razmeroma zmogljivi sistemi umetne inteligence, katerih osrednji namen je razbiranje in ustvarjanje besedil v naravnem jeziku. Takšni modeli so tvorbe besedila zmožni na podlagi zelo obsežnega nabora predhodnih besedil oz. besedilnih podatkov, na katerih so se učili. Med drugim se lahko uporabljajo za odgovore na vprašanja, pisanje zgodb ali pesmi, prevajanje, učenje in za številne druge namene.

Veliki jezikovni modeli so običajno zgrajeni z uporabo tehnik globokega učenja in nevronskih mrež (Raaijmakers 2022). Ti modeli se naučijo vzorcev in odnosov v besedilnih podatkih, na katerih se učijo, kar jim omogoča ustvarjanje skladnih in kontekstualno ustreznih odgovorov na dane pozive oz. glede na dana navodila. Pri tem se pridevnik ‚velik' nanaša predvsem na to, kako obsežen je nabor besedilnih podatkov. Gre za milijarde v nize povezanih besed, pridobljenih iz različnih besedil – knjig, člankov, spletnih strani in drugih besedilnih virov –, ki so na voljo na internetu ali so vsaj digitalizirana (učni podatki). Kljub tolikšnemu obsegu lahko zaradi povečanih zmogljivosti in zaobjema strojne opreme velike jezikovne modele izurimo v le nekaj tednih. Postopek vključuje izpostavljanje modela tej ogromni količini podatkov, pri čemer se model nauči napovedovati naslednjo besedo

[1] Prispevek je nastal v okviru raziskovalnega programa P6-0269 „Religija, etika, izobraževanje in izzivi sodobne družbe" in raziskovalnega projekta P6-24684 „Teologija, digitalna kultura in izzivi na človeka osredotočene umetne inteligence", ki ju finančno podpira Agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije, ter v okviru raziskovalnega projekta „Epistemic Identity and Epistemic Virtue: Human Mind and Artificial Intelligence" s podporo fundacije John Templeton (krovni projekt „New Horizons for Science and Religion in Central and Eastern Europe").

ali zaporedje besed (Ouyang idr. 2022). Model pridobi neke vrste razumevanje slovnice, semantike in jezikovnih struktur, vendar moramo biti pri rabi besede razumevanje zelo previdni (Bender in Koller 2020). Poleg učenja na podlagi osnovnih besedilnih podatkov je veliki jezikovni model običajno še nadgrajen, to je v postopku učenja uglašen oz. uravnan. Tu je model izpostavljen naboru pogovornih podatkov oz. besedil, ki so sestavljena iz parov vhodnih in izhodnih zaporedij: vhodno sporočilo je tisto, ki ga napiše en govorec, izhodno pa odgovor oz. odziv nanj. Takšen veliki jezikovni model se lahko potem uporabi za ustvarjanje odgovorov na vhodna sporočila. Ustvarjalci tovrstnih modelov lahko dodatno uravnavajo različne parametre, kot je npr. dolžina odgovora ali izogibanje rabi žaljivih besed. Veliki jezikovni model lahko potem integriramo v klepetalnega robota ali podobno okolje za pogovorno komunikacijo (Raaijmakers 2022).

Veliki jezikovni modeli se res učijo na ogromnih količinah besedilnih podatkov, kar jim pomaga razviti razbiranje jezikovnih vzorcev in v večini primerov ustvarjanje kontekstualno ustreznih odzivov. Vendar pa nimajo možnosti neodvisnega preverjanja dejstev ali potrjevanja pravilnosti podanih informacij. Zato sta natančnost in pravilnost odgovorov v veliki meri odvisni od kakovosti in natančnosti učnih podatkov, ki so jim bili modeli izpostavljeni. Ne gre torej za to, da bi takšni modeli lahko informacije neposredno preverjali. Tako je mogoče reči, da čeprav lahko podajajo koristne informacije in vpoglede, je ključnega pomena, da njihove odgovore preverjamo z zunanjimi viri. Občasno ti modeli ustvarjajo napačne ali nesmiselne odgovore, za katere se je uveljavil izraz halucinacije sistema. Z učenjem na velikih naborih podatkov z različnimi besedili se model ‚zgolj nauči' statističnih vzorcev in odnosov med besedami (ali deli besed). To mu omogoča napovedovanje, ki ga uporabi za ustvarjanje odgovora – na podlagi kontekstualnih namigov, ki se jih je naučil med usposabljanjem. Dobro je opozoriti, da čeprav veliki jezikovni modeli lahko ustvarijo na prvi pogled smiselne in verjetne odgovore, seveda nimajo pravega razumevanja ali zavesti.

Na koncu naj izpostavimo dva osrednja vidika, za katera je pomembno, da ju pri razmišljanju o velikih jezikovnih modelih upoštevamo. Prvi je, da gre res zgolj in samo za sestavljanje niza besed na podlagi statističnih verjetnostnih metod – za uporabo statističnih tehnik za napovedovanje naslednje besede v zaporedju. Eden od pristopov je uporaba N-gramskih modelov, ki ocenijo verjetnost besede na podlagi konteksta prejšnjih besed (npr. bigramski model zajame zgolj prejšnjo besedo, medtem kot trigramski model zajame dve prejšnji besedi). Model torej iz učnih podatkov zajame statistične vzorce in odnose med besedami (oz. natančneje, med žetoni/primerki, kar pomeni, da lahko en primerek tvori posamezna beseda ali pa model kot en primerek zajame dve besedi skupaj). Model nato zgradi n-grame in glede na kontekst n-grama računa verjetnost vsake možne naslednje besede. Takšna ocena temelji na pogostnosti pojavljanja n-grama in naslednje besede v celotnem naboru učnih besedil (Raaijmakers 2022, 93–95). Napovedana beseda postane del konteksta za napovedovanje naslednjih besed. Povedano na kratko, z analizo pogostnosti in vzorcev besednih zaporedij v učnih podatkih se jezikovni model nauči statističnih odvisnosti med besedami in to znanje nato upo-

rabi za napovedovanje naslednje besede v nizu. Glavni omejitvi sta nezmožnost zajemanja daljših vzorcev odvisnosti v odnosih med besedami, kontekstualnih odtenkov – oz. ustrezno razbiranje pomenskih odtenkov. Razvoj novejših modelov gre v smer čim večje odprave teh omejitev.

Drugi vidik je, da načelo zasnove teh modelov ni resnica. V tem smislu gre za fenomen, podoben temu, o katerem je že pred skoraj dvema desetletjema v eseju z naslovom „O nakladanju" razmišljal ameriški filozof in esejist Harry G. Frankfurt. Njegova teza je bila, da je zahodna družba vedno bolj prežeta s slednjim (v izvirniku je uporabljen angleški izraz za nakladanje – *bullshit* – in njegove tvorjenke). Pri nakladanju gre za zatrjevanje stvari, ki jih na neki način niti sami ne verjamemo. Ne gre za laž, saj nakladanje ostane nakladanje, če je resnično ali neresnično: nakladanje se na resnico ne ozira. Gre za prav poseben odnos do resnice, kjer je pomembna le zvestoba, iskrenost do želje po izjavljanju, medtem ko je resničnost izjav nepomembna. Skratka, gre za ravno prav odmerjeno mešanico nespoštovanja iskrenosti, nesmisla, nezainteresiranosti, zavajanja in samoprevare. Veliki jezikovni modeli nam nudijo oz. oblikujejo odgovor na naš poziv ali vprašanje povsem brez ozira na to, ali zanj sploh obstaja kakšen temelj. Npr. takšen model lahko vprašamo, kateri otok ima najraje, pa nam bo brez pomisleka odgovoril.

> »Model nima instinktivnih ali pridobljenih preferenc, kot jih imamo mi; prav tako nima telesa, čutil ali pripovednega spomina na avtobiografsko preteklost. Da bi njegovi odgovori ostali smiselni in specifični, mora pripraviti informativne odgovore na vprašanja, kot je ›Kateri je tvoj najljubši otok na svetu?‹, in slediti vsemu prej ustvarjenemu nakladanju, da bi bilo njegovo naslednje nakladanje dosledno. /…/ Ni nujno, da je nakladanje slabo. Gre za temeljni gradnik domišljijske igre, ki je osnova za leposlovje in snov pravljic, kar sta kulturna zaklada. Težava nastane šele, ko je oseba, ki je deležna takšne prevare, zavedena ali ko se kakovost diskurza zaradi nakladanja tako poslabša, da izgubimo občutek za stvarnost (kar je danes zelo zaskrbljujoče).« (y Arcas 2022, 185)

Prihodnji razvoj velikih jezikovnih modelov bo zelo verjetno prinesel napredek v smeri večjega obsega in njihove učinkovitosti, hkrati pa gredo težišča razvoja v smeri bolj pretanjenih nastavitev in zmožnosti prilagajanja teh sistemov (glede na jezik, različna področja delovanja, ciljne skupine uporabnikov, ipd.) – kar bo vključevalo tudi to, da si jih bodo uporabniki glede na svoje želje in potrebe lahko prilagodili sami. Sposobnost sprotnega učenja bo še povečala robustnost in prilagodljivost. Ena od smeri vključuje tudi multimodalne sisteme, kar pomeni vključevanje drugih modalitet – kot so slike, video in zvok – v jezikovne modele, zato bodo ti modeli znali razbirati in ustvarjati različne.

2.1 Tveganja velikih jezikovnih modelov

Preden se posvetimo pregledu etičnih izzivov, ki jih veliki jezikovni modeli odpirajo, na hitro preletimo še nabor prepoznanih tveganj, ki jih ti sistemi prinašajo. V

mnogih ozirih ta tveganja že vključujejo tudi etične izzive oz. so etični izziv sama po sebi. Sistematični pregled tveganj velikih jezikovnih modelov (Weidinger et al. 2022) ta razvršča na šest glavnih področij, in sicer: (1) diskriminacija, sovražni govor in izključevanje; (2) informacijska tveganja oz. tveganja zlorabe informacij; (3) tveganja napačnih ali lažnih informacij; (4) zlonamerna uporaba sistemov; (5) tveganje za nastanek škodljivih posledic na podlagi interakcije med človekom in računalnikom; (6) tveganje okoljske in socialno-ekonomske škode.

V okviru prvega področja tveganj je jasno, da veliki jezikovni modeli lahko delujejo tako, da tvorjeno in ponujeno besedilo spodbuja ali utrjuje različne predsodke in stereotipe. To lahko privede do slabšalnih upodobitev in nepravilne obravnave vrste marginaliziranih skupin, spodbujanja sovraštva in nasilnega vedenja ali pa do tega, da se izključujočost in marginalizacija pri nekaterih vrstah identitet še poglobita. Do nepravičnosti lahko pride tudi v primeru, ko so posamezne skupine ali identitete brez utemeljitve obravnavane privilegirano. Pojavlja se lahko sovražni govor (žaljivke, napadi na identiteto, grožnje), kar lahko posameznika ali skupine ogrozi ter jih izpostavi nevarnostim in psihološki škodi. V okvir tega področja spada tudi oblikovanje in delovanje družbenih norm, ki so lahko do identitet posameznikov ali skupin izključujoče, kar zanje pomeni nesorazmerno breme. Prav tako lahko izpostavimo izključevanje jezikovnih skupnosti, ki v naboru učnih podatkov niso zajete ali pa so zastopane slabše; to potem lahko pomeni izključevanje njihove perspektive ali njihove udeležbe pri uporabi obravnavanih sistemov.

Drugo področje zajema nevarnosti, povezane z razkritjem osebnih informacij ali informacij, občutljivih zaradi drugih razlogov (npr. varnost, konkurenčna prednost). Širjenje takšnih zasebnih ali občutljivih informacij lahko vodi do škodljivih posledic za posameznika, skupino ali širšo skupnost. Pri tem ne gre spregledati niti možnosti, da lahko veliki jezikovni model pravilno napove nekatere osebne podatke ali občutljive informacije na podlagi drugih dostopnih informacij – vključno z načini, ki v interakciji med ljudmi niso prisotni (lahko npr. napove duševno bolezen ali drugi osebno okoliščino posameznika, ki je ta sicer ne želi razkriti, na podlagi podatkov, ki sami po sebi niso pomenljivi).

Tretje področje tveganj zajema posredovanje napačnih oz. zmotnih, zavajajočih, nesmiselnih ali nekakovostnih informacij – in to ne da bi pri tem kakršen koli slab namen imel uporabnik sam. To lahko privede do posameznih materialnih, psiholoških ali drugih škod (zamislimo si lahko primer, ko nekdo vzame določeno zdravilo na priporočilo pogovornega robota, pri čemer je to zanj v trenutnem stanju v bistvu škodljivo), pa tudi do znižanja stopnje družbenega zaupanja in človekove avtonomije nasploh.

Četrto področje tveganj predstavlja zlonamerna uporaba velikih jezikovnih modelov, ki je povezana z informacijami. Ti modeli npr. omogočajo precej lažje in cenejše izvajanje dezinformacijskih kampanj (zavajanje javnosti, usmerjeno oblikovanje javnega mnenja) ali pa kampanj, ki razširjajo nerelevatne informacije za zakrivanje relevantnih informacij. Prav tako ti modeli lahko služijo kot orodje spletnih goljufij in prevar (povezanih npr. s krajo identitete). Vse to temelji na tem, da

zmorejo ti modeli brez posebnih stroškov ustvariti širok nabor besedil, ki so navidezno podobna besedilom človeških avtorjev in prepričljiva. Zlonamerna raba pa je lahko povezana tudi z ustvarjanjem škodljivih programskih orodij oz. kod, ki ogrozijo kibernetsko varnost. Nenazadnje v ta nabor tveganj spada tudi uporaba teh modelov za nezakonit nadzor in cenzuro, saj omogočajo hiter pregled ogromne količine besedilnih podatkov.

Peto področje tveganj zadeva morebitne škodljive posledice, ki izhajajo iz interakcije med človekom in sistemi umetne inteligence, ki so v nekaterih primerih lahko del širšega sistema oz. stroja (npr. robot za oskrbo bolnikov ali izobraževanje). Ta vrsta tveganj vključuje možnost, da zaradi podobnosti takšne interakcije z običajno interakcijo posameznik sistemu zaupa v meri, ki ni ustrezna; lahko pride tudi do poglobitve stereotipov in predsodkov (npr. pogovorni robot pomočnik ima že izbrano žensko ime ali pa mu ga izbere uporabnik). Prav tako lahko pri uporabniku pride do zmotne predstave, da imajo ti sistemi posebno in relativno trajno identiteto in da so zmožni sočutja (posebej je to relevantno pri pogovornih robotih, ki so namenjeni krepitvi duševnega zdravja ali družabništvu). Širše lahko prihaja do zlorab vzpostavljenega zaupanja in do tega, da operaterji teh sistemov zaupanje in druge podatke izrabljajo za namene prepričevanja ali motiviranja (npr. za nakup določenih izdelkov). Omeniti velja še, da kadar imajo ti sistemi določene cilje, lahko vzorce in strategije, ki predstavljajo nekakšno obliko manipulacije, vzpostavljajo sami.

Zadnje, šesto področje tveganj predstavljajo tveganja, povezana z okoljskimi in družbenoekonomskimi posledicami uporabe velikih jezikovnih modelov. Za sámo učenje in delovanje teh obsežnih sistemov je potrebna znatna količina energije in drugih virov. Zaradi porazdeljenosti virov in hitrosti avtomatizacije, ki jo omogočajo, lahko prihaja do poglobitve neenakosti in nesorazmerne porazdelitve bremen in koristi. Hkrati se odpira tudi vprašanje dela in obsega delovnih mest ter kakovosti teh delovnih mest. Nekatera področja dela so lahko še posebej ogrožena, pa tudi sam tehnološki razvoj je zaradi omejene dostopnosti teh tehnologij porazdeljen neenakomerno in bo verjetno tudi neenakomerno pospešen – kar bo poglobilo vrzel pri drugih oblikah razvitosti in kakovosti življenja (Weidinger idr. 2022).

3. Etični izzivi velikih jezikovnih modelov

Etični izzivi velikih jezikovnih modelov so v veliki meri povezani z že omenjenimi tveganji. Na podlagi teh prepoznanih tveganj se odpirajo globlja oz. bolj temeljna vprašanja (Green 2018). Prvi sklop etičnih izzivov lahko osredinimo na vidik informacij in podatkov ter njihove uporabe; povezani so z vidikom resnice in resnicoljubnosti. Sem spadajo primeri napačnih informacij in manipulacij (lažne vsebine), pa tudi problem zasebnosti in uporabe sicer resničnih informacij – v tem primeru gre med drugim za kršitev pravice do zasebnosti. Hkrati se odpira tudi vprašanje same uporabe učnih podatkov (avtorske pravice, povezane z besedili, na podlagi katerih se ti sistemi učijo podatkov). Tesno je s tem povezan tudi vidik predsodkov

in pravičnosti. Jezikovni modeli lahko okrepijo obstoječe pristranskosti, prisotne v podatkih za učenje, in tako odražajo družbene pristranskosti, povezane z raso, spolom ali drugimi občutljivimi lastnostmi (Deery in Bailey 2022). To lahko vodi k pristranskim ali nepravičnim rezultatom, ki krepijo in ohranjajo družbeno neenakost.

Naslednji sklop etičnih izzivov je povezan z vidikom transparentnosti, razložljivostjo delovanja in soglasjem uporabnikov. Transparentnost terja tako to, da uporabnik natančno ve, da je v stiku s sistemom umetne inteligence, kot tudi to, da razume, na kakšen način ta sistem deluje – in sicer v primeru, ko je tak sistem uporabljen kot priporočilni oz. odločitveni sistem (veliki jezikovni modeli se sicer neposredno za ta namen ne uporabljajo, vendar pa bi v prihodnosti to vlogo lahko privzeli). Pri uporabi teh sistemov je treba vzpostaviti oz. zagotoviti jasne mehanizme soglasja – uporabniki pa bi morali biti s podatki, ki se v interakcijah s temi sistemi zbirajo, seznanjeni (Miklavčič 2021).

Tretji sklop zajema vidike globalne pravičnosti ter nevarnosti neravnovesja moči in odvisnosti. Znanje in usmerjanje razvoja takšnih sistemov (skupaj z njihovo uporabo) je trenutno v rokah nekaj organizacij ali posameznikov, ki imajo svoj nabor motivacij, ciljev in idealov. Ta koncentracija moči lahko privede do neenakega dostopa in omejenega nadzora posameznikov in skupnosti nad sistemi umetne inteligence. Obenem druge subjekte to postavlja v odnos odvisnosti. Dodajmo še, da uvajanje velikih jezikovnih modelov lahko vodi v pomembne posledice za zaposlovanje oz. trg dela, posebej glede delovnih mest, ki vključujejo naloge, ki jih je mogoče s temi sistemi avtomatizirati. To sicer v zgodovini človeštva ni nov pojav, a odprto ostaja vprašanje, kakšen bo obseg teh posledic in hitrost njihovega širjenja – in ali smo na družbene in gospodarske posledice takšnih pretresov pripravljeni.

Četrti sklop izzivov zadeva interakcijo med človekom in sistemi umetne inteligence ter stroji. Veliki jezikovni modeli lahko spremenijo dinamiko človeške komunikacije in interakcije. Vse večja uporaba klepetalnih robotov in virtualnih pomočnikov, ki jih ti modeli poganjajo, lahko privede do spremembe načina, kako posamezniki s tehnologijo sodelujejo, kar pa lahko vpliva na družbene norme, zaupanje in kakovost medčloveških odnosov. V ospredju je vprašanje tehnologije in delegiranja vse bolj osebnega, miselnega in čustvenega dela sistemom umetne inteligence. To ima lahko pomembne etične implikacije, še posebej če preide na področja dela, ki vključujejo tudi npr. skrb ali sočutje (Dorobantu idr. 2022, 21–22).

Peti sklop izzivov zadeva vprašanje pristnosti, avtorstva, avtoritete in ustvarjanja znanja. Razširjanje ustvarjenih vsebin velikih jezikovnih modelov sproža vprašanja o avtorstvu in izvirnosti dela, ki ga je ustvaril človek. Pogosto je namreč težko razbrati, ali je vsebino ustvaril človek ali sistem umetne inteligence, kar lahko zmanjša vrednost in zaupanje, povezano s človeško ustvarjalnostjo in izražanjem (gre zlasti za besedilo, pri multimodalnih sistemih tudi ustvarjeno sliko, melodijo idr.). Kot primer lahko navedemo študijo, ki je povezana s področjem filozofije (Schwitzgebel idr. 2023), v okviru katere so raziskovalci jezikovni model izurili na delih Daniela Dennetta. Nato so Dennettu samemu postavili deset filozofskih vprašanj o njegovih stališčih. Enak niz vprašanj so večkrat zastavili tudi sistemu umetne inteligence in

zbrali vse odgovore. Zanimalo jih je, v kolikšni meri lahko vsaka skupina (strokovnjaki za Dennettovo delo, strokovna publika in laična javnost) prepozna, kateri odgovor je zares podal Dennett in katerega je ustvaril jezikovni model. Rezultati so bili presenetljivi, saj v mnogih primerih tudi sami strokovnjaki za Dennettovo delo večinsko niso izbrali pravih odgovorov. Predstavniki strokovne in splošne javnosti so medtem izbirali bolj ali manj zgolj z mero pravilnosti, ki jo doseže tudi golo ugibanje. Nadalje, razširjena uporaba velikih jezikovnih modelov in zanašanje nanje lahko vplivata na strukture ustvarjanja znanja in avtoritete. Ustvarjena vsebina se lahko dojema kot avtoritativna ali natančna zgolj zaradi načina delovanja – ne glede na njeno dejansko veljavnost ali verodostojnost. To lahko vpliva na distribucijo in dojemanje znanja v družbi. Vse to pa ima posledice tudi za izobraževalni proces in odpira vprašanja glede vloge človeških učiteljev, pomena kritičnega mišljenja in vrednosti lastne ustvarjalnosti. Pomembno je izpostaviti tudi, da veliki jezikovni modeli lahko ustvarjajo zelo prepričljiva in skladna besedila, kar vzbuja pomisleke glede dezinformacij, zavajanja in manipulacij. Vprašanje avtorstva pa spremlja tudi vprašanje odgovornosti za vsebino, ki jo ustvarjajo. Pri velikih jezikovnih modelih ne gre za besedilo, ki bi bilo primer običajnega pisanja: nima avtorja, ne gre za pričanje oz. pričevanje – ne gre za govorna dejanja. Kar pomeni, da se zastavlja vprašanje naštetih vlog oz. kdo naj jih pri tako tvorjenih besedil prevzame.

Šesti sklop predstavljajo vprašanja o sami etičnosti velikih jezikovnih modelov oz. implementaciji etičnega delovanja in o načinu te implementacije (Tolmeijer idr. 2020; Constantinescu in Crisp 2022). Pri velikih jezikovnih modelih – podobno kot pri drugih sistemih umetne inteligence – o etičnosti seveda ne moremo govoriti neposredno, saj jim ne moremo pripisati zavesti, namer, motivacije ali značaja. Vseeno pa se vprašanje etičnosti odpira že pri samem razvoju in upravljanju teh sistemov (npr. plačljiv terapevtsko-družabniški sistem je lahko zasnovan tudi na način, da poskuša karseda povečati čas, ki ga posameznik preživi v stiku s tem sistemom, pri tem pa ne zasleduje cilja pomoči posamezniku, ampak cilj povečanja višine plačila za storitev, ki jo nudi). Ta vprašanja se še poglobijo, ko govorimo o sistemih umetne inteligence, ki morajo ali bodo morali sprejemati etične odločitve v zapletenih situacijah, kot je to npr. pri avtonomnih vozilih ali zdravstveni oskrbi (Tolmeijer idr. 2020).

Kot zadnji sklop izzivov lahko izpostavimo te, ki so vezani na vidike kulture (prilaščanje, predstavljanje) ter jezika in jezikovne raznolikosti. Kot omenjeno, lahko veliki jezikovni modeli kulturne stereotipe in predsodke prevzamejo in okrepijo, saj ne posedujejo kulturne občutljivosti in razumevanja. Hkrati lahko pri vidiku jezika privedejo oz. prispevajo k ogroženosti jezikov in zmanjševanju jezikovne raznolikosti, če so viri in prizadevanja pri razvoju teh sistemov usmerjeni pretežno le v najbolj uporabljane jezike.

Za reševanje teh izzivov je potreben pristop z več deležniki, to pa vključuje tako raziskovalce, razvijalce, oblikovalce politik kot tudi širšo javnost. Vse to pa zajema tudi razvoj etičnih smernic za to področje – temelječih na premislekih o človekovem dostojanstvu, človekovih pravicah ter pomenu raznolikosti in bogastva človeške kulture in komunikacije.

3.1 Etične smernice in veliki jezikovni modeli

Obstoječe etične smernice oz. priporočila za sisteme umetne inteligence sicer niso posebej naravnani na velike jezikovne modele, vseeno pa lahko razberemo poglavitne etične omejitve razvoja in rabe teh sistemov. Etične smernice se nanašajo na različne sklope razvoja in uporabe velikih jezikovnih sistemov, in sicer od področja soglasja in transparentnosti delovanja (privolitev in obveščenost uporabnika, nadzor nad osebnimi podatki ter njihovim razširjanjem, spoštovanje zasebnosti, sprejemanje informiranih odločitev na podlagi generiranih vsebin ipd.), preko etičnih premislekov, vezanih na tvorjenje besedila oz. informacij (preprečevanje lažnih informacij, neresničnih vsebin oz. halucinacij; preprečevanje manipulacij, odgovornost za tvorjene vsebine, iskanje ravnotežja med ustvarjalnostjo teh sistemov in odgovornostjo za vsebine) do specifičnih vprašanj, povezanih s temami in vsebinami, ki so kočljive oz. sporne (npr. predsodki in pristranskost, vprašanje varnosti posameznika in družbe, seksualno eksplicitne vsebine ipd.).

Etične smernice za zaupanja vredno umetno inteligenco, ki jih je izdala Evropska komisija (2019), tako izrecno poudarjajo zlasti vidik transparentnosti in z njo povezane razložljivosti. Uporabniki morajo biti seznanjeni s tem, da so v stiku s sistemom umetne inteligence, pri čemer je treba odkrito navajati tudi zmogljivosti in namen sistemov umetne inteligence; razložljivost naj vsebuje tudi vidike sledljivosti in možnosti revidiranja – pri tem je zahtevana stopnja odvisna od okoliščin in resnosti posledic napačnega ali netočnega rezultata iz tega sistema. Smernice izhajajo iz načel spoštovanja človekove avtonomije, preprečevanja škode, pravičnosti in razložljivosti, izpostavljajo pa zahteve človekovega delovanja in nadzora, tehnične robustnosti in varnosti teh sistemov, zasebnosti in varovanja podatkov, preglednosti, raznolikosti, nediskriminacije in pravičnosti, družbene in okoljske blaginje ter odgovornosti (Evropska komisija 2019).

Posebno pozornost tem etičnim izzivom namenja tudi nastajajoči evropski „Akt o umetni inteligenci“, ki predstavlja prvo celovito zakonodajo za področje umetne inteligence. Poslanci in poslanke evropskega parlamenta so glede na predhodne osnutke že sprejeli pogajalska izhodišča, sprejem pa se pričakuje do konca leta 2023. Pomembno je, da ta zakonodaja velike jezikovne modele umešča med tako imenovane generativne sisteme, pri katerih tveganje ob uporabi ni visoko, lahko pa postane, če bi jih uporabili npr. za namen vplivanja na izide volitev ipd. Akt predvideva, da mora biti pri velikih jezikovnih modelih uporabniku jasno povedano, da so bile vsebine ustvarjene z umetno inteligenco, upravljalec sistema pa mora sprejeti ukrepe proti ustvarjanju nezakonitih vsebin. Prav tako osnutek zakona vključuje zahtevo, da morajo biti besedila, ki so zaščitena z avtorskimi pravicami in na katerih se je sistem učil, povzeta in javno dostopna (Evropski parlament 2023).

Glede na smer razvoja bo v prihodnje posebno pozornost treba posvetiti tudi bolj specifičnim oz. usmerjenim etičnim smernicam, še zlasti glede implementacije velikih jezikovnih modelov v robotske sisteme, ki bodo zmožni npr. skrbi za starejše in bolne (Miklavčič 2021) ali izobraževanja otrok (Yan idr. 2023). Hitro se razvija tudi področje družabniških sistemov ter uporaba velikih jezikovnih mode-

lov za terapevtsko dejavnost (Kraus, Seldschopf in Minker 2021), pri čemer še ni posebnih oz. obvezujočih etičnih smernic.

4. Teološki izzivi velikih jezikovnih modelov

Tudi teološki izzivi velikih jezikovnih modelov se bodo v veliki meri povezovali z že izpostavljenimi tveganji in etičnimi izzivi. Vseeno pa lahko omenimo nekaj dodatnih premislekov, posebej povezanih tudi z antropološkimi vidiki. Pri etičnih izzivih novih tehnologij in vizijah razvoja sveta se namreč lahko vedno vprašamo, na kakšni implicitni antropologiji temeljijo in ali je ta implicitna antropologija takšna, da našo človeškost pospešuje ali zavira (Dorobantu idr. 2022, 7).

Bog je ustvaril svet z besedo (Jn 1,1-4) – in ustvaril tudi človeka z jezikom oz. sposobnostjo govora. Sedaj pa smo mi ustvarili sisteme, ki imajo sposobnost tvorbe oz. vsaj posnemanja tvorbe jezika. Kot izhodiščno filozofsko vprašanje se zato zastavlja vprašanje pomena in posledic tega razvoja. Ena smer razmišljanja izpostavlja področje človeškega dostojanstva, edinstvenost človeka in stvarstva. Če lahko stroji ustvarijo človekovemu podoben jezik in izvajajo ustvarjalne naloge, ali to edinstvenost in vrednost človeka kot Božje stvaritve zmanjšuje? Edinstvenost človeških bitij je ena temeljnih predpostavk skoraj vseh kultur in religij, ne le krščanstva (Dorobantu idr. 2022). Vendar prav krščanstvo posebej poudarja tudi vidik, da je človek ustvarjen po Božji podobi (*imago Dei*) – to paradigmo je treba z vidika teološke antropologije ustrezno razumeti, sicer jo lahko razvoj umetne inteligence (zlasti splošne ali zmogljive umetne inteligence) postavi pod vprašaj. »Posledice za krščansko teologijo so globoke. Ideji človekove edinstvenosti in *imago Dei* sta v jedru naše antropologije in sta povezani z vsem drugim: nista vgrajeni le v protologijo, temveč tudi v kristologijo in eshatologijo. Pravzaprav je težko najti katero koli vejo teologije, ki ne bi bila povezana s konceptom Božje podobe.« (2022, 21) Dorobantu zato predlaga, da moramo posebej razmisliti o razlagi stvarstva po Božji podobi in se približati modelu, ki poudarja eshatološki vidik prihodnje usmeritve in odnosa z Bogom, ne pa sklop človeških značilnosti.

Nadalje ima razvoj velikih jezikovnih modelov pomembne posledice tudi za odnos med človekom in tehnologijo. Teološka vprašanja se lahko nanašajo na vpliv obsežnih jezikovnih modelov na človeške odnose in na razumevanje samega odnosa. Ti modeli lahko vplivajo tudi na sporazumevanje med ljudmi, empatijo in zmožnost smiselnega dialoga, kar posledično vpliva na razvoj pristnih človeških vezi. Ali veliki jezikovni modeli vodijo v smer pristnejših odnosov ali morda bolj poudarjajo individualizem in samozadostnost (2022)? Praktične dilema in vprašanja, ki se pri tem pojavljajo, so med drugim povezane s perspektivo oziroma razvojem robotskih duhovnikov in pogovornih sistemov, ki temeljijo na svetih besedilih. Pri obsežnih jezikovnih modelih gre obenem tudi za osiromašenje bogatega metaforičnega jezika svetih spisov. Ker so ti sistemi zmožni nadgradnje z vidiki virtualne ali obogatene resničnosti, lahko opozorimo tudi na naslednje: v navidezni resničnosti je vse virtualno, vera pa ne more biti zvedena na virtualni svet.

Pomenljiv je primer evharistije, saj sta hrana in hranjenje je nekaj, česar ni mogoče v celoti tehnologizirati; enako velja za evharistijo. V evharistiji moramo biti mi sami resnično tam – in Bog je resnično tam.

Eric Salobir in njegovi soavtorji (Dorobantu idr. 2022) poudarjajo, da sistemov umetne inteligence nismo samo ustvarili, temveč nas ta razvoj spreminja, pri čemer ostaja odprto vprašanje, ali na bolje ali ne. Teološke izzive umetne inteligence na splošno (ne zgolj obsežnih jezikovnih modelov) delijo v tri sklope, in sicer: (i) umetna inteligenca in naši odnosi z drugimi in svetom, (ii) umetna inteligenca in naš odnos z Bogom oz. presežnim ter (iii) umetna inteligenca in odnos do nas samih.

V okvir prvega sklopa spada izhodiščni premislek, da smo človeška bitja to, kar smo, prvenstveno preko odnosov z drugimi (Juhant in Strahovnik 2010). In kot smo omenili že zgoraj, delegiranje dela, ki je vse bolj osebno, miselno in čustveno delo, umetni inteligenci lahko privede po škodljivih posledic. Je res možno delegirati vse vrste dela, tudi npr. skrb, ljubezen, sočutje? Kako to razumeti v luči zapovedi, da moramo ljubiti Boga in bližnje (Dorobantu idr. 2022, 21–22)? Sistemi umetne inteligence nam lahko pomagajo delovati, ne morejo in ne smejo pa nas nadomestiti – posebej ne v prostoru odnosov. Hkrati je treba opozoriti tudi, da gre ena veja razvoja umetne inteligence v smer, da postanemo bolj ali manj pasivni (npr. odnosi z drugimi nam postanejo breme, ker je odnos s klepetalnim robotom toliko lažji; človeški odnosi potem za naš razvoj in človeškost niso več konstitutivni), kar vodi do razvoja šibkosti in hib, ne pa izboljšanja (11).

V okvir drugega sklopa lahko umestimo premislek, ali se preko razvoja umetne inteligence ljudje morda postavljamo v vlogo Boga in s pomočjo teh sistemov težimo k naši predrugačitvi in nadgraditvi (Žalec 2019; Globokar 2019) – morda celo nesmrtnosti. Z velikimi jezikovnimi modeli smo se tudi približali ustvarjenju sistema, ki lahko misli sam, kar odpira pomembna vprašanja naše moči ter pobožanstvenja (Dorobantu idr. 2022, 12). Pri tem lahko pripomnimo, da jezik ni zgolj nekaj instrumentalnega, ampak nekaj, kar ima pomembno neodnosno vrednost. Tu je še vidik izrivanja razsežnosti milosti, daru in božje previdnosti ter drugih vidikov naključja iz naših življenj (v smeri napovedljivosti in obvladljivosti vseh vidikov življenja), obenem pa razvoj v smeri ‚dataizma' (informacije kot temeljna podstat vsega) in mehanicističnega razumevanja stvarstva (11). V temelju mora zato ostati odnos s presežnim.

> »Medtem ko si je utelešenje Boga v virtualni resničnosti težko predstavljati, so verske izkušnje morda bolj obetavne, saj so že v izhodišču bolj subjektivne. In res so postale predmet iskanja tehnoloških nadomestkov. To iskanje pa lahko pomeni performativno samopotrjevanje: kajti ali je lahko kaj, česar Bog ni povzročil, izkušnja Boga? Naj Ray Kurzweil še tako fantazira o tehnološko povzročenih ›religioznih izkušnjah‹ (s stimulacijo možganov), to niso oblike osebnega srečanja z Drugim, niti prave oblike samotranscendence. Vsaka manipulacija kot taka izključuje osebno srečanje, če se takšna srečanja po svoji naravi morajo zgoditi v svobodnem podarjanju samega sebe. Ta brezplačna svoboda je še posebej potrebna, če je zadev-

> ni drugi Bog, saj Bog po svoji naravi presega vsako naše razumevanje in ga lahko srečamo le v njegovem lastnem daru milosti.« (39)

V okvir tretjega sklopa spadajo premisleki o vplivu in vlogi umetne inteligence v vsakdanjem življenju. Prav lahko se zgodi, da bodo ti sistemi odgovarjali na vsako našo željo in kaprico, zato se bomo razvijali v smeri samozadostnosti in lenobe. Trud je po drugi strani hvalevreden ne glede na posledice – npr. trud pri skrbi za bližnje, trud pri učenju ipd. Hkrati gre z napredkom sistemov umetne inteligence razvoj v smeri vse bolj zaznavnega občutka nelagodnosti in tesnobe, ki te sisteme spremlja. Umetna inteligenca lahko tako predstavlja grožnjo človekovi avtonomiji, ker razvoj kaže v smeri atropije naših sposobnosti. S prelaganjem odločitev ali delovanja na sisteme umetne inteligence obenem sproža tudi vprašanje naše odgovornosti (Dorobantu idr. 2022).

5. Zaključek

Jezik in jezikovne tehnologije so že od nekdaj del človekovega sveta in s tem tudi človeške domišljije. Vprašanje jezikovnih modelov se je na pomemben način pojavilo v kratki zgodbi „Devet milijard imen Boga“ (1953) A. C. Clarka. V zgodbi spremljamo menihe v tibetanskem samostanu, ki si prizadevajo našteti in zapisati vsa božja imena. Verjamejo, da je bilo vesolje ustvarjeno za ta namen in da bo po koncu naštevanja imen Bog svet kot svojo stvaritev izničil. Tako so menihi pred tremi stoletji ustvarili sistem in abecedo, s katero so izračunali, da lahko naštejejo vsa možna Božja imena in da jih je približno devet milijard (predpostavka je, da je ime lahko sestavljeno iz največ devetih znakov). Imena so začeli zapisovati ročno, ker pa je to zamudno opravilo, so na koncu najeli računalnik, ki lahko izpiše vse možne permutacije imen, in dva zahodna strokovnjaka, ki sta stroj programirala in upravljala. Upravljalca računalnika sta skeptična, vendar nalogo vseeno sprejmeta. Po treh mesecih, ko je izpis vseh imen skoraj dokončan, se ustrašita, da ju bodo menihi obtožili, da je podvig spodletel zaradi njiju, zato samostan tik pred koncem naloge zapustita. V zvezdnati noči, ko se ustavita na poti po pobočju gore, ki vodi do letališča, izračunata, da je stroj zelo verjetno svojo nalogo pravkar opravil ... in ko pogledata v nebo, vidita, da zvezde preprosto začnejo ugašati.

Na drugi strani je Jorge Luis Borges v zgodbi *Babilonska knjižnica* opisal knjižnico, sestavljeno iz medsebojno povezanih šesterokotnih sob ali knjižničnih celic, katerih sestav nima konca. Za naše premisleke je zanimiva njegova zamisel o vsebini knjig v tej knjižnici. Vsaka od njih ima štiristo deset strani s štiridesetimi vrsticami na vsaki strani in približno osemdesetimi znaki v vsaki vrstici. Knjige vsebujejo vse možne kombinacije vseh možnih besed ali zaporedij črk in ločil. Nobeni dve knjigi v knjižnici nista enaki. Po drugi strani pa je popolna, saj vsebuje »vse, podrobno zgodovino prihodnosti, avtobiografije nadangelov, zvesti katalog knjižnice, tisoče in tisoče lažnih katalogov, dokaz o lažnosti teh lažnih katalogov, dokaz o lažnosti pravega kataloga, gnostični Bazilidov evangelij, komentar tega evangelija, komentar komentarja tega

evangelija, pravo zgodbo o tvoji smrti, prevod vsake knjige v vse jezike, interpolacije vsake knjige v vse knjige, traktat o mitologiji Saksoncev, ki bi ga Beda Častitljivi lahko napisal (a ga ni), izgubljene Tacitove knjige.« (Borges 1998, 115) V zgodbi Borges opisuje, kako so pred stoletji bralci oz. prebivalci neke knjižnične celice našli knjigo, ki je vsebovala skoraj dve strani smiselnega besedila. Ključno pa je, da se zavedamo, da vse knjige obstajajo. Sprva se v zgodbi ob tem spoznanju pojavi občutek vznesenosti, ko se vsi počutijo, kot da imajo v lasti nek prvinski in skrivni zaklad, nato pa sledita razočaranje in depresija – ker so bile knjige, ki so bile smiselne, praktično nedosegljive ali jih ni bilo mogoče najti. Zdi se, da se s pojavom obsežnih jezikovnih modelov takšnemu stanju na neki način približujemo.

> »Gotovost, da je bilo vse že napisano, nas izniči ali pa nas naredi fantazmatske. Poznam okrožja, v katerih se mladi priklanjajo pred knjigami in kot divjaki poljubljajo njihove strani, čeprav ne znajo prebrati niti črke. Epidemije, heretični spori, romanja, ki se neizogibno sprevržejo v razbojništvo, so zdesetkali prebivalstvo. Mislim, da sem že omenil samomore, ki so vsako leto pogostejši. Morda me vara starost in strah, vendar sumim, da se človeška vrsta, edina vrsta, giblje po robu izumrtja. Pa vendar bo Knjižnica – razsvetljena, osamljena, neskončna, popolnoma negibna, oborožena z dragocenimi zvezki, brezpredmetna, nepokvarljiva in skrivnostna – obstala.« (118)

Če povzamemo: teološki izzivi, ki jih odpirajo veliki jezikovni modeli, se križajo z etičnimi vprašanji, povezanimi z umetno inteligenco, kar spodbuja ponovno ovrednotenje naših prepričanj o človeštvu in tehnologiji. Ena od ključnih skrbi se vrti okoli človekovega dostojanstva in edinstvenosti posameznikov kot Božjih stvaritev. Ker veliki jezikovni modeli ustvarjajo človekovemu podoben jezik, se pojavljajo vprašanja o vrednosti človeka. Koncept ustvarjenosti po Božji podobi (*imago Dei*), ki je v krščanski teologiji osrednjega pomena, zahteva ob napredku umetne inteligence poglobljeno prevpraševanje. Poleg tega obsežni jezikovni modeli vzbujajo skrb glede pristnih človeških vezi, obenem pa tudi verskih skupnosti, ki bi posredovanim odnosom sledile – povezovanje teh modelov z virtualno in razširjeno resničnostjo namreč postavlja pod vprašaj versko avtentičnost, vključno z obredi in praksami.

Reference

Bender, Emily M., in Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. V: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.463/ (pridobljeno 23. 4. 2023).

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major in Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? V: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. New York: Association for Computing Machinery. ACM Digital Library. https://dl.acm.org/doi/10.1145/3442188.3445922 (pridobljeno 23. 4. 2023).

Borges, Jorge Luis. 1998. *Collected Fictions.* New York: Viking.

Clarke, Arthur C. 1953. The Nine Billion Names of God. V: Frederik Pohl, ur. *Star Science Fiction Stories No. 1*, 195–199. New York: Ballantine Books.

Constantinescu, Mihaela, in Roger Crisp. 2022. Can Robotic AI Systems Be Virtuous and Why Does This Matter? *International Journal of Social Robotics* 14:1547–1557. https://doi.org/10.1007/s12369-022-00887-w

Deery, Oisín, in Katherine Bailey. 2022. The Bias Dilemma: The Ethics of Algorithmic Bias in Natural-Language Processing. *Feminist Philosophy Quarterly* 8, št. 3–4. https://ojs.lib.uwo.ca/index.php/fpq/article/view/14292 (pridobljeno 8. 5. 2023).

Dorobantu, Marius, Brian Patrik Green, Anselm Ramelow in Eric Salobir. 2022. Being Human in the Age of AI. Research gate. https://www.researchgate.net/publication/365945288_Being_Human_in_the_Age_of_AI?channel=doi&linkId=6389b3b82c563722f22e84df&showFulltext=true (pridobljeno 8. 5. 2023).

Dorobantu, Marius. 2021. Human-Level, but Non-Humanlike: Artificial Intelligence and a Multi-Level Relational Interpretation of the Imago Dei. *Philosophy, Theology and the Sciences* 8, št. 1:81–107. https://doi.org/10.1628/ptsc-2021-0006

Dorobantu, Marius. 2022. Strong Artificial Intelligence and Theological Anthropology: One Problem, Two Solutions. V: P. Jorion, ur. *Humanism and its Discontents: The Rise of Transhumanism and Posthumanism*, 19–33. New York: Palgrave/MacMillan.

Evropska komisija. 2019. *Etične smernice za zaupanja vredno umetno inteligenco.* Bruselj: Evropska komisija.

Evropski parlament. 2021. Briefing: Artificial intelligence act. Evropski parliament. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

Floridi, Luciano. 2023. AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology* 36, št. 15:1–7.

Frankfurt, Harry. 2005. *On Bullshit.* Princeton: Princeton University Press.

Globokar, Roman. 2019. Normativnost človeške narave v času biotehnološkega izpopolnjevanja človeka. *Bogoslovni vestnik* 79, št. 3:611–628. https://doi.org/10.34291/bv2019/03/globokar

Green, Brian Patrick. 2018. Ethical Reflections on Artificial Intelligence. *Scientia et Fides* 6, št. 2:9–31. https://doi.org/10.12775/setf.2018.015

Juhant, Janez, in Vojko Strahovnik. 2010. Pristno delovanje je vzajemnost in sodelovanje. *Bogoslovni vestnik* 70, št. 3:351–364.

Kraus, Matthias, Philip Seldschopf in Wolfgang Minker. 2021. Towards of Trustworthy Chatbot for Mental Health Applications. V: Jakub Lokoč, Tomáš Skopal, Klaus Schoeffmann, Vasileios Mezaris, Xirong Li, Stefanos Vrochidis in Ioannis Patras, ur. *MultiMedia Modeling*, 354–366. New York: Springer International Publishing.

Miklavčič, Jonas. 2021. Zaupanje in uspešnost umetne inteligence v medicini. *Bogoslovni vestnik* 81, št. 4:935–946. https://doi.org/10.34291/bv2021/04/miklavcic

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike in Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*:2203.02155.

Raaijmakers, Stephan. 2022. *Deep learning for natural language processing*. Shelter Island, NY: Manning.

Schwitzgebel, Eric, David Schwitzgebel in Anna Strasser. 2023. Creating a large language model of a philosopher. *ArXiv*:2302.01339.

The Future of Life Institute. 2023. Pause Giant AI Experiments: An Open Letter (22. 3. 2023). Future of Life. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (pridobljeno 4. 7. 2023)

Tolmeijer, Suzanne, Markus Kneer, Cristina Sarasua, Markus Christen in Abraham Bernstein. 2020. Implementations in Machine Ethics: A Survey. *ACM Comput. Surv.* 53, št. 6: članek 132.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving in Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In: *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022*. Seoul. ACM Digital Library. https://dl.acm.org/doi/10.1145/3531146.3533088 (pridobljeno 21. 7. 2023).

y Arcas, Blaise Agüera. 2022. Do Large Language Models Understand Us? *Daedalus* 151, št. 2:183–197.

Yan, Lixiang, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin in Dragan Gašević. 2023. Practical and ethical challenges of large language models in education: A systematic scoping review. *arXiv*:2303.13379.

Žalec, Bojan. 2019. Liberalna evgenika kot uničevalka temeljev morale: Habermasova kritika. *Bogoslovni vestnik* 79, št. 3:629–641. https://doi.org/10.34291/bv2019/03/zalec