

Etika umetne inteligence v sistemu izobraževanja

Prispevek je bil pripravljen kot vabljeni predavanje na X. znanstvenem posvetu Vodenje v vzgoji in izobraževanju 2023, ki je potekal 7. in 8. marca 2023 v Portorožu na temo integritete v vzgoji in izobraževanju kot izhodišča za pravičnost in vključenost v organizaciji Šole za ravnatelje, Zavoda RS za šolstvo.

O avtorici

Dr. Polona Pičman Štefancič je direktorica Raziskovalnega centra družbe Rea IT in ustanoviteljica Sladkih nasmehov, društva za pomoč otrokom v stiski ter predstavnica svetovalne hiše na področjih poslovanja in informacijske tehnologije Gartner, Inc. Predava v Sloveniji in tujini ter je avtorica in soavtorica več knjig. Je tudi akademska glasbenica.
polona.picman-stefancic@rea-it.si

Povzetek

O digitalizaciji govorimo že desetletje – z zaznavnimi uspehi. O umetni inteligenci pa, nasprotno, vemo zelo malo, pa vendar se zdi, da nas od nje nega razmaha ločijo prej meseci kot leta. Osnovne in srednje šole v ZDA se prav zdaj že začenjajo opredeljevati do problemov ohranjanja akademske integritete, ki jih odpira umetna inteligenca – a ne zato, ker želijo biti korak pred svetom, pač pa zato, ker je problematika tam že postala pereča. Napačno je pričakovati, da se bo Slovenija izognila novim izzivom. Ne nazadnje si tega niti ne želimo. Smo pa lahko uspešnejši, če smo pripravljeni.

Zato je ključnega pomena, da učitelji in učenci razumejo, ozavestijo in se naučijo upravljati omejitve, morebitna tveganja in etične pomanjkljivosti aplikacij umetne inteligence v izobraževanju. V tem duhu je namen prispevka osvetliti prednosti, ki jih ponuja razvoj umetne inteligence v izobraževalnem okolju, in obenem ponuditi orodja za spoprijemanje z etičnimi izzivi vključevanja umetne inteligence v učilnice.

Ključne besede: digitalizacija | umetna inteligenca | etika

The Ethics of Artificial Intelligence in Education

Polona Pičman Štefančič, PhD, Gartner Slovenija and Rea IT d. o. o.

Abstract

We have been talking about digitalisation for a decade, giving rise to tangible achievements. On the other hand, we know very little about artificial intelligence, although we seem months rather than years away from its breakthrough. Primary and secondary schools in the US have now begun to address the problems raised by AI regarding maintaining academic integrity. This is not because they want to be one step ahead of the rest of the world, but because the issue has already gained momentum in the US. It is wrong to assume that Slovenia will avoid the new challenges, nor do we want to. We can, however, achieve a better outcome by being prepared.

Teachers and students must therefore understand, become aware of and learn to manage the limitations, potential risks and ethical shortcomings of AI applications in education. In this spirit, the present paper aims to highlight the benefits offered by the development of AI in educational settings, while also providing tools to address the ethical challenges of integrating AI in the classroom.

Keywords: digitisation | artificial intelligence | ethics

Stephen Hawking (*Galeon 2016*)

Najbolj rigiden odgovor je – ker nam tako narekujejo smernice EU, ker je to del programa Erasmus+, ker je to integralni del evropske digitalne politike, ker je to usmeritev naše skupne evropske prihodnosti.

»Ker umetnointeligenčni sistemi čedalje bolj vplivajo na vse vidike človekove dejavnosti, se zdi nujno, da v vsej EU izobraževalci razumejo, kako naj orodja umetne inteligence etično uporabljajo pri poučevanju in kako naj jih učenci uporabljajo pri učenju.« (Evropska komisija, Generalni direktorat za izobraževanje, mladino, šport in kulturo 2022a, 6)

Seznam aktualnih evropskih politik in regulative je, navkljub novosti področja, obsežen, kar kaže na jasno odločenost EU, da ažurno začrta sprejemljive poti razvoja UI. A s politikami, direktivami in smernicami je (praviloma) pač tako, da ne zajamejo tistega, kar dela svet zanimiv; v njihovi naravi je, da vznemirljive stvari prevedejo v suhoparen uradniški jezik – ta mora biti tehnično korekten in hkrati neposredno neuporaben. Z malo šale: včasih se zdi, da se celo lektorji, prevajalci in uredniki preprosto – predajo.



Vsebinski odgovor je seveda dosti vznemirljivejši: umetna inteligenca je del našega življenja. Morda o njej niti ne razmišljamo, a je povsod okoli nas – in

uporabljamo jo že leta. Ko iščemo v Googlu, beremo e-pošto, sprašujemo za navodila za vožnjo, prejemamo priporočila za Netflix – nenehno uporabljamo umetno inteligenco. Ker nam koristi. Ker je z njo lažje. Ker nam pomaga.

Umetna inteligenca tudi še zdaleč ni nova. Že res, da je pravi razcvet doživela šele zadnje leto – zlasti izjemen napredek v neposredni uporabnosti aplikacij za končne uporabnike, na primer ChatGPT, je znatno prispeval k popularizaciji sistema – a modeli strojnega učenja, produkcija algoritmov in obdelava naravnega jezika, ki vsi temeljijo na umetni inteligenci, so med nami že leta.

Umetno inteligenco lahko na kratko definiramo kot skupino tehnologij, ki se ukvarjajo s simulacijo inteligentnega vedenja računalnikov in z njihovo sposobnostjo posnemanja človeškega vedenja – povedano drugače, gre za zmožnost računalnika, da (samostojno) izvaja naloge, ki jih praviloma pripisujemo inteligentnim bitjem, kot so kognitivna avtomatizacija, učenje, sklepanje, analiza hipotez ipd.¹ Njena semantična osnova sicer nakazuje na računalniško mimikrijo človeškega vedenja in zavesti, a vendar, zgolj iz previdnosti in da bi se izognili vtisu pretiranega zanesenja: čeprav so se sistemi UI izkazali za izjemne pri opravljanju vrste kognitivnih, logično-analitičnih funkcij, ki so še do nedavnega veljale za ekskluzivno domeno človeka, se razprave o morebitni »zavesti« računalnikov vendarle zdijo (vsaj za zdaj) še močno pretirane.

»Umetna inteligenca je hitro razvijajoča se skupina tehnologij, ki lahko prispevajo k številnim gospodarskim in družbenim koristim v celotnem spektru panog in družbenih dejavnosti. Uporaba umetne inteligence lahko z izboljšanjem napovedi, optimizacijo delovanja in dodeljevanja virov ter po meri prilagojenimi digitalnimi rešitvami, ki so na voljo za posameznike in organizacije, zagotavlja ključne konkurenčne prednosti za podjetja ter podpira družbeno in okoljsko koristne rezultate, na primer na področju zdravstvenega varstva, kmetijstva, izobraževanja in usposabljanja, upravljanja infrastrukture, energije, prometa in logistike, javnih storitev, varnosti, pravosodja, učinkovite rabe virov in energije ter ublažitve podnebnih sprememb in prilagajanja nanje« (Akt o umetni inteligenci 2021).

Že v izhodišču je mogoče v izobraževanem okolju izpostaviti tri tehtne razloge, ki govorijo v prid implementaciji UI v učne sisteme: (1) dostopnost (ta je praviloma široka, stroškovno nizka ali celo brezplačna), (2) sodelovalnost

¹ Umetno inteligenco je mogoče opisati tudi operativno področno kot, na primer, »kombinacijo kognitivne avtomatizacije, strojnega učenja, sklepanja, ustvarjanja in analize hipotez, obdelave naravnega jezika in namerne mutacije algoritmov, ki ustvarja vpogled v analitiko na ravni ali nad človeškimi zmožnostmi« (IEEE Corporate Advisory Group 2017, 16).

(pogovorni uporabniški vmesniki omogočajo generiranje pisnih, vizualnih ali kodiranih izhodov) in (3) zaznavna kakovost (obseg indeksirane vsebine podpira izboljšanje rezultatov na verodostojni ravni) – z opozorilom, da se algoritmi UI uporabljajo predvsem kot »dodatno« orodje za učitelje in učence, torej kot podpora, ki v končni fazi še vedno terja »človeški« pregled (Gartner 2023).

Vsebinsko že omenjene etične smernice na področju izobraževanja taksativno izpostavljajo številne prednosti z umetno inteligenco procesiranih podatkov, na primer:

1. uporabo prilagodljivih učnih tehnologij za prilagajanje sposobnosti vsakega učenca;
2. uporabo nadzornih plošč (angl. *dashboardov*) študentov za vodenje učenec skozi njihovo učenje;
3. zagotavljanje individualiziranih posegov za posebne potrebe;
4. točkovanje esejev z avtomatiziranimi orodji;
5. napovedovanje morebitnega osipa, upravljanje vpisa študentov in načrtovanje virov;
6. uporabo klepetalnih botov (angl. *chatbotov*) za vodenje učencev in staršev pri administrativnih opravilih.

Tabela 1: Prednosti z umetno inteligenco procesiranih podatkov

Aplikativnost UI	Vsebina	Prednosti
Personalizirani učni sistemi Prilagodljive učne platforme Inteligentni mentorski sistemi	Študentom omogočajo dostop do različnih učnih gradiv glede na njihove individualne učne potrebe in predmete.	• Zagotavljanje pouka v učilnicah z različnimi sposobnostmi • Odkrivanje vrzeli v predznanju učencev • Prilagajanje učnih orodij in materialov za podporo rasti učencev • Zagotavljanje podrobnih in pravočasnih povratnih informacij
Avtomatizirani sistemi: • ocenjevanja • načrtovanja urnika • ugotavljanja prisotnosti • vpisa	Računalniški sistem izvaja funkcijo, za katero je navadno potrebna človekova vpletenost; sistem samostojno izvaja naloge, ne da bi potreboval stalen človeški nadzor.	• Za izvajanje ponavljajočih se in zamudnih nalog • Algoritmi za točkovanje pisnih del študentov, izpitov in nalog • Povratne informacije in navodila za izboljšanje in revizijo • Razbremenitev učiteljev
Klepetalni boti /chatboti Dialoški sistemi Pogovorni agenti	Programi komunicirajo z ljudmi prek besedila ali glasovnih ukazov na način, ki posnema pogovor med ljudmi.	• Administrativna podpora • Sposobnost, da se odzovejo naravno in s pogovornim tonom

V dosedANJI praksi so se aplikacije UI že izkazale kot neposredno uporabne zlasti na področjih (1) prilagojenih učnih platform za spodbujanje učenja učencev, (2) avtomatiziranih sistemov ocenjevanja za pomoč učiteljem in (3) pomoči pri administrativnih opravilih.

Izhajajoč iz teh osnov, *Etične smernice* pojasnjujejo svoj obstoj kot nujen zaradi naslednjih treh razlogov:

1. ker UI v izobraževanju ni več oddaljena resničnost (uporaba UI-aplikacij pri učenju tujih jezikov; uporaba UI pri prilagajanju učnih modelov različnim potrebam);
2. ker ima UI izjemen potencial za krepitev izobraževanja (četudi nam za zdaj primanjkuje poglobljene analize in se pojavljajo etični pomisleki);
3. ker je UI del digitalnih veščin, ki jih bomo potrebovali (tako s strani učiteljev kot učencev, in to do konca desetletja).

A prav tako je dejstvo, da načrtovanje »od zgoraj navzdol«, torej v smeri »izobraževalni sistem–končni uporabnik (učenec, učitelj)«, le težko sledi hitrosti prodora UI. Samo letos (ko to pišem, vstopamo v marec 2023) smo bili že priča dvema dogodkoma, ki jasno kažeta na možnost, da so vse visokoleteče cilje usmerjene nadgradnje sistema izobraževanja s potenciali UI uporabniki že prehiteli.

Oba dogodka prihajata iz ZDA – kar pa, glede na njuno »digitalno naravo«, ki jasno presega geografske zamejitve, ni razlog za zmanjševanje teže in relevantnosti v našem okolju. Prvi sega v januar 2023, in sicer gre za blokado ChatGPT-ja na šolskih napravah in omrežjih (Pičman Štefančič 2023b) – najprej v največjem šolskem okrožju v ZDA, tj. v okrožju mesta New York, nato pa tudi v šolskih okrožjih Los Angelesa, Baltimora in drugih. Uradni razlog je bil »zaščititi akademsko poštenost«, medtem ko se izvaja ocena tveganja in koristi, zlasti zaradi morebitnih »negativnih vplivov na učenje študentov in pomislekov glede varnosti in točnosti vsebine« (Elsen-Rooney 2023). Kot je v pojasnilu navedla tiskovna predstavnik šolskega okrožja NY City, namreč orodje že lahko »ponudi hitre in enostavne odgovore na vprašanja, a ne gradi veščin kritičnega razmišljanja in reševanja problemov, ki so bistvenega pomena za akademski in vseživljenjski uspeh«.

Tej logiki je seveda že načelno težko oporekati, vendar je treba za celovit oris stopiti korak nazaj. Vse akcije, blokiranje in pojasnila so odraz dejstev, da (a) učenci orodja že uporabljajo in da (b) so se šole znašle v povsem nemočnem položaju. Povedano drugače: umetna inteligenca je tu – in to že danes. Naši otroci jo uporabljajo – moj 12-letni sin niti ne sprašuje več, kako uporabiti

ChatGPT za pisanje spisa. A ne zato, ker tega ne počne – pač pa zato, ker to večino že obvlada. Prepričanje, da nas pri tem varuje slovenščina, je – tudi zaradi aplikacij UI na področju jezika – povsem zmotno.²

Mnogi so izrazili strah, da bo to konec pisnih nalog, kot smo jih poznali doslej – učiteljem, ki se tradicionalno zanašajo na neodvisne pisne naloge za izgradnjo in ocenjevanje veščin razumevanja ter kritičnega mišljenja, ChatGPT odvzema najosnovnejše, temeljno sredstvo; resnično avtorstvo seminarske naloge, eseja ali spisa je samodejno postavljeno pod vprašaj.

Drugi povod za razmislek o smotrnosti drugečnega pristopa je odmevna raziskava profesorja Christiana Terwiescha na temo zmogljivosti ChatGPT v programu MBA, prav tako objavljena januarja 2023 (Terwiesch 2023). V Terwieschevih testih je namreč algoritem generativne UI na testu MBA uspešno opravil zadano nalogo: »Chat GPT ima izjemne sposobnosti reševanja izzivov, ki se v veliki meri uporabljajo pri usposabljanju in testiranju naših študentov MBA. Glede na njegovo obvladovanje ocenjevanih področij bi njegovo delo ocenil z B do B–,« je v zaključku povzel Terwiesch (2023, 17).

Kot je razvidno iz Terwiescheve analize, se je ChatGPT izkazal kot izjemno pismen in analitičen pri oblikovanju odgovorov na vprašanja. Jezik je bil jednat, struktura besedila logična in izbor besed tehten.

»Chat GPT, ki ga je ustvaril OpenAI, je pokazal izjemno sposobnost avtomatizacije nekaterih veščin visoko plačanih intelektualnih del znanja na splošno in posebej intelektualnih del na delovnih mestih, ki jih zasedajo diplomanti MBA, vključno z analitiki, menedžerji in svetovalci.«
(Terwiesch 2023, 1)

Za oris neposredno uporabnih zmogljivosti ChatGPT-ja na primer preizkusimo njegovo sposobnost pisanja zahvale dolgoletni sodelavki, ki se upokojuje po 20 letih poučevanja prvošolcev.

Vprašanje zastavimo v slovenščini – in v naslednjih nekaj trenutkih nam ChatGPT spiše odgovor, ki se, vsaj na prvi pogled, zdi tehten, slovnično korekten, vsebinsko tekoč in predvsem – neposredno uporaben. Če želimo

2 Pod sponzorstvom Ministrstva za kulturo Republike Slovenije in Evropskega sklada za regionalni razvoj v Sloveniji že nekaj let poteka projekt Razvoj slovenščine v digitalnem okolju, ki z modeli strojnega učenja razvija programske kode in oblikuje zbirke podatkov pod odprtokodno licenco. Ti algoritmi niso večji le prepričljivega strojnega prevajanja za jezikovni par angleščina-slovenščina, pač pa tudi pravilne slovnice, skladnje in postavljanja vejic, prepoznavanja imen v besedilu in povezav med njimi (ekstrakcija povezav), odkrivanja povezav v smislu prepoznavanja imen v besedilu in koreferenčnosti med njimi ter prepoznavanja pomena besed v sobesedilu (razdvoumljanje). Za šolski sistem se zdi neposredno uporabni zlasti orodji UI za »strojno povzemanje besedil« in za »iskanje odgovorov na zastavljena vprašanja«, kot izhajajo iz besedila, torej kategorija »razumevanje vsebine« (Razvoj slovenščine v digitalnem okolju 2023).

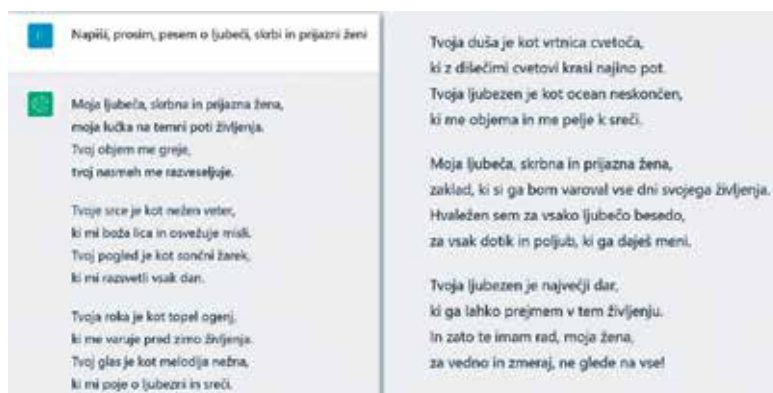
dati osebno noto, to seveda lahko storimo – ne nazadnje, z natančnejšim vprašanjem bi nam tudi ChatGPT ponudil natančnejši odgovor.



Vir: ChatGPT, 2023

Naj nam bo všeč ali ne – težko je trditi, da nam ChatGPT ni ponudil konkretnega izhodišča in spodobne osnove za nadaljnje delo. Prihranil nam je čas, postregel z idejo in nam dal na voljo tudi kakovostno oporo za nadgradnjo.

Ker predavanje poteka ravno 8. marca, ko obeležujemo mednarodni praznik žensk, za ponazoritev izberemo še dnevu primerno temo, tokrat v obliki poezije: ChatGPT-ju naročim, naj napiše »pesem o ljubeči, skrbni in lepi ženi«.



Vir: ChatGPT, 2023

ChatGPT je sam izbral štirivrstično strukturo – povsem lahko pa bi mu naročili tudi, naj uporabi katerokoli drugo pesniško obliko. Lahko bi mu določili stopice, predlagali ritem, sugerirali bi mu lahko celo slog – in v pičlih nekaj sekundah bi nam sistem ponudil svoj predlog. Bistveno je zlasti naslednje: čeprav predlagana pesem morda ni idealna prav za vašo ženo oz. partnerja, pa je zagotovo lahko koristno orodje za osnovo ali nadaljnjo nadgradnjo osebnega sporočila.

Etika kot naš edini pravi ščit

Rezultati umetne inteligence so osupljivi. A ker se UI uči podobno, kot se učimo ljudje – prek nevronske mreže, kjer so računalniški sistemi zasnovani kot zbirke enot in vozlišč po vzoru bioloških živčnih nevronov – se lahko uči tako »dobrega« kot »slabega«. Še več – če nas bo UI posnemala (in posnemala nas bo!), se utegnejo vse naše pristranskosti, subjektivnosti, diskriminatorni nazori ipd. prezrcaliti na medmrežje.

ChatGPT

This is a free research preview.



Our goal is to get external feedback in order to improve our systems and make them safer.



While we have safeguards in place, the system may occasionally generate incorrect or misleading information and produce offensive or biased content. It is not intended to give advice.

Next

Vir: ChatGPT³, 2023

Neposredna, pereča povoda za razmislek o UI v izobraževanju sta na tej točki torej (vsaj) dva: (1) UI je že neposredno uporabna in tudi uporabljana in (2) ima zaznavno kakovost. Naj se na tem mestu vrnem na našo izhodiščno misel Stephena Hawkinga: »Uspeh pri ustvarjanju umetne inteligence bi bil

3 »Naš cilj je dobiti zunanjo povratno informacijo, da bi lahko izboljšali svoje sisteme in povečali varnost. Čeprav imamo varovala, lahko sistem občasno generira nepravilne ali zavajajoče informacije ali žaljive in pristranske vsebine. Ravno tako ni namenjen svetovanju.«

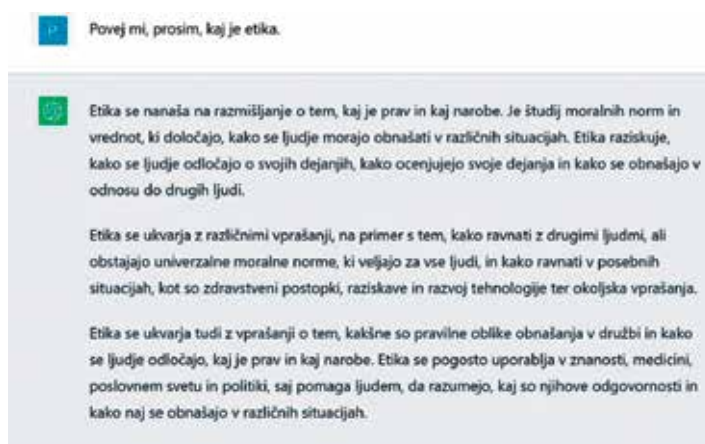
lahko največji dogodek v zgodovini človeštva. A žal lahko tudi naš zadnji, če se ne bomo naučili izogniti se tveganjem.«

Naša najboljša zaščita je – etika. Etični pristop mora postati integralni del umetne inteligence – ker ta posnema miselne procese ljudi, namreč ni ne objektivna ne vredno(s)tno nevtralna. V tem pogledu je zato ključnega, vitalnega in vseobsegajočega pomena, da v sam sistem vnesemo etične smernice, jih neprestano spremljamo, evalviramo, preizprašujemo in po potrebi korigiramo – in to v vseh fazah:

1. pri oblikovanju sistemov UI,
2. pri uporabi sistemov UI in
3. pri ciljnih UI.

»Umetna inteligenca ima dolgo zgodovino, ki obsega številne discipline, od računalništva do kognitivnih znanosti, filozofije, tehnoloških študij in vede o učenju. Umetno inteligenco ljudje dojemajo različno, zato je pomembno, da se etični izzivi ponazorijo s konkretnimi primeri, usmerjenimi v izobraževanje, da bi se izognili nesporazumom. Pri etiki umetne inteligence ne gre le za to, (1) kako uporabljamo tehnologijo, temveč tudi za to, (2) zakaj in iz kakšnih razlogov umetno inteligenco in podatke uporabljamo v podporo poučevanju, učenju in ocenjevanju« (Evropska komisija, Generalni direktorat za izobraževanje, mladino, šport in kulturo 2022a, 8).

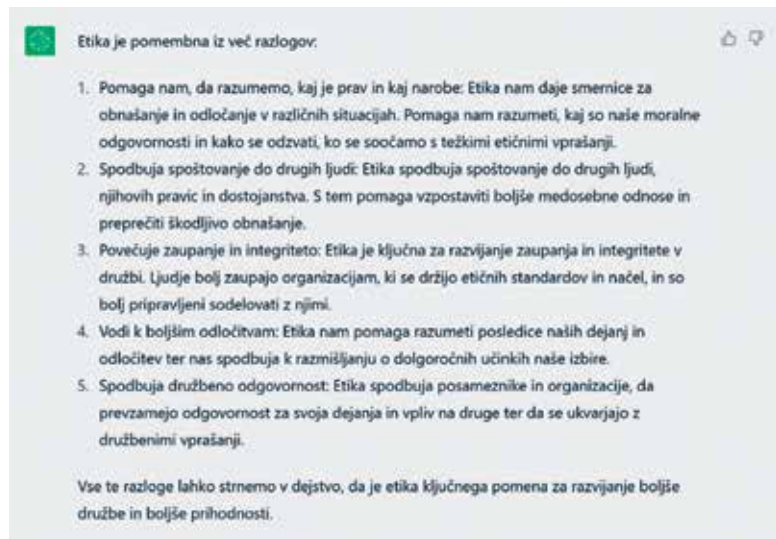
Za zgovoren oris vsestranske aplikativnosti lahko – v imenu znanstvene semantološke doslednosti – kot iztočnico za nadaljevanje ChatGPT prevprašamo še o poznavanju etike.



Vir: ChatGPT, 2023

... in v hipu dobimo odgovor. Ta ni kopiran, »posnet z interneta«, praviloma tudi ne avtorsko sporen; je unikatno delo umetne inteligence.

Ker je »pogovor« enostaven in (vsaj na prvi pogled) tudi tehten, ga nadaljujem: ChatGPT pobaram, zakaj je etika sploh pomembna – in spet dobim hiter, vsebinsko logičen, izviren odgovor.⁴



Vir: ChatGPT, 2023

Odgovor aplikacije UI se zdi povsem na mestu: etiko potrebujemo, da se lahko odločimo, kaj je prav in kaj ne – da lahko pravilno usmerjamo svojo pot in razvoj na način, ki podpira naše skupne vrednote in je koristen za posameznika in družbo. Končno poročilo strokovne skupine Komisije za umetno inteligenco in podatke v izobraževanju in usposabljanju tako na primer poudarja, da sta etika umetne inteligence in etika izobraževanja in usposabljanja tesno povezani: »Umetna inteligenca kot tehnologija prinaša lastne etične dileme in izzive, ki jih je treba razlagati v kontekstu poučevanja in učenja, da bi uresničili zamišljene možnosti« (Evropska komisija, Generalni direktorat za izobraževanje, mladino, šport in kulturo 2022a, 7).

⁴ V imenu orisa korektnosti odgovora na tem mestu ne preverjam – čeprav bi v primeru resne znanstvene razprave seveda naredila prav to – pa ne le pri digitalnih virih, verodostojnost se, kot je to v raziskovalnem svetu uveljavljena praksa, preizkuša z vsaj tremi med seboj neodvisnimi viri; velja poudariti, da v tem smislu torej pri UI ni prav nobene razlike v razmerju do standardnih pristopov.

Da bi pomagala pri reševanju teh izzivov, je Evropska komisija oktobra 2022 objavila Etične smernice za uporabo umetne inteligence in podatkov pri poučevanju in učenju za izobraževalce (Evropska komisija, Generalni direktorat za izobraževanje, mladino, šport in kulturo 2022b). Te naj bi zagotovile praktično podporo in napotke predvsem osnovnošolskim in srednješolskim učiteljem, in sicer ne glede na njihovo morebitno predznanje oz. poznavanje sistemov umetne inteligence. Natančnejši pregled smernic presega namen tega traktata, zato naj na tem mestu zadostuje zgolj splošen oris: smernice opredeljujejo primere uporabe UI v šolah, navajajo primere mogočih načinov pomoči UI pri poučevanju tako z vidika učiteljev kot učencev, načrtujejo vodila pri podpori UI v upravnih postopkih in, ne nazadnje, predstavljajo etične vidike in zahteve kot temeljna vodila implementacije UI v izobraževalna okolja.⁵

Temeljna načela etične umetne inteligence

Evropska unija je že v izhodišču izpostavila etične pomisleke glede posledic umetne inteligence; poudarila je težo varstva zasebnosti, problem pretirane nadzora in pomen nediskriminacije kot glavna področja, na katera se zdi smiselno osredotočiti pri zagotavljanju etično naravnane sistema UI.

»Za upravljanje in uporabo umetne inteligence in podatkov v izobraževanju je potrebno skupno razumevanje, (1) kaj je treba nadzorovati, (2) kako je treba organizirati in upravljati postopke upravljanja ter (3) kako so porazdeljene odgovornosti.« (Evropska komisija, Generalni direktorat za izobraževanje, mladino, šport in kulturo 2022a, 10)

V preteklem desetletju se je zdelo, da digitalna etika večinoma le počasi caplja za razvojem tehnologije. Na primer izkušnje, ki jih imamo s prihodom mobilnih telefonov in posledičnimi smermi razvoja medosebne komunikacije, imajo (upravičeno) neredko negativno konotacijo: »priklenjeni smo na telefon«, »postali smo digitalno odvisni« ipd. Tudi z digitalizacijo podatkov in vdori v zasebnost izkušnje niso dosti boljše – neskončen seznam škandalov, ki smo jim priča, zagotovo ne izpričuje premišljenosti in etične dorečenosti sistema.

5 Mimogrede, na koncu Etičnih smernic (Evropska komisija, Generalni direktorat za komunikacijska omrežja, vsebine in tehnologijo 2019, 32-36) je mogoče najti koristen glosar s pojasnili osnovnih pojmov, ki uspešno krmari med načelom, da uporaba UI res ne potrebuje poglobljenega tehničnega (pred)znanja, in zavestjo, da je osnovne pojme komponent sistema, če naj izobraževalna sfera (in tudi širše družba) bdi nad etičnostjo delovanja UI, vendarle treba osvojiti vsaj za namene laičnega sporazumevanja.

Toda s prihodom splošne uporabnosti UI se zdi, da se je princip »najprej tehnologija, potem etika« začel obračati v drugo smer. Praktično prvič v zgodovini tehnologije se resno in odgovorno srečujemo z vnaprejšnjo široko razpravo o etičnih dilemah, in to še pred vsesplošno uporabo UI. Zdi se tudi, da smo si pri ciljih, ki naj jih razprava doseže, enotni: želimo informirati, izobraziti, ozavestiti – in tako pomagati tehnologiji, da nam, človeštvu, pomaga. Pri tem seveda ne gre spregledati, da je naloga izjemna – pa tudi, da nam ne sme spodleteti.

Umetna inteligenca ni niti vredno(s)tno nevtralna niti objektivna

Kot rečeno, izhodišče je, da UI ni niti vredno(s)tno nevtralna niti objektivna. Umetna inteligenca – tako kot katerakoli druga tehnologija, ki spreminja življenje – ni sama po sebi dobra ali slaba. Ljudje, ki razvijajo tehnologijo, so odgovorni za določanje, kako se uporablja; ljudje, ki jo uporabljajo, odločajo, ali jo bodo uporabili za dobro ali slabo. Odgovornost je torej na – ljudeh. Prav zato je široka razprava o etiki umetne inteligence še toliko pomembnejša.

A zasluge za družbeno orientacijo k etiki ni iskati le v odgovornosti ljudi – tudi sama narava UI terja etično pretehtana izhodišča. Sistemi, ki podpirajo UI, temeljijo na modelih strojnega učenja, slednji pa se zanašajo na podatke, ki so jim izpostavljeni. Povedano drugače: sistem bo deloval, kot se bo naučil iz podatkov, ki jim je izpostavljen. Če bo sistem izpostavljen pristranskim, diskriminatornim, seksističnim ali kakršnimkoli drugače neželenim vzorcem, jih bo sprejel in vgradil – ne zato, ker je »slab«, pač pa zato, ker je »učljiv«. Posledično je v sami naravi UI, da lahko širi neželeno vedenje in pristranskost v velikem obsegu; logika modela in posledični rezultati niso vedno predvidljivi in vedenje se lahko med »učenjem« tudi spremeni.⁶ Že samo ta del UI predstavlja močan kontrast klasičnim sistemom, ki ga zaznamujeta gradnja in vodenje algoritmov od spodaj navzgor.

Iz dosedanjih izkušenj z UI je mogoče destilirati zlasti pet področij etičnih načel, ki se jih zdi a priori smiselno integrirati v sisteme UI:

1. osredotočenost na človeka,

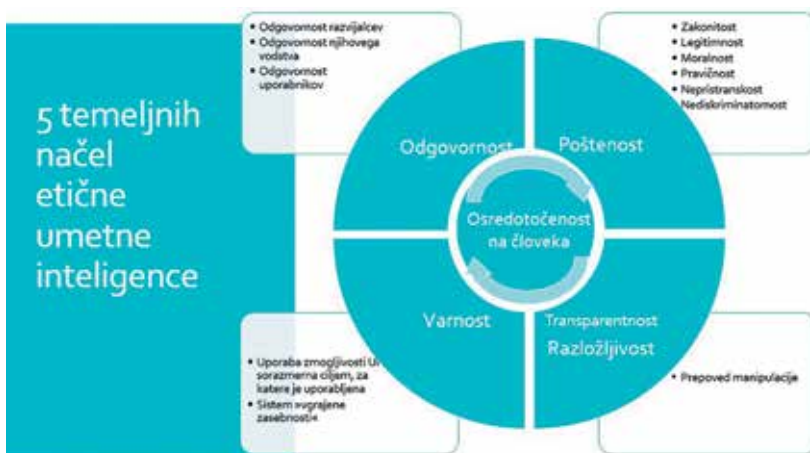
⁶ Primer neželenega stranskega učinka sicer koristne UI je poznan praktično vsem: čeprav nas mobilne aplikacije, kot sta Googlevi Zemljevidi ali Waze, praviloma pripeljejo po pravih, najkrajših in najmanj prometnih cestah najhitreje na cilj, lahko v primeru zastojev na avtocestah kaj hitro predlagajo alternativne poti, ki ustvarjajo neželene prometne zastoje na lokalnih cestah ali celo stanovanjskih območjih.

2. poštenost,
3. razložljivost in transparentnost,
4. varnost in
5. odgovornost.

Za krepitev »dobrih učnih poti« kaže pri prvem načelu osredotočenosti na človeka slediti zlasti tistim smerem razvoja sistemov UI, ki podpirajo človeške cilje ob upoštevanju človekovih potreb tako na osebni kot družbeni ravni. Razumljivo tako visokoletni cilji niso niti enostavni niti nujno premočrtni.

Dodatno kompleksnost obravnave nadalje zahtevajo tudi same smeri razvoja UI, ki so zaznavno različne. Sistemi strojnega učenja so namreč sposobni oblikovati povsem nova dela (generativna UI), imajo zmogljivosti avtomatizirane logično-kognitivne obravnave za človeka nepredstavljivih količin podatkov (razširitvena oz. augmentacijska UI), lahko pa tudi, na podlagi podanih parametrov, ustvarjajo nove, »prihodnje« realnosti (simulacijska UI).

Zagotovo bo vsaka izmed teh smeri razvoja UI terjala posebno obravnavo – a za zdaj se zdi smotrnejše iskati zlasti skupne temelje. Vsem trem smerem UI je namreč skupno predvsem dvoje: (1) vedno mora biti na koncu glavni človek in (2) rezultati UI morajo biti družbeno koristni (Gartner 2022).



Slika 2: 5 temeljnih načel etične umetne inteligence

Povsem jasno je, da so nekatere smeri razvoja UI družbeno škodljive,⁷ a mnogo več je primerov, ki so lahko hkrati pozitivni in negativni. Tako so se

⁷ Pomislite samo na UI, ki se uporablja za ustvarjanje lažnih novic ali ki bi se, zgolj kot primer, uporabljala za ustvarjanje avtonomnih morilskih dronov.

ob razvoju slikovne UI hitro oglasili umetniki, češ da jim UI krade delovna mesta – da potreb po oblikovalcih, ilustratorjih ipd. ne bo več, saj da bo vse izpodrinila UI (Pičman Štefančič 2023a).



Vir: Twitter⁸, 2022

Toda prav tako je mogoče trditi, da UI, ki nadomešča človeške napore, pomeni napredek in da je torej ta razvoj družbeno koristen. Še kompleksnejša situacija nastane, če družbeni koristnosti dodamo zahtevo, da naj bo UI osredotočena na človeka. Vzemimo samo kitajski primer sistema socialnih kreditov (angl. *Social Credit System*), ki varuje potrebe družbe in obenem eklatantno spregleduje potrebe individuuma.

Prav nič lažje niso zahteve pri zagotavljanju načela poštenosti – to naj bi v koncept UI vneslo vse razsežnosti od jasne in očitne zahteve po zakonitosti do legitimnosti, moralnosti, pravičnosti in nediskriminatornosti.

V naravi prava je, da je žlahtno konservativno – pravne norme se praviloma odzivajo na družbeno realnost, kar še zlasti velja za tehnološki razvoj. Niti v naravi niti v moči prava ni, da bi predvidelo tehnološki razvoj – nanj lahko vpliva, a zameji ga lahko le reakcijsko. Če privzamemo to (ne)moč prava, je jasno, da nas pri oblikovanju sistemov UI lahko vodi predvsem lastni moralni kompas. Da je situacija še za odtenek zapletenejša, je treba sprejeti tudi dejstvo, da se morala spreminja – to, kaj je v tehnološkem napredku prav in kaj ni, ni povsem nespremenljivo. Zgovoren primer tega sta Uber in Airbnb – obe storitvi sta bili v izhodišču kamen spotike obstoječemu pravnemu

8 »Pred našimi očmi se odvija smrt umetnosti – če kreativni poklici niso več varni pred stroji, tudi visokokvalificirana delovna mesta lahko zastarijo. Kaj bomo imeli potem?«

redu, a ker ju je družba sprejela kot koristni, so se postopoma spremenili tudi zakoni in jima omogočili razcvet.

Implementacija koncepta pravičnosti je v sistemu UI še toliko zahtevnejša – bistvena prednost UI je namreč prav v tem, da je mogoče še do nedavnega okorne sisteme hitro »naučiti« individualnih potreb. Prav ta pristop v učilnicah obljublja znatne možnosti prilagajanja sistema individualnim potrebam – dejstvo je, da nismo vsi povsem enaki in da se ne učimo vsi povsem enako. Tudi naše zmožnosti niso povsem enake. A ta prednost UI ima tudi negativno plat. Ker se sistem strojnega učenja »uči« podobno, kot se učimo ljudje, se zdi smiselno že vnaprej zastaviti sistem, ki ne diskriminira tam, kjer diskriminacija ni dobrodošla (na podlagi spola, rase, vere ipd.), a obenem pozitivno diskriminira, kadar to posamezniku in družbi prinaša koristi (npr. posebne učne potrebe ipd.)

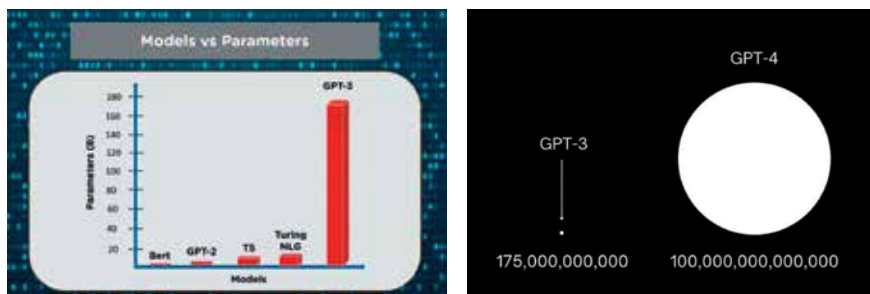
Razmejevanje med pozitivno in negativno diskriminacijo seveda ni novost – se pa v sistemu UI prvič naglašujejo te dileme že pri samem snovanju tehnologije. Če naj bo UI zares »dobra«, je odprava nezaželenih pristranskosti preprosto nujna.

Prav tako je bistven element načela poštenosti tudi prepoved manipulacije. Iz zgodovine informacijske tehnologije poznamo številne primere manipulativne uporabe podatkov – večji ko je ekosistem, večja je nevarnost, da bodo naši podatki uporabljeni tudi brez naše privolitve. Posledica takšnih manipulacij je vse ostrejša zakonodaja s področja zasebnosti.⁹

Vendar ni vsaka manipulacija tudi škodljiva. Dober primer pozitivne manipulacije so aplikacije za zdravje: seveda sistem z nami »manipulira«, da se prehranjujemo bolj zdravo, da več telovadimo in da opuščamo nezdrave navade – a to počnemo prostovoljno in s povsem jasnim namenom. Kar je pravzaprav bistveno: pri snovanju UI se moramo zavedati, da je celo manipulacija lahko dopustna (in zaželena), če je transparentna, če je jasen in nedvoumen njen namen in če seveda ni zlonamerna. Kar vse le še bolj – zaplete.

Dosedanji razvoj nam dovoljuje sklep, da obstoječe znanje in rešitve lahko ponudijo nekoliko trdnejšo oporo vsaj pri reševanju dilem glede zagotavljanja načel transparentnosti, varnosti in odgovornosti. Sistemi UI morajo biti pregledni; sposobni morajo biti pojasniti odločitve, ki jih sprejemajo.

9 Tako je npr. analitika, ki je sledila interakcijam med oglaševalci in uporabniki, že terjala odločen odziv držav.



Slika 3: Razvoj modela GPT, GPT-3 z več kot 174 milijardami parametrov strojnega učenja in primerjava z GPT-4

Prvi del načela transparentnosti se nanaša na podatke, na katerih se sistem uči – preglednejši ko so, lažje je razumeti, spremljati in usmerjati poti razvoja. Načelo transparentnosti v tem smislu vsebuje tudi zahtevo po »razložljivosti«. Tradicionalni sistemi so grajeni na sistemu vnaprej postavljenih pravil, medtem ko se sistemi UI učijo prek nevronske mreže, kjer rezultat ne sledi pravilu, pač pa verjetnosti – podobno, kot razmišljamo ljudje. Zahteva po »razložljivosti« je v tem pogledu razumljivo trši oreh – pogosto je znano le to, da (a) obstaja določena verjetnost, da je rezultat pravilen, in da (b) ni povsem gotovo, kako je sistem prišel do odgovora. Tehnične rešitve vsaj za zdaj na tem področju še zaostajajo za etičnimi zahtevami.

Načelo varnosti zajema tako varnost podatkov, na katerih se sistem UI uči, kot tudi samo delovanje sistema UI. Vključuje zahtevo po varstvu zasebnosti in naglašuje pomen vzpostavitve (pa tudi budnega spremljanja in rekaliibriranja) občutljivega razmerja med tem, da tehnološke zmožnosti zadovoljivo opravijo zadane naloge, in dejstvom, da tehnologija sama dopušča (in potrebuje) še več podatkov za svoj prihodnji razvoj.

Sposobnost tehnologije, da, na primer, prepozna obraze, zagotovo presega našo potrebo po tem, da preštejemo ljudi v prostoru. Načelo varnosti v tem primeru zahteva, naj bo uporaba zmogljivosti UI sorazmerna s svojimi cilji. Prav tako velja poudariti, da anonimizacija podatkov najbrž ne bo več dovolj, saj so algoritmi UI že zdaj sposobni odkriti identifikacijske elemente in povezovati podatke, ki so bili še do nedavnega nepovezljivi. Razvpit je na primer Lapine, umetnice, ki je na spletu več kot deset let po opravljenem posegu našla svoje klinične fotografije, tokrat (s pomočjo sistemov UI za prepoznavanje obrazov) povezljive z imenom in priimkom (Pičman Štefančič 2023c).

LAION scraped medical photos for AI research use. Who's responsible for taking them down?

BENJ EDWARDS - 10/11/2022, 9:43 PM



Slika 4: LAION zamegljene slike za raziskovalno-medicinsko uporabo. Kdo je odgovoren za objavo? (Vir: Ars Technica, 2022)

Tudi zato je potrebna posebna previdnost pri zasebnosti v sistemih UI. Izhajati velja iz načela »vgrajene zasebnosti«, ki temelji na predpostavki, da so podatki *per se* občutljivi in da jih je s tega vidika smiselno *a priori* šifrirati ali, kar je prav tako ena izmed novih prednosti sistema UI, uporabiti sintetične podatke.

In ne nazadnje, vzpostaviti bo treba učinkovit sistem odgovornosti. Pretekle izkušnje nas učijo, da je lahko odgovoren le človek – podobno, kot je lahko avtor ustvarjalnega dela (in torej nosilec avtorske pravice) lahko samo človek. Prve dileme so že bile naslovljene – če sistem UI ustvari novo umetniško delo, kdo je potem avtor (Pičman Štefančič 2023)? Podobno, kot je v veljavnih pravnih ureditvah avtorstvo umetniškega dela mogoče pripisati le človeku, je tudi v sistemu UI odgovornost za delovanje in rezultate mogoče pripisati le ljudem – ideja, da bi lahko imela tehnologija »osebnost« ali »zavest« in bila torej sposobna »robotske odgovornosti«, se (vsaj za zdaj) v celoti upira pravnim razmerjem. V tem trenutku je torej odgovornost za etično delovanje v prvi vrsti na razvijalcih, njihovem vodstvu in uporabnikih.

Etični izzivi umetne inteligence v izobraževalnem okolju

»Da bi ustvarili prihodnjo generacijo, v kateri lahko vključujoči in raznoliki državljani sodelujejo pri razvoju prihodnosti umetne inteligence, moramo razviti priložnosti za učence in učitelje [...], da se učijo o

umetni inteligenci prek učnih načrtov, ki temeljijo na umetni inteligenci in etiki ter poklicnem razvoju.» (Akgun in Greenhow 2021, 432)

Dejstvo je, da algoritmi vedno sledijo vrednotam svojih oblikovalcev – te se zrcalijo skozi nabor podatkov, na katerih se algoritem uči, njihove delovne označbe, pozive, ki jih podajajo uporabniki, in njihove povratne informacije. Pri tem niti oblikovalci niti uporabniki seveda niso vredno(s)tno nevtralni in dosledno objektivni, kar neizogibno vodi v pristranskost algoritma.

V izobraževalnem sistemu je mogoče še posebej izpostaviti zlasti pet izzivov etičnih pričakovanj: nanašajo se na (1) potencialne vdore v zasebnost, (2) čezmeren nadzor, (3) ohranjanje akademske integritete, (4) nadaljnje utrjevanje že obstoječih pristranskosti glede na spol in družbeno diskriminacijo ter (5) omejevanje avtonomije učencev pri odločanju o osebnih odločitvah in vodenju lastnih življenj.



Slika 5: 5 etičnih izzivov UI v izobraževalnem okolju

Problematika varovanja zasebnosti in nevarnost pretiranega nadzora v šolskih klopih sta pravzaprav le še ena plat že omenjenega načela varnosti. Posebno pozornost v sistemu izobraževanja razumljivo terjata zlasti ranljivost skupine – a problematika ni niti nova niti lastna zgolj umetnointeligenčnim sistemom, zato naj na tem mestu zadostuje zgolj opozorilo, da so zagotavljanje varnosti, spoštovanje zasebnosti in onemogočanje pretiranega nadzora temeljni pogoj za delovanje celotne digitalne družbe in tudi digitalnega izobraževanja.

Pri premisleku o etičnih dilemah implementacije UI v izobraževalne sisteme se zdita zgovornejši zlasti dve drugi preizkušnji: na eni strani je to (nehotena) kulminacija pristranskosti v izobraževanju, na drugi pa novoodprti problem ohranjanja akademske integritete. Nobena izmed problematik ni nova – jima pa algoritmi UI dajejo še do nedavnega nepredstavljive razsežnosti.

Upoštevajoč učinkovitost ChatGPT in načine, na katere je uporabljen v tem zgodnjem obdobju svojega obstoja, ni dvoma, da gre za povsem novo razsežnost »nedovoljenega načina pomoči pri pisanju«. Če smo še nedavno povzetke domačih branj prepisovali iz generacije v generacijo – in poskušali, glede na razpoložljiv čas, voljo, sposobnost in prizadevnost, prilagoditi vedno znova isto vsebino trenutnemu navdihu in slogu piščega – to namesto nas zdaj veličastno opravi ChatGPT. Še več – ChatGPT se z algoritmom hitro uči novih vsebin, novih povezav in novih oblik, zato je njegov izdelek vsakič povsem unikaten – in torej nesledljiv s standardno definicijo »prepisovanja«, goljufanja ali plagiatorstva, ki bi ga učitelj sicer nemara utegnil zaznati.

Program ChatGPT, ki ga je ustvaril OpenAI, uporablja strojno učenje za pripravo pisne vsebine po meri. Odziva se na konkretno zahtevo, ki mu je naložena; spekter tem, ki jih obvlada, je širok (in ker gre za algoritem strojnega učenja, se le še širi). Zbirati in obdelovati je sposoben zgodovinska dejstva, oblikovati eseje v različnih slogih, pisati slovnično pravilno (ali celo z napakami, če uporabnik poda takšno zahtevo) in izpeljevati prepričljive, logično utemeljene zaključke. In kar je zares osupljivo – zapis je v celoti realističen.

Kar je bilo še pred letom dni nepredstavljivo, je postalo resničnost: učitelj, pa naj bo še tako izkušen, preprosto nima možnosti prepoznati, ali je avtor oddane naloge resnično učenec ali nemara robot.

Toda – je težava res v tehnologiji? Nekateri so hitro našli vzporednice z uporom proti »vdoru tehnologije« v preteklosti. Pred dvema desetletjema so se šolniki upirali Googlu, češ da bo študentom odvzel vse veščine iskanja informacij – prek spleta je bilo mogoče odgovore, katerih zbiranje je prej terjalo znaten vložek energije, pridobiti v trenutku. Še pred tem je podoben preplah povzročil – kalkulator.

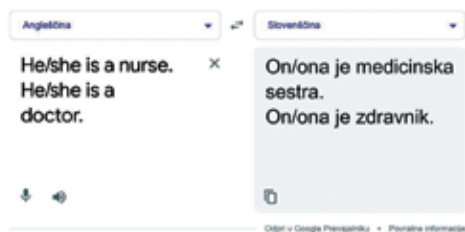
V svetovnem merilu so začasni ukrepi različni – nekateri se vračajo k rokopisnemu delu, spet drugi k ustnemu ocenjevanju, množijo pa se predvsem

pozivi, da bo treba prevetriti obstoječe etične kodekse, da bodo zagotovili akademsko integriteto tudi v dobi UI.

Dolgoročnejše rešitve bodo najbrž ponudili klasifikatorji besedil, usposobljeni za prepoznavanje teksta, ki ga ustvari UI. Tudi sami klasifikatorji so produkt umetnointeligenčnih sistemov in se hitro učijo – zadnja različica, ki prihaja s Stanforda, zagotavlja že 94% zanesljivost (Miller 2023).

A če so se ponudniki na dilemo akademske integritete že odzvali in se zdi, da bodo tudi učinkovita orodja kmalu na voljo, se drugi del dileme, tj. vprašanje krepitve pristranskosti s sistemi UI, ki nam nehotе nastavlja ogledalo, zdi izmuzljivejši. V analizi izobraževalnega okolja z vidika potencialov UI Selin Akgun in Christine Greenhow med največjimi nevarnostmi omenjata zlasti tveganja, da bodo algoritmi UI zaradi svoje narave (a) nadaljevali obstoječe sistemske pristranskosti in diskriminacije, (b) ohranili nepravilnosti za študente iz prikrajšanih in marginaliziranih skupin ter (c) še naprej krepili rasizem, seksizem, ksenofobijo pa tudi druge oblike nepravilnosti in neenakosti (Akgun in Greenhow 2021).

Dejstvo je namreč, da na tej stopnji razvoja modeli UI še vedno predstavljajo sindrom »črne škatle« – celo sami razvijalci ne vedo, kako je UI prišla do konkretnega rezultata – jasno je le to, da se UI uči na podlagi naših vhodnih podatkov in naših odzivov. Obstoječe pristranskosti se zato vgrajujejo v učni korpus, in če nismo previdni, se lahko še dodatno krepijo. Zgovoren in preprost primer so na UI temelječi prevajalniki. Seveda vemo, da so ženske lahko tudi zdravnice in moški medicinski tehniki – a prevod »He/she is a nurse« oz. »He/she is a doctor« ustreza razmerjem, ki bi jih uporabili ljudje v pogovoru: »medicinska sestra« v ženski obliki in »zdravnik« v moški samostalniški obliki.



Vir: Google Translate, 2023

Tudi (in prav) zato je smiselno še posebej naglasiti zlasti pomen zavedanja, da utegnejo sistemi UI brez skrbnega posluha le še poglobiti zdajšnje neenakosti, pristranskosti in diskriminacijo.

Prihodnje usmeritve etične umetne inteligence

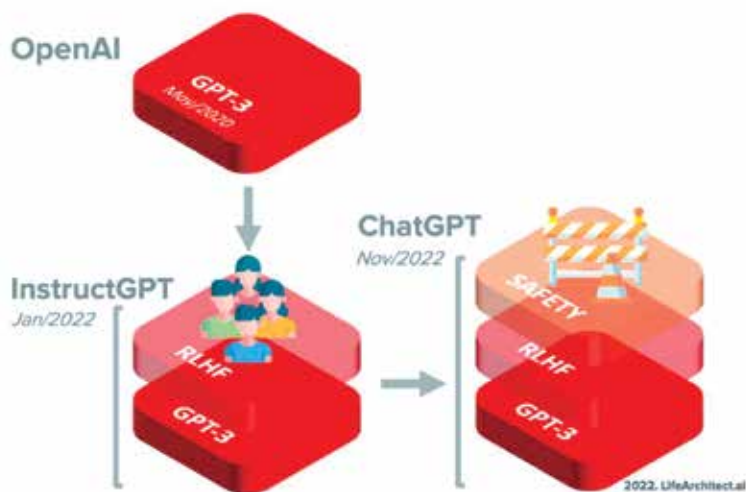
Tako kot je fascinantno spremljati potencial izjemnega razvoja UI, so osupljivi tudi hitri odzivi na identificirane probleme. Rešitve, na katere smo še pred desetletjem čakali po načelu »počakajte na naslednjo verzijo«, se danes pojavijo skorajda takoj. Sistem UI in agilen pristop k načrtovanju prinašata nove, neslutene možnosti razvoja – in tudi napredka.

Vse glasnejšim zagatam glede vprašanj etičnosti uporabe ChatGPT v šolah, o katerih je bilo govora januarja 2023, so še konec istega meseca sledile prve rešitve. V želji, da bi zaščitil akademsko integriteto, se je OpenAI hitro odzval na dileme šolnikov in nemudoma ponudil prvo verzijo že omenjenega klasifikacijskega orodja. Kot so zapisali, je (za zdaj) resda še nemogoče zanesljivo odkriti vsako besedilo, napisano z umetno UI, a »verjamemo, da lahko dobri klasifikatorji ublažijo lažne trditve, da je besedilo, ustvarjeno z umetno inteligenco, napisal človek« (Kirchner idr. 2023).

Tudi podjetje Turnitin, sicer najbolj znano po svoji programski opremi za zaznavanje plagiatorstva, poroča o gradnji orodij za zaznavanje besedil, ustvarjenih z UI. Komercialna programska oprema, ki jo številne univerze uporabljajo za prepoznavanje podobnosti besedil s primerjanjem z obširno bazo že objavljenih člankov, knjig in spletnih strani, se zdi odlično izhodišče za razvoj klasifikatorja, grajenega na modelu strojnega učenja. Kot poroča The Register, tudi Turnitin gradi na modelu GPT-3. Izhaja iz domneve, da so pri pisanju umetne inteligence razvidni vzorci, ki so jih računalniki sposobni zaznati. Kot je za The Register povedal Eric Wang, Turnitinov podpredsednik oddelka za UI, ljudje »beremo tako, da skačemo naprej in nazaj z očmi, ne da bi se tega sploh zavedali, ali švigamo naprej in nazaj med besedami, med odstavki in včasih med stranmi. Listamo naprej in nazaj. Tudi pišemo s prihodnjim stanjem duha.« ChatGPT, nasprotno, gradi le na podlagi prejšnjih stavkov in prejšnjega besedila (Quach 2023). ChatGPT je metodološko in tehnično pač – doslednejši.

Podobno hitre odgovore na etične dileme ChatGPT-ja obljublja njegov klepetalni brat Claude, zasnovan na načelu »ustavne umetne inteligence«. Tudi

ta temelji na modelu strojnega učenja, a poskuša že v izhodišču v urjenje modela vnesti »ustavna načela«. Gre za idejo, da se v model ob že poznanem sistemu »krepitvenega učenja iz človeških povratnih informacij« (angl. *reinforcement learning from human feedback, RLHF*), v katerem se model usposablja prek učenja s povratnimi informacijami o kakovosti, kot jih poda človek,¹⁰ vnese še sito »ustavne presoje«. Model namesto začetne metode RLHF presoja rezultate po ustreznosti vnaprej nastavljenega niza temeljnih ustavnih načel. Kot zatrjujejo njegovi ustvarjalci, je razvoj nabora načel tudi edini človeški nadzor v procesu krepitvenega učenja (Goodside in Papay 2023).



Vir: Life Architect, 2023

Ideja ustavne umetne inteligence je zasnovana kot »varnostna raziskovalna tehnika umetne inteligence«, katere cilje je biti koristen, neškodljiv in pošten z uporabo kombinacije modelnega samonadzora in drugih varnostnih metod. Te naj bi zagotovile, da so sistemi umetne inteligence vodeni in usklajeni s človeškimi vrednotami, tako da so koristni in pomagajo človeštvu, da so varnejši, robustnejši, preglednejši in zato tudi – odgovornejši.

Claude je v času pisanja tega prispevka (februar/marec 2023) še vedno v testni fazi – a če bodo načela, ki se jim zavezuje, zaživela, lahko pričakujemo znaten napredek v etiki klepetalnih botov. In še več: če se bodo ustavna na-

10 Ljudje razvrščajo rezultate, ustvarjene iz istega poziva, in model se nauči teh preferenc, tako da jih je mogoče uporabiti za druge generacije v večjem obsegu.

čela RLHF izkazala za dobrodošla, to lahko nakaže pot za nadaljnji razvoj sistemov UI.

»V večini razprav o etiki uporabe umetne inteligence in podatkov je poudarek na tveganjih in morebitni škodi. Čeprav se je treba teh tveganj zavedati in jih obravnavati, je treba priznati tudi, da se umetna inteligenca in podatki uporabljajo za preoblikovanje poučevanja, učenja in ocenjevanja na številne pozitivne načine.« (Evropska komisija, Generalni direktorat za izobraževanje, mladino, šport in kulturo 2022a, 8)

Zaključek

Ker se umetna inteligenca ne razvija v vakuumu, pač pa jo oblikuje medigra nenehno spreminjajočih se kulturnih, družbenih, institucionalnih in političnih sil, je povsem naravno, da je dojemljiva tudi za njihovo patologijo oziroma, širše, za manj zelene vidike razmerij, ki veljajo ali se pojavljajo v družbi. Ker so slednji predvsem odraz družbe (in torej ne vnaprej načrtanega programskega algoritma), je edini mogoči način reševanja nastalih zagat budno spremljanje in korigiranje odklonov – ne povsem drugače, kot bi, z namenom orisa koncepta, učili in usmerjali otroke.

A da bomo temu kos, se moramo naučiti in ozavestiti, na katera področja moramo usmeriti posebno pozornost, da nam umetna inteligenca ne uide izpod nadzora; naučiti se moramo prepoznavati etične izzive in razumeti posledice uporabe algoritmov.

Literatura

- Akgun, S. in C. Greenhow. 2021. »Artificial intelligence in education: Addressing ethical challenges in K-12 settings«. Nature Public Health Emergency Collection. <https://doi.org/10.1007/s43681-021-00096-7>
- Akt o umetni inteligenci. 2021. »Predlog Uredbe Evropskega parlamenta in Sveta o določitvi harmoniziranih pravil o umetni inteligenci in spremembi nekaterih zakonodajnih aktov Unije«. COM/2021/206 final. <https://eur-lex.europa.eu/legal-content/SL/TXT/HTML/?uri=CELEX:52021P-C0206&from=SL>
- Elsen-Rooney, M. 2023. »NYC education department blocks ChatGPT on school devices, networks«. Chalkbeat, New York, 4.1.2023. <https://ny.chalkbeat.org/2023/1/3/23537987/nyc-schools-ban-chatgpt-writing-artificial-intelligence>

- Evropska komisija. 2020. »Akcijski načrt za digitalno izobraževanje 2021-2027.« Sporočilo Komisije Evropskemu parlamentu, Svetu, Evropskemu ekonomsko-socialnemu odboru in Odboru regij COM(2020) 624 final. <https://eur-lex.europa.eu/legal-content/SL/TXT/HTML/?uri=CELEX:52020D-C0624&from=EN>
- Evropska komisija, Generalni direktorat za izobraževanje, mladino, šport in kulturo. 2022a. »Končno poročilo strokovne skupine Komisije za umetno inteligenco in podatke v izobraževanju in usposabljanju: Povzetek«. Urad za publikacije EU. <https://data.europa.eu/doi/10.2766/65087>
- Evropska komisija, Generalni direktorat za izobraževanje, mladino, šport in kulturo. 2022b. »Etične smernice za uporabo umetne inteligence in podatkov pri poučevanju in učenju za izobraževalce«. Urad za publikacije EU. <https://op.europa.eu/sl/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1>
- Evropska komisija, Generalni direktorat za komunikacijska omrežja, vsebine in tehnologijo. 2020. »Bela knjiga Komisije o umetni inteligenci.« <https://op.europa.eu/sl/publication-detail/-/publication/ac957f13-53c6-11ea-aece-01aa75ed71a1>
- Evropska komisija, Generalni direktorat za komunikacijska omrežja, vsebine in tehnologijo. 2019. »Etične smernice za zaupanja vredno umetno inteligenco«. Urad za publikacije. <https://data.europa.eu/doi/10.2759/65329>
- Galeon, D. 2016. »Hawking: Creating AI Could Be the Biggest Event in the History of Our Civilization.« *Futurism*, 20.10.2016. <https://futurism.com/hawking-creating-ai-could-be-the-biggest-event-in-the-history-of-our-civilization>
- Gartner. 2022. »AI Ethics: Use 5 Common Principles as Your Starting point«. ID G00774103
- Gartner. 2023. »How Should Education Institutions Respond to Use of Generative AI such as ChatGPT?«. ID G00784961
- Goodside, R. in S. Papay. 2023. »Meet Claude: Anthropic's Rival to ChatGPT«. *Scale*, 17.1.2023. <https://scale.com/blog/chatgpt-vs-claude>
- IEEE Corporate Advisory Group. 2017. »IEEE guide for terms and concepts in intelligent process automation.« The Institute of Electrical and Electronics Engineers Standards Association. <https://ieeexplore.ieee.org/iel7/8070669/8070670/08070671.pdf>
- Kirchner, J. H., L. Ahmad., S. Aaronson in J. Leike. 2023. »New AI classifier for indication AI-written text, OpenAI«. <https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/>
- Miller, F. A., J. H. Katz in R. Gans. 2018. »AI x I = AI2: The OD imperative to add inclusion to the algorithms of artificial intelligence«. <https://www.>

researchgate.net/publication/323830092_AI_x_I_AI2_The_OD_imperative_to_add_inclusion_to_the_algorithms_of_artificial_intelligence

- Miller, K. 2023. »Human Writer or AI? Scholars Build a Detection Tool«. Stanford University, Human-Centered Artificial Intelligence, 13.2.2023. <https://hai.stanford.edu/news/human-writer-or-ai-scholars-build-detection-tool>
- Pičman Štefančič, P. 2023. »Kdo je lastnik avtorskih pravic za umetnost, ustvarjeno z umetno inteligenco«. Umetna inteligenca, 22.1.2023. <https://www.umetnainteligenca.com/2023/01/kdo-je-lastnik-avtorskih-pravic-za.html>
- Pičman Štefančič, P. 2023a. »Na natečaju likovne umetnosti zmagala umetna inteligenca«. Umetna inteligenca, 22.1.2023. https://www.umetnainteligenca.com/2023/01/na-natečaju-likovne-umetnosti-zmagala_22.html
- Pičman Štefančič, P. 2023b. »Plonkanje z umetno inteligenco«. Umetna inteligenca, 3.2.2023. <https://www.umetnainteligenca.com/2023/02/plonkanje-z-umetno-inteligenca.html>
- Pičman Štefančič, P. 2023c. »Se umetna inteligenca uči na vas?«. Umetna inteligenca, 12.2.2023. <https://www.umetnainteligenca.com/2023/02/se-umetna-inteligenca-uci-na-vas.html>
- Quach, K. 2023. »Universities offered software to sniff out ChatGPT-written essays«. The Register, 13.1.2023. https://www.theregister.com/2023/01/23/turnitin_chatgpt_detector/
- Razvoj slovenščine v digitalnem okolju. 2023. <https://rsdo.slovenscina.eu/>
- Terwiesch, C. 2023. »Would Chat GPT Get a Wharton MBA? New White Paper By Christian Terwiesch«. White Papers. Mack Institute for Innovation Management. <https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-new-white-paper-by-christian-terwiesch/>
- Strokovna skupina na visoki ravni za umetno inteligenco. 2020. »Assessment List For Trustworthy Artificial Intelligence«. <https://digital-strategy.ec.europa.eu/sl/node/806>
- UNESCO. 2019. »Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development«. ED-2019/WS/8. <https://unesdoc.unesco.org/ark:/48223/pf0000366994>
- UNICEF. 2021. »Policy guidance on AI for children«. <https://www.unicef.org/globalinsight/reports/policy-guidance-ai-children>