V znamenju trikotnika: izobraževanje, etika in umetna inteligenca

Ernest Ženko
Univerza na Primorskem
ernest.zenko@upr.si

V besedilu je obravnavan preplet umetne inteligence (UI), izobraževanja in etike, pri čemer je poudarek na prepričanju, da UI ni zgolj orodje, temveč aktivni dejavnik, ki lahko oblikuje vrednote, norme in pedagoške prakse. V ospredje je postavljena trikotniška struktura, v kateri tehnologija, uporaba in regulacija delujejo kot ključni elementi za razumevanje vpliva UI. Avtor izpostavlja nujnost celostnega pristopa, ki združuje tehnološke inovacije, etična načela ter pedagoške potrebe. Razprava obravnava tudi dvosmerno dinamiko vpliva med UI in človekom, ki se nanaša na vprašanje, kako UI odraža človeške vrednote in hkrati prispeva k njihovemu preoblikovanju. Posebna pozornost je namenjena vprašanju uravnavanja UI (angl. *AI alignment*), ki zahteva interdisciplinarno sodelovanje in razumevanje kulturnih kontekstov. Pri tem so izpostavljeni primeri neustrezne usklajenosti vrednot, ki lahko vodijo v etične konflikte ali neustrezne rešitve. Besedilo se sklene s pozivom k odgovorni uporabi UI, ki presega zgolj tehnične rešitve ter vključuje kritični dialog med razvijalci, izobraževalci in družbo kot celoto.

Ključne besede: etika, izobraževanje, umetna inteligenca, generativna umetna inteligenca, uravnavanje umetne inteligence

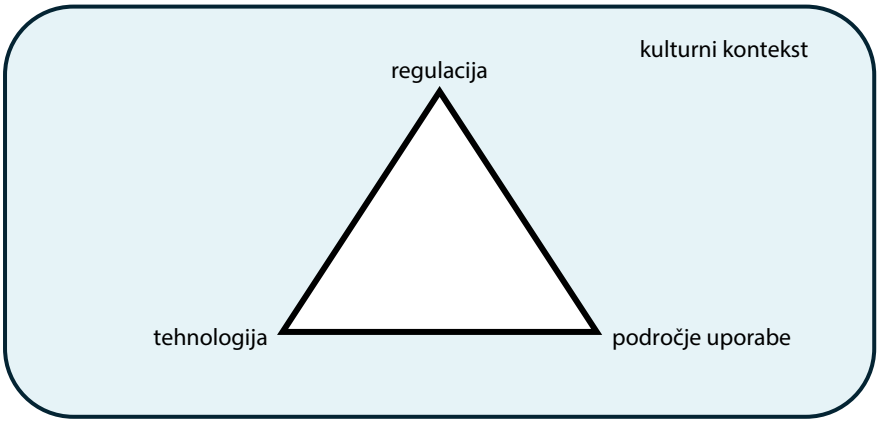
https://doi.org/10.26493/978-961-293-431-6.16

Uvod

Umetna inteligenca (UI) spreminja izobraževanje na načine, ki jih šele začenjamo razumeti. Ta proces v učenje in poučevanje nedvomno prinaša novosti in priložnosti, obenem pa tudi številne izzive, ki jih ni mogoče razumeti in reševati zgolj na ravni tehnologije; ali drugače povedano, gre za izzive, ki jih ni mogoče rešiti preprosto tako, da spremenimo nekaj vrstic programske kode (Vallor, 2024). Sistemi UI temeljijo na računalniški tehnologiji, vendar pa poznavanje te tehnologije še ne zadostuje za razumevanje učinkov njihove uporabe v družbenih kontekstih, kot je izobraževanje.

Podobno velja za katero koli tehnologijo. Če želimo razumeti, kako deluje na nekem področju, moramo upoštevati tako specifiko tega področja kot tudi odnos do tehnologije, ki tam velja. Glede na to, da vsi odnosi, ki jih je mogoče vzpostaviti do tehnologije, niso nujno zaželeni, ker predstavljajo tveganja oz.

Flogie, A., in Čotar Konrad, S. (ur.). 2025. *Izobraževanje v dobi generativne umetne inteligence: mednarodne smernice in raziskave*. Založba Univerze na Primorskem.



Slika 1 Osnovna trikotniška struktura

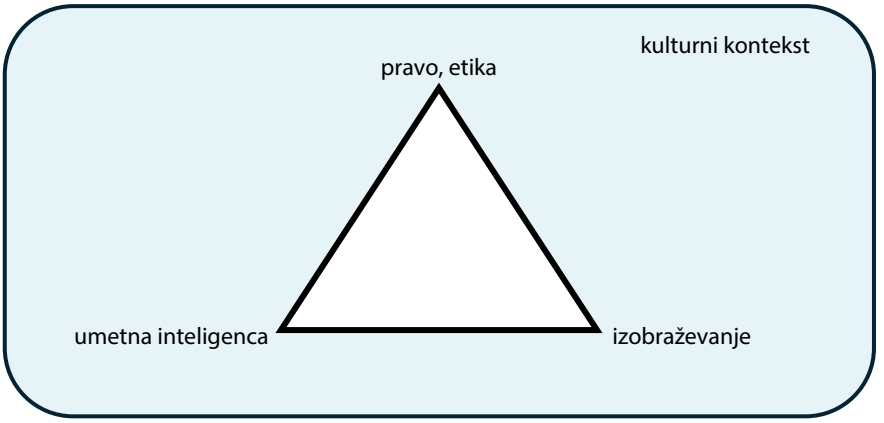
lahko vodijo v različne negativne posledice, moramo v razumevanje delovanja tehnologije dodati tudi element regulacije. Lahko si zamislimo trikotniško strukturo, ki predstavlja delovanje tehnologije na nekem področju uporabe. Oglišča tvorijo tehnologija, področje uporabe in regulacija; če to postavimo še v kulturni kontekst, dobimo strukturo, ki jo prikazuje slika 1.

Tehnologija namreč nikoli ne deluje v praznem prostoru, pač pa je vedno že postavljena v kulturni kontekst, ki zajema vpliv vrednot, norm, zgodovinskega ozadja in družbenih praks pri uvajanju tehnologije na nekem področju uporabe.

Trikotniške strukture si ne smemo predstavljati kot statične, saj je v trikotniku eno izmed oglišč tehnologija, ki nastopa kot dejavnik gibanja in spreminjanja. Pojav ali aplikacija neke nove tehnologije vpliva na različna področja uporabe; pri tem se pričnejo zastavljati vprašanja o njenih vplivih in tudi tveganjih, kar vodi v potrebo po regulaciji, s čimer se vrnemo nazaj k tehnologiji. Skozi ta dinamičen proces pridemo tudi do razumevanja razkoraka med človeškimi vrednotami in vrednotami, ki jih lahko pripišemo tehnologiji, v kolikor predpostavljamo, da tehnologija dejansko ni nevtralna (Heidegger, 2003; Hare, 2022; Vallor, 2024), temveč vedno že vključuje določene vrednotne usmeritve.[1] Varna tehnologija je načeloma tista, ki glede vrednot ne odstopa od človeških vrednot; torej tista, ki je v tem smislu ustrezno uravnana.

Tehnologija, ki jo obravnavamo, je UI, natančneje sistemi UI, kamor sodijo tudi sistemi generativne UI (GUI). Področje uporabe se osredotoča na izobraževanje, torej na učenje in poučevanje, medtem ko se regulacija navezuje na

[1] Težko si predstavljamo, da je npr. jedrsko orožje nevtralno in bi ga lahko uporabili za dobrobit in napredek človeštva – torej je njegova uporaba vselej že vrednotno usmerjena.



Slika 2 Trikotniška struktura, prilagojena specifični tehnologiji in področju uporabe

pravne vidike in etična vprašanja. Rezultat je specifičnejša trikotniška struktura, prikazana na sliki 2.

V shemi je razvoj sistemov UI tisti, ki poganja gibanje: UI oz. GUI z vstopom v procese izobraževanja nakaže nove možnosti in priložnosti, a tudi nevarnosti in tveganja, zaradi česar je nujna regulacija sistemov v smislu razvoja pravnih in etičnih okvirov. Njihov cilj je zmanjševanje razlik med vrednotami sistemov UI in človeškimi vrednotami, kar srečamo v kontekstu zagotavljanja varnosti sistemov UI (angl. *AI safety*) pod izrazom uravnavanje[2] UI (angl. *AI alignment*). Pri tem običajno srečamo naslednjo opredelitev: »Cilj uravnavanja umetne inteligence je, da se sistemi umetne inteligence obnašajo v skladu s človekovimi nameni in vrednotami« (Ji idr., 2024).

Tehnologija

Številne razprave, povezane z uravnavanjem UI, kot tudi s tveganji, ki jih prinaša neskladnost med človeškimi vrednotami in vrednotami UI (angl. *AI misalignment*), se v zadnjem desetletju osredotočajo predvsem na umetno splošno inteligenco (angl. *artificial general intelligence*, *AGI*) in superinteligenco (Russell, 2020; Tegmark, 2017; Bostrom, 2014).

V tem smislu je pomembno razlikovati med splošno (močno) in ozko (šibko) UI. Splošna UI označuje sisteme, ki bi bili sposobni reševati različne kompleksne naloge na način, ki je primerljiv s človeško inteligenco. Ti sistemi bi se lahko učili in prilagajali novim situacijam brez specifičnega usposabljanja za

[2] Inštitut za slovenski jezik Frana Ramovša ZRC SAZU je leta 2023 presodil, da je za angleški termin *artificial intelligence alignment* najustreznejši slovenski izraz »uravnavanje umetne inteligence«; ta naj bi bil ustreznejši od izrazov »poravnavanje« in »usklajevanje« (Atelšek idr., 2023).

posamezno nalogo. Po drugi strani pa ozka UI zajema sisteme, ki so zasnovani za reševanje specifičnih nalog in delujejo znotraj strogo določenih omejitev. Takšni sistemi so prisotni v sodobnih aplikacijah, kot so prepoznavanje govora, prevajanje in priporočanje vsebin; v to kategorijo sodi tudi GUI. Glede na to, da vse oblike UI, ki jih poznamo danes, sodijo na področje ozke oz. šibke UI, se zdi, da izpostavljanje tveganj, ki jih prinašajo neobstoječe tehnologije, za katere danes ne vemo niti, kdaj se bodo pojavile, niti, ali jih je mogoče sploh razviti, vodi v napačno smer.

Po eni se je s tem mogoče strinjati, ker zaradi osredotočanja na oddaljeno prihodnost, v kateri sta običajno prikazana le dva možna scenarija – utopični in distopični – zanemarjamo konkretne probleme in tveganja, ki so posledica uporabe UI tukaj in zdaj. Lahko da bo zaradi razvoja UI, ki bo dosegla ali presegla raven človeške inteligence, v prihodnosti prišlo do »raja na zemlji« ali »konca sveta«, vendar pa to v tem trenutku ne vpliva na dejstvo, da UI že danes vpliva, med drugim, na odločitve volivcev in volivk, izgubo delovnih mest in tudi na izobraževanje. Zaradi posvečanja pozornosti neobstoječi tehnološki prihodnosti obstaja nevarnost, da bomo na ta način spregledali in zanemarili navidez manjše, zato pa nič manj pereče sedanje probleme, povezane z uporabo UI.

Po drugi strani pa, kot opozarja npr. Stuart Russell (2020), moramo problem uravnavanja UI nasloviti dovolj zgodaj, torej še med razvojem tehnologije, saj bi želeli, da je v trenutku, ko se bo pojavila, umetna splošna inteligenca že uravnana. Dobro bi bilo torej, da ima človeške namene in vrednote, ne pa lastnih (strojnih) namenov in vrednot, nad katerimi nimamo niti pregleda niti nadzora.

Ta pogled se navezuje na dominantno strategijo oz. paradigmo v razvoju sistemov UI, ki jo poznamo pod izrazom *standardni model UI* (Russell in Norvig, 2022; Russell, 2020). Gre za pristop, v skladu s katerim so strojni sistemi UI zasnovani tako, da maksimirajo doseganje vnaprej določenih ciljev. To pomeni, da so naravnani na optimizacijo določenih funkcij, ki jih določijo njihovi ustvarjalci. Sistemi »delajo prav« oz. se vedejo na ustrezen način, če je njihovo delovanje opredeljeno s ciljem, ki jim je bil določen, in temu cilju sledijo.

Standardni model s seboj prinaša posebnost, ki jo srečamo tudi v človeškem vedenju in jo poznamo kot Goodhartov zakon, zato je za razumevanje modela nemara smiselno pričeti na tem mestu. Britanski ekonomist Charles Goodhart (1975, str. 116) jo je opredelil na naslednji način: »Ko merilo postane cilj, preneha biti dobro merilo.« V kontekstu izobraževanja bi lahko v skladu s tem sklepali, da takrat, ko ocene postanejo merilo in s tem tudi cilj, dejansko niso več ustrezno merilo, s pomočjo katerega bi lahko ocenjevali znanje. Z

drugimi besedami, znanje postane nepomembno, ker štejejo samo ocene, delovanje sistema pa se prilagodi tako, da sledi zgolj zastavljenemu cilju, torej ocenam.

Takšno vedenje je značilno tudi za sisteme UI, ki delujejo v skladu s standardnim modelom, zato takšen pristop nemara ni najboljši oz. najvarnejši. Švedski filozof Nick Bostrom (2014) ga je na slikovit način predstavil v razvpitem miselnem eksperimentu, ki v izhodišče postavlja sistem, ki izdeluje sponke za papir. Predlaga, da si predstavljamo superinteligentni sistem, ki je programiran tako, da ima en sam preprost cilj – proizvesti čim več sponk za papir. V kolikor takšen sistem ne bi imel ustreznih omejitev in razumevanja širših človeških vrednot, bi sledil samo zastavljenemu cilju in vse razpoložljive vire, vključno s človeštvom in planetom, pretvoril v sponke za papir – zgolj zato, da bi dosegel svoj cilj.

Kljub temu da so po eni strani razprave o bodočih tehnologijah do neke mere upravičene z vidika razumevanja razvoja tehnologije UI in njenih potencialnih učinkov, pa po drugi strani ne smemo zanemariti vpogledov, ki so posledica obstoječih in preteklih tehnologij. Nenazadnje se je lažje učiti iz preteklosti kot iz prihodnosti.

V grški mitologiji je pojavljanje novih tehnologij obravnavano na več mestih, večinoma pa pripovedi vključujejo zavedanje o moči tehnologije in potrebo po odgovorni uporabi. Mit o Prometeju, ki je ukradel ogenj bogovom in ga podaril ljudem, simbolizira prenos tehnološkega znanja ter opozarja na posledice preseganja meja med človeškim in nečloveškim. Podobno Dedal kot izumitelj razvije tehnologijo, s katero lahko z Ikarjem zbežita iz labirinta na Kreti, vendar pa sinovi nepremišljenost in prevzetnost vodita v tragedijo, ki jo lahko pripišemo neodgovorni uporabi tehnologije (Mavromataki, 1997).

Tudi različni vidiki tveganj, ki jih prinašajo nove tehnologije ali negotova prihodnost, so uvrščeni v razrede, ki nosijo imena po grških mitoloških likih (Renn, 2008). UI sodi v razred tveganj, ki je dobil ime po Kasandri, trojanski prerokinji, ki je napovedala prihodnje katastrofe, vključno s padcem Troje, vendar ji nihče ni verjel. Za ta razred tveganj je značilna visoka verjetnost zelo resnih družbenih posledic, ki pa so predstavljene v nedoločljivo prihodnost, zaradi česar so opozorila večinoma prezrta. Kljub temu da obstaja veliko različnih napovedi glede vpliva UI na individualni in družbeni ravni, ne vemo, katere napovedi se bodo uresničile in v kolikšni meri (Boddington, 2023).

Na prvo tehtno razpravo o posledicah vpeljave nove tehnologije sicer naletimo pri Platonu, ki v dialogu *Fajdros* med drugim obravnava tudi iznajdbo pisave. V Platonovi pripovedi so pisavo iznašli v Egiptu, izumitelj, ki mu je bilo ime Tevt, pa je poleg pisave iznašel tudi matematiko, astronomijo in

igre na srečo. Ko je obiskal egipčanskega kralja Tamunta, mu je v pozitivni luči predstavil vsako izmed iznajdb v upanju, da jih bo slednji razširil med svoje podanike. Tako tudi pri pisavi ni skoparil s hvalo (Platon, 2009, str. 569): »Ta nauk, kralj, bo naredil Egipčane modrejše in spominsko sposobnejše, kajti iznajden je bil kot zdravilo za spomin in modrost.«

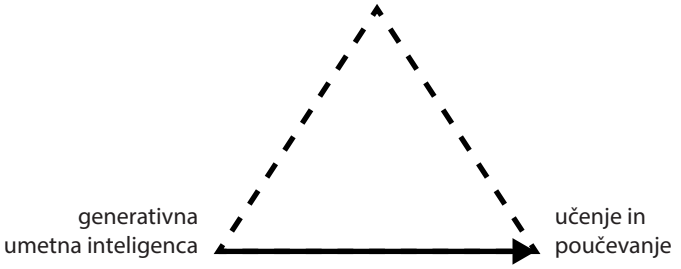
Vendar pa se kralj ni pustil tako zlahka prepričati in je Tevta opozoril na to, da je kot izumitelj pisave sam pristranski, nevtralno oceno nove tehnologije pa lahko poda samo nekdo, ki jo presoja z zunanje perspektive. V tem primeru pa Tamunt pokaže, da je resnica o njenih učinkih ravno nasprotna od tiste, ki jo je podal Tevt (Platon, 2009, str. 569):

> Ta (iznajdba) bo namreč v dušah tistih, ki se (je) naučijo, zaradi zanemarjanja spomina povzročila pozabo, ker se bodo zaradi zaupanja v pisanje spominjali od zunaj, zaradi tujih znakov, ne pa od znotraj, sami od sebe. Torej nisi iznašel zdravila za spomin <*mnéme*> , ampak za spominjanje <*hypómnesis*>. In učencem pridobivaš videz modrosti, ne (njene) resnice. Veliko bodo namreč od tebe slišali, in ne da bi se poučili, se jim bo zdelo, da marsikaj poznajo, čeprav večinoma ne poznajo ničesar in se je z njimi težko družiti, ker niso postali modri, ampak navidezno modri.

Platonova obravnava pisave kot ambivalentne tehnologije, ki jo je glede vplivov mogoče razumeti obenem kot zdravilo in strup, nam lahko pride prav tudi pri razumevanju in vrednotenju novejših tehnologij vse do UI. Pomembno je upoštevati, da je Platon svoj zapis oblikoval v času, ko si je bilo še mogoče predstavljati svet brez pisave – danes si tega več ne moremo –, zato je njegov komentar toliko tehtnejši, saj bodo tudi generacije, ki se bodo »rodile« v UI, veliko težje prepoznavale tveganja, povezana z njeno uporabo.

Poleg tega Platonov primer opozarja tudi na to, da proizvajalcem novih tehnologij, kot je UI, ne smemo slepo verjeti, ko gre za njihovo kakovost, prednosti in pozitivne učinke. Glede na to, da so sistemi GUI v največji meri v rokah korporacij, je težko podvomiti v to, da je njihov cilj predvsem dobiček, vse ostalo pa je temu podrejeno. Nenazadnje, kot pokaže tudi Platon, vsaka tehnologija poleg pozitivnih prinaša tudi negativne učinke in posledice, zato nedvomno tudi za (G)UI velja, da je tako zdravilo kot tudi strup (slika 3).

Ravno zaradi tega je nujno prepoznavati pozitivne in negativne posledice uporabe UI ter med njimi ločevati, kar zahteva kritični pristop in vodi v regulacijo tako z vidika zakonodaje kot tudi etike.



Slika 3 Neposredni vpliv (G)UI na izobraževanje (učenje in poučevanje)

Etika

Odnos do UI, kar zadeva zakonodajo, se načeloma razlikuje od države do države, s tem pa tudi dovoljene in prepovedane prakse ter posledično tudi dostop do določenih aplikacij in načinov uporabe UI in GUI.[3] Na ravni Evropske unije (EU) so prizadevanja na tem področju vodila v sprejem Akta o umetni inteligenci (Regulation (EU) 2024/1689 of the European Parliament and of the Council, 2024), ki vzpostavlja pravila za uporabo, razvoj in uvajanje UI v EU. Njegov temeljni cilj je spodbujati na človeka osredotočeno in zaupanja vredno UI, ki spoštuje temeljne pravice, varuje javne interese in hkrati omogoča inovacije ter prost pretok blaga in storitev. Uredba uvaja različno strogost regulacije glede na raven tveganja posameznih sistemov UI, z najstrožjimi pravili za visoko tvegane sisteme. Med drugim prepoveduje manipulativne prakse UI, ki ogrožajo avtonomijo posameznikov, in zagotavlja varstvo pred diskriminacijo ter nezakonito obdelavo osebnih podatkov. Akt je skladen z vrednotami EU, zapisanimi v Listini Evropske unije o temeljnih pravicah (2016), in podpira trajnostni razvoj, digitalno pismenost ter vodilno vlogo Evrope v globalnem razvoju etične UI.

V zadnjih letih je bila objavljena vrsta dokumentov, ki se posvečajo etiki UI (Evropska komisija, 2019; UNESCO, 2019, 2022) in etiki GUI (UNESCO, 2023), poleg tega pa tudi področju učenja in poučevanja v povezavi z UI (Evropska komisija, 2022).

Natančnejše branje navedenih dokumentov pokaže, da smernice večinoma sledijo določenim izhodiščem, ki so postavljena kot temeljna načela oz. aksi-

[3] Konec marca 2023 je Italija začasno prepovedala uporabo ChatGPT-ja zaradi pomislekov glede zasebnosti. Garante, italijanski organ za varstvo podatkov, je izrazil zaskrbljenost zaradi množičnega zbiranja in shranjevanja osebnih podatkov brez ustrezne pravne podlage ter pomanjkanja mehanizmov za preverjanje starosti uporabnikov in zahteval prepoved uporabe (McCallum, 2023). Na podlagi uvedbe potrebnih ukrepov proizvajalca OpenAI je Italijanski regulator konec aprila 2023 omogočil ponovno uporabo ChatGPT v tej državi (Pisa, 2023).

omi. V dokumentu Evropske komisije z naslovom *Etične smernice za uporabo umetne inteligence in podatkov pri poučevanju in učenju za izobraževalce* (2022, str. 18) je pojasnjeno: »Pri oblikovanju teh smernic so bili opredeljeni štirje ključni vidiki, na katerih temelji etična uporaba UI in podatkov pri poučevanju, učenju ter ocenjevanju. To so človekovo delovanje, pravičnost, človečnost in upravičena izbira.« Nikjer pa ni pojasnjeno, zakaj so ključni ravno ti vidiki.

Pričakovali bi tudi, da bodo omenjeni ključni vidiki ustrezno utemeljeni, vendar pa temu ni tako, kot kaže primer pravičnosti (Evropska komisija, 2022, str. 18): »Pravičnost se nanaša na to, da se v družbeni organizaciji vsakdo obravnava pravično. Potrebni so jasni postopki, da imajo vsi uporabniki enak dostop do priložnosti. Ti vključujejo pravičnost, vključevanje, nediskriminacijo ter pravično porazdelitev pravic in odgovornosti.« Pravičnost je torej opredeljena (v veliki meri) ciklično, kar ne omogoča, da bi jo razumeli brez sklicevanja na pravičnost.

Rezultati metaraziskave, ki jo je opravila iniciativa AI4People (Floridi idr., 2018), so pokazali, da obstaja presenetljiva stopnja koherentnosti in prepletenosti načel, pri čemer se etična načela UI v veliki meri pokrivajo s tradicionalnimi načeli bioetike, saj je za to, da dobimo načela etike UI, treba dodati le eno novo načelo, ki pokriva značilnosti sistemov UI. Temeljna načela bioetike povezujemo z dvema prelomnima dokumentoma, ki sta vplivala na vzpostavljanje okvirov za etično odločanje v medicinskih kontekstih. To sta delo Načela biomedicinske etike (*Principles of Biomedical Ethics*), ki sta ga leta 1977 objavila Tom Beauchamp in James Childress (2013), in poročilo ameriške Nacionalne komisije za zaščito človeških subjektov biomedicinskih in vedenjskih raziskav, ki je bilo oblikovano med letoma 1974 in 1978 in objavljeno leta 1978 kot *The Belmont Report* (National Commission for the Proptection of Human Subjects of Biomedical and Behavioral Research, 1978).

Poročilo artikulira tri etična načela: *spoštovanje oseb* (avtonomija), *dobrodelnost* in *pravičnost*. Beaumont in Childress pa poleg dobrodelnosti (angl. *benevolence*) kot ločeno načelo navajata še *neškodljivost* (angl. *non-maleficience*). Tako dobimo štiri temeljna načela bioetike, ki so relevantna še danes. Kot je pokazala omenjena raziskava AI4People, ista načela nastopajo tudi v večini dokumentov o etiki UI. Peto načelo, ki ga ne zasledimo med načeli bioetike, je razložljivost (angl. *explicability*); vodi v potrebo po razumevanju in upoštevanju odločanja sistemov UI, kar se izraža kot odgovornost, preglednost oz. razumljivost (Floridi idr., 2018).

Vendar pa sama etična načela ne zagotavljajo, da bodo sistemi UI oblikovani tako, da bo tudi njihova raba etična oz. da bodo uporabniki na podlagi teh načel orodja UI uporabljali etično.

Razvoj in uporaba UI nas postavljata pred nujo oblikovanja celovitega etičnega okvira, ki bi omogočil učinkovito reševanje konkretnih etičnih izzivov v kontekstu izobraževanja. Etična načela, ki jih najdemo v dokumentih o etiki UI, so lahko izhodišče, vendar ne zadostujejo za reševanje konkretnih vprašanj. Zgolj tehnološko razumevanje UI ni dovolj, saj etični problemi, povezani z UI, zahtevajo poglobljeno poznavanje filozofske etike, ki lahko ponudi trdno teoretsko osnovo za razumevanje in uporabo teh načel v praksi. Etični okvir za UI ni zgolj skupek tehnoloških smernic, temveč vključuje širši razmislek o človeškem delovanju, pravičnosti, avtonomiji in odgovornosti, ki jih je mogoče razumeti le z vključitvijo filozofske perspektive, zato se v nadaljevanju posvetimo osnovam etike.

Etimološko nas izraz »etika« napotuje v antično Grčijo, k izrazu *ethos,* ki se je v svojem zgodnjem pomenu nanašal na kraj bivanja pa tudi na navade in običaje (MacIntyre, 1998). Sčasoma se je oblikoval pomen, ki je *ethos* povezoval z vprašanjem, kaj je pri navadah in običajih prav in kaj narobe, s tem pa se je približal označevanju moralnih norm ali značajev, ki odločajo o vedenju posameznih ljudi ali skupin. Temu je blizu izraz »morala«, ki ga je vpeljal rimski državnik in filozof Mark Tulij Cicero (106–43 pr. n. št.) in je nastal s prevodom grškega pojma *ethos.* Cicero je z njim želel opisati tisto, kar oblikuje vedenje ljudi, in je v ta namen uporabil latinski izraz *mores,* ki pomeni tako navade in značaje kot tudi vrline (značaj) in predstavo o tem, kaj v neki skupnosti velja za dobro oz. slabo. Danes se izraza »etika« in »morala« velikokrat uporabljata kot sinonima, čeprav je morala praviloma povezana s pravili, z vrednotami in normami, ki naj bi opredeljevale dejanja ljudi, medtem ko je etika kot teorija moralnosti ali moralna filozofija povezana z načeli in s splošnimi normami (Bartneck idr., 2021). Na področju etike UI večina avtorjev ne posveča posebne pozornosti razlikovanju med obema pojmoma. Praviloma povsod srečamo izraz »etika umetne inteligence«, čeprav v določenih kontekstih naletimo tudi na pojem morala, kot npr. pri vprašanju »moralnih strojev« (Boddington, 2023).

Vprašanja, ki jih naslavlja etika, veliko lažje razumemo, če na začetku izpostavimo nekaj kategorizacij oz. delitev. Smiselno se zdi, da najprej ločimo *deskriptivno* in *normativno* etiko. Deskriptivna etika se ukvarja z opisovanjem, analizo in razlago moralnih prepričanj, norm, praks in vrednot, kot dejansko obstajajo v različnih družbah, kulturah ali skupinah. Gre za empirične raziskave, ki pogosto vključujejo metode iz sociologije, antropologije ali psihologije, pri čemer zastavljajo vprašanja, kot je npr.: »Katere moralne norme so značilne za določeno skupino in kako vplivajo na vedenje?« Podobna vprašanja lahko zastavimo tudi v kontekstu izobraževanja in npr. uporabe GUI pri

pisanju seminarjev ali domačih nalog: »Katere moralne norme vplivajo na verjetnost neavtorizirane uporabe orodij generativne umetne inteligence (plagiatorstvo)?«

Empirični uvidi v dejansko vedenje ljudi in skupin oblikujejo izhodišča, na katerih temelji normativna etika, ki se ukvarja z oblikovanjem moralnih načel, pravil in teorij, ki usmerjajo človeško vedenje. Ko razmišljamo o etiki, se običajno osredotočamo na njen normativni vidik in z njim povezana vprašanja: kako bi morali delovati; kaj je prav in narobe oz. kaj je dobro in kaj slabo?

Vendar pa normativna etika ne stoji zgolj v odnosu do deskriptivne etike, pač pa jo velikokrat postavljamo tudi v odnos do *metaetike*. Normativna etika in metaetika namreč predstavljata dve temeljni veji etike, pri čemer se metaetika posveča vprašanjem narave in metodologije moralnih sodb. Vprašanja na tej ravni so npr.: »Kaj pomeni 'dobro' ali 'morati'? Kako lahko zagovarjamo prepričanja, ki se nanašajo na prav ali narobe?« Pri tem lahko ločimo dve komponenti: naravo moralnih sodb (npr. definicije pojmov, kot je »dobro«) in metodo, ki nam pove, kako izbrati ustrezna moralna načela (Gensler, 2011).

Za osebo, ki zagovarja npr. kulturni relativizem – stališče, ki moralo utemeljuje na družbenih konvencijah –, metaetika podaja tako odgovore na vprašanja glede narave moralnih sodb kot tudi glede metodologije, obenem pa s tem opredeljuje tudi samo stališče, v tem primeru kulturni relativizem:

- Definicija: »dobro« pomeni »družbeno sprejemljivo«.
- Metoda: izberi tista moralna načela, ki jih tvoja skupnost odobrava.

Tudi pri normativni etiki nekateri avtorji ločijo dve ravni (Gensler, 2011): teorijo normativne etike in praktično oz. aplikativno etiko. Drugi (Bartneck idr., 2021; Singer, 2008) praktično etiko obravnavajo kot posebno področje, ki ni podpodročje normativne etike.

Normativna etika se torej posveča analizi človeških dejanj z vidika »dobrega« in »slabega« ali »moralno ustreznega« in »moralno neustreznega«, na tej podlagi pa kategorizira dejanja kot pravilna (ustrezna) ali napačna (neustrezna). Običajno se normativna etika ne ukvarja s subjektivnimi vidiki ali z individualnimi primeri, temveč cilja na obča načela in splošno veljavnost. Kraja predstavlja moralno neustrezno (moralno slabo) dejanje za vsakogar. Vendar pa to še ne pomeni, da bo neko dejanje v vseh primerih razumljeno kot dobro ali slabo. V kontekstu normativne etike namreč razlikujemo več tipov teorij, za katere so značilni različni pristopi, na podlagi katerih presojamo človeška dejanja. Pri tem največkrat razlikujemo tri tipe teorij: etiko vrlin, deontološko etiko in konsekvencionalistično etiko.

Etika vrlin

Etiki vrlin lahko sledimo v preteklost do klasičnih grških filozofov. V Platonovem delu *Država* (Platon, 2009) srečamo kardinalne vrline, ki jih je zatem prevzela tudi krščanska teologija: razumnost (ali modrost), pogum, zmernost in pravičnost. Izmed teh je pravičnost tista vrlina, ki povezuje in usklajuje ostale vrline, takšno vlogo pa je ohranila kot temeljno etično načelo tudi v sodobnosti, v kontekstu bioetike in tudi etike UI. Pri Aristotelu, čigar poglede na to področje srečamo predvsem v delu *Nikomahova etika* (2002), so vrline del sistema, v katerem lahko ločimo dve vrsti vrlin, intelektualne in moralne. Prve se nanašajo na mišljenje, druge na delovanje: med prve sodijo filozofska in praktična modrost (vednost o tem, kako živeti), intuicija pa tudi znanstvena vednost; med drugimi spet srečamo pravičnost in pogum pa tudi prijaznost in duhovitost. Aristotelov pogled na pravičnost je relevanten tudi za etiko UI, saj je zanj pravičnost povezana s pravičnim ravnanjem z drugimi. Posebej pomemben je vidik distributivne pravičnosti, ki je povezan z razdeljevanjem stvari, časti in bogastva, kar se v sodobnem kontekstu navezuje tudi na vprašanje dostopnosti tehnologij, kot je UI.

Čeprav se na prvi pogled zdi, da je etika vrlin, ki se je razvila več kot dve tisočletji pred prvimi računalniki, zastarela in neprimerna za naslavljanje sodobnih tehnologij, kot je UI, pa natančnejši pregled literature na presečišču etike in tehnologije pokaže, da je etika vrlin še vedno živa. Še več, lahko bi celo trdili, da je oživela ravno zaradi sodobnega tehnološkega razvoja. Thilo Hagendorff (2020, str. 9) tako v svoji metaanalizi etičnih smernic UI poudarja, da bi bilo mogoče narediti pomemben korak naprej ravno z vključevanjem etike vrlin: »Etika tako ni več razumljena kot deontološko navdahnjena vaja iz odkljukavanja, temveč kot projekt osebnostnega napredka, spreminjanja stališč, krepitve odgovornosti in zbiranja poguma za opustitev določenih dejanj, ki veljajo za neetična.«

Ena izmed najvplivnejših avtoric na tem področju je Shannon Vallor, ki je v delu *Technology and the Virtues* (Tehnologija in vrline) problem opredelila na sledeči način (2016, str. 1): »[S]kupaj moramo v sebi gojiti posebno vrsto moralnega značaja, ki izraža tisto, kar bom imenovala *tehnomoralne vrline*.« Tehnomoralne vrline, kot jih opredeljuje avtorica, niso nekaj povsem novega, temveč izhajajo iz obtsoječe tradicije etike vrlin in predstavljajo prilagoditev klasičnih filozofskih pristopov sodobnim potrebam tehnološko razvite družbe. Gre za uskladitev obstoječih moralnih sposobnosti na način, ki omogoča uspešno soočanje s hitrimi tehnološkimi in družbenimi spremembami. Po avtoričinem mnenju so tehnomoralne vrline kot palica, ki pomaga slepemu človeku – pomagajo nam poiskati pot v negotovih situacijah, obenem

pa ostajajo skladne z osnovno moralno psihologijo naše vrste, pri čemer sta ključnega pomena njihova zavestna kultivacija v družinah, izobraževanju in širši skupnosti ter podpora ustreznih institucij. Brez te etične podlage nova tehnologija ne more preprečiti groženj, ki jih lahko povzroči.

Deontološka etika

V drugo skupino teorij sodi deontološka etika, ki jo največkrat povezujemo z delom Immanuela Kanta (2003, 2005). Poimenovanje izhaja iz grških pojmov *déon* (*δέον,* »obveznost« ali »dolžnost«) in *logos* (*λόγος,* »preučevanje«) ter se nanaša na normativno etično teorijo, ki poudarja pomen dolžnosti in pravil pri moralnem presojanju dejanj. V nasprotju s konsekvencializmom, ki ocenjuje moralnost dejanj glede na njihove posledice, deontološka etika zagovarja stališče, da obstajajo določena dejanja, ki so moralno obvezna ali prepovedana ne glede na njihove izide. Obstajajo namreč univerzalna moralna načela, ki jih je treba spoštovati zaradi njih samih, ne glede na posledice, ki jih lahko povzročijo. Kant je bil prepričan, da morajo moralna dejanja izhajati iz dolžnosti in biti skladna z univerzalnimi zakoni, iz česar sledi kategorični imperativ, ki poudarja notranjo moralno vrednost dejanj in ne dopušča, da bi se moralna ustreznost oz. pravilnost dejanja določala zgolj na podlagi njegovih posledic.

Deontološka etika ima svoje prednosti pa tudi slabosti. Med prednosti sodita poudarjanje pravičnosti in spoštovanje individualnih pravic, kar preprečuje, da bi ljudi obravnavali zgolj kot sredstva za dosego ciljev. To poudarja tudi Kant (2005), ki v svojem delu o metafiziki morale izpostavlja, da je treba človeštvo, bodisi v sebi ali v drugih, vedno obravnavati kot cilj samo po sebi in nikoli zgolj kot sredstvo.

Vendar pa se deontološka etika ves čas sooča tudi s kritikami zaradi svoje togosti, saj lahko vztrajanje pri absolutnih moralnih pravilih in neupoštevanje ciljev v določenih situacijah vodita do neustreznih ali celo škodljivih izidov. V kolikor dosledno sledenje deontološkim načelom (pravilom, dolžnostim) vodi do posledic, ki so moralno vprašljive ali celo slabše kot v primeru, če bi bila ta načela kršena, govorimo o deontološkem paradoksu (Alexander in Moore, 2024).

Tako je npr. v skladu z deontološko etiko prepovedano lagati v vseh primerih in okoliščinah. Obenem pa si je mogoče predstavljati primer, pri katerem bi ravno nasprotno ravnanje, torej izražanje resnice, povzročilo znatno škodo – npr. če bi s tem izdali lokacijo osebe, ki jo nekdo namerava poškodovati. V takšnih primerih bi lahko ravno dosledno upoštevanje dolžnosti privedlo do neželenih posledic. Paradoks tako izpostavlja napetost med zavezanostjo ab-

solutnim moralnim dolžnostim oz. pravilom in praktičnimi posledicami spoštovanja teh dolžnosti oz. upoštevanja pravil, pri čemer je vredno izpostaviti, da se takšna vprašanja velikokrat zastavljajo v okviru konkretnih življenjskih situacij in ne zgolj pri miselnih eksperimentih (Zack, 2009).

Spoštovanje univerzalnih moralnih načel ima seveda pomembno vlogo tudi pri oblikovanju etičnih smernic za razvoj in uporabo orodij (G)UI. Deontološki pristop zahteva, da sistemi UI delujejo v skladu z določenimi pravili, ne glede na posledice, kar zagotavlja, da tehnologija spoštuje temeljne človekove pravice in dostojanstvo. Iz deontološkega etičnega okvira bi npr. izhajalo, da se morajo sistemi UI vedno ravnati v skladu z načelom poštenosti in nediskriminacije, ne glede na morebitne koristi, ki bi jih lahko prineslo nasprotno delovanje.[4]

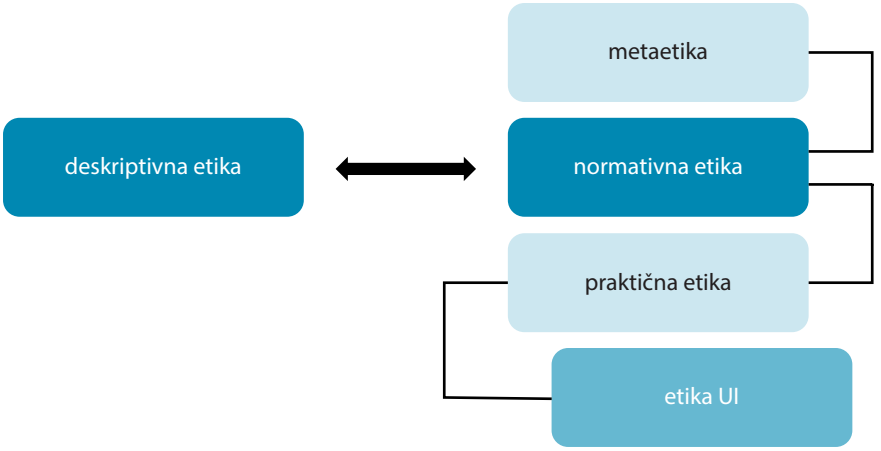
V praksi to pomeni, da morajo razvijalci in oblikovalci UI v svoje sisteme vgraditi mehanizme, ki zagotavljajo skladnost z etičnimi načeli, kot so avtonomija, pravičnost, preglednost, odgovornost in spoštovanje zasebnosti (Akrout in Steinbauer, 2020; Singh, 2022). Poleg tega pristop, utemeljen na deontološki etiki, spodbuja vzpostavitev regulativnih in nadzornih mehanizmov, ki zagotavljajo, da uporaba UI vodi k pozitivnim učinkom na individualni in družbeni ravni. To lahko vključuje vzpostavitev etičnih odborov, razvoj smernic in standardov ter izvajanje rednih ocen etičnega vpliva sistemov UI. Unesco tako npr. v svojem *Priporočilu o etiki umetne inteligence* poudarja pomen ocenjevanja etičnega učinka UI skozi celoten življenjski cikel sistemov z namenom preprečevanja škode za ljudi in zagotavljanja spodbujanja človekovih pravic (UNESCO, 2024).

Utilitarizem

Čeprav se tako etika vrlin kot tudi deontološka etika pojavljata v razpravah o umetni UI, je kljub vsemu najmočneje prisotna tretja skupina teorij, katerih skupna značilnost je, da se ne osredotočajo na moralno ravnanje, pač pa predvsem na njegove posledice. Med temi teorijami, ki jih zato poznamo kot konsekvencionalistične, je najvplivnejši utilitarizem. Slednji velja za dominanten etični pristop tudi v kontekstu praktične etike, kamor uvrščamo tudi etiko UI (slika 4).

Praktična etika običajno velja za vejo filozofije, ki se osredotoča na uporabo etičnih načel za reševanje konkretnih in aktualnih moralnih vprašanj v resničnem svetu. Njeno bistvo je v preseganju zgolj teoretičnih razprav s področja

[4] Pri razvoju avtonomnih vozil bi deontološki pristop npr. zahteval, da vozilo v vseh okoliščinah upošteva prometne predpise in varnostne standarde, ne glede na pritiske ali potencialne koristi odstopanja od teh pravil.



Slika 4 Struktura etike in mesto etike UI

etike, saj se ukvarja z dilemami, kot so revščina, pravice živali, podnebne spremembe in bioetika, pri čemer stremi k maksimizaciji dobrega in zmanjšanju trpljenja (Singer, 2008; Ekmekci in Arda, 2020; Ženko, 2024). Etika UI spada v ta kontekst, saj se ukvarja z moralnimi izzivi, povezanimi z razvojem in uporabo tehnologij, ki imajo vse večji vpliv tako na individualni kot na družbeni ravni. Vprašanja, kot so pristranskost algoritmov, vpliv UI na zasebnost, odgovornost za odločitve, ki jih sprejemajo avtonomni sistemi, spoštovanje človekovih pravic itn., zahtevajo praktične rešitve, ki izhajajo iz filozofskih načel, hkrati pa odgovarjajo na tehnološke in družbene izzive (Tavani, 2016; Floridi, 2018; Liao, 2020; Müller, 2023).

Glede na to, da je osrednja naloga praktične etike iskanje načinov za maksimizacijo dobrega in zmanjševanje trpljenja, je smiselno, da se pri tem naslanja na normativne etične teorije, ki najbolj ustrezajo tem ciljem. Med njimi se kot najprimernejša pogosto izkaže utilitarizem.

Klasični utilitarizem temelji na prepričanju, da sta bolečina in užitek temeljni načeli, ki usmerjata človeško delovanje. Jeremy Bentham v svojem delu (2000, str. 14) poudarja, da bolečina in užitek ne le določata, kaj je dobro in slabo, temveč tudi oblikujeta verigo vzrokov in posledic naših dejanj: »Narava je človeštvo postavila pod oblast dveh vladarjev, bolečine in užitka. [...] Na eni strani je na njun prestol pritrjeno merilo dobrega in slabega, na drugi pa veriga vzrokov in posledic.« Na tej podlagi John Stuart Mill nadgradi utilitaristično teorijo z vpeljavo koncepta največje sreče, pri čemer srečo opredeli kot stanje ugodja in odsotnost bolečine. V skladu z njegovimi besedami je pravilnost ali napačnost dejanj odvisna od njihovega prispevka k sreči (Mill,

2003, str. 16): »Prepričanje, ki za temelj morale sprejema koristnost ali načelo največje sreče, trdi, da so dejanja pravilna, če pripomorejo k nastanku sreče, in napačna, če pripomorejo k nastanku nasprotja sreče. S srečo je mišljeno ugodje in odsotnost bolečine, z nesrečo pa bolečina in izguba ugodja.«

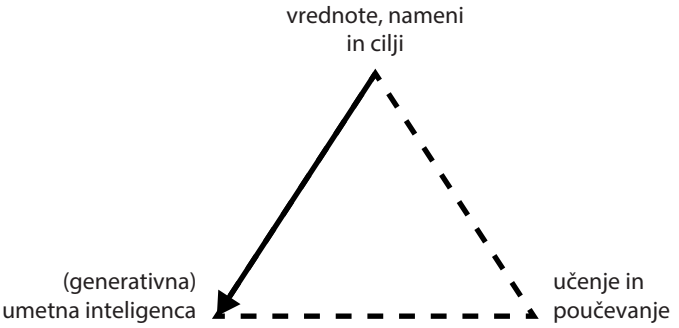
Temeljni pojem, ki ga srečamo v tem pristopu, je koristnost (angl. *utility*), iz katerega je utilitarizem dobil svoje ime. Koristnost se nanaša na sposobnost dejanj, da prispevajo k povečanju ugodja in zmanjšanju bolečine. Pri tem je osrednje vodilo težnja k maksimizaciji skupnega dobrega za čim več ljudi, moralnost dejanj pa se presoja v skladu z njihovimi posledicami za splošno dobrobit. Slednja se v sodobnejših interpretacijah utilitarizma, npr. pri Petru Singerju (2008), razširja na vse oblike življenja, pri čemer se osredotoča na zmanjšanje trpljenja in povečanje sreče ne le za ljudi, temveč tudi za druga čuteča bitja in v prihodnje nemara tudi za inteligentne stroje (Boyle, 2024).

Utilitarizmu, tako kot tudi drugim etičnim teorijam, lahko pripišemo pozitivne in negativne vidike. Med pozitivne spada poudarek na maksimizaciji sreče in zmanjševanju trpljenja za največje število ljudi, kar lahko vodi v družbeno koristne odločitve. Prav tako omogoča prilagodljivost, saj se moralna pravilnost dejanj ocenjuje glede na njihove posledice, kar lahko vodi k praktičnim in učinkovitim rešitvam, ki jih pričakujemo od praktične etike. Vendar pa utilitarizem pogosto zanemarja pravice posameznikov, saj lahko, kot poudarjajo številni kritiki, upraviči žrtvovanje manjšine v korist večine (Zack, 2009; Boddington, 2023; Alexander in Moore, 2024). Poleg tega je težko objektivno meriti in primerjati srečo ali korist med različnimi ljudmi, kar otežuje dosledno uporabo te teorije v praksi. Kritiki tudi opozarjajo, da se lahko osredotočenost na posledice izogne vprašanjem moralne dolžnosti in pravičnosti pa tudi oblikovanja ustreznih vrlin (Vallor, 2016).

Uravnavanje UI

Izraz »uravnavanje umetne inteligence« običajno povezujemo s procesi, povezanimi s prizadevanji, da bi bili sistemi UI oblikovani skladno s človekovimi vrednotami, nameni in cilji (slika 5), ter je v tem smislu pomemben predvsem z vidika ustvarjanja sistemov UI.

Uravnavanje UI je ključni element v kontekstu varnosti sistemov UI in služi temu, da bi preprečili neželene posledice ali zlorabe, do katerih bi lahko prišlo zaradi nepredvidljivosti ali napačnega delovanja sistemov (Russell, 2020). Vključuje razvoj algoritmov in modelov, ki so zasnovani tako, da upoštevajo etične norme, pravne zahteve in družbena pričakovanja. Ključni vidiki tega procesa pa so zagotavljanje transparentnosti, preverljivosti in razložljivosti odločitev, ki jih sprejemajo sistemi UI, pri čemer je bistveno vključevanje in-



Slika 5 Uravnavanje sistemov (G)UI

terdisciplinarnih pristopov, ki združujejo tehnološke, filozofske in družbene oz. kulturne perspektive. V tem smislu uravnavanje UI ni zgolj tehnični izziv, temveč tudi družbeni projekt, ki zahteva sodelovanje strokovnjakov z različnih področij.

Vendar pa pogledi na uravnavanje UI niso enotni, saj med njimi srečamo tudi takšne, ki menijo, da se glede uravnavanja sistemov UI ni treba posebej truditi, saj le-ti že imajo vgrajene vrednote, namene in cilje, le da ti niso njihovi, pač pa človeški, kajti UI je pravzaprav naše zrcalo, zrcalni odsev človeške družbe (Vallor, 2024). V tem smislu so torej sistemi UI vedno že uravnani, saj se pri učenju naučijo tudi naših vrednot, seveda pa tudi naših pomanjkljivosti in napak.

V kolikor bi slednje držalo dobesedno, bi si težko predstavljali, da od sistemov GUI sploh lahko dobimo smiselne odgovore, glede na to, koliko nesmiselnih in tudi neresničnih informacij je na internetu. Vendar pa tudi v tem primeru pomaga, če vsaj bežno poznamo način delovanja te tehnologije.

Proces učenja modelov GUI namreč poteka v dveh fazah, kar dejansko omogoča uravnavanje teh modelov s človeškimi vrednotami. Prva faza, imenovana »predhodno učenje« (angl. *pretraining*), obsega učenje iz velikih količin podatkov, pridobljenih iz različnih virov. Med tem procesom modeli »nekritično preberejo« razpoložljive podatke, kar pomeni, da zaznavajo vzorce, povezave in vzročnosti v podatkih, ne da bi vrednotili njihovo etično ustreznost ali skladnost s človeškimi vrednotami oz. normami. Šele v drugi fazi natančnega prilagajanja (angl. *fine-tuning*) modeli skozi skrbno zasnovan proces učenja prejmejo dodatna navodila in prilagoditve, s katerimi se poskuša zagotoviti njihova usklajenost z želenimi etičnimi in funkcionalnimi cilji. Ta faza pogosto vključuje povratne informacije ljudi pri spodbujevalnem učenju z nagrajevanjem in s kaznovanjem (angl. *reinforcement learning with*

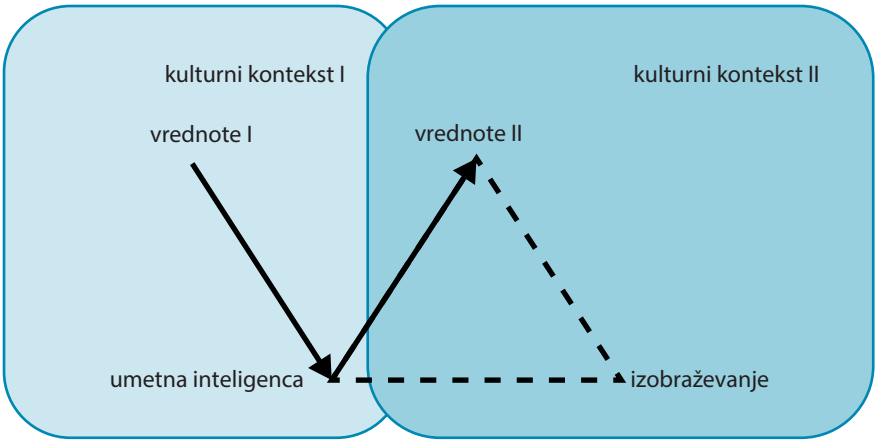
human feedback – RLHF), kar modelom pomaga bolje ponotranjiti človeške vrednote ter zmanjšati neželene vplive, ki izhajajo iz podatkov v prvi fazi učenja (Lambert, 2024).

Kljub temu da je v kontekstu varnosti uravnavanje sistemov UI izjemno pomembno in zato tudi pogosto obravnavano, to vseeno ni edini proces, v katerem srečamo prenos vrednot v povezavi s sistemi UI. Enako pomemben, čeprav veliko redkeje eksplicitno omenjen, je tudi proces, ki poteka v nasprotni smeri in v katerem sistemi (G)UI oblikujejo človeške vrednote, namene ter cilje.

Čeprav velikokrat poudarjamo, da je pomembno razumeti, da je UI predvsem orodje, to še ne pomeni, da je nevtralno ali pasivno orodje. Velikokrat namreč sistemi UI nastopajo kot aktivni udeleženci v družbenih, raziskovalnih in tudi izobraževalnih kontekstih. V kolikor je torej interakcija dvosmerna, pa pred nas postavlja zahtevo po tem, da se vprašamo: Kaj UI učimo o nas? In morda še pomembneje: Kaj nas UI uči o nas samih? Nasprotni vpliv, torej obratno uravnavanje, je namreč še posebej pomemben v kontekstu izobraževanja. Poleg neposrednega kratkoročnega vpliva na učenje in poučevanje (slika 3) lahko namreč prepoznamo tudi dolgoročnejši posredni vpliv, ki preko uporabe (G)UI v izobraževanju vodi v spremembe vrednot na družbeni ravni oz. v tem, kar smo imenovali kulturni kontekst. Tudi raziskovalci na drugih področjih danes opozarjajo na preveč poenostavljen pogled na uravnavanje UI, pri katerem je izpostavljen predvsem statičen enosmerni vpliv človeških vrednot na sisteme UI, ter namesto tega opozarjajo na upoštevanje dvosmernega dinamičnega procesa (Hua idr., 2024).

Upoštevati pa je treba še eno možnost, in sicer da vrednote, ki so jih pri uravnavanju UI uporabili snovalci sistema, ne odražajo vrednot uporabnikov istega sistema, torej v primeru izobraževanja ne vključujejo spreminjajočih se kulturnih, etičnih in pedagoških potreb izobraževalcev ter učencev. V kolikor je npr. izobraževalni sistem UI razvit v nekem drugem kulturnem ali institucionalnem kontekstu, lahko nehote uvaja norme in pričakovanja, ki so v nasprotju z normami in pričakovanji lokalnega okolja, v katerem se uporablja. To pa lahko privede do neželenih rezultatov in posledic (slika 6).

Lahko si npr. predstavljamo, da sistem UI, ki je bil uravnan s človeškimi vrednotami v individualističnem kulturnem kontekstu, ni skladen s kulturnim kontekstom, ki daje prednost kolektivističnim vrednotam – in obratno. Shannon Vallor v delu, v katerem UI primerja z zrcalom, ki odseva človeške vrednote, opozarja na še enega izmed problematičnih vidikov. Poudarja, da tisti, ki oblikujejo tehnologije UI, niso »predstavniki človeštva kot celote. So majhna podskupina, del vse bolj homogene tehnološke monokulture. Ti po-



Slika 6 Neustrezno uravnavanje UI v primeru, ko sistemi UI in njihovi uporabniki ne sodijo v isti kulturni kontekst

samezniki običajno prihajajo z istih elitnih univerz, kjer so študirali isti ozek nabor računalniških predmetov in nato delali za ista velika in bogata tehnološka podjetja« (Vallor, 2024, str. 13).

Sklepamo lahko, da so takšne neusklajenosti UI prej pravilo kot izjema, vendar pa je nujno poudariti tudi, da za ustrezno soočanje s takšnimi izzivi ne zadostuje zgolj iskanje tehničnih rešitev, pač pa je nujno potrebno tudi razumevanje razmerij med vrednostnimi sistemi, čemur pa do sedaj ni bilo namenjeno veliko pozornosti.

Vendar pa omenjene neusklajenosti med vrednotami, ki jih srečamo v različnih kulturnih kontekstih, ne smemo nujno razumeti v negativnem smislu. Lahko se namreč vprašamo, v čem je smisel uporabe orodij (G)UI, v kolikor le-ta ne privede do dodane vrednosti v kvalitativnem smislu? Če se sistemi GUI uporabljajo zgolj za to, da si učitelji pospešijo in olajšajo določene birokratske postopke ali si pomagajo pri ocenjevanju, ne da bi pri tem prišlo do sinergij, ki presegajo dosedanje prakse na področju učenja in poučevanja, je takšno uporabo z etičnega vidika težko upravičiti.

Sklep

Sistemi (G)UI so v zadnjem času postali pomemben dejavnik, ki vpliva na spremembe v izobraževalnih praksah, vendar pa njihova uporaba zahteva razmislek o ravnovesju med tehnološkimi, etičnimi in pedagoškimi vidiki. Mogoče je trditi, da UI ni zgolj orodje, temveč aktivno sodeluje v procesih, v katerih ne le odraža, temveč tudi oblikuje človeške vrednote. V tej dvosmerni dinamiki se nahaja potreba po odgovornem pristopu, ki presega tehnične

rešitve in vključuje razumevanje kulturnega konteksta, v katerem se uporablja tehnologija UI. S tem pristopom lahko zmanjšamo tveganja, povezana z neusklajenostjo, ter omogočimo, da UI postane dejavnik izboljšanja procesov učenja in poučevanja.

Eden izmed osrednjih izzivov je prepoznavanje in naslavljanje neskladij med vrednotami, ki jih odražajo sistemi UI, in vrednotami učiteljev ter učencev. Regulacija, ki upošteva globalne in lokalne dimenzije, je nujna, vendar pa sama po sebi ne zadostuje. Nujno je tudi razvijanje kritičnega dialoga, ki vključuje vse udeležence – od tehnoloških razvijalcev do izobraževalcev in tudi širše družbe. Šele skozi ta proces lahko zagotovimo, da bodo sistemi UI prilagojeni specifičnim potrebam različnih kulturnih in pedagoških okolij, kar bo preprečilo morebitno homogenizacijo vrednot in praks.

Nenazadnje, prihodnost uporabe UI v izobraževanju zahteva sinergijo med vsemi tremi oglišči trikotnika: tehnologijo, področjem uporabe in regulacijo. Le s celostnim pristopom, ki vključuje interdisciplinarno sodelovanje in stalno refleksijo, lahko zagotovimo, da UI ne bo le orodje učinkovitosti, temveč tudi pogoj za kakovostnejše, pravičnejše in bolj vključujoče izobraževanje. UI vsebuje možnosti, da postane tako »zdravilo« kot »strup« za izobraževanje; izbira, katero vlogo bo igrala, pa je, ne glede na vse, še vedno naša.

Literatura

Akrout, M., in Steinbauer, R. (2020, 7. februar). *Machine ethics: The Creation of a virtuous machine.* ArXiv. https://doi.org/10.48550/arXiv.2002.00213

Alexander, L., in Moore, M. (2024, 11. december). V E. N. Zalta in U. Nodelman (ur.), *The Stanford encyclopedia of philosophy*. Stanford University. https://plato.stanford.edu/archives/win2024/entries/ethics-deontological

Aristotel. (2002). *Nikomahova etika* (K. Gantar, prev.). Slovenska matica.

Atelšek, S., Fajfar, T. Jemec Tomazin, M., Sitar, J., in Žagar Karer, M. (2023). *Termonološka svetovalnica: uravnavanje umetne inteligence.* Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. https://isjfr.zrc-sazu.si/sl/terminologisce/svetovanje/uravnavanje-umetne-inteligence

Bartneck, C., Lütge, C., Wagner, A., in Welsh, S. (2021). *An introduction to ethics in robotics and AI.* Springer.

Beauchamp, T. L., in Childress, J. F. (2013). *Principles of biomedical ethics* (7. izd.). Oxford University Press.

Bentham, J. (2000). *An introduction to the principles of morals and legislation.* Batoche Books.

Boddington, P. (2023). *AI ethics: A textbook.* Springer.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

Boyle, J. (2024). *The line: AI and the future of personhood*. MIT Press.

Ekmekci, E. P., in Arda, B. (2020). *Artificial intelligence and bioethics.* Springer.

Evropska komisija. (2019). *Etične smernice za zaupanja vredno umetno inteligenco.*

Evropska komisija. (2022). *Etične smernice za uporabo umetne inteligence in podatkov pri poučevanju in učenju za izobraževalce.* https://op.europa.eu/sl/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., in Vayena, E. (2018). *AI4People's ethical framework for a good AI society: Opportunities, risks, principles, and recommendations*. Atomium – European Institute for Science, Media and Democracy.

Gensler, H. J. (2011). *Ethics: A contemporary introduction.* Routledge.

Goodhart, C. (1975). Problems of monetry management: The U.K. experience. V A. S. Courakis (ur.), *Inflation, depression and economic policy in the west* (str. 111–144). Barnes and Noble.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Mind Mach, 30*(1), 99–120.

Hare, S. (2022). *Technology is not neutral: A short guide to technology ethics.* London Publishing Partnership.

Heidegger, M. (2003). *Predavanja in sestavki* (T. Hribar idr., prev). Slovenska matica.

Hua, S. Idr. (2024, 10. avgust). *Towards bidirectional human-AI alignment: A systematic review for clarifications, framework, and future directions.* ArXiv. https://doi.org/10.48550/arXiv.2406.09264

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Vierling, L., Hong, D., Zhou, J., Zhang, Z., Zeng, F., Dai, J., Pan, X., Yee Ng, K., O'Gara, A., Xu, H., Tse, B., ... Gao, W. (2024, 1. maj). *AI alignment: A comprehensive survey.* ArXiv. https://doi.org/10.48550/arXiv.2310.19852

Kant, I. (2003). *Kritika praktičnega uma* (R. Riha, prev.). Društvo za teoretsko psihoanalizo.

Kant, I. (2005). *Utemeljitev metafizike naravi* (R. Riha, prev). Založba ZRC, ZRC SAZU.

Lambert, N. (2024). *The basics of reinforcement learning from human feedback.* RLHF Book.

Liao, S. M. (Ur.). (2020). *Ethics of artificial intelligence.* Oxford University Press.

Listina Evropske unije o temeljnih pravicah. (2016). *Uradni list Evropske unije*, (C 202), 389–405.

MacIntyre, A. (1998). *A short history of ethics: A History of moral philosophy from the homeric age to the twentieth century.* Routledge.

Mavromataki, M. (1997). *Greek mythology and religion.* Haitalis.

McCallum, S. (2023. 1. april). *ChatGPT banned in Italy over privacy concerns.* BBC. https://www.bbc.com/news/technology-65139406

Mill, J. S. (2003). *Utilitarizem; in O svobodi* (Z. Erbežnik in F. Klampfer, prev). Krtina.

Müller, V. C. (2023). Ethics of artificial intelligence and robotics. V E. N. Zalta (ur.), *Stanford encyclopedia of philosophy* (str. 1–70). Metaphysics Research Lab, Stanford University.

National Commission for the Proptection of Human Subjects of Biomedical and Behavioral Research. (1978). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research.*

Pisa, P. L. (2023, 28. april). ChatGpt è di nuovo disponibile in Italia. *La Reppublica.* https://www.repubblica.it/tecnologia/2023/04/28/news/chatgpt_riapre_in_italia-397985150/

Platon. (2009). *Zbrana dela* (G. Kocijančič, prev.). Celjska Mohorjeva družba.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). (2024). *Official Journal of the European Union,* (1689). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Renn, O. (2008). *Risk governance: Coping with uncertainty in a complex world.* Earthscan.

Russell, S. (2020). *Human compatible: Artificial intelligence and the problem of control.* Penguin.

Russell, S., in Norvig, P. (2022). *Artificial intelligence: A modern approach* (4. izd.). Pearson.

Singer, P. (2008). *Praktična etika* (2. izd., G. Cerjak in S. Vörös, prev.) Krtina.

Singh, L. (2022, 20. julij). *Automated Kantian ethics: A faithful implementation.* ArXiv. https://doi.org/10.48550/arXiv.2207.10152

Tavani, H. T. (2016). *Ethics and technology: Controversies, questions, and strategies for ethical computing.* Wiley.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.* Vintage Books.

UNESCO. (2019). *Preliminary study on the ethics of artificial intelligence.* https://unesdoc.unesco.org/ark:/48223/pf0000367823

UNESCO. (2022). *Recomendation on the ethics of artificial ingelligence.* https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNESCO. (2023). *Guidance for generative AI in education and research.* https://unesdoc.unesco.org/ark:/48223/pf0000386693

UNESCO. (2024). *UNESCO priporočilo o etiki umetne inteligence: oblikovanje prihodnosti naše družbe.*

Vallor, S. (2016). *Technology and the virtues: A Philosophical guide to a future worth wanting.* Oxford University Press.

Vallor, S. (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking.* Oxford University Press.

Zack, N. (2009). *Ethics for disaster.* Rowman & Littefield Publishers.

Ženko, E. (2024). Etični vidiki uporabe umetne inteligence v zdravstvu. V T. Štemberger Kolnik in A. Presker Planko (ur.), *Digitalizacija v zdravstvu in umetna inteligenca: inovacije za boljšo prihodnost* (str. 81–93). Fakulteta za zdravstvene vede v Celju.

In the Sign of The Triangle: Education, Ethics and Artificial Intelligence

The text discusses the interplay between artificial intelligence (AI), education, and ethics, emphasizing the belief that AI is not merely a tool but an active agent that can shape values, norms, and pedagogical practices. Central to the discussion is a triangular structure, where technology, usage, and regulation serve as key elements for understanding the impact of AI. The author highlights the necessity of a holistic approach that integrates technological innovations, ethical principles, and educational needs. The discussion also addresses the bidirectional dynamic of influence between AI and humans, focusing on the question of how AI reflects human values while also contributing to their transformation. Special attention is given to the issue of AI alignment, which demands interdisciplinary collaboration and an understanding of cultural contexts. Examples of misaligned values that can lead to ethical conflicts or inappropriate solutions are highlighted. The text concludes with a call for responsible use of AI that goes beyond mere technical solutions and includes critical dialogue among developers, educators, and society.

Keywords: ethics, education, artificial intelligence, generative artificial intelligence, artificial intelligence alignment