RAZISKOVALNA UMETNA INTELIGENCA IN STANDARDI TRANSPARENTNOSTI

BORUT TRPIN
Univerza Ludwig-Maximilian München, München, Nemčija
borut.trpin@lrz.uni-muenchen.de

Sinopsis Umetna inteligenca se vedno pogosteje uporablja v raziskovalne namene, bodisi kot podporno orodje bodisi kot povsem avtonomen vir znanstvenih spoznanj. Tovrstna uporaba umetne inteligence v raziskovanju odpira vprašanje, kakšne standarde transparentnosti moramo od nje zahtevati. Po študiji primera nato sledi sklep, ki se osredinja na dejstvo, da moramo od raziskovalne umetne inteligence zahtevati večjo transparentnost oz. razložljivost kot od ljudi.

Ključne besede:
umetna inteligenca,
znanstvene metode,
standardi transparentnosti,
spoznavni standardi,
filozofija znanosti,
spoznavna teorija

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.4
ISBN 978-961-286-675-4

RESEARCH ARTIFICIAL INTELLIGENCE AND STANDARDS OF TRANSPARENCY

BORUT TRPIN
Ludwig-Maximilians-University Munich, Munich, Germany
borut.trpin@lrz.uni-muenchen.de

Keywords: artificial intelligence, scientific methods, standards of transparency, epistemological standards, philosophy of science, epistemology

Abstract The use of artificial intelligence in research is increasing, be it as a supporting tool that complements scientists or as a completely autonomous source of scientific findings. Such usage of artificial intelligence in research raises a question regarding the standards of transparency that we request of it. After considering a case study it is established that higher standards of transparency or explainability should be requested for artificial intelligence in research than for human scientists.

https://doi.org/10.18690/um.ff.11.2022.4 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Uvod

Umetna inteligenca že nekaj časa ni zgolj predmet raziskovanja, temveč tudi sama poganja raziskave – bodisi kot orodje, ki pomaga pri obdelavi podatkov (npr. v fiziki; Radovic et al. 2018), vse pogosteje pa tudi kot povsem avtomatizirano orodje, ki je zmožno samo formulirati hipoteze, jih testirati in privesti do novih odkritij (glej npr. King et al. 2009). Uporaba umetne inteligence oz. vsaj določenih tovrstnih metod sega praktično v vsa raziskovalna področja, kjer imamo opravka z analizo večje količine podatkov, npr. od fizike, biologije, kemije pa tudi do filozofije (Grim in Singer 2020), kjer računalniške metode vse pogosteje služijo v podporo argumentaciji. Uporablja se tudi pri analizi filozofskih problemov.

Pri vsem tem velja opozoriti, da nastane pri uporabi umetne inteligence poseben izziv zaradi njene (ne)transparentnosti, oziroma z drugimi besedami, da imamo pri umetni inteligenci pogosto zgolj delni vpogled v postopke, ki usmerjajo njeno procesiranje, včasih pa niti tega ne (v primeru t. i. popolne črne škatle) ter posledično ne razumemo oz. niti ne moremo razumeti, kaj točno vodi do rezultatov, ki jih na tak način pridobimo. Ravno zahteve po večji transparentnosti so privedle do vedno večjega razmaha tako imenovane razložljive umetne inteligence (angl. *explainable artificial intelligence*), kjer vsaj z določenimi približki umetno inteligenco skušamo osmisliti in jo tako približati našemu razumevanju.

Vprašanje, ki se torej postavlja samo po sebi, je, kakšni bi morali biti standardi transparentnosti za umetno inteligenco v primerjavi s standardi transparentnosti, ki jih pričakujemo ob človeških odločitvah. S transparentnostjo pri tem metaforično ciljamo na to, ali imamo vpogled v razloge za določene odločitve ali ne (pri tem se naslanjamo na angleška pojma *opacity* in *transparency*, pojma neprosojnosti in transparentnosti, ki sta postala ustaljena v tovrstni nedobesedni rabi).

V splošnem obstaja več pogledov, ki jih lahko grobo razdelimo v dve skupini oz. tri, pri čemer tretja ne uživa prave podpore. Po eni strani se zdi, da bi morali biti standardi transparentnosti, ki jih pričakujemo od umetne inteligence, enaki kot standardi transparentnosti, ki jih pričakujemo od ljudi. Od ljudi ob razlagi odločitev ne moremo pričakovati razlage delovanja možganov, kvečjemu intencionalne razlage v mentalističnem jeziku (tj. do neke odločitve je prišlo, ker *verjamemo*, *mislimo*, *pričakujemo* ipd., da je nekaj tako in tako; glej tudi Dennet 1987 ter Zerilli et al. 2019 ter Günther in Kasirzadeh 2021 za diskusijo standardov transparentnosti na tej

podlagi). Kljub razvoju nevroznanosti namreč pri ljudeh popolnega vpogleda vseeno nimamo in določene plasti mišljenja nam tako ostanejo zakrite. Posledično, tako npr. Zerilli et al. (2019), tega ne moremo pričakovati niti od umetne inteligence.

Po drugi strani se zdi, da bi od umetne inteligence morali zahtevati višje standarde transparentnosti. Vsaj v principu imamo namreč lažji vpogled v algoritme kot pri ljudeh (če ne gre za primere popolnih črnih škatel), hkrati pa so algoritmi sami ključni za razumevanje odločitev. Poleg tega algoritme vsaj v osnovi zasnujemo ljudje (za še več argumentov v podporo višjega standarda transparentnosti glej Günther in Kasirzadeh 2021).

Zgolj kot možnost omenimo še tretjo varianto – od umetne inteligence bi načeloma lahko sprejeli tudi nižjo raven transparentnosti, kot jo pričakujemo od ljudi. Ta opcija je razumljivo bolj ali manj brez podpore, bi pa jo lahko zagovarjali v okviru določenega instrumentalizma v smislu, če nam umetna inteligenca zanesljivo služi, potem nam ni treba razumeti zakaj. Da to ni dobra ideja, nam kažejo primeri, kjer umetna inteligenca pri, denimo, vizualni prepoznavi odpove zaradi ljudem težko razumljivih razlogov (glej npr. Nguyen, Yosinski in Clune 2015 za primer v povezavi z vidnim prepoznavanjem ter Finlayson et al. 2017 za primer težav v medicini).

Čeprav puščam debato glede standardov transparentnosti umetne inteligence v splošnem ob strani, bom v pričujočem prispevku trdil, da moramo vsaj v okviru znanstvenega raziskovanja od umetne inteligence zahtevati višje standarde transparentnosti. To je tako, ker pri znanstvenem raziskovanju stremimo k zanesljivim spoznanjem, ki nam omogočajo opis, razlago, predvidevanje ter nadzor nad raziskovanimi pojavi. Če razlaga sklepov umetne inteligence umanjka, lahko hitro pristanemo v varljivem občutku znanja, ki to ni. Prav zmožnost razlage je namreč temelj za razumevanje (čeprav to debato puščam ob strani, za nasprotno stališče glej npr. Lipton 2009). Tak primer bi bil npr., če bi s pomočjo umetne inteligence prišli do neke ugotovitve in to zagovarjali kot rezultat znanstvene raziskave, dasiravno ne bi razumeli, zakaj sklep dejansko drži. Tako bi se kasneje lahko izkazalo, da je 'odkritje' zgolj napačen rezultat, ki je temeljil na pristranskem algoritmu. A zdi se, da tudi če bi šlo za pravilen rezultat, bi brez razumevanja procesa, ki je vodil do odkritja, šlo zgolj za naključno spoznanje, ki ne zadošča kriterijem znanstvenega odkritja.

Seveda je treba omeniti, da znanstvenih spoznanj, ki nastanejo na podlagi umetne inteligence, ne sprejemamo brez dodatnega preverjanja – podobno kot to velja tudi za znanstvena odkritja sicer. Vseeno pa gre pri sklepih na podlagi raziskovalne umetne inteligence, kot imenujem umetno inteligenco, ki se uporablja v raziskovalne (znanstvene) namene, za drugačen pristop kot pri človeškem raziskovanju. Pri slednjem se zdi, da v celotnem postopku – od snovanja hipotez do zbiranja in analize podatkov – implicitno stremimo k smiselnemu stapljanju raziskovanja s korpusom že obstoječega znanja. Po drugi strani gre pri umetni inteligenci pogosto za algoritmično iskanje povezav, ki jih razumemo le do neke mere. To lahko vodi tudi do odkritja nesmiselnih povezav oz. prepoznave vzorcev, ki so zgolj plod naključja. Kot pokaže Vigen (2015) v sicer drugačnem kontekstu, lahko ob analizi dovolj podatkov odkrijemo raznorazne povezave, ki več kot očitno niso vzročno-posledične kljub statistično visoki korelaciji. Eden od bolj humornih primerov je npr. izjemno visoka korelacija (98,9 %) med številom ločitev v ameriški zvezni državi Maine ter povprečno porabo margarine po osebi v ZDA med leti 2000 in 2009. Na podoben način bi načeloma lahko nesmiselne vzorce iz velike količine podatkov izluščila tudi umetna inteligenca.

Kot bom prikazal s študijo primera iz računalniške filozofije (angl. *computational philosophy* oz. filozofije, kjer računalniške metode igrajo osrednjo vlogo kot metoda filozofske argumentacije), lahko zanašanje na navidez smiselne rezultate oz. rezultate, ki jih ne razumemo dovolj, vodi v zmotne zaključke. S tem bom tudi poudaril potrebnost višjega standarda transparentnosti pri raziskovalni umetni inteligenci oz. vsaj potrebo po večjem pretresanju robustnosti tovrstnih rezultatov.

2 Raziskovalna umetna inteligenca

Preden se posvetimo problemom glede transparentnosti raziskovalne umetne inteligence, je smiselno pogledati, kaj raziskovalna umetna inteligenca sploh je in kako se uporablja. Najprej omenimo, da se raziskovalno umetno inteligenco pogosto precej sinonimno omenja kot umetno inteligenco, strojno učenje in globoke nevronske mreže. Pri tem sicer ne gre za popolne sinonime, a vse tri trenutno običajno temeljijo na obdelavi velike količine podatkov. Obstajajo sicer tudi določeni hibridi, kot npr. globoko učenje, ki označuje kombinacijo strojnega učenja in globokih nevronskih mrež.

Prav to osredotočanje na obravnavo velike količine podatkov in iskanje povezav in vzorcev v njih je za človeško razumevanje problematično. Kot trdita Pearl in Mackenzie (2018), nas ne zanima zgolj, *kako* pride do določenih spoznanj, temveč tudi oz. predvsem *zakaj* pride do njih. Torej nam pri razumevanju ne gre zgolj za iskanje vzorcev, temveč si želimo vpogleda v vzročno-posledične odnose, ki so za človeško razumevanje pojavov ključni.

Kljub temu pa metode umetne inteligence, sploh v vzponu je strojno učenje, že igrajo pomembno vlogo tako rekoč na vseh znanstvenih področjih. Npr. v biologiji in kemiji lahko spremljamo že povsem avtonomne eksperimente (King et al. 2009; Häse, Roch in Aspuru-Guzik 2019), svojo vlogo pa tovrstne metode igrajo tudi v digitalni humanistiki in jezikoslovju (npr. pri preučevanju vejic v slovenščini; Holozan 2013) in morda presenetljivo tudi v filozofiji (npr. agentska optimizacija v simulacijah; Douven 2020), če omenimo le nekaj primerov.

Celostnega pregleda na tem mestu ne moremo izvesti, saj je umetna inteligenca v raziskovanju zelo razširjena, smiselno pa si je nekoliko pobliže pogledati vsaj en primer. Melnikov et al. (2018) opisujejo strojno učenje v okviru fizike v kvantnem laboratoriju. V kvantnih eksperimentih imamo namreč težave z različnimi razredi prepletenosti (angl. *entanglement classes*). Melnikov in kolegi so zato uporabili t. i. projektivni simulacijski sistem, ki je zasnoval kompleksne fotonične kvantne eksperimente, ki so nato proizvedli visokodimenzionalna multifotonska stanja prepletenosti. Sistem, ki temelji na umetni inteligenci, se je torej naučil proizvesti raznolika stanja prepletenosti in izboljšal učinkovitost njihove realizacije. V tem procesu je avtonomno (ponovno) odkril eksperimentalne tehnike, ki sicer postajajo standard tudi v eksperimentih moderne kvantne optike. Pri tem je zanimivo, da tega raziskovalci niso eksplicitno zahtevali ali sistema naučili, temveč je umetna inteligenca skozi proces učenja to odkrila sama. Na podlagi tega lahko sklenemo, da lahko pričakujemo, da bo raziskovalna umetna inteligenca v prihodnosti igrala pomembno ustvarjalno vlogo, saj so se tovrstni sistemi očitno zmožni sami učiti in odkrivati raziskovalne tehnike ter na podlagi obstoječih eksperimentov oblikovati smiselne hipoteze, ki so vredne eksperimentalnega preverjanja.

Pa vendar: v kolikor tovrstne tehnike razumemo (npr. ker je sistem odkril že znane tehnike), gre seveda za primere, kjer lahko potrdimo, da je dosežek umetnega sistema izjemen. Kaj pa, v kolikor bi sistem odkril tehnike, ki nam še niso znane? Podobno se lahko vprašamo, kaj bi storili v primeru, da bi nam v preučevanje ponudil

hipoteze, ki se zdijo z našega stališča nerazumne in nevredne preučevanja. Kako jih lahko vzamemo za smiselne? Konec koncev je znan problem umetne inteligence tudi to, da potencira pristranskosti iz podatkov, kar vodi do raznolikih etičnih zagat v povezavi s tako imenovano algoritmično poštenostjo (npr. kako se znebiti rasističnih posledic algoritmov, ko umetna inteligenca obdeluje podatke o ljudeh; glej npr. Kleinberg, Ludwig, Mullainathan in Ramachan 2018).

Tako kot v primeru izogibanja algoritmične (ne)poštenosti in pristranskosti ob uporabi umetne inteligence v družbenih odločitvah je torej tudi pri uporabi umetne inteligence v znanosti ključno, da imamo dovolj visoko raven razumevanja, ki pojasni, kaj je vodilo do kakšne odločitve. Raven transparentnosti mora biti visoka prav zaradi izogibanja pristranskosti in zaradi dostopa do tega, kako poteka procesiranje v ozadju, pa tudi višja kot pri razlagi človeških odločitev. Posvetimo se torej še vprašanju glede standardov transparentnosti.

3 Standardi transparentnosti raziskovalne umetne inteligence

Umetna inteligenca je vse bolj prisotna na vseh področjih in kot smo lahko videli tudi v raziskovalne namene. Andras et al. (2018) zaradi vedno večje prisotnosti izpostavijo pomen človeškega zaupanja v umetno inteligenco in izpostavijo, da je pri tem pomembno, da so umetna inteligenca oz. njeni zaledni procesi razložljivi, kar je pogosto tudi sinonim za transparentnost oz. prosojnost umetne inteligence. Jasno se zdi, da so potrebe po transparentnosti umetne inteligence večje, ko govorimo o pomembnih vprašanjih s praktičnimi posledicami, manj pa, ko govorimo npr. o procesih, ki usmerjajo umetno inteligenco do usvojitve in uspešnega igranja igre *go* (Silver et al. 2017).

Trenutno najboljši algoritmi običajno temeljijo na metodah strojnega učenja. Ti algoritmi so zasnovani na tak način, da sami prepoznavajo skrite vzorce v podatkih in zasnujejo natančna predvidevanja o podatkih v neki domeni, ki še niso bili odkriti. Kot opozarjata npr. Günther in Kasirzadeh (2021), ta natančna predvidevanja za sabo prinesejo izjemno kompleksnost algoritmov, ki so posledično spoznavno precej nedosegljivi (netransparentni) tudi za same snovalce teh algoritmov. Tako nam umanjka razumevanje razlogov, ki vodijo do rezultatov umetne inteligence. Ali lahko na podlagi tega zahtevamo višje standarde transparentnosti od umetne inteligence kot od ljudi?

Vprašanje ima tudi praktične posledice, saj s tem namreč na nek način zaviramo razvoj umetne inteligence. Ko zahtevamo, da snovalci algoritmov poskrbijo, da je procesiranje razložljivo oz. transparentno in vsaj v principu razložljivo, hkrati izločimo oz. bistveno otežimo razvoj takih algoritmov, ki bi bili sicer skoraj povsem v svojevrstni črni škatli (npr. v primeru globokih nevronskih mrež).

Pa smo sploh upravičeni do zahteve po višjih standardih transparentnosti v primeru umetne inteligence? Zerilli et al. (2019) argumentirajo, da ne. Trdijo, da od umetne inteligence ne moremo zahtevati več kot od ljudi – standardi transparentnosti morajo biti v obeh primerih enaki, in sicer lahko v obeh primerih zahtevamo zgolj razlage na intencionalni ravni (Dennett 1987), v okviru katere ljudje izrazimo svoje praktične razloge za določeno odločitev, umetna inteligenca pa za svoje 'odločitve'. 'Odločitve' sicer navajam v navednicah, saj gre za mentalistično izrazoslovje, v praksi pa bi to izgledalo tako, da npr. pojasnimo, zakaj je raziskovalna umetna inteligenca predlagala določeno raziskovalno hipotezo na podlagi tega, da je prepoznala določene vzorce v preteklih eksperimentalnih podatkih in iz tega izračunala, da je velika verjetnost za nova odkritja v preverjanju predlagane nove hipoteze.

Kar želim poudariti, je to, da v tem primeru umanjka razlaga, kako točno je raziskovalna umetna inteligenca prišla do tega sklepa. Umanjka nam torej razlaga, na podlagi katerih algoritmov oz. zakaj natančno je prišlo do formulacije te hipoteze. Če si izposodim Marrov (1982) besednjak – umanjka nam razlaga na nivoju algoritmične oz. implementacijske analize, čeprav imamo morda dostop do funkcionalne ravni. Zerilli et al. (2019) trdijo, da to ni težava, saj tega ne pričakujemo niti pri človeškem odločanju.

Na drugi strani imamo stališče, da moramo od umetne inteligence (in s tem tudi od raziskovalne umetne inteligence) zahtevati več. Tak primer sta že omenjena Günther in Kasirzadeh (2021), ki svojo argumentacijo utemeljita na dveh ključnih točkah. Po eni strani se zdi, da vsaj pri določenih algoritmičnih odločitvah ključno vlogo igra oblika (oz. dizajn). Kot primer podata letalsko nesrečo letala Boeing 737 Max 8, ki je leta 2017 strmoglavilo in povzročilo smrt 189 ljudi, ki so bili na krovu. Do težave je prišlo, ker je računalniški sistem za podporo pri letu v preveliki meri temeljil na enem specifičnem senzorju, ki se je okvaril. Šlo je torej za napako na ravni oblike oz. dizajna umetne inteligence – prekomerno je temeljila na podatkih iz enega senzorja. Na intencionalni ravni do težave ni prišlo – sistem je 'verjel', da ravna pravilno, a ob napačni predpostavki, da so bili vhodni podatki iz senzorja zanesljivi. Za zanašanje

na umetno inteligenco je, tako Günther in Kasirzadeh (2021), ključno razumevanje ne zgolj intencionalne razlage delovanja (kot jo pričakujemo od ljudi), temveč tudi algoritmične in oblikovne, torej na čem temeljijo 'odločitve' (česar od ljudi ne pričakujemo, saj imamo omejen vpogled v možgane). V kolikor bi bilo to možno, bi enake standarde transparentnosti tako ali tako vzpostavili tudi za ljudi. Če nekdo nekaj stori zaradi diagnosticirane možganske lezije, pri razlagi vedenja prav tako ne sledimo (zgolj) razlagi prepričanj, ki so osebo vodila, temveč upoštevamo tudi nevrološko specifiko osebe.

Čeprav se zdi, da so tovrstna vprašanja za raziskovalno umetno inteligenco morda irelevantna, temu ni tako. Vzemimo npr. primer strojnega učenja v okviru medicine. V kolikor poznamo obliko algoritmov (oz. še natančneje, v kolikor je ne poznamo), lahko algoritme zlorabimo za napačno diagnosticiranje (oz. spregledamo, da je bil algoritem zaveden zaradi ljudem načeloma nerazumljivih razlogov). Finlayson et al. (2019) tako izpostavljajo, da bi se lahko tovrstne algoritme v okviru medicinskega diagnosticiranja zlorabilo kot orožje. Zloraba kot del napada je vsekakor realna možnost tudi pri raziskovalni umetni inteligenci širše, sploh v kolikor govorimo o aplikativnih raziskavah. Zamislimo si lahko, da bi lahko napadalec npr. dodal očesu nevidne piksle v vizualne podatke, ki jih obdeluje umetna inteligenca, ter tako zavrl konkurenco, npr. v farmakološkem razvoju.

Pustimo špekulacije ob strani, pri obvladovanju rezultatov umetne inteligence je poznavanje oblikovne ravni algoritmov ključno vsaj zaradi tega, da razumemo, ali so rezultati zanesljivi ali ne. Tudi če izključimo možnost zunanjih napadov, se lahko primer pretiranega zanašanja na eno samo orodje (senzor) skoraj direktno preslika iz strmoglavljenega letala na znanstveno raziskavo, ki pretirano temelji na okvarjenem ali zgolj neprimernem senzorju kot viru podatkov za strojno učenje. Čeprav so v tem primeru posledice manjše, so v primeru povsem novih raziskav tudi težje za prepoznanje (in lahko vodijo v slepo raziskovalno ulico).

Drugi primer, kjer se zdi, da moramo zahtevati drugačne standarde od umetne inteligence kot od ljudi, je, tako Günther in Kasirzadeh (2021), ko govorimo o popolnih črnih škatlah, pri katerih nimamo vpogleda v procesiranje (Creel 2020). V tem primeru torej ne moremo govoriti niti o intencionalni razlagi, saj nam niti ta ni dostopna. Razumevanje, zakaj se je sistem umetne inteligence odločil, kakor se je, povsem umanjka. V takem primeru, tako se vsaj zdi, brez obširnega testiranja robustnosti rezultatov sploh ne moremo govoriti o spoznavni zanesljivosti. Pri tem

velja izpostaviti tudi povezavo med analizo robustnosti in razlagalnim mišljenjem, kot jo je pred kratkim v nekoliko drugačnem kontekstu vzpostavil Schupbach (2018).

Prav zato, ker od znanosti torej zahtevamo visoke spoznavne standarde, da lahko govorimo o zanesljivem spoznanju oz. znanju, se torej zdi smiselno, da od raziskovalne umetne inteligence zahtevamo višjo transparentnost, kot jo zahtevamo pri razlagi človeškega znanstvenega raziskovanja. Raziskovalna umetna inteligenca se torej zdi primer, kjer je debata o dvojnih standardih transparentnosti še bolj enostavna kot pri umetni inteligenci na splošno – v znanosti so zahteve po razumljivosti enostavno integralne in posledično višje.

4 Študija primera

To nas privede do študije primera iz računalniške filozofije oz. filozofije, ki pri svojem raziskovanju temelji na računalniških metodah. Na podlagi tega primera bomo namreč lahko videli, zakaj je pomembno, da imamo vpogled ne le v to, kaj sledi na podlagi algoritmov, temveč da pri raziskovalni umetni inteligenci potrebujemo tudi razumevanje pogosto precej kompleksnih zalednih procesov. Pri tem se bomo kritično naslonili na nedavno objavljen članek o t. i. ekološki racionalnosti razlagalnega sklepanja (Douven 2020). Pri ekološki racionalnosti gre za oznako, da je pri presoji upravičenosti določenega tipa sklepanja treba upoštevati ne toliko, ali je nek način sklepanja smiseln kot tak, temveč ali je smiseln v nekem določenem okolju. Kot v primeru škarij je pri razumevanju mišljenja pomembno upoštevati dve stvari – mišljenje na eni strani (kot eno rezilo) in strukturo okolja na drugi (kot drugo rezilo metaforičnih Simonovih škarij; glej npr. Simon 1956).

Douven (2020) v svojem prispevku na podlagi t. i. agentske optimizacije (angl. *agent-based optimization*) oz. posebnega tipa računalniške simulacije pokaže, da je razlagalno mišljenje ekološko racionalno, saj lahko v določenem kontekstu privede do optimalnega izkupička. Pri tem avtor razlagalno mišljenje razume v verjetnostnem smislu kot adaptacijo bayesovskega učenja, ki posebej preferira najboljše razlage, ekološkost sklepanja pa skuša pokazati v simulaciji zdravniške diagnostike na primeru simulirane enote intenzivne nege.

Pri zasnovi izhaja iz psiholoških eksperimentov, ki kažejo, da se v opisnem smislu ljudje poslužujemo tega, da preferiramo najboljšo razlago oz. da najboljšo razlago smatramo kot (bolj verjetno) resnično od slabših razlag. Nato na podlagi skladnosti s psihološkimi raziskavami, predvsem pa na podlagi podatkov iz računalniških simulacij v agentski optimizaciji, skuša pokazati, da so standardni argumenti v podporo racionalnosti in normativni prednosti bayesovskega učenja pretirani. Njegova agentska optimizacija (računalniška simulacija in analiza simuliranih podatkov) namreč vodi do sklepa, da v primeru časovnih pritiskov, ki nastanejo npr. pri diagnostiki kritično poškodovanih bolnikov in bolnic, bayesovsko učenje evolucijsko odpade na račun bolj adaptivnih razlagalnih oz. sorodnih nebayesovskih principov sklepanja (Goodovo in Popprovo učenje). Pa je temu res tako oz. ali se pri argumentaciji res lahko tako enostavno zanesemo na rezultate razmeroma kompleksne računalniške simulacije?

Kot se izkaže, ko njegovo simulacijsko študijo nekoliko prilagodimo, rezultati niso pretirano robustni, saj lahko po manjši adaptaciji scenarija oz. zaledne kode pridemo do povsem neskladnih sklepov. Avtorjeva študija torej, tako bom vsaj trdil, zgolj slučajno podpira argumenta, da je razlagalno sklepanje ekološko racionalno in da bayesovskega učenja vsaj v kontekstu časovnih pritiskov ni, ker ni dovolj adaptivno.

Predpostavimo, da v simulacijo namesto povsem zanesljivih diagnostičnih testov (kot jih v svojih simulacijah uporablja Douven 2020) uvedemo teste, ki so varljivi. Ta sprememba je smiselna, saj se tudi sicer ne moremo povsem zanašati na svojo zaznavo oz. zunanje vire informacij, ki jih dobimo preko instrumentov ali pa na podlagi pričevanj. Rezultati v tem primeru vodijo v čudno situacijo, kjer različna razumevanja tega, kaj pomeni nezanesljivost testov, vodijo do rezultatov, ki jih lahko sprejmemo kot podporo v prid različnim načinom sklepanja. Pomembno pri tem je, da določena nezanesljivost testov pokaže celo, da bayesovsko sklepanje, torej način sklepanja, ki ima manj parametrov in bi zato moral biti manj adaptiven (če bi trditve iz osnovnega prispevka držale), lahko celo vodi do agentske optimizacije in evolucijske prevlade. Iz tega lahko (nasprotno kot sva s soavtorico najprej sklepala; glej Trpin in Plementaš 2021) razberemo, da je pri interpretaciji simulacij osnovnega članka (Douven 2020) prišlo do zmote, ker je avtor navidezno prepričljive rezultate interpretiral kot podporo svoji argumentaciji, dejansko pa je šlo za težavo na oblikovni ravni. Algoritem, ki je v ozadju, namreč ne kaže tega, kar bi pričakovali, tega pa ne opazimo, ker se nam rezultat zdi smiseln.

Za boljše razumevanje si velja ta primer pogledati bistveno bolj podrobno. V osnovi gre pri raziskavi za epistemološko dilemo. Obstaja namreč več pravil sklepanja, ki jih lahko uporabimo kot vodilo dobrega mišljenja. Ali je kakšen izmed (verjetnostnih) načinov sklepanja boljši od drugih?

Pri tem se moramo seveda najprej vprašati, kaj sploh je merilo dobrega mišljenja. Dve merili, za kateri se zdi, da ju velja upoštevati, sta sledeči: (i) resničnost prepričanj in (ii) hitrost snovanja prepričanj. Prvo merilo lahko upoštevamo tako, da pogledamo, v kolikšni meri določeno pravilo sklepanja vodi do resničnih prepričanj – več in močnejša prepričanja o resničnih propozicijah, kot jih imamo na podlagi določenega pravila sklepanja, bolje za pravilo sklepanja kot tako. Drugo merilo pa nam pomaga, ko ocenjujemo, kako hitro lahko zmanjšamo svojo negotovost – tem hitreje, tem bolje za pravilo.

V idealnih razmerah bi obe merili izpolnjevali hkrati, torej bi imeli pravilo sklepanja, ki hitro vodi do resničnih prepričanj. Rezultati iz literature kažejo, da to ni tako: pravila, ki so dobra po prvem merilu, so običajno slabša po drugem (ter obratno; Douven 2013; Trpin in Pellert 2019). Osnovno vprašanje pri filozofski oceni načinov (verjetnostnega) sklepanja je vezano na to, katero pravilo je najboljše v podpori merila (i) in merila (ii) ter kako lahko ta dva zaželena cilja (hitrost in natančnost sklepanja oz. mišljenja) spravimo v ravnovesje, sploh če upoštevamo, da viri informacij niso nujno povsem zanesljivi ali so nemara celo zavajajoči.

Če torej upoštevamo, da so podatki lahko nezanesljivi oz. nemara celo zavajajoči, moramo zasnovo računalniških simulacij, na katerih temelji Douvenov (2020) argument, spremeniti, saj sam te opcije ne upošteva. Pri tem sva se v nedavni analizi (Trpin in Plementaš 2021) oprla na analizo učenja iz t. i. delnih laži in različnih stopenj zaupanja (glej Trpin, Dobrosovestnova in Götzendorfer 2020), saj nam ta pristop prinaša formalizacijo zavajajočih in nezanesljivih virov informacij.

Prav tovrstni pomisleki – da bi lahko bil vir informacij varljiv podobno, kot so varljive delne laži – so vodili do adaptacije simulacij, ki jih je izvedel Douven (2020). Njegovo raziskavo lahko razdelimo v dva dela: najprej pokaže, da se pravila sklepanja (specifično: bayesovsko, razlagalno, Goodovo in Popprovo pravilo sklepanja) razhajajo glede omenjenih meril dobrega mišljenja (natančnost in hitrost). V drugem delu nato predlaga način, kako lahko ti dve merili uravnovesimo in preko selekcijske

optimizacije ugotovimo, katero je najboljše v določenem okolju (torej v smislu ekološke racionalnosti verjetnostnih pravil sklepanja).

Pri tem si je zamislil simulirano enoto intenzivne nege, v kateri zdravniki oz. zdravnice rešujejo paciente. Pri tem imajo tri možnosti: bodisi naredijo pravilno ali napačno odločitev glede posega, oz. če niso prepričani, kaj je narobe s pacientom oz. pacientko, ne naredijo nobenega posega. Pri tem se skozi čas spreminja verjetnost preživetja pacienta oz. pacientke, ki je odvisna tudi od posega. Kasneje kot pride do posega, manjša je verjetnost preživetja. Podobno pravilen poseg zviša verjetnost preživetja, napačen pa jo zniža. Če posega ni, je verjetnost preživetja vmes med pravilnim in napačnim posegom v določenem trenutku.

Douven preko simulacije, ki temelji na tem principu, pokaže, da je verjetnostno sklepanje na najboljšo razlago boljše pravilo sklepanja kot bayesovsko sklepanje. Čeprav gre za bolj tvegano pravilo (večja nevarnost zmote), je ravno zato tudi hitrejše in v primeru intenzivne enote prevlada v bitki med časom in natančnostjo. Drugače rečeno, čeprav glede natančnosti ni optimalno pravilo, je dovolj natančno, da zaradi hitrosti prevlada. Specifično v svojih simulacijah si je zamislil 200 zdravnikov oz. zdravnic. Od tega jih na začetku 50 sklepa na bayesovski, 50 na razlagalni, 50 na Popprov in 50 na Goodov način sklepanja. Vsak od njih ima 100 simuliranih pacientov in pri vsakem izvaja diagnostične teste, da ugotovi njihovo bolezen (100 testov). Ko na podlagi testov zasnuje prepričanje o bolezni, se odloči za poseg, ki je lahko pravilen ali nepravilen (oz. ga sploh ni, če bolezni ne more dovolj zanesljivo diagnosticirati). Na koncu preverimo, kolikšna je bila verjetnost preživetja pri posameznem zdravniku oz. zdravnici, in najboljših 100 od 200 podvojimo, ostale pa izbrišemo iz simulirane populacije. Ta postopek nato ponavljamo za 100 generacij zdravnikov oz. zdravnic in na tak način izvedemo t. i. agentsko optimizacijo.

Pri tem velja opozoriti, da simulacije temeljijo na tem, da so diagnostični testi povsem zanesljivi, čeprav v praksi temu ne bi bilo nujno tako. Zato se zdi smiselno, da kombiniramo uvide iz raziskav delnega laganja in delnega zaupanja ter jih uporabimo v analizi agentske optimizacije, ki poteka enako kot opisana s simuliranimi zdravniki oz. zdravnicami in pacienti oz. pacientkami. Razlika je torej ta, da so testi v spremenjenih simulacijah lahko nezanesljivi oz. delno zanesljivi in da zdravniki oz. zdravnice opazujejo, ali so testi sovpadali z opazovanimi simptomi ter na tak način kalibrirajo tudi svoje zaupanje v teste. Tak postopek vodi v

nepričakovane rezultate, ki na prvi pogled kažejo na to, da so različna pravila sklepanja bolj primerna za različna okolja (tako npr. Trpin in Plementaš 2021), dejansko pa na oblikovno pomanjkljivost opisanih simulacij, ki zaradi kompleksnosti dejansko ne prikažejo tega, kar naj bi – tj. premoči določenih pravil sklepanja.

Nasprotno, kot bi morda pričakovali, se po tej adaptaciji izkaže, da verjetnostno sklepanje na najboljšo razlago ni nujno najboljši način oz. da v agentski optimizaciji v določenih razmerah izpade. Poglejmo, denimo, rezultate simulacij, v katerih imamo teste, ki so stalno zavajajoči na način, ki je analogen enostavnemu laganju (test pokaže odsotnost nekega simptoma, če je bolezen taka, da je simptom bolj verjetno prisoten kot ne). V tem primeru testi za pacienta, ki ima bolj verjetno nek simptom kot ne (ker ima takšno simulirano bolezen), test vedno pokaže, da simptom ni prisoten. Zdravniki oz. zdravnice, ki po kalibraciji zaupanja in seznanitvi s takimi testi sklepajo na najboljšo razlago simptomov, skozi proces agentske optimizacije prevladajo. Do tukaj so torej rezultati skladni s tem, kar je v osnovni različici trdil Douven (2020).

A zgodba se s tem ne zaključi: če testi zavajajo na način, ki je podoben rokohitrskemu laganju (najprej nekdo testira, ali je simptom prisoten, in potem na testu pokaže nasprotno od dejanskega stanja), potem v teku optimizacije skozi generacije vseeno prevladajo razlagalni zdravniki oz. zdravnice, a ne s popolno prevlado – velik delež populacije optimalnih agentov namreč vsebuje tudi te, ki sklepajo na t. i. Goodov način sklepanja.

Najbolj zanimivi rezultati pa se pojavijo, če so testi zavajajoči na jasnovidni način: v kolikor je pri pacientu prisoten simptom, test pokaže, da simptoma ni. Po verifikaciji testa simulirani zdravniki oz. zdravnice nato kalibrirajo zaupanje v zanesljivost testa. Izkaže se, da v tem primeru skozi optimizacijo prevladajo bayesovski agenti, čeprav je bayesovsko pravilo sklepanja najmanj adaptivno (ima en parameter manj kot ostala tri simulirana pravila sklepanja (za več podrobnosti glej Trpin in Plementaš 2021).

Na podlagi teh rezultatov bi torej lahko sklepali, da nam simulacije predlagajo pluralističen pristop k pravilom sklepanja. V kolikor imamo opravka z nezanesljivostjo enega tipa, je bolj smiseln en, v kolikor z drugačno nezanesljivostjo testov pa drug način sklepanja. Tak bi bil vsaj sklep, če bi sledili načelom ekološke racionalnosti oz. tega, da je za oceno smiselnosti nekega načina sklepanja treba upoštevati skladnost z okoljem, v katerem se uporablja. Na podlagi tega bi se nato

lahko vprašali, kako prepoznati značilnosti okolja (oz. nezanesljivosti testov), da bi uporabili smiselno strategijo sklepanja oz. kdaj je smiselno preklapljati različne strategije.

A tak sklep bi bil preuranjen – kar nam pokažejo ti rezultati, je namreč ravno nasprotno: težava je v zasnovi samih računalniških simulacij. V osnovni izvedbi so namreč rezultati pokazali, da so v kontekstu, kjer ocenjujemo tudi hitrost sklepanja, bayesovsko sklepanje vedno izpodrinili drugi principi. To se je avtorju (Douven 2020) zdelo razumljivo, saj imajo ti drugi principi več parametrov, oziroma ti drugi principi lahko upoštevajo več vidikov situacije in so se zato zmožni bolje adaptirati. A kot pokažejo rezultati zgoraj omenjene adaptacije, temu ni nujno tako – v kolikor uporabimo nezanesljivost virov analogno z jasnovidnim laganjem, prevlada v situaciji bayesovsko sklepanje, čeprav je manj adaptivno.

Težava je torej v tem, da omenjena računalniška simulacija ne meri tega oz. ne prinaša podpore v prid argumentaciji, ki jo pričakujemo, ker pretirano temelji na tem, kako je zasnovana zanesljivost simuliranih testov. Prav v tem pa se pokaže tudi potreba po večjem standardu transparentnosti: avtor (Douven 2020) je namreč javno delil programsko kodo, ki je v ozadju omenjene študije. V kodi sicer ni težav, nam pa omogoča testiranje različnih variant (po adaptaciji) in posledično pretresanje, ki pokaže, ali so rezultati zanesljivi ali ne. Izkaže se torej, da koda na intencionalni ravni dela, kar naj bi (ni programskih težav), ne dela pa tega, kar njen avtor misli, da dela, saj je zasnovana na tak način, da rezultati temeljijo na vkodiranih domnevah glede zanesljivosti vira informacij. Tega pri običajnih znanstvenih raziskavah, kjer v igri ni metod umetne inteligence oz. vsaj simulacijskih metod v širšem smislu, ne zahtevamo oz. pričakujemo. Prav omenjeni rezultati pa nam kažejo, da to upravičeno zahtevamo od tako kompleksnih pristopov, kot je omenjena optimizacijska študija.

5 Zaključek

Čeprav smo se v pričujočem prispevku bolj poglobljeno posvetili enemu primeru, v katerem se pokaže, da je transparentnost v primeru raziskovalne umetne inteligence pomembna in da upravičeno pričakujemo tudi vpogled v to, kako so algoritmi v njenem ozadju zasnovani, velja izpostaviti, da je tovrstnih kritik več in da pri tem nismo prvi. Zgolj v filozofiji imamo npr. več primerov, kjer je pretresanje algoritmov v ozadju nekega argumenta privedlo do odmevnih kritik. Tak primer je npr. diskusija glede t. i. principa, da raznolikost prevlada nad zmožnostjo (skupine raznolikih ljudi

naj bi bile kognitivno bolj uspešne kot skupine bolj zmožnih, a manj raznolikih ljudi; glej Hong in Page 2004 za izvorni argument, ter npr. Thompson 2014 za eno od kritik) oz. glede epistemske delitve dela (glej Weisberg in Muldoon 2009 za izvorni argument ter Thoma 2015 za kritiko domnev iz njunega računalniškega modela). Podobno velja tudi onstran filozofije v znanosti širše (pri čemer lahko tudi že omenjeno diskusijo Honga in Pagea 2004, ter kritike njunega prispevka štejemo kot plod družbene vede in ne (samo) filozofije).

Problem torej, kot smo lahko videli, ni v tem, da raziskovalna umetna inteligenca vodi do novih spoznanj, ki jih sicer morda ne bi sami odkrili – to lahko kvečjemu pozdravimo. Zahtevati pa moramo visoko raven transparentnosti in vpogled v to, kako so zasnovani algoritmi, ki so v njenem zaledju. Le na tak način namreč lahko uvidimo, do kakšne mere so rezultati dejansko zanesljivi, pa čeprav imamo pri rezultatih, ki so nastali z 'naravno' inteligenco, nižje zahteve, ker vpogleda v drobovje pač ne moremo zahtevati. To do neke mere tudi pojasni, zakaj smo v znanosti tako rigidni glede zahteve po opisu metodologije. Zahtevamo namreč, da so rezultati zanesljivi in ponovljivi oz. v parafrazi ponarodele pesmi: za znanost je dobro le najboljše.

Viri in literatura

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. T., Urquhart, N. in Wells, S. (2018). »Trusting intelligent machines: Deepening trust within socio-technical systems«. *IEEE Technology and Society Magazine*, 37(4), str.76–83.

Creel, K. A. (2020). »Transparency in complex computational systems«. *Philosophy of Science*, 87(4), str. 568–589.

Dennett, D. C. (1987). *The Intentional Stance*. Massachusetts: MIT Press.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L. in Kohane, I. S. (2019). »Adversarial attacks on medical machine learning«. *Science*, 363(6433), str. 1287–1289.

Grim, P. in Singer, D. (2020). »Computational philosophy«. V: Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (pomlad 2020). URL = https://plato.stanford.edu/archives/spr2020/entries/computational-philosophy/.

Hong, L. in Page, S. E. (2004). »Groups of diverse problem solvers can outperform groups of high-ability problem solvers«. *Proceedings of the National Academy of Sciences*, 101(46), 16385–16389.

Douven, I. (2013). »Inference to the Best Explanation, Dutch Books, and Inaccuracy Minimisation«. *Philosophical Quarterly*, 63, str. 428–444.

Douven, I. (2020). »The ecological rationality of explanatory reasoning«. *Studies in History and Philosophy of Science Part A*, str. 1–14.

Günther, M. in Kasirzadeh, A. (2021). »Algorithmic and human decision making: for a double standard of transparency«. *AI & SOCIETY*, str. 1–7.

Häse, F., Roch, L. M. in Aspuru-Guzik, A. (2019). »Next-generation experimentation with self-driving laboratories«. *Trends in Chemistry*, 1(3), str. 282–291.

Holozan, P. (2013) »Uporaba strojnega učenja za postavljanje vejic v slovenščini«. *Uporabna informatika*, 21(4), str. 196–209.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., Sparkes, A., Whelan, K. E. in Clare, A. (2009). »The automation of science«. *Science*, 324(5923), str. 85–89.

Kleinberg, J., Ludwig, J., Mullainathan, S. in Rambachan, A. (2018). »Algorithmic Fairness«. *AEA Papers and Proceedings*, (108), str. 22-27.

Lipton, P. (2009). »Understanding without explanation«. V de Regt, H. W, Leonelli, S. in Eigner, K. (urd.), *Scientific Understanding: Philosophical Perspectives*. Pittsburgh: University of Pittsburgh Press, str. 43–63.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Melnikov, A. A., Nautrup, H. P., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A. in Briegel, H. J. (2018). »Active learning machine learns to create new quantum experiments«. *Proceedings of the National Academy of Sciences*, 115(6), str. 1221–1226.

Nguyen, A., Yosinski, J. in Clune, J. (2015). »Deep neural networks are easily fooled: High confidence predictions for unrecognizable images«. V *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 427–436.

Pearl, J. in Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Radovic, A., Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A. A., Terao, K., in Wongjirad, T. (2018). »Machine learning at the energy and intensity frontiers of particle physics«. *Nature*, 560(7716), str. 41–48.

Schupbach, J. N. (2018). »Robustness analysis as explanatory reasoning«. *The British Journal for the Philosophy of Science*, 69(1), str. 275–300.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel T. in Hassabis, D. (2017). »Mastering the game of Go without human knowledge«. *Nature*, 550(7676), str. 354–359.

Simon, H. A. (1956). »Rational choice and the structure of the environment«. *Psychological Review*, 63(2), 129–138.

Thoma, J. (2015). »The epistemic division of labor revisited«. *Philosophy of Science*, *82*(3), str. 454–472.

Thompson, A. (2014). »Does Diversity Trump Ability?« *Notices of the AMS*, 61(9), str. 1024–1030.

Trpin, B. in Pellert, M. (2019). »Inference to the best explanation in uncertain evidential situations«. *The British Journal for the Philosophy of Science*, 70(4), str. 977–1001.

Trpin, B., Dobrosovestnova, A. in Götzendorfer, S. J. (2020). »Lying, more or less: a computer simulation study of graded lies and trust dynamics«. *Synthese*, 199, 991–1018.

Trpin, B. in Plementaš, A. M. (2021). »The ecological rationality of probabilistic learning rules in unreliable circumstances«. V Strle, T., Trpin, B., Rebernik, M. in Markič, O. (urd), *Zbornik 24. mednarodne multikonference Informacijska družba: Zvezek B - Kognitivna znanost*, str. 51–55.

Vigen, T. (2015). *Spurious Correlations*. Hachette Books, New York in Boston.

Weisberg, M. in Muldoon, R. (2009). »Epistemic Landscapes and the Division of Cognitive Labor«. *Philosophy of Science*, 76(2), str. 225–52.

Zerilli, J., Knott, A., Maclaurin, J. in Gavaghan, C. (2019). »Transparency in algorithmic and human decision-making: Is there a double standard?« *Philosophy & Technology*, 32(4), str. 661–683.