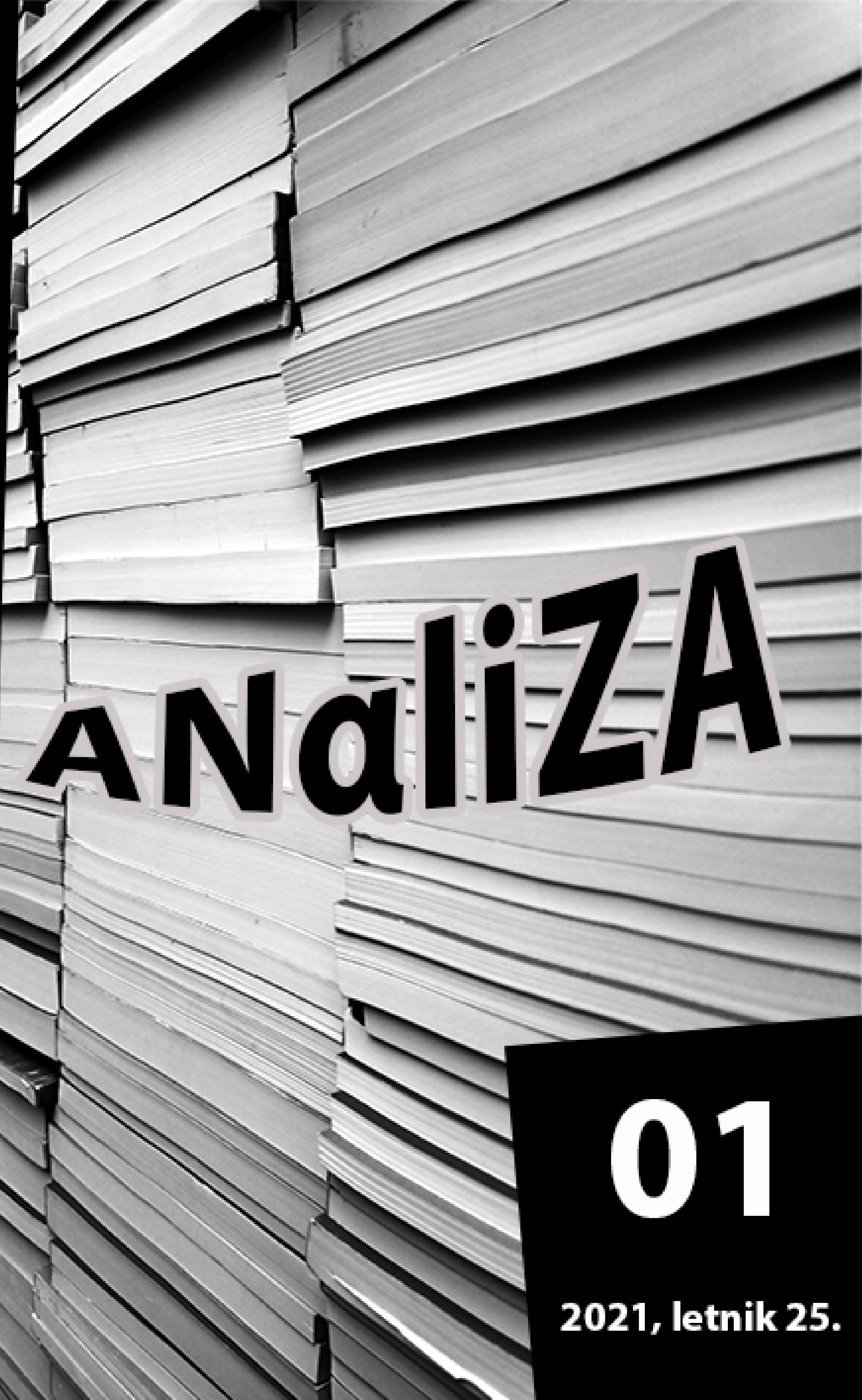
Časopis za kritično misel



ANaliZA

01

2021, letnik 25.

ANaliZA

Društvo za analitično filozofijo in filozofijo znanosti

Kazalo

Miselni eksperimenti v preteklosti in prihodnosti

Človek, resnica in moralna odgovornost

Miselni eksperimenti v preteklosti in prihodnosti

Danilo Šuster

Univerza v Mariboru

Prediktor, fatalizem in miselni eksperimenti[1]

V članku analiziram kratko zgodbo *Kaj se pričakuje od nas* (Chiang, 2005) kot primer filozofskega miselnega eksperimenta, ki domnevno dokazuje, da ni svobodne volje. Pri tem uporabim členitev po stopnjah, ki jo je za miselne eksperimente predlagal Miščević (2012: 2017): najprej scenarij in hipoteza, potem model in razumevanje, vživetje in izpeljava posledic, spontani intuitivni odgovor, uvedba variacij. Miselni eksperiment nas vodi do tega, da v domišljiji 'vidimo' in razločimo pojmovne komponente tam, kjer jih običajno ne. *Prediktor* uvede ločitev smeri časa in smeri vzročnosti. Ta obrat smeri vodi do kolizije našega doživljanja časa in sebe kot vzročnega dejavnika v času. Dejanski red našega delovanja je vzročni, izkustveni red pa je časovni. Pa vendar miselni eksperiment ne uspe dokazati, da naše izbire niso svobodne, saj temelji na zmotni logiki argumentov za fatalizem.

Ključne besede: svobodna volja, fatalizem, vzročnost, čas, miselni eksperiment.

[1] Besedilo je nastalo s podporo raziskovalnega program št. P6-0144, ki ga je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije.

1.

Sodobni severno-ameriški pisatelj kitajskih korenin Ted Chiang ne objavlja veliko, a je prejemnik številnih nagrad na področju znanstvene fantastike. Velja za perfekcionista, ki lahko leta 'medi' eno samo kratko zgodbo. *Zgodba tvojega življenja* (*Story of Your Life*), po kateri je bil posnet znani film *Prihod* (*Arrival*, 2016), je tako 'mukoma' nastajala štiri leta (po preučevanju lingvistike in piljenju stavka za stavkom). Njena tema je sicer jezik, a glavno filozofsko vprašanje, značilno za vrsto Chiangovih novel, je starodavna dilema svobodne volje in časa: če že veš, kaj se bo zgodilo, kako boš potem doživljal svoje bodoče odločitve in izbire? Chiang je eden redkih sodobnih avtorjev, ki poudarja, da znanstvena fantastika vsebuje besedo 'znanost', saj raziskuje meje spoznanja: loteva se miselnih eksperimentov in raziskuje filozofska vprašanja[2]. Prav to je tudi značilnost prevedene kratke zgodbe, spet miselni eksperiment o času in svobodni volji. Objavljena je bila v reviji *Nature* in citirana celo v resni znanstveni literaturi kot primer miselnega eksperimenta o »mašini za napovedovanje odločitev«, ki bi jo bilo morda mogoče realizirati na kak drug način (Haynes, 2011: 172).

Zgodba je v svojem minimalizmu lepa ponazoritev uporabe miselnih eksperimentov, pa tudi zmotne 'logike' fatalizma in naše naravne uročenosti z 'nespremenljivo' preteklostjo, ki tako kot mitološka Meduza razteguje svoje kačaste lovke naprej v prihodnost. Kdor ji pogleda v oči, konča v stanju 'akinetičnega mutizma'. Morda pa to sploh ni končna poanta zgodbe in gre za nekakšnega vsevednega ironičnega Pripovedovalca, ki se pri sebi hahlja in ironizira človeško 'zaslepljenost'? Ampak tu me zanima filozofija, ne pa literarna kritika in navadna proza, ne pa več-značnost literarnega besedila. Naj takoj razkrijem svoje karte: glavni miselni eksperiment je neuspešen, saj sklep, da zamišljena napravica dokazuje, da nimamo svobode volje, ne sledi.[3] Toda v ozadju enostavnega miselnega eksperimenta se skriva še eno vprašanje: ali naše *izkustvo* vnaprejšnje vednosti nujno vsebuje elemente fatalizma kot neke življenjske drže? Tu je Chiang bolj prepričljiv.

2.

Miselni eksperimenti sodijo v tisti del filozofske orodjarne, ki je najbliže umetniškemu delu. Če je metafora sanjsko delo jezika, kot pravi Davidson, potem so miselni eksperimenti laboratorijske vaje domišljije. Gre za konkretne in nazorne scenarije, ki so hipotetični ('protidejstveni') in zelo premišljeno zgrajeni. Ko se 'odvijejo' v naši domišljiji ali ko v mislih preigramo korake, h katerim

[2] Prim. Clark, 2015.

[3] Podobno zgrešene se mi zdijo Haynesove izpeljave, da rezultati Libetovih eksperimentov ovržejo 'naivno' prepričanje v svobodno voljo, a to je druga zgodba – sicer povezana z idejo Prediktorja, kot je opazil Haynes. O Libetu več na koncu članka.

smo napoteni, nas končni rezultat poskusa vodi do nekega intuitivnega sklepa. Cilji so lahko različni, toda v filozofiji gre običajno za nek jasno določen spoznavni smoter – potrditev ali ovržbo neke teorije ali hipoteze. Pravi eksperiment je pogosto neizvedljiv, razlogi so lahko fizikalni tako kot v naši zgodbi (pošiljanje signala v času *nazaj*), 'banalno' finančni ali pa tehnološki in etični.

Vzemimo morda najbolj znani miselni eksperiment s področja svobode in odgovornosti. Splošno sprejeto načelo je načelo alternativnih možnosti (NAM): »Oseba je moralno odgovorna za to, kar je storila samo, če bi lahko ravnala drugače.« Če ste zapadli v stanje akinetičnega mutizma in nimate več sposobnosti premikanja, ne morete biti odgovorni za to, da se niste udeležili sestanka hišnega sveta. Frankfurt (1969) je predstavil nazoren scenarij, ki naj bi naše intuicije preusmeril do sklepa, da NAM ne drži. V malo posodobljeni različici: oseba A smrtno sovraži osebo B in načrtuje njen umor. Toda tudi oseba C sovraži B in si želi njene smrti, a želi ostati skrita v ozadju. Recimo še, da je C spreten nevrokirurg in manipulator, ki vstavi v možgane osebe A neko *napravico*, ki spremlja A-jevo možgansko aktivnost. Če znaki kažejo, da bo oseba A po svoji presoji izpeljala prvotno namero, napravica miruje. Če pa znaki kažejo, da bo oseba A odstopila od svoje namere, napravica sproži intervencijo (morda stimulira ustrezni del možganov), ki privede do tega, da se A spet odloči za umor. No, izkaže se, da intervencija ni potrebna, saj se A zaradi svojih razlogov odloči in izpelje načrtovani umor. Intuicija o tem konkretnem scenariju nam pravi, da je A ravnal svobodno, po lastni volji in je zato moralno odgovoren, čeprav ni mogel ravnati drugače.

Če to posplošimo – začeli smo s hipotezo (NAM), zamišljeni scenarij pa nas privede do tega, da jo zavrnemo. Hipotetični scenarij je zaradi tehnoloških in etičnih razlogov neizvedljiv. Pogled v 'drobovje' tega scenarija pa nam razkrije še nekaj pomembnih značilnosti miselnih eksperimentov. Nazorni scenarij sproži nekakšno mentalno 'destilacijo', na ognju domišljije se prvotna pojmovna 'zmes' razloči na komponente. Ko odstranimo možnost alternativ, preostali 'kondenzat' še vedno zadošča, da osebi pripišemo moralno odgovornost za njeno dejanje. Navsezadnje gre za *namerno* dejanje – oseba je izvor dejanja, z njim se poistoveti, če bi *njeni* razlogi govorili proti, bi se mu odpovedala. Miselni eksperiment nas vodi do tega, da v domišljiji 'vidimo' in razločimo pojmovne komponente tam, kjer jih običajno ne (prim. tudi Miščević, 2012: 202).

Poskusimo zdaj razčleniti zgodbo po fazah miselnega eksperimenta, kot jih predlaga Miščević (2017). V prvi fazi tvorec miselnega eksperimenta predstavi zamišljeni scenarij in zastavi vprašanje. Spomnimo so: Prediktor je enostavna igralna napravica, sestoji iz gumba in svetila, ki vedno zasveti sekundo *preden* pritisnete na gumb. Pojasnilo je znanstveno-fantastično: v Prediktorju je krogotok z negativnim časovnim zamikom, ki pošilja signal v času nazaj. *Hipoteza,* ki je, morda ironično, predstavljena precej apodoktično, se skriva v tezi, da »Prediktorji *dokazujejo,* da ni take stvari kot je svobodna volja.« Zato bo začetno vprašanje: »Ali Prediktor dokazuje, da svobodna volja ne obstaja?« V naslednji fazi si bralec predoči situacijo in zgradi model. Razumeti mora, zakaj, glede na zgradbo Prediktorja, ni načina, kako bi ga pretentali: »Ne glede na to, kaj naredite, luč vedno zasveti *preden* pritisnete na gumb.« Ko scenarij razumemo, začne delovati naša mentalna 'mašinerija' (vživetje, razmišljanje, izpeljava posledic itd.): »implikacije nespremenljive prihodnosti prodrejo v zavest.« Očitno te implikacije načnejo naše izkustvo svobodne volje, saj ugotovimo, da »naše izbire niso pomembne.« V zadnji fazi potem sledi naš intuitivni odgovor: »Prav res, Prediktor dokazuje, da ni svobodne volje.«

Tudi Chiangov miselni eksperiment temelji na *ločitvi* pojmovnih komponent, za katere se nam zdi samoumevno, da sodijo skupaj. Filozofsko jedro je razmislek o pojmovanju časa, vzročnosti in našega delovanja. Vzročnost ima neko značilno smer – dogodek vzrok ima vzročno prioriteto, vzročnost nekako 'teče' od vzroka (pritisk gumba) do učinka (blisk), vzroki učinkujejo in vodijo do učinkov, nikdar obratno (prim. Tooley, 1987: 179). Naravno se nam zdi, da je smer vzročnosti smer časa: vzročnost je že za Huma relacija, v kateri so vzroki časovno *pred* učinki. Relacija med vzrokom in učinkom je asimetrična, ker je relacija »biti časovno predhoden« asimetrična, čas pa 'teče' samo naprej. Prediktor izzove domnevno apriorni status te povezave: ali ni (vsaj logično) možno 'vnazajšnje' povzročanje, v katerem učinek (blisk) nastopi časovno *pred* svojim vzrokom (pritisk gumba)?

Kar nekaj filozofskih teorij poskuša razložiti usmerjenost vzročne relacije, ne da bi jo takoj pojmovno utemeljili na časovnem zaporedju dogodkov. V starejših teorijah nastopa pojem moči – vzrok poseduje neko moč, ki privede do spremembe v drugi stvari (učinek). Drugi predlagajo bolj antropomorfna merila – nastop vzroka *pojasni*, zakaj nastopi učinek, ne pa obratno (npr. Horwich, 1987). Po teoriji manipulacije ima dogodek *V* vzročni primat pred dogodkom *U*, če lahko manipuliramo z dogodkom *V* in na ta način privedemo ali preprečimo nastop dogodka *U*. Po teh pojasnilih je ujemanje časovnega in vzročnega reda samo kontingentna značilnost dejanskega sveta ali naše pojmovne sheme. In Chiang s pomočjo izmišljene napravice, ki pošilja signal v času nazaj, preobrne in s tem *razloči* obe smeri: vzročni red signala teče od gumba do bliska, časovni red pa od bliska do gumba.

Nekateri so dokazovali, da je to nemogoče, ampak dovolimo si ta miselni izlet. Chianga ne zanima, kako je možen krogotok z negativnim časovnim zamikom, ampak raziskuje vprašanje, kako mi v svojem delovanju *doživljamo* to ločitev. Kaj se zgodi z našim doživljanjem svobodnih izbir, ko sta si smer časa in smer vzročnosti v nasprotju? Pripovedovalec pravi, da »Ne glede na to, kaj naredite, luč vedno zasveti, preden pritisnete na gumb« in zato, domnevno, naše izbire niso pomembne. Pravil igre ne moremo prekršiti, ne moremo pritisniti gumba, ne da bi se pred tem pojavil blisk, točno eno sekundo pred tem. Blisk nas, tako rekoč, 'prisili' v premik prsta. Naš prst je kot marioneta na vrvici bliska, svoboda volje pa je iluzija.

Ampak *zakaj*, natanko? Vzemimo neko enostavno dejanje: poštar stoji pred vašimi vrati in pritisne na gumb zvonca. Zakaj? Ker vam *želi* izročiti poštno pošiljko, saj je *prepričan*, da ste doma (prej sta se dogovorila po mobilnem telefonu) in vam *namerava* pošiljko izročiti (to je njegov *cilj*, navsezadnje je to njegova služba). Po tej, zelo vplivni Davidsonovi analizi, poštarjevi *razlogi* (želje, prepričanja, namere, cilji) *povzročijo* njegovo dejanje. Nekateri takšni vzročni analizi sicer oporekajo, ampak v razpravi o Prediktorju bom privzel, da je namera pritisniti na gumb nujni del tistega, kar *povzroči* dejanje. Poštarjevo dejanje je *svobodno*, če bi se lahko odločil drugače, oblikoval drugačno namero, zaradi katere bi izbral drugačno ravnanje (vsi imamo izkušnje z 'lenimi' poštarji, ki iz takšnih ali drugačnih razlogov ne pozvonijo in vam samo vržejo obvestilo v nabiralnik).

Zdaj se počasi pokažejo razpoke v domnevnem 'dokazu', da nimamo svobodne volje. Ali zares nimamo *nadzora* nad tem, kaj se bo zgodilo? Naša namera pritisniti na gumb je nujni del tistega, kar *povzroči* dejanje. Kadar nekaj naredim, ker to *nameravam* storiti, potem je moje dejanje odvisno od mojih razlogov (želja in prepričanj, s katerimi se poistovetim) in v tem smislu ga (vsaj) doživljam kot svoje svobodno dejanje. Moj prst ni samo marioneta na vrvici Prediktorja, njegovi premiki so vzročno odvisni od *mojih* odločitev. Lahko bi se odločil drugače in oblikoval drugačno namero, zaradi katere bi izbral drugačno ravnanje. Preberimo še enkrat besedilo:

> Če *poskusite* pritisniti gumb, ne da bi videli blisk, se blisk svetlobe takoj pojavi in ne glede na to, kako hitro se premikate, nikdar ne pritisnete gumba, preden mine ena sekunda. Če čakate na blisk z *namero*, da bi se kasneje odpovedali pritisku na gumb, potem se blisk nikoli ne pojavi.

Toda *poskus* pritiska gumba brez bliska vendar vključuje tudi enostavno *namero* pritisniti gumb. In če čakate na blisk z *namero*, da bi se pritisku odpovedali, potem očitno *ne* nameravate pritisniti gumba. Z drugimi besedami: pritisk gumba in z njim povezan blisk svetlobe sta vzročno odvisna od vaših namer. Res je, časovni in vzročni red sta obrnjena: blisk je časovno *pred* pritiskom na gumb in vzročno *za* njim. Ampak v opisanem scenariju je namera vzročno *pred* obema, morda tudi časovno. To se vsaj zdi smiselna interpretacija: blisk je v časovno negativnem zamiku, povezan s stanji gumba, psihološka stanja igralca pa ne. Namere igralca ne *registrirajo* položaja gumba, ampak ga povzročajo na klasičen način: učinek časovno sledi vzroku.

No, človeška psihologija je včasih zapletena in način oblikovanja namer ni vedno transparenten. Denimo, da je igralec v stanju negotovosti, potem se pojavi blisk, igralec (zelo hitro) oblikuje namero in pritisne na gumb. Morda to ni nemogoče. V znamenitih Libetovih eksperimentih se oseba *zave* želje po premiku približno 200 milisekund pred aktivacijo mišice, ki sproži premik prsta (prim. Šuster: 2007, 307). Toda tudi v tem primeru je pritisk gumba pod mojim nadzorom, določajo ga moja volja in moji razlogi. Morda se res odločim zaradi bliska, ampak blisk me ne *prisili* v to odločitev. Zakaj potem ne gre za svobodno dejanje? Ali ne gre za zmotno sklepanje *post hoc ergo propter hoc*: »Izbira ne vodi (časovno) do bliska. Torej izbira ne vodi (vzročno) do bliska?« Premagati 'sistem' bi pomenilo *nemogoči* upor proti svoji *lastni* volji: nameravati pritisniti na gumb (to naznači blisk, ki registrira *učinek* naše namere), ampak *obenem* tega tudi ne nameravati (saj želimo pretentati Prediktorja).

Lahko bi se odločil drugače in oblikoval drugačno namero? Toda, ko se pojavi blisk, ne moreš ravnati drugače, kot da pritisneš na gumb! Nobene izbire nimaš. In dokler bliska ni, ne moreš pritisniti na gumb, nam prišepetava pripovedovalec. Ali ni to strašljivo?

Morda med pojavom bliska in pritiskom na gumb zares ni alternativ, ampak, prvič, imeli smo jih *pred* tem in drugič, kot kažejo Frankfurtovi primeri, še vedno ravnamo po lastni (svobodni) volji. Zakaj nas blisk ne 'prisili' v pritisk gumba? Ker gumb v vsakem primeru nameravamo stisniti, frustriran je le naš poskus privesti do gumba *brez* bliska. Tak poskus pa je obsojen na neuspeh, ne glede na to, kdaj se pojavi blisk: *pred* pritiskom na gumb ali *po* pritisku na gumb. Vzemimo namesto Prediktorja zelo navadni *Retrodiktor*: luč zasveti eno sekundo *po* tem, ko pritisnete na gumb. Retrodiktor je izjemno dobro narejen, s pomočjo novih materialov, supersonične tehnologije in toka nevtrinov, tako da je nemogoča kakršnakoli zunanja intervencija. Ko enkrat pritisnete na gumb, je glede na zakone narave fizikalno nemogoče, da se ne bi čez eno sekundo pojavil blisk svetlobe. Oziroma nemogoče vsaj v tistem smislu, v katerem je nemogoče, da ne bi krogotok v Prediktorju, ko enkrat pritisnete gumb, vodil do tega, da se sekundo *pred* tem pojavi blisk (tudi tu izključimo možnost napak v materialu in 'vnazajšnjih' intervencij, ki bi blokirale signal). Retrodiktor poroča o preteklosti, pove vam, da ste sekundo *pred* bliskom pritisnili na gumb. Izredno nezanimiva napravica za igranje, morda bi pritegnila šestmesečne dojenčke. Ampak tudi Retrodiktorja ne moremo pretentati: ne moremo pritisniti gumba, ne da bi se *za* tem pojavil blisk, točno eno sekundo kasneje. In nemogoče je, da se pojavi blisk, ne da bi *pred* tem pritisnili na gumb. Pravil igre ne moremo prekršiti, vendar mislim, da nihče ne bi trdil, da Retrodiktor dokazuje, da nimamo svobodne volje! Ne v enem, ne v drugem primeru vaša svoboda ne more biti v tem, da z neko magično močjo prekinete vzročno nujno zvezo med pritiskom na gumb in bliskom, ampak v tem, da po *lastni* izbiri pritisnete na gumb.

Prediktor (signal 'nazaj') v paru z Retrodiktorjem (signal 'naprej') ponazarja še eno od značilnosti miselnih eksperimentov. Miselni eksperiment, posebej to velja za 'močne' in prelomne, zelo pogosto vodijo do variacij, majhnih ali večjih sprememb v scenariju, ki sklep okrepijo, razširijo na druga področja ipd. ali pa ga ovržejo (prim. Miščević, 2017: 121). Retrodiktor je proti-eksperiment, ki nam kaže, da je intuicija o nesvobodi, ki temelji na nujni zvezi med (časovno predhodnim) bliskom in (časovno kasnejšim) gumbom, zavajajoča. Še ena variacija bi bila 'Časovna turistka': denimo, da nekoč v daljni prihodnosti odkrijejo in izkoriščajo možnost potovanja v času. V naključnem pogovoru vas zgovorna turistka iz prihodnosti preseneti z napovedjo, da boste, danes še zakrneli samec in velik ljubitelj poznega vstajanja, čez šest let in pol vsako jutro sprehajali labradorca po imenu Maks. Tako kot je blisk signal iz prihodnosti (ena sekunda), je ta napoved poročilo iz prihodnosti (šest let in pol). Ali je lahko takšno resnično poročilo razlog za depresivno ugotovitev, da vaše izbire niso pomembne? Zaradi napovedi ne morete ravnati drugače kot da spremenite svoj življenjski slog? Pa saj boste vendar *sami* izbrali življenjsko pot pasjeljubca med različnimi alternativami, ki jih prinaša življenje.

3.

Prediktor zavaja, ker usmerja naše intuicije k napačnemu odgovoru na začetno vprašanje. Pripovedovalec predpostavlja, da naše izbire niso pomembne, a to bi moral biti šele naš intuitivni odgovor. No, v obrambo lahko navedemo, da so tudi najbolj slavni filozofski miselni eksperimenti deležni podobnih očitkov (prim. Miščević, 2017: 115). In tu gre, navsezadenje, za umetniško prozo, ne pa filozofsko razpravo. Pripovedovalec spretno manevrira med opisovanjem domnevnih dejstev in poročanjem o *učinku* Prediktorja na igralce. Miselni eksperiment črpa prepričljivost iz našega *doživljanja* vzročnosti in časa v našem vsakdanjem delovanju. Smer našega delovanja v našem izkustvu sovpada s smerjo časa, ki je smer vzročnosti in tega, kar pod našim nadzorom. Preteklost je vzročno izolirana, ne moremo je spremeniti. Celo v filozofski literaturi najdemo obrambo intuicije, da časovno *predhodni* blisk, nad katerim igralec domnevno nima nadzora, nekako *izsili* kasnejši pritisk na gumb.

Gre za standardni filozofski ugovor proti sami zamisli o vnazajšnjem povzročanju, t. i. 'argument preklica' (angl. *bilking argument*, prim. Horwich, 1987: 91–92). Uporabimo kar Prediktorja: blisk (B) je v zgodbi pred pritiskom na gumb (G), čeprav gre za vzročni učinek pritiska na gumb. Potem, pravi ta argument, je vsaj v načelu mogoče, da *interveniramo* v toku dogodkov in po nastopu bliska preprečimo kasnejši nastop G. S tem uničimo zvezo med B in G: blisk nastopi brez gumba, zato (kasnejši) G ne more biti vzrok za (predhodni) B. Če pa, ko enkrat nastopi B, ne moremo več *preprečiti* kasnejšega pritiska na gumb, potem to dokazuje (?), da je prvi dogodek (blisk) vzrok drugega – časovni in vzročni red sovpadeta, zamisel o vnazajšnji vzročnosti pa je nekoherentna.

A zakaj le? Chiang se s tem ne ukvarja, žanr znanstvene fantastike mu omogoča, da prekrši nek (domnevni) zakon narave in potem v zgodbi razdeluje posledice. Ampak zamisel sploh ni nekoherentna. Če ne moremo preprečiti pritiska na gumb, zakaj naj bi to pomenilo, da je potem blisk časovno predhodni vzrok in ne obratno? To je sicer čudno, ne more pa biti izključeno *a priori*. Faye (2018) lepo opozori, zakaj se nam to zdi *čudno*: ker razmišljamo o vnazajšnji vzročnosti kot pojavu, ki ga lahko neposredno nadziramo na običajen način s svojimi vsakdanjimi dejanji. Ampak to se nam zdi nemogoče, naše izbire vplivajo in določajo prihodnje dogodke, nikdar preteklih. Preteklosti je izven vzročnega 'dosega' naših dejanj, v tem smislu je fiksna (angl. settled) in neogibna. Prediktor privede do kolizije našega doživljanja časa in dojemanja sebe kot vzročnega dejavnika z obratom časovne in vzročne smeri. Dejanski red našega delovanja je vzročni (od namere do gumba, od gumba do bliska), naš izkustveni red pa je časovni (od bliska do gumba). Končni proizvod teh pojmovno-kemičnih reakcij pa je *fatalizem*.

Fatalizem v vsakdanji rabi označuje neko življenjsko *držo*, v katero zapademo, ko začnemo verjeti, da ničesar, kar lahko danes naredimo, ne more spremeniti tega, kar se *bo* zgodilo. Spomnimo se nesrečnega Ojdipa – usojeno je bilo, da bo ubil svojega očeta in se poročil z lastno materjo in ne glede na vse njegove izbire in poskuse, da bi to preprečil, se je prerokba uresničila. Tako kot preteklost tudi dogodki, ki šele prihajajo, niso pod našim nadzorom in so za nas neogibni, pravi fatalist. V našem doživljanju se lahko takšno prepričanje manifestira kot stoična 'vdanost' v usodo. Chiang dramatično opisuje bolj pogubne odzive, ko domnevne »implikacije nespremenljive prihodnosti prodrejo v zavest.« Nekateri ljudje, ki zaradi Prediktorja ugotovijo, da njihove odločitve niso pomembne, se ne lotevajo več nobenih spontanih dejavnosti in tretjina igralcev konča »v stanju akinetičnega mutizma, v neke vrste budni komi.«

Filozofi danes uporabljajo izraz *fatalizem* drugače, kot oznako za *stališče*, da nimamo moči, da bi storili karkoli drugega od tistega, kar dejansko storimo: »Fatalizem /…/ je teza, da je logična ali pojmovna resnica, da nihče ne more ravnati drugače kot dejansko ravna, gola zamisel dejavnika, ki ima pred sabo alternativne poteke dejanj je protislovna« (van Inwagen, 1983: 23). Ker gre za logično resnico, je stališče znano tudi kot *logični* determinizem. *Navadni* determinizem je teza, da v danem trenutku iz dejstev o preteklosti sveta po *zakonih* narave sledijo vsa dejstva o prihodnosti sveta (prim. Šuster, 2007). V vsakem trenutku univerzuma obstaja natanko ena fizikalno možna prihodnost. Determinizem dopušča, da bi preteklost lahko bila drugačna, prav tako zakoni narave (fiksnost enega ali drugega ni logična in pojmovna resnica). Za logičnega determinista pa je vsak dogodek v preteklosti, sedanjosti in prihodnosti vedno že 'fiksiran' samo zaradi zakonov logike ter narave pojmov resnice in časa.

Od kod ta zamisel? Pri Aristotelu (*De interpretatione*: 18b23) najdemo znameniti primer jutrišnje pomorske bitke. Do nje bo prišlo ali pa ne, to nam pove 'čista' logika (zakon izključene tretje možnosti). Propozicija, da bo jutri pomorska bitka, je tako (danes, dan pred tem) resnična ali pa neresnična. Ampak, če je, denimo, (danes) resnično, da se bitka zgodi, potem je bilo to resnično tudi že včeraj. Propozicije v času ne spreminjajo svoje resničnostne vrednosti in resnice ne moremo spremeniti v neresnico. To velja tudi za resnične propozicije o bodočih dogodkih, zato v *nobenem* trenutku časa ni resnično, da bitke ne bo. Toda potem je tudi v preteklosti vedno že bilo resnično, da bitka bo. Preteklost pa je fiksirana, nihče ne more narediti, da bi bila drugačna. Torej tudi (pretekle) resnice tega, da jutri bo pomorska bitka, ne moremo spremeniti in je zato bitka *neogibna*. Enak razmislek pokaže, da je tudi propozicija, da jutri pomorske bitke ne bo, če je *ta* resnična, neogibna. Ali eno ali drugo: bodoči potek dogodkov je neodvisen od naših namer in odločitev.

Mislim, da strašljivost Prediktorja temelji na logiki argumenta za fatalizem. Zamislimo si igralca pred Prediktorjem, ki razmišlja o tem, kaj naj naredi in kaj sploh lahko naredi. Po vsakdanji fenomenologiji naše izbire vplivajo in določajo prihodnje dogodke, nikdar preteklih, ti so izven 'dosega' našega delovanja. Za poljubni trenutek *t* v prihodnosti igralca potem velja: ali bo pritisnil na gumb v tem trenutku, ali ne. Ampak, če bo pritisnil na gumb, se bo sekundo pred tem pojavil blisk in *napovedal* njegovo dejanje. In če ne bo pritisnil na gumb, potem bliska ne bo. Toda sekunda pred *t* je v preteklosti glede na njegov gib, pred trenutkom njegovega delovanja, zato v trenutku *t* ne more več vplivati na to, kaj se bo zgodilo *pred* tem. Resnične napovedi 'iz preteklosti' ne moremo spremeniti v neresnično, zato je njegovo dejanje neogibno.

Privlačno, ampak zmotno. Za aristoteljanskega fatalista ima resnična *napoved* (pred bitko) enak metafizični status kot resnično *poročilo* (po bitki). Oboje je fiksno, ničesar ni, kar bi kdorkoli lahko naredil, da bi eno ali drugo spremenil. Podobno pri igralcu: v trenutku pritiska na gumb ne more narediti ničesar, kar bi vplivalo na preteklost (pojav bliska). Pojav bliska je resnična *napoved* pritiska na gumb, zato lahko razmišljanje igralca prikažemo v skladu z logiko pomorske bitke:

> Če je napovedano, da bo igralec pritisnil na gumb, potem je neogibno, da bo pritisnil na gumb. Če pa ni napovedano, da bo pritisnil na gumb, potem je neogibno, da ne bo pritisnil na gumb. Ampak pritisk na gumb je napovedan ali pa ni napovedan (blisk se pojavi ali pa se ne pojavi). Torej je neogibno, da bo igralec pritisnil na gumb ali pa je neogibno, da ne bo pritisnil na gumb.

Igralec v resnici nima izbire, svobodna volja pa je iluzija, pravi Pripovedovalec. Naj bo 'Np' oznaka za »Neogibno je, da *p*.« V najbolj splošnem smislu gre za neko nujnost, ki označuje našo nemoč: nezmožnost vplivanja, nesposobnost spreminjanja, odsotnost izbire ali kak podoben pojem, ki označuje nesvobodo glede tistega, kar izraža stavek 'p' ali o čemer je propozicija, da *p*. Naj 'B' označuje »Napovedano je, da bo igralec čez eno sekundo pritisnil gumb,« 'G' pa »Igralec pritisne na gumb.« Razmišljanje lahko potem logično prikažemo na dva načina (prim. recimo Priest, 2017: 39-44):

PRVI	DRUGI
1. B → N G 2. ~B → N ~G 3. B v ~B Torej, 4. N G v N ~G	1. N (B → G) 2. N (~B → ~G) 3. B v ~B Torej, 4. N G v N ~G

V PRVEM primeru je sklepanje veljavno, ampak prva in druga premisa sta sporni. V DRUGEM primeru pa sta premisi sprejemljivi, le da sklep *ne* sledi.

Zakaj je neresnična prva premisa, v kateri trdimo, da iz resnice napovedi sledi njena nujnost? Razprava o bodočih kontingencah ima častitljivo zgodovino, h kateri pa tu ne bom nič dodajal in spoštovanja vredno 'tehnično' obravnavo v sodobnosti (večvalentne logike, supervaluacija), ki se ji bom v glavnem izognil (prim. Šuster, 2014). Mislim, da je iz zgornje razprave jasno, kaj je narobe s prvo premiso igralca (PRVI vzorec):

> Če je napovedano, da bo igralec pritisnil na gumb, potem je neogibno, da bo pritisnil na gumb.

Ali smo zares nemočni glede tega, ali bomo pritisnili na gumb? Spomnimo se: pritisk gumba in z njim povezan blisk svetlobe sta odvisna od igralčevih namer. Zato je to dejanje pod njegovim nadzorom, določajo ga njegova volja in njegovi razlogi. Prva (in po analogiji druga) premisa v PRVEM vzorcu je neresnična. Ampak, ali ni tako, da blisku *nujno* sledi pritisk na gumb, gre za neogibno zvezo? Res je, ampak potem gre za DRUGI vzorec in premiso:

> Neogibno je, da, če je napovedano, da bo igralec pritisnil na gumb, potem bo pritisnil na gumb.

Ta premisa je resnična, ampak sklepanje ni veljavno. Vzemimo spet Retrodiktorja, denimo, da pritisnem na gumb. Zato velja tudi: »Neogibno je, če pritisnem na gumb, potem sekundo kasneje blisne svetloba.« Ampak iz tega ne sledi sklep, da je blisk svetlobe neogiben: lahko bi se odločil drugače in na gumb ne bi pritisnil (to je, navsezadnje, odvisno od moje volje). Premisi (»Pritisnem na gumb«, »Neogibno je, če pritisnem na gumb, potem sekundo kasneje blisne svetloba«) sta resnični, sklep (»Neogibno je, da blisne svetloba«) pa ne. Kot bi sklepali:« Imam sina in nujno, če imam sina, potem imam otroka. Torej je nujno, da imam otroka.« *Non sequitur*!

Kaj pa če gre za drugačen vzorec prenosa nemoči?

1. N B
2. N (B → G)

Torej,

3. N B

Gre za znameniti 'modalni katapult', ki 'izstreli' začetno neogibnost na sklep prek vezi, ki so same neogibne (prim. Šuster, 2009). Recimo, pri Retrodiktorju: Neogibno je, da pritisnem na gumb in neogibno, če pritisnem na gumb, potem sekundo kasneje blisne svetloba. Torej je neogibno, da blisne svetloba. Razprava o tem načelu (znanem kot *Beta*) je razvejana, a se ji lahko izognemo, saj je jasno, da je prva premisa tako pri Retrodiktorju kot pri Prediktorju sporna. Zakaj bi bilo neogibno, da se pojavi blisk? To je neogibno samo, če je pritisk na gumb nujen glede na njegovo vzročno zgodovino, torej, če se igralec ne more odločiti drugače. Ampak o tem sploh ni bilo govora – rekli smo, da je to dejanje odvisno od igralčevih namer. Če bi *te* bile neogibne, potem bi nas res lahko začeli skrbeti za našo svobodo.

Prav to, po mnenju nekaterih, dokazuje znameniti *dejanski* eksperiment, še ena variacija Prediktorja, imenujmo jo Libet-Prediktor. Oseba naj bi svobodno izvedla neko enostavno dejanje – recimo premaknila prst. Meritve pokažejo, da pred dejanjem nastopi električna aktivnost v tistem področju možganov, ki sproži aktivacijo mišice, ki potem vodi do premika prsta. Toda ta aktivnost nastopi že 550 ms pred samim dejanjem, oseba pa se, po teh meritvah, zave svoje namere šele 150-200 ms pred premikom prsta. Torej *pravi* vzroki dejanj niso naše namere, saj se vzročna veriga sproži, preden se sploh zavemo odločitve, da dejanja izvedemo (prim. Libet, 2007)? Tudi ta *pravi* eksperiment je vodil do trditev, da je odsotnost svobodne volje zdaj *znanstveno* dejstvo, ne samo filozofska teorija (Harris 2012) in smešnih naslovov v popularnih medijih: »Moji možgani so me prisilili, da sem to storil« (angl. *My Brain Made Me Do It*). Kot da so moji možgani nek sadistični trener, ki me priganja k delanju sklec, ne pa substanca moje istovetnosti. In daleč od tega, da je odsotnost svobodne volje zdaj znanstveno dejstvo (prim. Šuster, 2007: 318–321). Eksperimenti pokažejo samo, da obstaja nek vzorec aktivnosti pred odločitvijo, ne pa, da je bila *odločitev* sprejeta preden smo se mi tega zavedali (prim. Mele, 2014). Interpretacija obeh eksperimentov, fiktivnega Prediktorja in dejanskega Libet-Prediktorja zahteva nekaj filozofskega pojmovnega dela!

4.

Pretekla resnica je del preteklosti (tudi, če je ta samo eno sekundo nazaj) in preteklost je fiksna, se glasi *mantra* fatalizma. Toda resnična (čeprav pretekla) napoved je rezultat naših odločitev in blisk je samo 'senca', ki jo *naše* namerno dejanje meče v času nazaj. Podobno, mislim, velja za glavni argument fatalizma: resnica je odvisna od sveta in ne obratno, resnična napoved je samo senca, ki jo tisto, o čemer je, 'meče' v času nazaj (prim. Šuster, 2014). Torej *ne*, Prediktor ne dokazuje, da svobodna volja ne obstaja, glavni miselni eksperiment je neuspešen.

Pa vendar, miselni eksperiment nas opozarja, da poleg smeri časa in smeri vzročnosti obstaja tudi smer našega delovanja, ki je v našem doživljanju istovetna s smerjo časa. Prihodnost je odprta in pod našim nadzorom, preteklost je fiksna in vzročno izolirana. To je vir psihološke prepričljivosti zgodbe. Predstavljajte si, da se vrata odprejo, šele potem jih odklenete; da lonček s kavo zavre, šele potem ga pristavite na ogenj; da se najprej pojavijo črke na računalniškem monitorju, šele potem jih natipkate itd. Najmanj, kar lahko predvidimo, je to, da takšno časovno obrnjeno zaporedje vodi do psiholoških težav, ki jih Chiang tako dramatično prikaže. Če veš, kaj se bo zgodilo, zakaj ne poskusiš tega preprečiti, to je naša naravna reakcija, ki je tudi jedro argumenta 'iz preklica'. In če ne moreš, potem ti je *usojeno*, nimaš kaj. Chiang je prepričljiv v fenomenologiji fatalizma, a tudi psihologiji zasvojenosti – vsi poznamo razne viralne napravice, da ne govorimo o največjem sodobnem viru obsedenosti s pritiskanjem na gumbe. Danes bi Prediktor bil aplikacija za pametne telefone, ki jih leta 2005 še ni bilo.

Zdaj se počasi izriše drugačna začetna hipoteza. Ne več, ali Prediktor dokazuje, da nimamo svobodne volje, ampak vprašanje, ali nas *doživljanje* ločitve smeri časa in smeri vzročnosti v našem delovanju vodi do fatalizma kot življenjske drže? Ločitev obeh smeri je za nas vir nelagodja, celo filozofi niso imuni, kot kažejo argumenti za domnevno pojmovno nemožnost vnazajšnjega povzročanja. Po argumentu preklica, denimo, moramo 'nenadni' pojav črk na monitorju interpretirati kot vzrok za to, da jih kasneje natipkate. Ali ni to tako, kot da bi bili naši prsti samo marionete na vrvicah usode? Za preverjanje te druge hipoteze bi moral biti scenarij bogatejši, lučka, ki zasveti sekundo, preden pritisnete na gumb ni dovolj.

Vzemimo zato še eno različico Prediktorja, imenujmo jo *Probadiktor*. Tokrat gre za napravico, kjer je vgrajeni krogotok probabilističen: brez pritiska na gumb ni bliska, toda pritisk na gumb samo v polovici primerov povzroči blisk svetlobe, mehanizem pa je povsem naključen in nepredvidljiv. Dodajmo še negativni časovni zamik – kadar se pojavi blisk, sekundo kasneje igralec 'nujno' pritisne na gumb. Še vedno velja: če čakate na blisk z namero, da bi se kasneje odpovedali pritisku gumba, potem bliska ni. Ampak, če poskusite pritisniti gumb, ne da bi videli blisk, vam to v polovici poskusov uspe, v polovici pa ne. Torej ni več res, da ne glede na to, kaj naredite, luč vedno zasveti, preden pritisnete na gumb. Kako bo igralka doživljala takšno igro?

Ne pri Prediktorju ne pri Probadiktorju igralka ne more privesti do situacije, v kateri se pojavi blisk, ona pa po eni sekundi ne pritisne na gumb. Kot rečeno gre za blokado, ki je *notranja* samemu scenariju: nameravati pritisniti gumb (blisk registrira učinek namere), ampak obenem tega tudi ne hoteti. Igralka bo frustrirana, ampak mislim, da je v ozadju tipična psihološka razklanost: nekaj nameravamo, čeprav tega nočemo. Če prevlada občutek nemoči in blokade, lahko konča v fatalizmu. Ampak tako kot pri Prediktorju: vedno ji uspe pritisk na gumb, kadar to (enostavno) namerava. V polovici primerov ji pri Probadiktorju uspe tudi bolj zapletena namera pritisniti gumb brez predhodnega bliska, pri Prediktorju ji to nikdar ne uspe. Popolnega nadzora pa še vedno nima: nemogoče je napovedati, kdaj bo njena (zapletena) namera uspela in kdaj ne. Tudi tu lahko prevlada občutek, da gre za igro usode ali neko 'silo', ki se z njo samo poigrava. Prediktor in Probadiktor kažeta, da obrat časovne in vzročne smeri lahko privede do naše *izkušnje* nemoči pred tem, kar prihaja. A to še ne pomeni, da naše izbire niso odvisne od naših namer in odločitev.

Predictor, Fatalism, and Thought Experiments

A short story by Ted Chiang (2005) is analysed as an example of a philosophical thought experiment allegedly proving that there's no such thing as free will. I explore the structure and stages of thought experiments as proposed by Miščević (2012: 2017): design, the question, understanding, tentative conscious production, immediate spontaneous answer, variations, and generalizations. Another characteristic is a certain mental "chemistry": in imagination we separate conceptual traits that normally go together. Predictor separates the causal order from the temporal order of events. This inversion leads to dramatic collision between our experience of time and self-agency. The actual order of our action is causal, but our experiential order is temporal. Nevertheless, the thought experiment fails in its demonstration that our choices don't matter. Its logical core is a fallacious argument for fatalism.

Keywords: free will, fatalism, causality, time, thought experiment.

Literatura

Chiang, T. (2005). »What’s expected of us.« *Nature*, 436, str. 150.

Clark, T. (2015). »The Perfectionist: Ted Chiang’s science fiction wins piles of awards. When he publishes, which is hardly ever«. *The California Sunday* Magazine (20.julij 2020). URL = https://story.californiasunday.com/ted-chiang-scifi-perfectionist/.

Faye, J. (2018). »Backward Causation«. V Zalta, E. N. (ur), *The Stanford Encyclopedia of Philosophy* (izdaja poletje 2018). URL = <https://plato.stanford.edu/archives/sum2018/entries/causation-backwards/>.

Frankfurt, H. (1969). »Alternate Possibilities and Moral Responsibility.« *Journal of Philosophy*, 66, str. 829–839.

Harris, S. (2012). *Free Will*. New York: Free Press.

Haynes, J. (2011). »Beyond Libet: Long-Term Prediction of Free Choices from Neuroimaging Signals«. V Dehaene, S., Christen, Y. (urd), *Characterizing Consciousness: From Cognition to the Clinic? Research and Perspectives in Neurosciences*. Berlin: Springer, str. 161–174.

Horwich, P. (1987). *Asymmetries in Time*. Massachusetts: The MIT Press.

van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.

Libet, B. (2007). »Ali imamo svobodno voljo?« V Šuster, D. (ur), *O svobodni volji: od Leibniza do Libeta*. Maribor: Aristej, str. 303–317.

Mele, A. (2014). *Why Science Hasn’t Disproved Free Will*. Oxford: Oxford University Press.

Miščević, N. (2012). »Political Thought Experiments from Plato to Rawls«. V M. Frappier, L. Meynell, R. Brown (urd.), *Thought Experiments in Science, Philosophy, and the Arts*. New York: Routledge, str. 191–206.

Miščević, N. (2017). »In Defense of the Twin Earth – The Star Wars Continue«. *European Journal of Analytic Philosophy*, 12(2), str. 111–130.

Priest, G. (2017). *Logic. A Very Short Introduction*.2nd. Oxford: Oxford University Press.

Šuster, D (ur). (2007). *O svobodni volji: od Leibniza do Libeta*. Maribor: Aristej.

Šuster, D. (2009). »'Modalni katapult' in argument za nezdružljivost«. *Analiza*, 13, str. 5–43.

Šuster, D. (2014). »Med resnico in neresnico.« *Analiza*, 18, str. 19–56

Tooley, M. (1987). *Causation*. Oxford: Clarendon Press.

Ted Chiang

Prevedel Danilo Šuster

Kaj se pričakuje od nas[1]

To je opozorilo. Prosim, da ga pazljivo preberete.

Do sedaj ste gotovo že videli Prediktorja, do trenutka, ko to berete, so jih prodali milijone. Za tiste, ki ga še niste videli, gre za majhno napravo, nekako velikosti daljinca za odpiranje vrat vašega avtomobila. Njegovi edini značilnosti sta gumb in veliko zeleno LED svetilo. Lučka zabliska, če pritisnete na gumb. Bolj natančno, luč zasveti eno sekundo preden pritisnete na gumb.

Večina ljudi pravi, da je občutek, ko prvič poskusite, tak, kot da bi igrali neko čudno igro, v kateri je cilj ta, da pritisnete gumb po tem, ko ste videli blisk in to igro je enostavno igrati. Toda, ko poskusite prekršiti pravila igre, ugotovite, da tega ne morete. Če poskusite pritisniti gumb, ne da bi videli blisk, se blisk svetlobe takoj pojavi in ne glede na to, kako hitro se premikate, nikdar ne pritisnete gumba, preden mine ena sekunda. Če čakate na blisk z namero, da bi se kasneje odpovedali pritisku na gumb, potem se blisk nikoli ne pojavi. Ne glede na to, kaj naredite, luč vedno zasveti preden pritisnete na gumb. Nobenega načina ni, kako pretentati Prediktorja.

Srce vsakega Prediktorja je krogotok z negativnim časovnim zamikom, ki pošilja signal v času nazaj. Vse razsežnosti te tehnologije bodo jasne kasneje, ko bodo doseženi negativni zamiki večji od ene sekunde, ampak to opozorilo ni namenjeno temu. Takojšnji problem je dejstvo, da Prediktorji dokazujejo, da ni take stvari kot je svobodna volja.

Vedno so obstajali argumenti, ki so dokazovali, da je svoboda volje iluzija, nekateri so temeljili na trdi fiziki, drugi na čisti logiki. Večina se strinja, da so ti argumenti neovrgljivi, toda nihče zares ne sprejema njihovih sklepov. Naše izkustvo svobodne volje je premočno, da bi ga z argumenti lahko ovrgli. Potrebujemo dokaze in natanko to nam ponuja Prediktor.

To, kar se tipično zgodi je, da se oseba obsedeno igra s Prediktorjem nekaj dni, ga razkazuje prijateljem in preizkuša razne načine, kako napravo pretentati. Včasih se zdi, da oseba izgubi zanimanje, a nihče ne more pozabiti, kaj naprava pomeni – v naslednjih tednih

[1] Prevod dela Teda Chianga, 2005, »What's expected of us«, *Nature,* 436, 150. DOI:https://doi.org/10.1038/436150a.

implikacije nespremenljive prihodnosti prodrejo v zavest. Nekateri ljudje po tem, ko ugotovijo, da njihove izbire niso pomembne, zavračajo kakršnokoli izbiranje. Kot nekakšen bataljon pisarjev Bartelby se sploh ne lotevajo več nobenih spontanih dejavnosti. Sčasoma morajo tretjino tistih, ki se igrajo s Prediktorjem, hospitalizirati, saj se nič več ne ne hranijo sami. Končajo v stanju akinetičnega mutizma, v neke vrste budni komi. Z očmi bodo sledili gibanju, občasno spremenili svoj položaj, ampak nič drugega. Sposobnost premikanja ostaja, motivacije pa več ni.

Preden so se ljudje začeli igrati s Prediktorji je bil akinetični mutizem zelo redek, posledica poškodbe sprednjega cingularnega področja možganov. Zdaj pa se širi kot kognitivna kuga. Ljudje so špekulirali o misli, ki uniči misleca, neki neizgovorljivi Lovecraftianski grozi ali Gödlovskemu stavku, ki sesuje človeški logični sistem. Izkaže pa se, da je paralizirajoča misel tista, ki jo vsi poznamo: misel, da svobodna volja ne obstaja. Enostavno pa ni bila škodljiva, dokler niste začeli vanjo verjeti.

Zdravniki skušajo razpravljati s pacienti, dokler se ti še odzivajo na razgovor. Vsi smo pred tem živeli srečna, dejavna življenja, tako pravijo, čeprav tudi takrat nismo imeli svobodne volje. Zakaj bi se karkoli spremenilo? »Nobeno dejanje prejšnjega meseca ni nič bolj svobodno izbrano od tistega, za katerega ste se danes odločili,« bi lahko rekel zdravnik. »Še vedno se lahko tudi zdaj obnašate na enak način.« Pacienti pa brez izjeme odgovorijo: »Ampak zdaj vem.« Nekateri pa nikdar več nič ne rečejo.

Nekateri argumentirajo, da dejstvo, da Prediktor vodi do takšne spremembe v obnašanju pomeni, da vendarle imamo svobodno voljo. Avtomata se ne da zastrašiti, samo svobodno-misleče bitnost se lahko. Dejstvo, da nekateri individuumi zapadejo v akinetični mutizem, drugi pa ne, samo poudari pomembnost izbire.

Ampak na žalost je takšno razmišljanje zmotno: vsako obnašanje je združljivo z determinizmom. Dani dinamični sistem lahko pristane v bazenu atrakcije in se umiri na neki fiksni točki, drugi dinamični sistem pa lahko svoje kaotično obnašanje nadaljuje in nadaljuje, ampak oba sta popolnoma deterministična.

To opozorilo vam posredujem nekako eno leto iz vaše prihodnosti; gre za prvo daljše sporočilo, ki je sprejeto, ko so krogotoki z negativnimi zamiki v obsegu megasekudne uporabljeni za izgradnjo komunikacijskih naprav. Sledila bodo druga sporočila o drugih temah. Moje sporočilo je naslednje: pretvarjajte se, da imate svobodno voljo. Zelo pomembno je, da se obnašate, kot da so vaše odločitve pomembne, čeprav veste, da niso. Realnost ni pomembna, pomembno je to, v kar ste prepričani in prepričanje v laž je edini način, kako se izogniti budni komi. Civilizacija je zdaj odvisna od samo-prevare. Mogoče je vedno bila.

In vendar vem, da je, ker je svoboda volje iluzija, vnaprej določeno, kdo bo zapadel v akinetični mutizem in kdo ne. Ničesar ni, kar bi kdorkoli lahko storil glede tega; ne morete izbirati, kako bo Prediktor na vas učinkoval. Nekateri boste podlegli, nekateri ne in moje pošiljanje tega sporočila teh razmerij ne bo spremenilo. Zakaj sem potem to storil?

Ker nisem imel izbire.

Niko Šetar

Univerza v Mariboru

Prihodnost miselnih eksperimentov

Ta članek naslavlja vprašanja povezana s trenutno (sodobno) in prihodnjo rabo miselnih eksperimentov kot metode znanstvenega raziskovanja. V ta namen se najprej obrnemo na splošno veljavnost miselnih eksperimentov, kar naredimo skozi preučevanje kritik, katerih cilj je diskreditacija znanstvene in spoznavne vrednosti miselnih eksperimentov, ter s pregledom odgovorov na tovrstne kritike kot tudi drugih argumentov v prid miselnim eksperimentom, kot so recimo proti-zmote.

Po tem uvodu primerjamo miselne eksperimente z drugimi raziskovalnimi metodami, natančneje empiričnimi eksperimenti in računalniškimi modeli. Poskušali bomo pokazati, da so lahko empirični eksperimenti vsaj toliko zmotni kot miselni eksperimenti, slednji pa imajo tudi določene postopkovne lastnosti, zaradi katerih nam lahko nudijo podatke, ki z drugima dvema omenjenima metodama niso dostopni, kar jim omogoča, da ohranijo svoj status kot veljavna, neodvisna raziskovalna metoda.

Ključne besede: miselni eksperimenti, empirični eksperimenti, računalniško modeliranje, raziskovalne metode

1. Uvod

Miselni eksperiment (v nadaljevanju ME) je raziskovalna metoda, ki se na različnih področjih tako v filozofiji in znanosti uporablja, kadar je izvedba klasičnega, empiričnega eksperimenta (v nadaljevanju EE) nemogoča iz takšnih ali drugačnih razlogov. Teorija ME – sama po sebi precej sodoben pojav – se sooča z definicijo ME v, njihovo strukturo in rabo, kot tudi z odgovarjanjem na kritike, ki trdijo, da ME zastarela ali neveljavna metoda raziskovanja, ki se prej sklicuje na fantazijo kot resničnost.

V danem članku nas predvsem zanimajo te kritike. V odseku 2 si bomo tako najprej ogledali klasične ugovore na ME, od problema sklicevanja na intuicijo pri izvajanju ME, do problemov povezanih z introspekcijo in subjektivnostjo, pa do denimo sklicevanja na vsakdanji jezik. Pri vseh ugovorih bomo povzeli tudi odgovore, ki branijo veljavnost ME. V odseku 3 bomo najprej primerjali ME z EE ter poskušali vzpostaviti zakaj sta EE in ME enostavno dve različni metodi raziskovanja, pri čemer nobena nima absolutne prednosti pred drugo, marveč imata obe svoje prednosti in slabosti. Nato se bomo obrnili k novejšemu konceptu – računalniškemu modeliranju (v nadaljevanju RM). Slednje postaja v zadnjih dveh desetletjih vse bolj popularno in do neke mere začenja nadomeščati tako ME kot EE. Skladno s tem se pojavljajo teorije, da bi RM utegnili v celoti nadomestiti ME, saj so natančnejši in bolj objektivni. Poskušali bomo pokazati, da imajo tako kot EE tudi RM svoje prednosti pred ME, a imajo tudi slednji svoje prednosti pred RM, zaradi česar do popolne nadomestitve nikakor ne more priti.

V odsekih 2 in 3 se osredotočamo na znanstveno orientirane ME, kot so npr. Galileova padajoča telesa, Maxwellov demon, ipd., medtem ko bomo v krajšem 4. odseku orisali tudi nadaljnjo funkcijo filozofskih (političnih in etičnih) ME in zakaj je nadomestitev le-teh z RM še bolj neverjetna kot nadomestitev znanstvenih ME. V sklepu ugotavljamo, da ME ostajajo legitimno in uporabno metodološko orodje v svoji klasični vlogi, poleg tega pa v sklopu RM pridobijo še dodatne stranske funkcije.

2. Klasični argumenti proti miselnim eksperimentom

V tem odseku bomo pregledali nekaj klasičnih argumentov proti veljavnosti in uporabnosti ME kot znanstvene metode ter jih poskusili tudi ovreči in pokazati, da ne ogrožajo ne trenutnega statusa, ne prihodnosti ME.

2.1. Problem intuicije

Pogosto se ME očita, da se pretirano zanašajo na intuicijo avtorja oziroma izvajalca. Sorensen (1992) problemu intuicije v okviru ME nadane izraz 'vudu epistemologija,' verjetno najbolje poznan primer takšnega očitka pa je Dennetova (1984) kritika Searlove (1980) Kitajske sobe, kjer ME slavno označi za 'intuicijske črpalke'. Kar Dennet dejansko trdi je, da so ME kot je Kitajska soba formulirani tako, da na dan privlečejo intuitivno smiselne, a končno nepravilne odgovore ali izide, medtem ko pomembnejši odgovori oziroma rezultati ME ostanejo spregledani, ker niso intuitivni.
Drugi problem je tudi, da naj bi bile epistemološke in metafizične hipoteze ME zastavljene tako, da jih ne moremo ignorirati ali prilagoditi, kar vodi v krožnost ME, zaradi česar so slednji neveljavni (Virvidakis, 2011) Predpostavke posameznih eksperimentov so pogosto določene tako, da ne izvirajo iz nekega splošnega konsenza, marveč iz subjektivne sodbe ali intuicije avtorja ME. Modalne izjave, ki se pojavljajo kot (pogosto intuitivne) premise ME, po mnenju mnogih filozofov 'visijo v zraku', v smislu, da se njihova možnost ne utemeljuje v ničemer, razen v izjavi sami. Poleg tega se te premise, kot tudi splošni konteksti ME, pogosto navezujejo na možne svetove, ki izhajajo iz intuicij posameznika in so do te mere različni od našega, da implikacije zaključkov niso relevantne za naš svet.

Že Dennett (1984) sam trdi, da je lahko uporaba intuicije za obravnavo kompleksnih scenarijev lahko uporabna saj nam pomaga »videti gozd, ne le dreves«, problem pa je, da je intuicije zelo enostavno (nenamerno) zlorabiti.

Virvidakis (2011) postopa v smeri tega, kako se tovrstni zlorabi intuicij izogniti, oziroma kako bi utegnila potekati odgovorna raba intuicij v ME. Virvidakis izpostavlja, da Kant odgovarja na Leibnizove miselne eksperimente glede teorije prostora kot definiranega s strani predmetov, ki se v njem nahajajo. Kant meni, da je narava razmišljanja o prostoru fenomenalne narave, saj gre za prostor kot ga mi izkušamo. Leibnizova redukcija prostora na sistem korelacij in interakcij med predmeti, ki bivajo v prostoru je po njegovem mnenju neupravičena, saj s tem zanemarja epistemske intuicije, ki jih imamo o prostoru brez sklicevanja na stvari, ki se v njem nahajajo. Enako velja za Descartesa (1641), ki je v Meditacijah trdil, da mu nič ni lažje dostopno kot njegov lasten um. Kant (1998) temu ugovarja, češ da Descartes ni dojel, da mu princip *Cogito ergo sum* ne nudi vpogleda v nič povezanega z njegovim umom ali jazom, razen tega, da enostavno biva (obstaja). Za kakršnekoli ugotovitve, ki bi podprle njegovo domnevo, da dejansko pozna lasten um, bi bila potrebna transcendentalna refleksija o naravi sebe in upoštevanje epistemskih pogojev, ki mu omogočajo tovrsten dostop. Ti epistemski pogoji so izpolnjeni skozi dosledno upoštevanje empiričnih spoznanj in izkušenj, ki ustrezno dopolnijo apriorna prepričanja, na katerih slonijo premise in hipoteze, vsebovane v miselnih eksperimentih. Kantovi pogoji za ustrezno izvajanje ME slonijo na njegovi teoriji modalnosti, ki se lahko povzame kot sledi:

1. Karkoli ustreza formalnim pogojem izkustva (v skladu z intuicijami in koncepti) je možno.
2. Kar je povezano z materialnimi pogoji izkustva (ali zaznave) je dejansko.
3. Česar povezavo z dejanskim se določa v skladu s splošnimi pogoji izkustva, je (obstaja) nujno. (Virvidakis, 2011: 139)

V kolikor so ta tri pravila upoštevana pri vzpostavljanju modalne premise ME, je ta veljaven. Torej, v kolikor so tudi intuicije, uporabljene v ME, skladne s temi pravili, so rabljene odgovorno. Medtem ko je to samo (relativno dober) primer, je glavna poanta tega odgovora (kot tudi Dennettove lastne izjave) zgolj to, da je potrebno preprečiti zlorabo intuicij in intuicijskih črpalk, ter jih v raziskovanju uporabiti odgovorno.

2.2. Problem introspekcije

John Locke (v Sorensen, 1992) zagovarja introspekcijo kot metodo na podlagi samorazpoznavnosti (ang. *self-intimation)* mentalnih stanj – za vsako mentalno stanje obstaja tudi zavedanje, da imamo to mentalno stanje. Leibniz temu principu ugovarja, češ da vsako mentalno stanje zahteva mentalno stanje o zavedanju prvega, in tako naprej do neskončnega regresa, kar pomeni, da je polna introspekcija nemogoča, in je torej naš dostop do lastnih mentalnih stanj nezadosten. Sorensen (1992) izpostavlja, da temu mnenju botruje tudi sodobna psihologija, ki dokazuje, da se ljudje redkokdaj zavedajo lastnih mentalnih stanj, ki so privedla do neke odločitve ali dejanja. To vodi do vprašanja, ali je predmet ME dostopen? Srž tega problema se nanaša na to, da so scenariji, vsebovani v miselnih eksperimentih, mentalna stanja, ki so predmet istega regresa kot introspektivna mentalna stanja.

Tudi če temu ni tako, ostaja vprašanje, kako introspekcija poteka na fizičnem nivoju? ME pogosto zahtevajo nivo intuicije (npr. Brown, 2011), ki za mnoge ni sprejemljiv, ter zahtevajo fizikalistično razlago pojava, v istem smislu kot oko in čutne celice v njem razlagajo pojav vida. Vprašanje delovanja introspekcije se pojavlja tudi izven fizikalističnega dvoma – Auguste Comte (v Sorensen, 1992: 24–25) denimo trdi, da samoopazovanje zahteva razdelitev zavesti osebe na tako radikalen način, da bi šlo za dve osebi v istem umu, kar pa (razen v okviru določenih psiholoških motenj) ni mogoče.

Če je introspekcija dejansko možna, moramo odgovoriti na vprašanja o njeni naravi. Prvo izmed teh je, ali je introspekcija konstantna in zvesta. Sorensen se ponovno obrača na Comteja, ki trdi da introspekcija vodi do različnih rezultatov pri različnih osebah – je torej intrakonstantna (tj. pri isti osebi konstantna), ne pa interkonstantna (tj. konstantna za različne osebe), le redko pa tudi zvesta, saj praviloma meri subjektovo stališče do propozicije, ne pa česarkoli o propoziciji sami. Na konstantnost ME meče luč dvoma to, da so zgodovinsko mnogi uporabljali iste miselne eksperimente za dokazovanje nasprotnih stališč. Peter Unger (1984) poudarja tudi, da so ME odvisni od družbenega položaja eksperimentatorja. Poleg tega lahko rečemo, da je introspekcija redko celovita, saj je nagnjena k izpuščanju ali dodajanju informacij, ki izhajajo iz naših implicitnih predobstoječih prepričanj. ME seveda prav tako izpustijo in dodajo informacije, ki se zdijo za ME (ne)relevantne, kar pogosto (problematično) izpustijo pa je kakršenkoli konkreten zaključek – predvsem filozofski ME so nagnjeni k temu, da so njihovi sklepi v celoti prepuščeni interpretaciji. Naš poskus odgovora na to kritiko bo precej redukcionističen. Omejili se bomo namreč na vprašanje, ali je ME sploh oblika introspekcije v pravem pomenu besede? Kljub temu, da je znotraj psihologije, kognitivne znanosti in filozofije definicij introspekcije cela kopica, lahko rečemo, da je skupna točka vsem, nekako slovarska definicija ta, da gre za postopek opazovanja in preučevanja lastnih mentalnih stanj. Kadar gradimo ME, načrtno gradimo neko mentalno stanje, za kar seveda potrebujemo zavedanje o tem mentalnem stanju. V kolikor bi potrebovali tudi zavedanje o tem zavedanju, pa zavedanje spet o onem zavedanju, bi Leibnizov argument o neskončnem regresu držal, a temu ni tako. Za izvedbo ME ne potrebujemo tovrstne 'popolne' introspekcije. Sorensenov pomislek o tem, da psihologija nakazuje, da se ljudje redko zavedajo lastnih mentalnih stanj, je v kontekstu ME prav tako odveč. Domneva, da bi lahko neka oseba izgradila ME brez da bi se zavedala mentalnih stanj, ki vznikajo v tem procesu, je absurdna – katerega dela postopka se ne bi zavedala? Mar se denimo Galileo ne bi zavedal zakaj si predstavlja padajoča telesa? Ali se ne bi zavedal, da si jih predstavlja? Ali se ne bi zavedal kako naj bi potekalo njuno gibanje? Se naj ne bi zavedal rezultata? Katerakoli izmed teh možnosti bi pomenila, da ME sploh ne more obstajati. Enostavno ker ta obstaja in ker ga je Galileo lahko izvedel, lahko vemo, da se je Galileo zavedal potrebnih mentalnih stanj tekom izvajanja ME.

Tudi Comtejevo (v Sorensen, 1992) stališče o delitvi zavesti je zelo vprašljivo. Smiselno bi bilo, če bi bili ME namenjeni dejanskemu preučevanju podzavestnih procesov ali globokih mentalnih stanj subjekta, ki izvaja ME, a temu ni tako. Gre za preučevanje scenarija (ali mentalnega modela), vsebovanega na mentalnem stanju (vprašanje ali gre za eno ali več mentalnih stanj bomo na tem mestu zanemarili). Razdelitev zavesti ni potrebna, ker ne preučujemo drugih aspektov naše zavesti, ampak uporabljamo zavest za preučevanje nečesa zunanjega tako, da si to predstavljamo znotraj zavesti. Padajoče telo in fizikalne predpostavke za njim niso na noben način del naše zavesti v takšnem smislu, da bi lahko rekli da gre za preučevanje naše zavesti same.

Nazadnje naj omenim še drugi ugovor, ki ga Sorensen (1992) povzema po Comtu, in sicer ugovor iz nekonsistentnosti. Če še malo vztrajamo pri Galileovih padajočih telesih, si je ponovno težko predstavljati, da bi dve osebi, ki izvajata isti ME pod istimi epistemskimi pogoji prišli do drugačnih (subjektivnih) zaključkov. V kolikor do slednjih dejansko pride, gre za napako pri individualni izvedbi ME, bodisi zaradi predsodka, pomanjkanja znanja, ali kakšnega drugega škodljivega faktorja. A to ne pove ničesar o naravi, sestavi ali veljavnosti ME samega, marveč zgolj odseva kompetentnost izvajalca. Druga zgodba je pri ne-znanstvenih miselnih eksperimentih, kjer pa je subjektiven, nekonsistenten rezultat dejansko željen izid – problem vozička denimo ni namenjen temu, da bi vsak izvajalec prišel do istega zaključka.

2.3. Določeni drugi problemi

V tem delu v celoti povzemamo po Sorensenu (1992), ki v svoji knjigi *Thought Experiments* nudi podrobno razdelavo očitkov ME (v naslednjem odseku pa tudi odgovore na slednje).

Najprej je omeniti problem zvestobe. ME merijo med drugim tudi naše prepričanje o nečem, in ne tistega nečesar samega po sebi. Ta prepričanja lahko vodijo v to, da kadar so predobstoječa prepričanja skupna širšemu krogu izvajalcev ME, lahko ME izvajalce prepriča v svojo veljavnost in pravilnost sklepov, čeprav temu dejansko ni tako. Pojavlja se tudi problem pristranskosti tako individualnega izvajalca, kot družbenih pristranskosti, kar lahko vpliva na sklep, ki sledi iz ME. V izrazito katoliški družbi bo, denimo, rezultat ME s slavnim violinistom mnogo pogosteje v nasprotje splava, kot je bil originalni namen, saj bodo pristaši obveznosti nadaljevanja nosečnosti aktivno iskali načine izpodbijanja analogije med violinistom in zarodkom. Še ena težava je, da so nekateri protidejstveni scenariji tako oddaljeni od (tudi možne) realnosti, da bi bilo za to, da se jih sploh upravičeno obravnava, potrebno vzpostaviti širok, podroben, in morda celo nemogoč kontekst. Z roko v roki z vsem tem gre tudi, da ME pogosto poenostavijo stvari do točke absurda – Sorensen pravi, da na tak način, kot bi obliko Francije obravnavali kot popolnoma šest kotno.

Na podlagi vsega opisanega si lahko zastavimo vprašanje, ali je implikacija, da zamisljivost neke situacije pomeni njeno možnost, popolnoma zgrešena? Sorensen problem podrobno analizira – dovolj podrobno, da je njegova obravnava izven obsega tega članka – vendar pa zaključi, da je zgornja implikacija statistično upravičena; pogosteje si bo nekdo zamislil neverjeten, a možen scenarij, kot pa popolnoma nemogočega.

Pogosto se ME sklicujejo tudi na vsakdanji jezik; vendar pa obstajajo polemike tudi glede tega, ali je tovrsten sklic upravičen. Sklic na vsakdanji jezik loči med empiričnimi vzorci, ki opisujejo opazljive ponavljajoče se pojave, in jih upravičeno induktivno posplošijo, ter logičnimi implikacijami, ki temeljijo na deduktivnem sklepanju iz definicij. Nekateri skeptiki v celoti zanikajo veljavnost sklica na vsakdanji jezik, saj trdijo, da je jezikovno znanje 'navadnega' posameznika omejeno, in je sklic na vsakdanji jezik podobno nesmiseln, kot spraševati slehernika o medicini ali astrofiziki. Sorensen jih odpiše kot radikalne, in se poglobi v zmernejši skepticizem o tem sklicu. Slednji trdi, da je vsakdanji jezik prestavljen preozko ali preširoko. Preozka je na primer definicija angleške besede *bachelor*, ki zaznamuje nekoga, ki je samski po lastni izbiri, a predpostavlja, da je alternativa, da sledi družbenemu standardu, in se poroči. Pri tem izvzema primere samskosti po lastni izbiri iz drugih razlogov, denimo duhovnikov ali pa homoseksualcev v okoljih, kjer je homoseksualnost stigmatizirana ali celo prepovedana. V okviru sklica na vsakdanji jezik pride do 'semantičnega spusta' v miselne eksperimente, pri čemer se vprašanja o stvareh reducirajo na vprašanja o tem, kako govorimo o teh stvareh. V primeru radikalnih predpostavk ME lahko o njih govorimo in jih obdelujemo oz. obravnavamo, vendar pa so brez dejanskega pomena. Obstaja tudi kritika glede informativnosti ME. Carl Hempel (1965) denimo definira, da so lahko ME teoretični, izpeljani deduktivno na podlagi konkretno definiranih principov, ali pa intuitivni. Večina jih po njegovem mnenju leži nekje na tem spektru, a se nagibajo bolj v intuitivno smer – to pomeni, da imajo takšni ME hevristično vrednost pri spoznavanju, ne pa tudi razlagalne vrednosti ali vrednosti v pridobivanju dejanskega znanja. Nadalje trdi, da je miselne eksperimente bodisi mogoče preveriti z empirično eksperimentacijo in jih potrditi, ali pa ne. V kolikor je to mogoče, je ME dejansko nepotreben, saj bi lahko v prvi vrsti izvedli empiričnega; ME tako v splošnem nimajo nikakršne znanstvene vrednosti.

Potrebno je preučiti tudi zmote, ki pritičejo slabim oziroma neveljavnim miselnim eksperimentom. Pogost vzrok neveljavnosti je sklicevanje na mit – slednjega Sorensen označuje kot lažno prepričanje, ki pa ni logična zmota. Biologi pred Darwinom so denimo zagovarjali kreacionizem zaradi korenitega verskega prepričanja in hipotezo podpirali z zapletenostjo živih bitij. ME, ki se sklicuje na mit, bo neveljaven, a Sorensen opozarja, da so tudi EEi lahko osnovani na mitu (tj. lažnem prepričanju), tako so v devetnajstem stoletju izvajale kemijske eksperimente na podlagi hipoteze, da so vsi kemijski elementi sestavljeni iz atomov vodika. ME je lahko neveljaven tudi brez zmote zaradi sklica na mitologijo. Prav tako lahko neveljavnost izhaja iz metodološke zlorabe, kjer lahko denimo filozof uporabi ME kot miselno opornico, ter z njim krožno (ali drugače zmotno) podpre obstoječo hipotezo, ali pa uporabi ME, kjer obstaja bolj zanesljiva metoda dokazovanja ali podpiranja hipoteze, recimo EE, anketna raziskava ali smiseln deduktiven sklep iz zakonov in aksiomov. Pri teh problemih odgovora ne bomo iskali, saj ga bržkone tudi ni. Dovolj je, da sprejmemo, da se ti problemi pojavljajo tudi pri EE in RM, ter niso omejeni na ME.

2.4. O zmotah in protizmotah

Zmotnih ME, kot so recimo ME, ki temeljijo na pristransko zastavljenih hipotezah, zmotno zastavljenih perspektivah, itd., na tem mestu ne bomo vključevali. Razlog za to je precej enostaven: zmoten je lahko prav vsak ME, pa tudi prav vsak EE, se pravi zmotnost ne pove ničesar o veljavnosti ME kot raziskovalne metode na splošno. Prvič jih ne morejo razveljaviti tiste zmote, ki se nanašajo tudi na empirične eksperimente, saj bi morali s tem zanikati tudi veljavnost EE. Drugič, previden raziskovalec lahko preuči, kje se utegnejo pojaviti zmote, in eksperiment osnuje tako, da do zmot ne bo prišlo, s čemer ustvari veljaven ME. Če mu to ne uspe, pa je njegov, in na podlagi tega *samo njegov,* ME neveljaven.

Po drugi strani Sorensen (1992) trdi, da je večina kritik proti ME dejansko protizmot; te so značilnosti ME ali logične strukture, ki na prvi pogled delujejo zmotno, zaradi česar jih mnogi obravnavajo kot zmote, kar pa dejansko niso. Sorensen poglavje odpre s primerom, da se protidejstvena 'kaj če' vprašanja, ki odpirajo miselne eksperimente, pogosto zavračajo na podlagi njihove hipotetične narave, pri čemer implikacija, da so hipoteze neveljavne, v najboljšem primeru temelji zgolj na krožnem argumentu, ki sloni na definiciji hipoteze kot še ne dokazane predpostavke. Seveda naj omenimo da protizmote govorijo v dobro ME in so uperjene proti neutemeljenim kritikam slednjih.

Prvi sklop protizmot so protizmote oddaljenosti, ki se v osnovi sklicujejo na to, da naj bi bila oddaljenost predpostavk, na katerih temeljijo ME, od našega sveta, zmotna. Prvi razlog za tovrstno dojemanje je trivialna interpretacija. V okviru te naj bi bile 'bizarne' predpostavke ME takšne, da so resnične, ko jih interpretiramo trivialno, a sporne, ko jih interpretiramo konkretno. Sorensen pravi, da je potrebno ločiti med trditvijo, da je »dejanje predpostavljanja p bizarno« in idejo, da je »predpostavljen p, ki je bizarna predpostavka« (Sorensen, 1992: 277). Če si predstavljamo bizaren možen svet, v katerem se dogaja nekaj, kar v našem svetu ni mogoče, nismo zagrešili nikakršne zmote – kvečjemu bi jo v primeru, če bi predpostavili, da se nekaj nemogočega dogaja v našem svetu, brez da bi ustrezno prilagodili kontekst.

Naslednja je nemogoča interpretacija. Pri njej večinoma gre za to, da si opazovalec ME napak tolmači, da so njegove predpostavke logično nemogoče, kadar so zgolj praktično nemogoče (o tem smo nekaj povedali že prej, zato na tej točki ne bomo razlagali globlje).

Tretja je interpretacija nedostopnosti, ki se pojavlja večinoma, kadar ME vsebuje izredno nenavadne elemente – pri tem se lahko zanemarja analogijska vloga nenavadnega in nedostopnega scenarija (kot denimo, če bi nedostopnost očitali eksperimentu z violinistom), lahko gre za mnemonične figure, ki pripomorejo vizualizaciji ME, v končni fazi pa bi lahko šlo tudi za neupravičeno absurdne in nedostopne scenarije – a Sorensen poudarja, da je to redek primer, absurdni elementi običajno niso bistveni za končni rezultat eksperimenta.

Sledi retorična verzija oddaljenosti, pri kateri gre v osnovi za lažno obtožbo, da ME vsebuje pristranskost do estetske vrednosti, ki smo jo opisali zgoraj, s čemer se ME razveljavlja na podlagi napačne interpretacije nenavadnih elementov kot nepotrebnih ali neupravičenih. Podoben je tudi ugovor iz nazadovanja, ki ME očita, da lahko s pretiranim poglabljanjem v opisne podrobnosti, ki niso neposredno relevantne za eksperiment, slednjega privedejo do točke, ko gre le še, kot pravi Sorensen, za 'utopičen eskapizem' brez dejanskih implikacij za karkoli v svetu.

Zadnji v tej kategoriji je ugovor iz semantičnega holizma, pri katerem naj bi šlo za jezikovno manevriranje do takšne mere, da je jezik uporabljen v miselnem eksperimentu, z namenom opisa praktično nemogočih in bizarnih scenarijev, tako oddaljen od splošnih komunikacijskih konvencij, da so izjave, ki jih vsebuje, semantično prazne. Redukcija do absurda takšnega primera bi bila, da bi ME opisali tako, da bi z njim preko jezikovnih reinvencij dokazali, da je dva plus dva dejansko pet, tako da bi besedo 'pet' pripisali vrednosti štiri. Sorensen trdi, da je to protizmota zato, ker avtor ME z jezikom ne more možnosti spremeniti v gotovost, v primeru zgornje redukcije bi to pomenilo le, da imamo svet, v katerem sta besedi 'pet' in 'štiri' zamenjani, ne pa tudi v katerem bi veljala kakršnakoli absurdna propozicija, kot da ima denimo 'dva in dva' vrednost pet.

Sorensen nadaljuje še s tremi primeri protizmot, ki pa stojijo izven okvira protizmot oddaljenosti. Prva med temi izoliranimi protizmotami je nekaj, kar Sorensen poimenuje »čudno noter, čudno ven« (angl. *strangeness in, strangeness out*) (Sorensen, 1992: 284). Gre za zakoreninjeno pričakovanje, da morajo biti vzroki kvalitativno povezani z učinki, npr. preprost vzrok mora imeti preprost učinek. Nasprotniki ME tako občasno izolirajo čudno predpostavko ME, jo postavijo v (po nepotrebnem) čuden kontekst, nato pa izpeljejo neveljaven, čuden zaključek, ki sledi predpostavki večinoma po principu krožnosti.

Sledi zmota opazovalca, kjer se izvajalcu ME očita, da s tem, da predpostavlja, da je nek zaznavni čut subjekta v eksperimentu v okvari, dela eksperiment neizvedljiv, kajti če subjekt ne zaznava možnega sveta eksperimenta, ga ne more niti opazovalec. Zakaj je to protizmota in ne dejanska zmota je jasno vidno, če si ogledamo vlogo okvarjenega zaznavnega čuta v miselnem eksperimentu obrnjenega barvnega spektra.

Nazadnje Sorensen predstavi protizmoto Kabuki, poimenovano po japonskem slogu gledališča, v katerem odrska tehnika rekvizite premika medtem, ko predstava teče, kar gledalci, ki s slogom niso seznanjeni, pogosto smatrajo za slabo izvedbo predstave. Gre za zavračanje ME na podlagi 'rekvizitnih' sredstev, ki jih ti uporabljajo – Sorensen pravi, da je takšno zavračanje enako temu, če bi študent fizike zavrnil reševanje naloge, v katerem mora izračunati kot rimskega katapulta, da zadane sovražnikovo obzidje, na podlagi tega, da potovanje po času ni mogoče, ali pa ker ni vojak.

3. Miselni eksperimenti in druge metode

Ta odsek primerja ME z drugimi metodami raziskovanja, ki prav tako temeljijo na posnemanju oziroma modeliranju naravnih pojavov. EE temeljijo na izgradnji fizičnega modela in empiričnem testiranju hipoteze, računalniški modeli (RM) pa na simulaciji naravnega pojava skozi virtualni model. Med znanstveniki na različnih področjih pogosto najdemo pristranskost v prid EE, ki naj bi bili bolj otipljivi, bolj objektivni in bolj zanesljivi, v zadnjem času pa se pojavljajo stališča, ki trdijo, da bodo RM nadomestili ME, saj omogočajo izredno podrobno modeliranje z veliko količino spremenljivk, s katerimi lahko operiramo.

3.1. Empirični in miselni eksperimenti

Da lahko začnemo preverjati, ali so EE res nadrejeni ME, moramo najprej pogledati, ali je ME sploh eksperiment, ter kakšne so podobnosti in razlike med obema kategorijama.

Sorensen (1992) trdi, da so ME eksperimenti v prvi vrsti v taksonomskem smislu, kljub navideznim razlikam med enimi in drugimi – navidezne lastnosti pa niso zadostne za določitev ME kot eksperimentov na podoben način kot fenotip kita ni zadosten za določitev kita kot sesalca, če naj njegov videz primerjamo s tipičnimi primeri kopenskih, štirinožnih, kosmatih sesalcev. Avtor ME primerja z miselnim šahom, pri katerem igralca ne posedujeta dejanske šahovnice, marveč igrata zgolj z besednimi oznakami premikov figur. Pri tem metodologija ni pomembna – ni pomembno namreč, ali obstaja neka Platonska šahovnica, na kateri se udeležujejo poteze obeh igralcev (Brown, 2011), ali pa oba igralca izvajata mentalno simulacijo igre – relevantno je samo, da igra poteka kljub temu, da ne poteka v klasični, fizični obliki. Temu je analogen tudi potek ME.

Tako pri miselnih kot empiričnih eksperimentih, pravi Sorensen, naivni induktivizem ne vzdrži. Rezultat vsakega eksperimenta so »surovi podatki, ki se uporabljajo za preverjanje teorij« (Sorensen, 1992: 233), vsakršna induktivna posplošitev rezultate enega miselnega ali empiričnega eksperimenta na (novo) teorijo pa je zmotna. Naslednji skupni lastnosti Sorensen pravi 'brkljanje' – tako kot se EE pogosto reproducirajo z majhnimi prilagoditvami v instrumentih, spremenljivkah ali postopku, da bi se izpostavile njihove pomanjkljivosti ali vključile nove hipoteze reprodukterja, se tudi v miselnih eksperimentih prilagaja kontekst, spremenljivke, postopek itd. To velja tudi za filozofske miselne eksperimente: številni utilitaristični in proti-utilitaristični ME, od problema vozička do loterije preživetja, so 'prebrkljane' oblike enega in istega osnovnega eksperimenta, ki v osnovni obliki primerja vrednost enega proti več življenj z različnimi permutacijami. Slednje nakazuje tudi na to, da imajo ME nek standarden format, ki navadno zaobjema kontekst, specifičen problem in akterjevo dejanje v okviru tega problema – tudi EE imajo standarden format, ki obsega med drugim postavitev opreme, vpeljavo in izključevanje spremenljivk, ter sprožitev in spremljanje poteka eksperimenta. Tako EE kot ME vodijo do nepričakovanih učinkov: Louis Pasteur je denimo skozi dokazovanje, da življenje ne nastane spontano v izoliranem okolju, pomotoma dokazal tudi, da je razkroj biološki in ne kemijski postopek. Podobno je Gilbert Harman (1986) pri poskusu podrobnejše razlage Gettierjevega ugovora na teorijo znanja kot upravičenega resničnega prepričanja prišel do tihe premise, da so protidokazi, katerih se ne zavedamo, prav tako ovira pri pridobivanju znanja iz poznane evidence, ki so jo kasneje pograbili mnogi epistemologi (še več, njegov ME se v kontekstu njegovega prvotnega namena smatra kot neuspešen).

Sorensen nadalje trdi, da je večina razlik med miselnimi eksperimenti in empiričnimi eksperimenti, ki bi prve utegnila diskvalificirati, zgolj navideznih. Prva med njimi je, da se ME ne replicirajo, kot se to počne z empiričnimi eksperimenti. Sorensen pravi, da to kratko malo ni res, pri čemer ni potrebna nobena argumentacija, marveč moramo le pogledati »petindvajset let napredovanja Gettierjevih protiprimerov, petnajst let vse bolj zapletenih problemov vozička, ali zadnjih deset let primerov Zemlje dvojčice (Sorensen, 1992: 249). Prav tako ne drži, da ME ne upravljajo s spremenljivkami – zmotna razlika izhaja iz tega, da je upravljanje spremenljivk v miselnih eksperimentih apriorno, kot denimo šahisti upravljajo s spremenljivko nasprotnikove naslednje poteze. Nadalje ne drži niti, da ME niso kvantitativni. Nekateri so eksplicitno kvantitativni, denimo Fisherjev eksperiment (1930), s katerim razlaga uravnoteženo naravno razmerje med številom pripadnikov obeh splov s tem, da si predstavlja variabilna razmerja (tj. svet v katerem je znatno več moških ali znatno več žensk). Spet drugi so kvantitativni, a apriorni, ker se nanašajo na neskončne (ali neskončno majhne) količine, npr. Heronov eksperiment (v Sorensen, 1992: 250), da se predmet na površini brez trenja prične premikati, ko površino nagnemo za poljubno (neskončno) majhen kot.

Razlike, ki dejansko obstajajo med enimi in drugimi eksperimenti gredo v prid miselnim eksperimentom. Noben miselni eksperiment namreč ne zahteva več kot enega eksperimentatorja, kar preprečuje tudi spore glede hierarhije med eksperimentatorji, kot tudi noben ME ne loči med teoretikom in izvajalcem eksperimenta. Rezultati empiričnih eksperimentov so lahko izprijeni zaradi sreče (ali nesreče), če pride do naključne spremembe neodvisne spremenljivke ali česa podobnega – to pri miselnih eksperimentih ni mogoče iz popolnoma očitnih razlogov. ME tudi ne potrebuje opreme in končno, ME so moralno trivialni. EE so lahko predmet etike, ko so za njihovo izvedbo potrebni živalski ali človeški testni predmeti.

Nadalje želimo pokazati, da je ena izmed podobnosti tudi ta, da EE enako enostavno podležejo nekaterim težavam opisanim v odseku 2.3. kot ME. Zgoraj smo povzeli, da se v ME utegne primeriti, da izvajalec izpelje nepravilne sklepe, a zaradi predobstoječih prepričanj ali gole želje po rezultatih verjame v pravilnost tega sklepa – a to še zdaleč ni omejeno na ME. Šimundić (2013) denimo opisuje 'mučenje podatkov' (ang. *torturing the data*) v medicinskih raziskavah, kar pomeni deljenje rezultatov v različne podkategorije, dokler ne postanejo statistično pomembni za svojo podkategorijo, a so v splošnem nepravilni in praktično popolnoma neuporabni. Poleg tega primera avtorica opisuje tudi druge pristranskosti v zbiranju, obdelavi in interpretaciji podatkov v EE, ki lahko vodijo v neveljavnost eksperimenta. Pojavljajo se tudi implicitne pristranskosti, ki so drugače motivirane, denimo to, da se za raziskave na živalih pogosto uporabljajo samo samci, saj so ti cenejši in pogosto manjši in manj agresivni, kar vodi v kvalitativno nepopolne rezultate, saj se lahko pogosto samice na iste spremenljivke odzivajo popolnoma drugače (Wald in Wu, 2010). Tudi poenostavitev ni omejena na ME, saj je v mnogih laboratorijskih eksperimentih, povezanih z biologijo, fiziko, in ostalimi 'trdimi' znanostmi pogosto nemogoče poustvariti pogoje, v katerih bi se nek postopek ali pojav odvijal v naravi.

3.2. Računalniško modeliranje in miselni eksperimenti

Zgoraj smo omenili, da je računalniško modeliranje (naprej RM) in izvajanje računalniških simulacij na teh modelih metoda, ki se z napredkom računalniške tehnologije vse pogosteje uporablja kot dopolnitev, pa tudi kot nadomestilo EE. Ta za zgodovino znanosti relativno nova metoda ponuja možnost poustvaritve izjemno kompleksnih naravnih pojavov na mikro in makro nivoju, od simulacije trkov delcev in zgradbe molekul, pa do modelov, ki simulirajo delovanje črnih lukenj, razvoj galaksije, itd. Ob tem omogoča tudi vnos velike količine spremenljivk in selektivno upravljanje slednjih. Pomislek, da utegnejo RM zamenjati tako ME kot tudi EE, se tako pogosto pojavlja in na prvi pogled ne izgleda zgrešen.

V kakšnem odnosu pa so ME in RM? El Skaf in Imbert (2012) opredelita ME kot način raziskovanja konceptualnega orodja in razvijanja teoretične podlage, RM pa kot načine zagotavljanja teoretičnih razlag. EE, po drugi strani, le malo pripomorejo k tema dvema funkcijama. Vendar pa avtorja poudarjata, da je osnovna funkcionalna shema vseh treh metod popolnoma enaka, kar opišeta v sledečih korakih (povzeto):

1. Osnovna dejavnost, usmerjena v odgovarjanje na vprašanja.
2. Vključitev scenarija: dejanskega ali izmišljenega, ki vsebuje zakone fizike, pravila, zunanje spremenljivke, ipd. parametre. Scenarij je vrstni opis (ang. *Type-description*), ki vključuje določeno količino podatkov o prikazani situaciji ali pojavu.
3. Razvoj scenarija: preiskava vsebine scenarija skozi analitično logiko, numerično analizo, fizični proces, ali kakšno drugo metodo. Lahko gre za računalniške (RM), mentalne (ME), fizične (EE), ali podobne procese.
4. Rezultati scenarija: relevantni, ne-poljubni podatki, ki lahko vodijo v odgovore na začetna vprašanja.
5. Znanstveni zaključki skozi interpretacijo rezultatov. (El Skaf in Imbert, 2012: 3454-3455)

Na tovrstno funkcionalno izomerijo se osredotoča tudi Arcangeli (2018), ki predvsem poudarja, da se za opredelitev razmerja med ME in RM pogosto pretirano zanašamo na problem identitete, tj. ali so RM vrsta ME, ME vrsta eksperimenta na splošno itd., kar vodi v pristransko preferenco do binarne analize in zmotno stališče da so RM nujno vrsta ME in jih bodo nadomestili kot neke vrste 'naslednja generacija'. Prednost stališča v smeri funkcionalne izomerije je tudi da, da se osredotoča na postopek prej kot na rezultate, kar preprečuje tudi tvorbo pristranske preference do EE nad ME (in RM) kot spoznavno nadrejenih.

Beisbart (2012) pa denimo išče podobnosti med ME in RM s pomočjo Nortonove (1996) argumentacijske hipoteze; tj. miselni eksperiment je logični argument, ki je zanesljiva metoda znanstvenega raziskovanja v kolikor je ta argument zdrav. Po Breisbartovi analogiji je algoritem, ki vodi simulacijo na RM, enostavno serija logičnih izjav v programskem jeziku, končni rezultat pa sklep, ki sledi iz teh izjav. Ponovno gre za vrsto funkcionalne izomerije, le da izhaja iz drugega izhodiščnega pristopa.

Nujno pa potrebujemo tudi opredelitev razlik med RM in ME; namreč če razlik ni in funkcionalna izomerija preraste v funkcionalno identiteto, potem lahko ponovno trdimo, da RM *je* oblika ME, in da lahko torej v celoti nadomesti ME kot njihova sodobnejša različica.

Arcangeli (2018) identificira osnovno strukturno razliko: avtorica trdi, da vsake vrste eksperiment poteka na dveh nivojih: proizvodnja (kar vključuje model, obravnavo oz. postopek in rezultat) in prestavitev (obdelava podatkov in interpretacija obdelanih rezultatov). Medtem, ko je ta razlika pri RM in EE jasna, saj se proizvodnja izvaja v sklopu fizičnega ali računalniškega postopka, je v ME ločevanje med nivojema mnogo bolj zapleteno, saj oboje poteka *in mente* izvajalca. Razlika je zgolj v tem, kdaj se izteče mentalna simulacija in začne obdelava iz nje pridobljenih rezultatov.

El Skaf in Imbert (2013) ilustrirata razlike med vsemi tremi obravnavanimi metodami na primerih poiskusov empiričnih in računalniških poustvaritev Maxwellovega demona. Pred samimi izvedbami sta bila že v prvi polovici prejšnjega stoletja predlagana dva praktična modela. Smoluchowski je denimo predlagal vratca na vzmet, ki se premaknejo, kadar se pritisk nanje poviša zaradi gostote delcev, kar sproži lastno Brownovsko gibanje vratc, zaradi česar spustijo delce v 'napačno' smer. Feynman (Feynman et al., 1977) je predlagal nekoliko bolj zapleten mehanizem z zobatim kolesom in zaskočno vzmetjo, ki sta nastavljena tako, da se sistem lažje zasuče v 'napačno' smer in prenese hladne (počasne) delce v toplo komoro.

V letu 1997 Kelly in drugi izvedejo EE, ki vključuje kompleksno sintetično molekulo, ki se v sistemu obnaša kot Feynmanov sistem z zaskočno vzmetjo. Molekulo zaradi strukture dela, ki predstavlja zaskočno vzmet, manjši delci lažje zavrtijo v smer urinega kazalca kot v nasprotno smer. Ko so molekulo postavili tako, da so jo počasnejši delci premikali v smer urinega kazalca, so res zabeležili nekaj nepričakovanih energetskih maksimumov, ko so delci prekršili drugi zakon termodinamike in prešli v komoro s hitrejšimi delci, vendar pa se je izkazalo, da povprečna razporeditev ostaja enaka kot v naravnem sistemu. Skordos in Zurek (2003) ustvarita RM Smoluchowskijevega predloga, na podlagi katerega nato izvedeta simulacijo Maxwellovega ME. Zaključek je enak kot pri opisanem EE – čeprav nekatere počasnejše molekule preidejo v drugo sobano in na mikroskopskem nivoju prekršijo drugi zakon termodinamike, je povprečna razporeditev naravna, torej na makroskopskem nivoju zakon ni prekršen. To potrjuje tudi hipotezo, ki jo predpostavljajo Maxwell in kasnejši izvajalci njegovega ME – da je zakon zgolj statistično veljaven, a je statistično veljaven absolutno, se pravi na makroskopskem nivoju ga ni mogoče prekršiti.

Chandrasekharan in drugi (2012) izpostavljajo, da nam RM in ME oboji nudijo nova znanja brez novih empiričnih podatkov, kot tudi, da oboji preverjajo protidejstva (ang. *counterfactuals*) – ME na podlagi konkretnih (mentalno modeliranih) elementov, RM pa na podlagi spremenljivk, vpeljanih v simulacijo – a beležijo pomembne razlike. Glavna izmed teh je, da so RM manj transparentni kot ME in zahtevajo več sprotnega vpogleda, nadzorovanja in spreminjanja mehanizma modela in simulacije. Kljub temu pa obstajajo mnoge prednosti RM – ta deluje poljubno, a nudi nove podatke, s pomočjo katerih je mogoče izpopolniti EE. Vzemimo primer preučevanja umetno vzgojenega nevronskega omrežja *in vitro.* V podrobnosti eksperimenta se ne bomo podajali; pomembno je, da sta vzporedni izvedbi istega osnovnega eksperimenta kot EE in kot RM prikazali, kako lahko v RM preverjamo učinek vpeljave poljubnih novih spremenljivk, dokler te ne privedejo do želenega rezultata. To nam omogoči da taiste spremenljivke vpeljemo v EE, ter hitro in zanesljivo potrdimo ali ovržemo ali res privedejo do željenega rezultata. Raziskovanje v RM se začne že pri izgradnji modela, saj je slednjega mogoče dopolniti z raznimi elementi, ki v preteklih teoretičnih ali empiričnih raziskavah vodijo k želenim rezultatom, ter s tem povečati možnost uspeha tako samega RM kot tudi morebitnega EE, ki ga RM dopolnjuje. Če gre za dodajanje poljubnih elementov, lahko model rekonstruiramo z manjkajočimi elementi in identificiramo tisti specifični element ali kombinacijo elementov, ki je privedla do želenih rezultatov.

RM lahko obravnavajo kompleksnejše naravne pojave in nudijo razlago in vpogled v podrobnosti teh pojavov do mere, do katere ME ne pridejo. Chandrasekharan et al. (2012) opisujejo relacijo med ME in RM s sklepom, da RM omogočajo eksternalizacijo kritičnega mišljenja v znanosti in širijo kognitivno kapaciteto za mentalno simulacijo tako, kot teleskopi širijo zaznavno kapaciteto za videnje. To pomeni, da RM ne nadomestijo ME, temveč jih pomembno dopolnjujejo. ME še zmeraj služijo kot podlaga za osnovno idejo pri izgradnji RM, preden so ti dejansko izgrajeni v računalniškem sistemu, ter sodelujejo pri selektivni vpeljavi spremenljivk – ME predvidi katera spremenljivka bi utegnila pripeljati do želenega rezultata, RM to teoretično preveri na podrobno simuliranem scenariju, z EE dodatno preverimo empirično. ME prav tako še zmeraj igrajo vlogo pri interpretaciji rezultatov RM in EE.

Mnogi so glede vloge ME mnogo bolj optimistični. El Skaf in Imbert (2013) denimo zaključujeta, da dejstvo, da so ME, RM in EE v primeru makroskopskih implikacij Maxwellovega demona z različno metodologijo privedli do istih rezultatov pomeni, da so vse tri metode znanstveno relevantne in funkcionalno izmenljive. Pri tem opozarjata, da funkcionalna izmenljivost ne pomeni epistemološke izmenljivosti; ker ima vsaka izmed zgornjih metod svoj postopek sklepanja in upravičevanja vsaka od njih nudi vpogled v podatke o predmetu raziskave, ki z drugima dvema metodama niso dostopni. Medtem, ko ME lahko še naprej služijo kot samostojna raziskovalna metoda je njihova dodatna uporabnost tudi njihova vloga v pripravi in interpretaciji RM in EE, podobno kot trdijo tudi Chendrasekharan et al. (2012).

Dodatna prednost ME je tudi to, da jih lahko ponovi kdorkoli, tako na nivoju proizvodnje kot na nivoju predstavitve, brez da bi zato potreboval kakršnokoli računalniško ali laboratorijsko opremo – zaradi tega dejstva samega po sebi je miselni ME nenadomestljiv (Arcangeli, 2018).

Najpomembnejša prednost ME pa je verjetno ta, ki jo opisuje Beisbart (2012). Avtor trdi, da so RM kvantitativno široki in kvalitativno ozki (podrobno obdelajo veliko količino spremenljivk in mikro-scenarijev), ME pa kvantitativno ozki in kvalitativno široki (v celoti obdelajo en splošen scenarij). Uporaba obeh metod se zelo razlikuje, ME bo namreč za določene vrste scenarijev vedno bolj smiselna metoda – »noben znanstvenik ne bi izvajal računalniške simulacije za pridobitev rezultata, ki ga lahko nemudoma izpelje v mislih« (Beisbart, 2012: 426). ME je edinstven tudi v tem, da lahko obravnava tako splošne scenarije, da njegovi rezultati privedejo do konceptualnih sprememb, za kar so RM prezapleteni in premalo transparentni; težko si je denimo predstavljati, da bi neko poljubno premikanje spremenljivk v RM lahko privedlo denimo do Galileove teorije o padajočih telesih ali EPR (Einstein-Podolski-Rosen) ugovora na kopenhagensko interpretacijo kvantne fizike (Beisbart, 2012).

Na koncu tega sklopa želimo na kratko pokazati tudi, da tudi RM niso imuni na težave iz odseka 2, enostavno zato ker se tovrstna negativna kritika rada pojavlja vzlic pozitivnim upravičitvam njihove veljavnosti in uporabnosti. Poprej smo omenjali pristranskosti, ki se pojavljajo v ME in EE – tovrstne pristranskosti se pojavljajo tudi v RM. V etiki digitalizacije se sicer primarno ukvarjamo z algoritemsko diskriminacijo, a tudi v znanstvenih RM pride do slednji analognih pojavov. V okviru algoritemske diskriminacije Hacker (2017) opisuje pristranske učne podatke in neenakopravne temeljne resnice. Pri slednjih gre za parametre, vnesene v algoritem, ki izhajajo iz zmotnih posplošitev – v primeru RM bi lahko šlo denimo za vnos osnovnih parametrov, ki temeljijo na zmotnih izsledkih EE; vnesli bi lahko denimo podatke pridobljene skozi študije primarno odraslih samcev laboratorijskih podgan (na primeru EE smo to problematiko omenili zgoraj), medtem ko bi bil naš RM usmerjen v simulacijo nekega pojava na celotni podganji populaciji (tj. tudi samic in mladičev), kar bi lahko pripeljalo v nepopolne rezultate. Pristranski učni podatki, po drugi strani, so podatki, ki jih uporabljamo za učenje programov, ki izvajajo RM. V kolikor so učni podatki pristranski, bo potencialno vsak RM, ki ga izvaja algoritem s temi učnimi podatki, vodil v neveljavne rezultate. Prav tako lahko selektivno vpeljevanje spremenljivk, ki vodijo v želene rezultate, vodi v podobne pojave, kot je prej opisano 'mučenje podatkov' v EE; tj. primeri se lahko, da v RM vpeljemo poljubno kombinacijo spremenljivk, ki vodijo v rezultat, podoben želenemu, ki ga nato interpretiramo tako, da ustreza naši hipotezi.

4. Ne-znanstveni miselni eksperimenti

Nobena izmed podobnosti z EE, ki jih povzemamo po Sorensenu v odseku 2, se ne nanaša izključno na znanstvene miselne eksperimente, marveč tudi na filozofske. Pravzaprav sta obe obliki neodtujljivo povezani, kar lepo ilustrira Shoemakerjev primer (v Sorensen, 1992): Shoemaker (v Sorensen, 1992: 11–12) je v prvi vrsti študiral nevrofiziologijo ljudi, ki so preživeli izgubo ene možganske poloble, v okviru česar je izvajal tako empirične kot (znanstvene) miselne eksperimente, vendar pa je njegov najznamenitejši ME ta, v katerem se vpraša, kaj bi se zgodilo, ko bi gospodu Smithu izrezali desno polovico možganov, in jo presadili v novo telo. Shoemakerjeve pretekle raziskave so pokazale, da bi lahko oba preživela relativno normalno, a vprašanje je, kdo izmed njiju bi bil gospod Smith? Gre za vprašanje identitete, ki je eksplicitno filozofske narave, a je tesno povezano z oziroma celo izpeljano iz znanstvenega ME. Kljub temu pa nas v tem odseku prej zanima status ME, ki so eksplicitno filozofski, kot sta denimo analogija z violinistom in tančica nevednosti.

Filozofski ME niso nič bolj občutljivi na probleme iz odseka 2.3 kot znanstveni ME, saj uporabljajo isto metodologijo, vsebujejo le drugačne protidejstvene scenarije (Sorensen, 1992; Brown, 2011). Pravzaprav so na nekatere izmed teh problemov manj občutljivi: v mnogih filozofskih ME je denimo subjektivna introspekcija želena, ne pa problematična. Rawlsova tančica nevednosti zahteva preučitev lastnega, subjektivnega stališča do možnih izidov; prav tako denimo analogija z violinistom zahteva preučitev subjektivnega stališča z introspekcijo. Posplošitev in objektivni zaključki sledijo šele na nivoju interpretacije rezultatov.

Vprašanje, ki ostaja je: kakšen je odnos med ME in EE ter RM. Jasno razvidno je, da so mnogi politični in etični ME neizvedljivi v empirični obliki. Tančica nevednosti je denimo praktično neizvedljiva, analogija z violinistom pa bi lahko bila izvedljiva, a bi bila izvedba tovrstnega eksperimenta neetična. Kar pa se tiče ideje filozofskih RM pa je zadeva nekoliko kompleksnejša. Shulzke (2014) vidi distopične video igre kot vrsto političnih RM. Avtor trdi, da lahko skozi video igre odlično simuliramo korupcijo, kriminal, 'zle' institucije, zmotne in zatiralske ideologije ter režime, itd. Dinamika in razsežnost sveta v video igrah omogočata, da lahko simuliramo vsak aspekt družbe z nekim distopičnim elementom; npr. kako tiranija vpliva na življenje v višjih političnih krogih, na podeželju, v *slumih* mest, itd. Vse to omogoča igralcu, da kritično razmisli o svetu, vsebovanem v video igri, in se do njega opredeli.

Vendar pa video igra ni dejanski RM političnega ali etičnega scenarija. Funkcionalna shema igre je drugačna od funkcionalne sheme RM. V RM vnašamo spremenljivke in preverjamo, katere vodijo do želenega rezultata. V video igrah so edine sproti (v razvoju scenarija) vnesene spremenljivke dejanja igralca, ki pa vplivajo na izid igre na enoznačen, vnaprej določen način – izid video igre ne more voditi k novemu znanju. Poleg tega leži težava tudi v tem, da je primarni namen video igre zabava, torej je popolnoma legitimno, da je želeni izid nekega posameznega igralca popolna distopija, tiranija itd. Analogno s tem bi bilo, da v znanstveni RM, ki išče neko sestavino v zdravilu za raka, vnašamo spremenljivke, za katere upamo, da bodo pospešile razvoj raka.

Prav zaradi tega, ker večina filozofskih ME kliče po introspekciji in subjektivni interpretaciji scenarija, je te scenarije najverjetneje nemogoče simulirati s pomočjo RM, saj ti preveč eksternalizirajo dan scenarij. Poleg tega je težko predstavljivo, kako bi se denimo etični RM izvajal samostojno, kot se znanstveni RM; kaj natanko bi bil učinek simulacije, v kateri voziček Philippe Foot znova in znova povozi enega človeka? Kako bi to doprineslo kakršnokoli novo znanje? Zaradi njihove odvisnosti od človeškega faktorja introspekcije in subjektivne interpretacije brez težav trdimo, da so etični ME metodološko omejeni na ME.

5. Sklep

V tem prispevku smo pokazali, da raba intuicij ni nujno metodološko slaba in zmotna, čeprav obstajajo mnoga tovrstna stališča. Preprečiti je treba zgolj zlorabo intuicij, kar je dosegljivo z zastavitvijo določenih pogojev za izvajanje ME in rabo intuicij – primeroma smo izpostavili Virvidakisov (2011) kantovski pristop, vendar pa ta brez dvoma ni edini, ki lahko pomaga preprečiti intuicijske zmote pri izvedbi ME.

Nadalje smo trdili, da so ugovori proti ME, ki temeljijo na problematiki introspekcije sami po sebi neupravičeni, saj se opirajo na preozko definicijo introspekcije in (zmotno) predpostavljajo, da introspekcija nujno pomeni subjektivnost.

Izpostavili smo nekaj drugih problemov, ki se pogosto očitajo ME, vendar ugotavljamo, da se ti problemi ali njim analogni problemi prav tako enostavno očitajo EE, saj gre za inherentne probleme, ki izvirajo iz človeškega faktorja pri vsaki raziskovalni metodi, in jih je potrebno odpraviti na nivoju raziskovalca, ne na nivoju metode.

Poleg tega smo po Sorensenu (1992) povzeli, da je marsikatera zmota, ki se očita ME, v bistvu proti-zmota oziroma psevdo zmota, se pravi gre za nek veljaven element ME, ki pa se smatra kot zmoten zaradi njegove abstraktne oz. na videz neznanstvene narave, oziroma zaradi pristranskosti kritika proti ME.

Preučili smo relacije med ME in drugimi metodami ter ugotovili, da se EE v veliki meri soočajo z enakimi problemi kot ME. Pri tem ne zanikamo, da obstajajo nekatere težave, ki so specifične zgolj za metodo ME, a na drugi strani obstajajo tudi določene težave, ki so specifične samo za EE. Tako lahko trdimo, da gre za enakovredni metodi, vsaka izmed katerih ima svoje prednosti in slabosti. Pri izvajanju obojih je za pridobitev relevantnih in veljavnih rezultatov potrebna ista mera previdnosti in odgovornosti.

Podobno velja za ME in RM: slednji so bolj primerni za nekatere podrobne, ozko usmerjene znanstvene raziskave, medtem ko so prvi boljša metoda za širše, konceptualno usmerjeno raziskovanje. Prav tako ME igrajo pomembno vlogo pri izgradnji in interpretaciji RM. Poudarek si zasluži tudi ugotovitev, da so RM primerni samo za znanstvene eksperimente, ne pa tudi za politične ali etične, ki ostajajo strogo v domeni ME.

The Future of Thought Experiments

The paper addresses the questions of current (contemporary) and future use of thought experiments as a method of scientific research. To do so we must first address the overall validity of thought experiments, which is done by examining criticisms that aim to discredit thought experiments' scientific and epistemic value and overviewing responses to those criticisms as well as other defences in favour of thought experiments, such as the notion of anti-fallacies.

After doing so, we compare thought experiments to other research methods, namely empirical experiments and computer models. We will argue that empirical experiments are faulty to at least the same degree as thought experiments, while the latter have certain procedural properties that allow them to provide data that is inaccessible by the other two methods and therefore retain their status as a valid, independent research method.

Keywords: thought experiments, empirical experiments, computer modelling, research methods

Literatura

Arcangeli, M. (2018). »The Hidden Links between Real, Thought and Numerical Experiments« *Croatian Journal of Philosophy,* 18(52), str. 3–22.

Beisbart, C. (2012). »How can computer simulations produce new knowledge?« *European Journal for Philosophy of Science*, 2, str. 395–434.

Brown, J. R. (2011). *The Laboratory of the Mind.* New York: Routledge.

Chandrasekharan, S., Nersessian, N. in Subramanian, V. (2012). »Is This the End of Thought Experiments in Science«. V Frappier, M., Mynell, L. in Brown, J. R. (urd.): *Thought Experiments in Science, Philosophy, and the Arts.* London: Routledge, str. 239–260.

Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting.* Cambridge, MA: MIT Press.

Descartes, R. (1641). »Meditations on First Philosophy«. *Internet Encyclopedia of Philosophy* (16. junij 2021). URL = https://yale.learningu.org.

El Skaf, R., & Imbert, C. (2012). »Unfolding in the empirical sciences: experiments, thought experiments and computer simulations«. *Synthese,* 190(16), str. 3451–3474.

Feynman, R. P., Leighton, R. B. in Sands, M. (1977). *The Feynman Lectures on Physics, volume 1.* Boston: Addison-Wesley Publishing.

Fisher, R. A. (1930). »Sexual Reproduction and Sexual Selection & Natural selection and the sex-ratio«. V Fisher, R. A.: *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press.

Harman, G. (1986). »Moral Explanations of Natural Facts - Can Moral Claims Be Tested Against Moral Reality?« *Southern Journal of Philosophy*, 24 (priloga.).

Hempel, C. (1965). *Aspects of Scientific Explanation.* New York: Free Press.

Kant, I. (1998). *Critique of Pure Reason.* Cambridge: Cambridge University Press.

Kelly, T. R., Tellitu, I. in Sestelo, J. P. (1997). »In search of molecular ratchets« *Angewandte Chemie International Edition in English,* 36 (17), str. 1866–1868.

Norton, J. D. (1996). »Are thought experiments just what you thought?« *Canadian Journal of Philosophy,* 26, str. 233–366.

Shulzke, M. (2014). »The Critical Power of Virtual Dystopias« *Games and Culture,* str. 1–20.

Skordos, P. A. in Zurek, W. H. (2003). »Maxwell's demon, rectifiers, and the second law: Computer simulation of Smoluchowski's trapdoor«. *American Journal of Physics*, 60(10), str. 876–882.

Sorensen, R. A. (1992). *Thought Experiments.* Oxford: Oxford University Press.

Šimundić, A. (2013.) »Bias in research«. *Biochemia Medica,* 23 (1), str. 12–15.

Unger, P. (1984). *Philosophical Relativity*. Minneapolis: University of Minnesota Press.

Virvidakis, S. (2011). »On Kant's Critique of Thought Experiments in Early Modern Philosophy«. v Ierodiakonou, K. in Roux, S. (urd.): *Thought Experiments in Methodological and Historical Contexts.* Leiden: Brill, str. 127–144.

Wald, C. in Wu, C. (2010). »Of Mice and Women: The Bias in Animal Models«. *Science*, 327(5973), str. 1571–1572.

Aljoša Toplak

Univerza v Ljubljani

Kočljivi primer Wittgensteinovega paradoksa

V tem članku bom predstavil nov primer, skozi katerega lahko preučujemo in predvsem poučujemo t. i. Wittgensteinov paradoks.[1] Ko ga je pred skoraj 50 leti formuliral Saul Kripke (1982), se je zanašal na precej abstraktne primere paradoksa, s čemer je seveda želel pokazati, da še idealizirani matematiki niso varni pred problemom, ki ga zastavlja. S tem pa ni uspel poudariti, kako konkreten in vsakdanji je za človeka pomen tega paradoksa. Dalje pa je relevantna literatura precej suhoparna in paradoks kar kliče po temu, da nanj »obesimo nekaj mesa«. To dvoje je namen tega članka.

Ključne besede: Kripke, Wittgenstein, paradoks, filozofske raziskave, sledenje pravilom.

1 Osramočena komentatorja

Tale prispevek bo morda všeč navdušencem nad nogometom, vsekakor pa ni mišljeno, da bi bilo razumevanje ideje odvisno od poznavanja nogometa. Eno izmed številnih pravil igre bo namreč služilo kot pisan prikaz Wittgensteinovega paradoksa. Skratka, da kar preidemo k naši zgodbi, leta 2010 se je v Južnoafriški republiki odvijalo svetovno prvenstvo v nogometu. Domača ekipa se je na otvoritveni tekmi pomerila z Mehiko, tekmo pa je zaznamoval incident v 37. minuti, ko je sodnik gostom razveljavil gol. Komentatorska veterana, Martin Tyler in Efan Ekoku, sta bila osupla – sodnik je gostujočemu napadalcu sodil prepovedani položaj, ampak vendarle, sta pravila komentatorja, ni bilo nikakršnega prepovedanega položaja! Milijonom gledalcem po vsem svetu sta razglasila, da je sodnik storil neverjetno napako[2]. »Obrambni igralec je stal na črti gola!« je klical Ekoku. »Kako bi prišlo do prepovedanega položaja, ko pa je obrambni igralec stal na črti gola!« Mar sodnik sploh ne ve kaj to pomeni, da je nekdo v prepovedanemu položaju? Kako je lahko tako zgrešil svojo sodbo? »Neverjetno, popolnoma neverjetno!« je dodal komentator. »Sodnik jim je ukradel tekmo.«

Tukaj pa pride do zadrege. Izkazalo se je namreč, da je bil v resnici komentator tisti, ki se je motil. Tako se je postavilo vprašanje – a komentatorja sploh vesta kaj je to, prepovedani položaj? Ju

[1] V literaturi tudi Kripke-Wittgensteinov, Kripkensteinov ali preprosto paradoks sledenja pravil. Odlomke iz Kripkejevega vplivnega dela, v katerem je formuliral idejo, najdete v *Analiza*, 15(4) (2011).

[2] Komentatorsko dvojico je prenašala ameriška platforma ESPN, relevanten izsek si lahko ogledate na kanalu YouTube pod senzacionalističnim naslovom »Best referee decision ever ! commentators gone wrong !!« ('farham salam', 2012).

desetletja ekstenzivnega ukvarjanja s tem športom niso naučila dejanskih pravil nogometa? Kako sta lahko tako zgrešila svojo sodbo?

2 Učenje pravil

K primeru se bomo vrnili. Še prej se nam pojavi vprašanje – kako se učimo pravil? Poglejmo nekaj odgovorov na to vprašanje. Sveti Avguštin v svojih *Izpovedih* opisuje učenje jezika[3] z naslednjimi besedami: »Ko so odrasli imenovali nek predmet in se hkrati obrnili k njemu, sem to zaznal in dojel, da je bil predmet poimenovan z izgovorjenim /…/« (Wittgenstein, 2014: §1). S tem stavkom Avguštin na kratko povzema idejo, da se skozi uporabo jezika učimo, kaj besede označujejo. Ampak mu lahko verjamemo, da je že pri prvemu poskusu »zaznal in dojel« kaj so mu odrasli želeli povedati?

Da bi pokazal, kako učenje jezika le ni tako preprosto, je Quine (1960) skoval miselni eksperiment, ki ga lahko povzamemo z naslednjo (vsekakor izmišljeno!) anekdoto: ko so prvi Evropejski kolonisti prišli v Avstralijo, so srečali aborigine ter pokazali na čudnega sesalca, ki v trebušnemu žepu nosi mladička in skače naokoli. Vprašali so jih, kaj je to. Aborigini so rekli »kenguru«. In pod tem imenom danes to žival vsi poznamo. Ampak, dodaja anekdota, beseda v jeziku aboriginov pomeni »ne razumem«.

Poanta miselnega eksperimenta je, da ko odrasli okoli Avguština pokažejo na stol in rečejo neko besedo, ni že samo po sebi jasno, ali beseda na splošno označuje predmet, ali pa označuje njegovo barvo, obliko, trdnost, umetelnost, ali pa morda gre za označevanje lastnine v smislu »moje,« ali »nedotakljivo!«. Avguštin bo potreboval mnogo več kot osamljeno besedo, da bo dojel tisto, kar so mu odrasli želeli povedati. Beseda v izolaciji nima pomena, kot ugotavlja de Saussure (1959), saj lahko hkrati označuje vse ali nič, dokler pač ni drugih besed (in z njimi celotnega jezikovnega sistema), ki bi jo dopolnjevale. Ko tistemu okroglemu sadežu rečem »jabolko«, zares ne vem na kaj se beseda nanaša, če rečem »jabolko« tudi stolu in nebu. Treba je razumeti, na kaj se beseda *ne* nanaša. Ne na oranžen citrus, saj je ta »pomaranča«, in ne na podolgovat sadež, ki je »hruška«. Samo, ko bom nehal vsemu praviti »jabolko«, bo beseda pridobila nek konkreten pomen.

Tak pogled pa nas pusti v zadregi. Če za razumevanje besed že potrebujem jezik, kako sem se ga kadarkoli sploh naučil? Wittgenstein (2009: §83) predlaga, da si jezik zamislimo kot igro podajanja žogice. Nekdo mi vrže žogo in se nasmeji; žogico pospravim v žep in grem. Nekdo drug bo jo vrgel nazaj, spet kdo drug bi jo brcnil, igra pa postane smiselna šele, ko oba bolj ali manj veva, kaj

[3] Jezik je seveda sistem številnih pravil; po eni strani vsebuje pravila, kako tvorimo stavke, po drugi pravila, kako je treba tolmačiti posamezne besede. Ko odgovarjamo na to, kako se učimo jezika, odgovarjamo na to, kako se učimo pravil.

počneva. Zmotno bi si bilo predstavljati, da je ta igra strogo definirana. Žogo si lahko podajava vsak dan, ko pa pride tretja oseba, se bova spogledala. Ničesar ni v najini dosedanji igri, kar bi določalo to, kaj je treba v situaciji storiti. Vse je odprto. Celotna igra »obvisi v zraku«. In če mislimo, da vemo kaj od nas pričakuje drugi, te domneve nismo dobili iz igre, ki smo jo z njim igrali.

Tak je tudi jezik; gledamo se, poslušamo, domnevamo, kolebamo, se lovimo, dokler nas ne zagrabi ideja, da se razumemo. Ko je Sveti Avguštin zaslišal besedo »stol«,[4] jo je asociiral s predmeti v sobi (lahko bi jo tudi s čem drugim, kot so abstrakcije in čustva, a recimo, da je dojel pomen iztegnjenega kazalca[5]). Zdaj si lahko predstavljamo, da Avguštin sicer ni mogel dojeti, da beseda označuje stol, lahko pa je to domneval. Lahko je sprva mislil, da beseda označuje mizo, pa je potem razvidel, da to ni res, ali pa je pravilno ocenil, da označuje stol, a je mislil, da zraven tega označuje še vse ostale stvari, na katerih lahko sedimo; stvari kot sta divan in klop. Za uspešno učenje je potrebno ponavljanje; v stilu *trial and error* (»poskušanja in napak«). Sčasoma, ko z ljudmi delimo svet in v njem sodelujemo, se o jeziku zmeraj več učimo. Kdaj pa se ga *naučimo*? To je Wittgensteinov paradoks; nikoli!

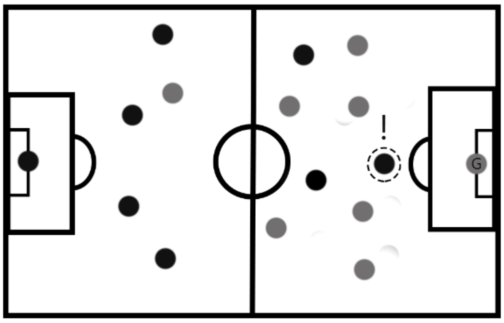
3 »Offside«

Kmalu bo jasno, zakaj bi naj paradoks vodil do tako presenetljivega sklepa; poglejmo še en primer učenja, namreč to, kako se je večina med nami naučila pravil nogometa. Pri tem pa se osredotočimo zgolj na pravilo prepovedanega položaja oz. »offside«, katerega nepoznavanje je v tolikšno zadrego spravilo komentatorja tekme v uvodu članka. Nekoč smo vsi to besedo prvič slišali, namreč »offside«, pa nam še ni mogla veliko pomeniti. Pri tem je redkokdo vzel v roke pravilnik FIFA ali slovar, da bi si besedo razjasnil. Njenega pomena smo se praviloma naučili skozi uporabo: z gledanje tekme.

Prvič smo gledali nogometno tekmo in opazili, da piščalka igro včasih prekine – po navadi takrat, ko se kateri izmed igralcev zvrne po tleh. Včasih pa igro prekine, na videz kar tako, da napadalci razočarano upočasnijo korake in zamahnejo. Kmalu spoznamo, da se to zgodi, ko se žoga z dolgo podajo približa golu. Včasih pa tudi s kratko podajo. Ampak zmeraj zgolj takrat, ko je napadalec žogo čakal preblizu gola.

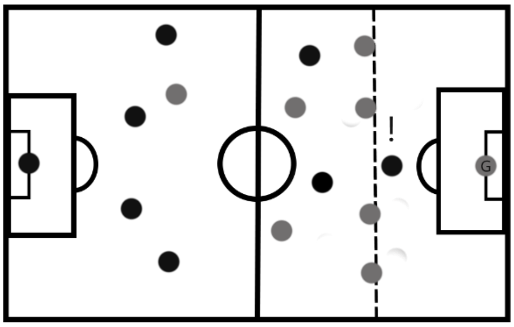
[4] Menda je Sveti Avguštin v resnici zaslišal latinsko besedo »cathedra«.

[5] Zanimiva študija gluhonemih otrok, ki se naučijo znakovnega jezika, razlaga da otroci praviloma ne dojamejo pomena iztegnjenega kazalca. Če ga položiš na svoje prsi, s tem misliš »jaz«. Pogosto pa otroci pokažejo na svoje prsi, ko mislijo »ti«. Preprosto kopirajo gesto in sprva ne uvidijo, da je njen pomen odvisen od tega, kdo jo naredi (Pettito, 1987).



Slika 1: Napadalec je morda žogo čakal preblizu nasprotnikovega gola?

Pravzaprav je včasih napadalec tudi daleč od gola, pa piščalka vseeno zapiska. Tako nenadoma – kot otroci, ki odkrijejo, da se beseda »jaz« zmeraj nanaša na tistega, ki izreče »jaz«[6] – dojamemo, da ni zgolj napadalčeva pozicija na igrišču tista, ki izzove piščalko. Temveč gre za *relacijo* napadalca do obrambnih igralcev nasprotne ekipe. Piščalka zapiska, ko napadalec žogo čaka za zadnjim obrambnim igralcem. In to je »offside«; namen pravila je onemogočiti, da bi napadalci žogo preprosto čakali na »napačni strani« obrambne črte. Če se črta premakne naprej, se mora tudi napadalec umakniti z njimi.



Slika 2: Napadalec je morda žogo čakal za obrambno črto?

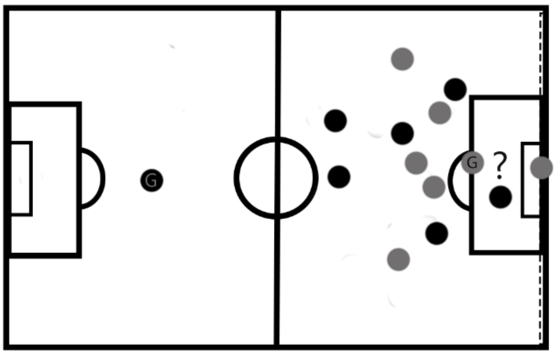
Nekako tako smo vsi pridobili intuitivni uvid v to, kako je treba razumeti pravilo. S takšnim razumevanjem pravila bomo lahko leta in leta pravilno sodili ter se hvalili s svojimi izkušnjami v nogometu. Potem pa bo nenadoma prišlo do situacije, ki bo pod vprašaj postavilo vse, kar smo kdaj mislili, da vemo o nogometu. Obstali bomo kot evropski kolonisti, ki so jim stoletja dolga izročila zapovedovala, da so »vsi labodi beli«. Zagledali bomo črnega laboda.

4 Črni labod: JAR proti Mehiki

Pogled v pravilnik FIFA bo razkril, da je pravilo vseeno bolj zapleteno – a so situacije, pri katerih bo to igralo vlogo, dovolj redke. Dovolj redke, da se lahko desetletja ukvarjamo z igranjem

[6] Čemur rečemo tudi »deiktični« zaimek, njegov pomen je relativen, odvisen od tega, kdo je govorec.

nogometa ali komentiranjem tekem in svojega dotedanjega razumevanja pravila ne bomo rabili popravljati. Vso življenje bomo govoriti o »offside«, pa bomo mislili, da govorimo o isti stvari kot ostali profesionalci. In tako bi lahko govorili še ves preostanek življenja, če ne bi naleteli na tisto redko situacijo:

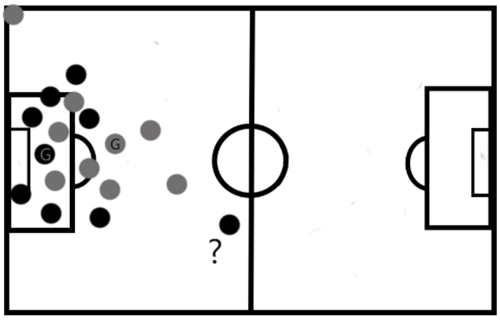


Slika 3: »Obrambni igralec stoji na črti gola!« kliče komentator.

Lahko bi sklepali takole; če je obrambna črta tam, kjer stoji zadnji obrambni igralec, potem seveda prepovedanega položaja ne more biti, ko obrambni igralec stoji na črti gola. Kje bi potem stal napadalec, v golu? Pri tem sklepanju nekako uspemo spregledati golmana, kot da nam predstavlja izjemo; tudi on je obrambni igralec! Redko se zgodi, da bi golman stopil *pred* svoje soigralce. Zato ga tudi nekako odmislimo iz pravila, večinoma je za našo presojo nerelevanten. Pravilo pa ne trdi, da črto prepovedanega položaja določa zadnji obrambni igralec – temveč *pred*zadnji. Ker je golman skoraj vedno *zadnji*, »zadnji branilec« pa skoraj vedno *predzadnji* igralec ekipe, na to pozabimo. Tako uporabljamo simplificirano pravilo, ki je skoraj vedno za prakso dovolj dobro. Redko imamo priložnost, da uvidimo njegovo nezadostnost.

Komentator bi to mogel vedeti, pa ni. Kot večina izmed nas, se nogometa ni učil iz knjige. Učil se je skozi gledanje in igranje. In ponotranjil to, kar se je intuitivno naučil. In dokler ni srečal črnega laboda, je (morda?) upravičeno mislil, da so vsi beli.

Dalje še gre reči, da niti z navedenim dodatkom nismo zaobjeli celotnega pravila – kot kaže pravilnik FIFA, je treba sprejeti *še več* podrobnosti. Lahko si zamislimo usodno tekmo, kjer v zadnjih zdihljajih ekipi gresta na vse ali nič – vsi v napad na eni strani, vsi v obrambo na drugi. Potem pa dobi podajo napadalec, ki je *vsem* za hrbtom:



Slika 4: »Obrambni igralec stoji na črti gola!« kliče komentator.

Tukaj ni prepovedanega položaja, »offside« velja šele na nasprotniki polovici; branilci obrambne črte ne morejo poriniti vse do nasprotnikovega gola. Pa je naslikana situacija tako zelo redka, da včasih res ne vemo, ali je k pravilu to res potrebno dodajati. Saj (skorajda) nikoli ne bo relevantno!

5 Wittgensteinov paradoks

Zdaj pa si zamislimo svet, kjer ni avtoritativnega pravilnika. Kako se bodo sodniki zmenili med sabo? Ne, pravzaprav je treba vprašati naslednje. Kako bodo vedeli, o čem govorijo, ko govorijo o »offside«?[7] Tako kot se o »offside« nismo učili iz uradnega pravilnika, se tudi ostalih besed ne učimo iz knjige; kot mali Avguštin poslušamo odrasle in ugibamo, domnevamo, kolebamo, lovimo pomene in iz besed sestavljamo smiselne celote. Tako lahko velik del svojega življenja uporabljamo pojme, ki so simplificirane različice tega, kar mislijo ostali – kot komentator, ki si je »offside« zamišljal preprosteje od ostalih profesionalcev. Ali kot človek, ki misli, da je vse, na kar se lahko usedeš, stol. Po drugi strani pa lahko ljudje okoli nas uporabljajo pojme, ki so zgolj plitka upodobitev tega, kako jih razumemo mi. Vso življenje se lahko strinjamo, kako razsoditi tekmo, pa ne bomo prišli do trenutka spoznanja, da se le ne strinjamo v celoti. Skupaj bomo govorili o »offside« in situacija nam bo dala misliti, da se razumemo. Z besedami Wittgensteina, v ključnemu paragrafu §201: »Naš paradoks je bil tale: pravilo ne more določati nobenega načina delovanja, saj je sleherni način delovanja mogoče uskladiti s pravilom« (Wittgenstein, 2014: §201). Paradoks je v tem (tako se vsaj zdi), da ni ničesar, kar bi nas obvarovalo pred večnim skepticizmom. Že jutri se lahko zbudimo kot Blake Ross, 30-letni ustvarjalec Mozille Firefox, ki je ob zajtrku v časopisu prebral, da je neka ženska po operaciji izgubila »zmožnost videti slike v glavi« (Ross, 2016). Čudno, je pomislil. Ne bi morali proslavljati tega, da je tega sploh kdaj bila zmožna? Zakaj se nihče drug temu ne čudi? Ross je bil kasneje diagnosticiran z afantazijo, čemur tudi pravimo »slepota notranjega očesa«. Vso življenje je Ross domneval, da se ostali ljudje, medtem ko opisujejo svoj mentalni svet, zanašajo na vizualne prispodobe. Nikoli ni pomislil na to, da ljudje v mislih dejansko

[7] Po slavni kratki zgodbi Raymond Carverja »O čem govorimo, ko govorimo o ljubezni?«, kjer imajo sogovorniki tako različno mnenje o tem, kaj je ljubezen, da se lahko začnemo spraševati, a z besedo načelno sploh mislijo isto stvar.

kaj vidijo. Seveda je domneval, da so vse besede, ki nakazujejo na obstoj vizualne domišljije, zgolj metafore. In, kot si lahko predstavljamo, bi lahko vso življenje živel v prepričanju, da tako pač je. Besede je mogoče interpretirati na več načinov. In mnogo načinov smiselno razlaga to, kar nam govorijo drugi ljudje.

Paradoks tukaj še zmeraj razlagam skozi precej abstraktno pravilo »offside« in specifične mentalne fenomene, govor o katerih je že v izhodišču izmuzljiv. A vse to je zgolj ilustracija. Kripke pravi, da naš skeptični problem velja za vso smiselno rabo jezika. Na koncu je vsaka »nova uporaba besede korak v neznano; vsakršno sedanjo intenco bi lahko interpretirali tako, da bi se ujemala s čimerkoli, za kar bi se odločili. Tako ne more biti niti ujemanja niti nasprotovanja. Prav to je povedal Wittgenstein v §201« (Kripke, 2011: 83). Kar mora paradoksu dati zgolj še večjega zagona, je dejstvo, da Wittgenstein verjetno ni želel povedati prav tega, in se literatura, da bi se izognila tej okoliščini, pri obravnavi paradoksa sklicuje na Kripkejevega Wittgensteina oz. Kripkensteina (Candlish in Wrisley, 2019). Paradoks namreč predstavi Kripke, ki je mislil, da zgolj razlaga, kar nam je želel povedati Wittgenstein. A morda se niti nista razumela, v svoji obravnavi razumevanja. Kaj drugega bi bila lepša geneza našega paradoksa?

6 Zaključek

Naša komentatorja si lahko spomin osvežita z vestnim prebiranjem pravilnika. A kaj naj storita z vsem ostalim besediščem, na razumevanju katerega temelji celo pravilnik, ki ga prebirata? Kaj naj storita, brez avtoritete, ki bi zapovedovala, kako razumeti besede? »[V]saka interpretacija visi v zraku skupaj z interpretiranim; ne more mu služiti kot opora. Zgolj interpretacije ne določajo pomena.« (Wittgenstein, 2014: §198). Kaj naj stori mali Avguštin, ko mu odrasli pravijo »Cathedra, cathedra!«, pa ni prepričan, ali mislijo stol – ali divan? Kaj naj stori afant, ko mu meditacijski trener razlaga, naj si predstavlja puščavo? Kaj naj storimo mi, ko nam sinestet Kandinski pripoveduje o zvenu barve?

To je jedro našega paradoksa. Osmisliti razmerje jezika do duha ter duha do drugega duha. So naši duhovi kot hrošči v škatlah, v katere lahko pogledamo samo mi, drugi pa ne (Wittgenstein, 2009: §293); ali pa ti hrošči letajo v prostoru med nami in jih beremo na telesu drugega? Je jezik forma naše misli, ali zgolj bled odtis, ki ga pošljemo v svet? Predstavljeni paradoks terja celostne odgovore.

An Unlucky Case of Wittgenstein's Paradox

Here I present a new case study through which we can think and teach about the Wittgenstein's paradox. When it was formulated by Saul Kripke almost 50 years ago, it relied on pretty abstract paradox examples – which goes to show that not even idealised mathematicians are safe from the posed problem. However, the cases did not succeed in emphasising how concrete and relevant the meaning of the paradox is for our lives. Furthermore, the literature on the paradox is quite dry and calls for refreshment. Those two are the objectives of this article.

Keywords: Wittgenstein, Kripke, paradox, philosophical investigations, rule-following.

Literatura

Candlish, S., in G. Wrisley. (2019). »Private Language«. V Zalta, E. N., *The Stanford Encyclopedia of Philosophy* (izdaja jesen 2019). URL = <https://plato.stanford.edu/archives/fall2019/entries/private-language/>.

'farham salam' (2012) »Best referee decision ever ! commentators gone wrong !!« *Youtube* (21. december 2021). URL = https://www.youtube.com/watch?v=l9EF-BDWxTs.

Kripke, S. (1982). *Wittgenstein on Rules and Private Language*. Cambridge, Massachussetts: Harvard University Press.

Kripke, S. (2011).«Wittgenstein o pravilih in zasebnem jeziku: izbrani odlomki iz istoimenske knjige«“ *Analiza*, 15 (4), str. 75–92.

Pettito, L. (1987). »On the autonomy of language and gesture: evidence from the acquisition of personal pronouns in American Sign Language«. *Cognition*, 27(1), str. 1–52.

Quine, W. V. O. (1960). »Chapter 2: Ontological relativity«. V *Word and Object*. Massachusetts: MIT Press, str. 26–68.

Ross, B. (2016). »Aphantasia: How It Feels To Be Blind In Your Mind«. *University of Exeter* (21. december 2021). URL = https://medicine.exeter.ac.uk/media/universityofexeter/medicalschool/research/neuroscience/docs/theeyesmind/Blake_Ross_April_2016_facebook_post_Aphantasia.pdf.

de Saussure F. (1959). »Chapter III«. V *Course in General Linguistics*. New York: The Philosophical Society, str. 7–16.

Wittgenstein, L. (2009). *Philosophical Investigations*. Hong Kong: Wiley-Blackwell.

Wittgenstein, L. (2014). *Filozofske raziskave*. Prevedel Erna Strniša. Ljubljana: Krtina.

Človek, resnica in moralna odgovornost

Human, truth, and moral responsibility

Marina Bajić

Univerza v Mariboru

Religious "Knowledge" as a Post-Truth Concept?

The word "post-truth" first surfaced when Donald Trump became president of the United States back in 2016. Since then, it has gone on to have many definitions and meanings, such as "where some feel emboldened to try to bend reality to fit their opinions", "a deliberately complicated relationship with the truth", or the simple Oxford Dictionary one: "relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief."

Though many people have taken on many angles on the post-truth subject, I have decided to focus on its relation to religion, or more specifically, religious "knowledge", which is defined in the following way: "religion *per se* is created by God but religious knowledge is human-made. The sacred law is divinely created but its understanding is a human enterprise." It seems that, almost by definition, religious "knowledge" falls under the category of post-truth, which is why I use the term loosely; "knowledge" implies "truth", and, in fact, it might be quite the opposite.

Leaning heavily on Morteza Hashemi and Amir R. Bagherpour's *A Theory of Evolution of Religious Knowledge in a Post-Revolutionary Iran: And a New Frontier for Sociology of Knowledge*, I examine how various definitions of the term post-truth can play into our understanding of religious "knowledge", focusing mainly on Islam, since it is the religion that Hashemi and Bagherpour focus on – that is not to say, of course, that these findings could not be applied to any of the other religions.

Keywords: religion, religious knowledge, post-truth, Islam, post-revolutionary Iran

1 Introduction

Morteza Hashemi’s and Amir R. Bagherpour’s *A Theory of Evolution of Religious Knowledge in a Post-Revolutionary Iran: And a New Frontier for Sociology of Knowledge* focused on dealing with the concept of post-truth as an answer to Abdolkarim Soroush’s riddle about post-revolutionary Iran. In this paper, I will start by examining post-truth as a concept and summarizing what Hashemi and Bagherpour wrote about in their work. Leaning on their work, I will go on to argue how divergent readings of religious texts can pose a problem, especially in a religiously-governed country; I will predominantly be using examples from Islam, since Hashemi and Bagherpour’s work focuses on this specific religion. In the end, I will argue how there could be a possibility that we can start viewing religious knowledge as a post-truth concept – if it is, indeed, possible.

2 A Short Look at Post-Truth

In order to have a clear understanding of what Hashemi and Bagherpour are discussing, it is best to start by looking at what post-truth truly means. In their case, Hashemi and Bagherpour define their understanding of post-truth as:

> By post-truth, we do not mean the irrelevance of truth, or post-factual politics, which are the popular understandings of the term after the rise of Trump to power in the United States /…/ The way we see it, post-truth is about a deliberately complicated relationship with the truth, which neither takes it for granted for one side of the debate nor considers it inaccessible. (Hashemi and Bagherpour, 2018: 72).

Indeed, the term post-truth first cropped up when Donald Trump became president of the United States back in 2016. The word was even chosen as the word of the year by Oxford Dictionaries. They defined is as “relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief.”

Lee McIntyre builds upon that definition even further by saying: “many see post-truth as part of a growing international trend where some feel emboldened to try to bend reality to fit their opinions, rather than the other way around.” (McIntryre, 2018: 5–6). McIntyre goes on to further argue that it is a simple process of falsification, when the person knows that what they are saying is wrong, yet they try to present it as truthful anyway. They also argue that post-truth can sometimes appear as someone believing in something, despite it being untruthful: “This is when self-deception and delusion are involved and someone actually believes an untruth that virtually all credible sources would dispute.” (McIntyre, 2018: 9).

The more modern problem the “post-true” world is currently facing is also the one accurately described by Hannah Arendt:

> To the citizens’ everchanging opinions about human affairs, which themselves were in a state of constant flux, the philosopher opposed the truth about those things which in their very nature were everlasting and from which, therefore, principles could be derived to stabilize human affairs. Hence the opposite to truth was mere opinion, which was equated with illusion, and it was this degrading of opinion that gave the conflict its political poignancy; for opinion, and not truth, belongs among the indispensable prerequisites of all power. (Arendt, 1977: 233)

Hashemi and Bagherpour do not understand post-truth in the same way when it comes to their arguments. Their main focus lies in “*deliberate problematic relationships with truth*” (Hashemi and Bagherpour, 2018: 72). However, it is still relevant to understand these classical approaches to post-truth when discussing their work, to understand the main definition of post-truth along with the problem of truth and opinion, as described by Arendt, and the expansion of the definition by McIntyre.

Of course, this is not a complete overview of the complex problem of post-truth (that would require a separate essay on its own), but just a few definitions, used in a broader sense, which can help us begin understanding the phenomenon.

3 Post-Truth Society in a Post-Revolutionary Iran

> On 16 January 1979, the last king of Iran (a.k.a. the Shah) fled the country for the last time. The new leader, Imam Khomeini, triumphed, and his revolutionary followers established a new political system in Iran which has transformed the Middle East in the ensuing decades. (Hashemi and Bagherpour, 2018: 73)

Khomeini’s goal was to “transform Iran into a theocratically ruled Islamic state” (Encyclopaedia Britannica, 2021b). He seemed to have succeeded in his vision by banning Western music and alcohol, women were required to wear a hijab; this also saw the return of punishments according to Islamic law.

Hashemi and Bagherpour decided to study an overlooked aspect of this revolution, which they best describe as “an epistemological sea change that has shaped post-revolutionary Iran. The discussions around relativism, religion and what we call *post-truth society* are at the centre of this upheaval.” (Hashemi and Bagherpour, 2018: 73)

The main problems that can be seen arising from a state governed predominantly through religion are those to do with non-religious questions. Hashemi and Bagherpour list some of those, such as how would this new regime support science and technology, or its approach to freedom (Hashemi and Bagherpour, 2018: 75). It always seems to boil down to questions of science and freedom.

Hashemi and Bagherpour pose some interesting questions, though the most interesting by far is the following: “What should be the government’s policy in approaching divergent readings of religious doctrines?” (Hashemi and Bagherpour, 2018: 75). “Divergent readings of religious doctrines” are a problem which many religions, not just Islam, have faced throughout history. “Thou shalt not suffer a witch to live” (*Holy Bible, The New International Version,* 1979, 1984, 2011: Exodus 22:18) is probably the most notable example of how a wrong interpretation of religious text can have disastrous consequences (read: Salem Witch Trials). How can that relate to post-truth then?

If we consider post-truth to be acting in accordance with personal beliefs rather than facts, then Witch Trials are a perfect example of divergent readings of religious doctrines mixing with a post-truth society (even though they were unaware of being in a such society, of course), more specifically of Arendt’s post-truth society where opinion and personal beliefs were used as prerequisites of power. Most people (though we could imagine there were some among them who were zealous believers) likely knew there is no such thing as witches, they were merely presented with the opportunity to take revenge on a neighbor that refused to sell a piece of their land, or a lover who did not return their affections.

The question is, could something like this have happened in post-revolutionary Iran? It is true that Khomeini brought back the more conservative traditions; for example, the Shah’s rule saw women attending universities, not being required to wear a hijab, slowly gaining more and more equality – Khomeini’s rule cancelled out all of it. The answers, as Hashemi and Bagherpour suggest, could be found in the works of Ali Shari’ati, a man who “laid the foundation for the Iranian revolution of 1979” (Encyclopaedia Britannica, 2021a).

3.1 Ali Shari’ati on Knowledge and Religion

It is first prudent to understand the view Shariati held about epistemology: “Consequently, Shari’ati argues that man's knowledge is co-ordinate with what he feels to be the interaction among God, man and nature” (Akhavi, 1988: 408). Therefore, we can assume that Shariati was one who believed that our knowledge only operates in a religious framework.

Another striking thing about Shari’ati, which can be observed in Akhavi’s helpful summary of Shariati’s views, is the following:

> The common man plays a critical role in selecting the leaders of the community. /…/ Not only that, but Shariati even declares outright that the people are charged with the mission of the Prophet and the imams until the return of the Mahdi (Hidden Imam)! The people will maintain close links with their leaders and in so doing will 'secure the government of knowledge ... just as Plato had urged.' (Akhavi, 1988: 416)

By believing the people hold the power to shape their own destiny, it is easy to see how this type of thinking could have influenced the people of Iran to rebel against a government whose process of Westernization was endangering their religion and culture; an opinion that Shari’ati also shared (Encyclopaedia Britannica, 2021a).

The most problematic view he held, however, was one that is closely related to our post-truth problem: “But at the same time, he has warned impatient parliamentary deputies - especially over the questions of property ownership, land and trade - that ‘if the Guardian Council says something is against the Islam, then it is against Islam’” (Akhavi, 1988: 421). That was precisely the problem I had pointed out with the example of the Salem Witch Trials (where we can consider priests and judges, who ultimately passed the sentence if the woman should be burned as a with or not, as the Guardian Council): how can we be sure that what, in this case, the Guardian Council says something that is against Islam is truly against Islam, or whether it is something closer to what Arendt had been saying about politics and opinion?

4 Religious Knowledge

Hashemi and Bagherpour continue by taking the theory of Abdolkarim Soroush as a basis for their arguments: “/…/ [w]hat fascinated me most was the details and intricacies of the differences in interpretation” (Hashemi and Bagherpour, 2018: 76). Soroush discussed how different people interpret religious texts differently and how we should distinguish between religion and religious knowledge. The main distinction that he makes here is “that religion per se is created by God but religious knowledge is human-made. The sacred law is divinely created but its understanding is a human enterprise.” (Hashemi and Bagherpour, 2018: 77). The question I pose here deals with the problems dealt before: how can we, despite differences in interpretation, provide accurate knowledge? It is not completely impossible, especially if the interpretation is based on something like a guide, which can help us find the best way of interpreting a text. The problem arises when interpretation and post-truth become intertwined in a sense that post-truth is understood as interpretation – later, I will expand upon that further.

4.1 Post-Islamism

Post-Islamism is “a recognition that politics rather than religion provides for welfare in this life” (Eteraz, 2007: para. 1). This is a simple solution to the problem of incorrect interpretations of religious texts, something that has come into practice in most first-world countries. Post-Islamism has spread throughout Iran as well: “Post-Islamists in Iran, regardless of their internal differences, have been pursuing political reform. For them, religious reform is a path to political reform” (Mojahedi, 2016: 54).

Something that is interesting to note in Mojahedi’s article on post-Islamism is one of the goals they list as part of the post-Islamite movement in Iran:

> Although they want to bring about political reforms, they seek political goals through religious reforms. Their focus, therefore, has been on giving alternative readings of the hadith and the Quran in line with modern values of democracy and human rights. (Mojahedi, 2016: 55)

This notion further supports the idea that religion *can* be interpreted in a way that fits today’s modern society, it just seems that *it is not*, perhaps because it goes in line with Arendt’s notion that opinion is a prerequisite of all power: one’s interpretation of a religious text (in this case the Quran) can be done so that it supports one’s personal interest, keeping them in power. If it can go one way, however, it can go the other: a more modern reading of the hadith and the Quran would fit better in today’s society, because this society, regardless of how simplistic this sounds, has much considerably relative to the society in which the Quran “came to be”.

4.2 Problems With Religious Knowledge

The problem that religious knowledge faces is rooted in Soroush’s definition of religion and religious knowledge: religion is created by God, religious knowledge by people. If we put aside the flaw that I have described earlier (interpretation can be manipulated) and to which we will return later, we encounter another problem: the existence of God. If there is no proof of God’s existence, this renders religion meaningless, therefore rendering religious knowledge meaningless because it stems from something that is already meaningless.

The problem with this statement is simple: religion is not intrinsically bad. For the people who believe in any God, it only matters that God exists for them, but it does not matter if God exists for us. Hashemi and Bagherpour summarize one of the theories Plantinga defended: “The belief in God is properly basic, and does not require any attempt to discover other propositions and statements as more fundamental bases for it” (Hashemi and Bagherpour, 2018: 78). As Plantinga himself said:

> The fact is, the vast majority of the world’s people do believe in God, or something like that, so it’s not that God is hidden in the sense that nobody knows about Him, all kinds of people do I suppose, as I say, the vast majority of the world’s people. God isn’t as plain to us as other people, let’s say, or as, I don’t know, trees and houses and material objects, but why think that he would have to be? (Plantinga, 2013: 1:02)

The basis of religion is faith, faith in God’s existence and God’s teaching (it does not matter if it is a Muslim, Christian or any other kind of God or Gods) and that is not intrinsically bad. So, when does it become “bad”?

The problem is the subjectivity of interpretation. The example of how bad interpretation can go wrong is the one I have given before of the Salem Witch Trials, where dozens of people were killed because someone interpreted the religious text as witches (witches being people who have allied themselves with the devil and can therefore perform magical acts) actually existing among us. It becomes an even larger problem when religion and state become intertwined, meaning religious views are turned into law. Even if the majority of the country shares the same religious beliefs, religious knowledge cannot be a basis for state laws, because:

a) It is based on a deity whose existence has not yet be proven.
b) Subjective interpretation does not equal to objective truth.
c) Subjective interpretation can stem from opinion, which also cannot be mistaken for truth.

As Peter Carmichael suggests, we could instead substitute religious knowledge with philosophy, or rather educating our minds to come up with the answers we seek:

> The proof here of God's existence being (if it is a proof) a great *coup*, a stroke of genius, the means to it must be eminently recommended to us; but as these are strictly philosophic means, it will therefore be philosophy, not religious usage, that will be recommended to us as a vehicle, so far, of ‘salvation’: not action but understanding, not following a pastor like sheep but educating our minds. (Carmichael, 1949: 55).

4.3 Secularization as a Possible Solution

As I have mentioned before, religious knowledge cannot be considered as a basis for law, especially when it is, almost as a post-truth concept, subject to subjective interpretation, which can often lead to disastrous consequences. However, this is hard to imagine for us (Europeans), because Europe had, on several fronts, already managed to separate itself from the church, which is not the case in many countries today. As Hashemi and Bagherpour write:

> The problem was that, by the 1980s, empirical data was increasingly proving that, although the theory [the state is separate from the church] is still applicable to Western Europe, it is not valid in many other parts of the world /…/ Moreover, the Iranian revolution was indeed a real-world counterexample to the theory. (Hashemi and Bagherpour, 2018: 80)

Hashemi and Bagherpour then continue by saying: "In its simplest form, the notion of the secular suggests that there is a place for making decisions over public policy which is not affected by diverging value judgments and religious motivations" (Hashemi and Bagherpour, 2018: 80). That would be in accordance with the point I was trying to make earlier.

5 Religious Knowledge as a Post-Truth Concept?

The last point that is still left up for discussion is whether we could actually consider religious knowledge as a post-truth concept. It is an interesting question, since concepts such as religion and religious knowledge are something that are, in fact, so old, yet the concept of post-truth is still relatively new. The question is whether it is actually possible to combine these two concepts?

In this case, of course, we use the term "knowledge" loosely, since it does not pertain to our general idea of what knowledge is (justified true belief). As Hashemi and Bagherpour defined it, "religious knowledge is human-made. The sacred law is divinely created but its understanding is a human enterprise" (Hashemi and Bagherpour,2018: 77), meaning that in this case, when we speak of "knowledge", we speak of the understanding of supposedly divine texts through human understanding and interpretation (which is not necessarily always accurate). Therefore, I use the term "knowledge" when it comes to religious knowledge somewhat loosely.

As mentioned before, religious knowledge has its basis in "the word of God" – religion, that this or that God had created. We may leave aside the fact that God has not yet been proven to exist, as belief in God is not intrinsically bad (as mentioned before). In this paper, I have mainly been discussing religious knowledge in connection to the "word of Gof" that is the Quran (though, in principle, these arguments could also work in connection to the Bible), as Hashemi and Bagherpour's focus was on Iran, a predominantly Islamic country.

The question they posed at the beginning, the one that I have already cited, is: "What should be the government's policy in approaching divergent readings of religious doctrines?" (Hashemi and Bagherpour, 2018: 75). I have been discussing these "divergent readings" in bits throughout the paper, but I would like to focus more on them in the next section.

5.1 Divergent Readings of Religious Doctrines in Islamic Countries

As someone who has had the opportunity to travel to an Islamic country (Tunisia), I have had the privilege of experiencing how one country can be so divided within the same religion (it should also be worth noting that Tunisia is considered the most progressive country in Africa): in the North, where the capital city of Tunis is located, women dress however they wish and do not need to cover themselves, and everything is modernized; in the South (especially on the island of Djerba, where I have stayed), women still cover themselves and a more traditional lifestyle is lead. This is not to say that the country is not considerably more conservative than any European country, it is just interesting to see how, within one country, such different lifestyles are led. Why, then, is it okay for half the country’s people to live a more modernized lifestyle, while the rest still live very traditionally, if their one religion has the same “rules”? And, going even further, how is it possible that different countries with the same religion have different religion-based laws, if the “rules” are the same?

Samina Ali, in their Ted Talk for University of Nevada, gives a helpful explanation about how women should dress according to the Quran, which is a good example to further support my argument that religious knowledge could be considered a post-truth concept. They start by explaining that the prophet Muhammad had originally instructed women to dress similarly so they would not get assaulted – women of status were usually left alone, but slaves were often assaulted. This, of course, sparked outrage, as slave women could not possibly dress the same as women of status, so the solution they came up with about how women should dress were the following two points: “A woman’s function in society, her role – what we might consider her job – and the society’s specific customs.” (Ali, 2017: 3:11). They later go on to explain that the Quran never specifies which body parts a woman should cover, and their intentional vagueness is so that a woman could choose how to dress according to the two points mentioned above. As for the term *hijab*, it is meant as a literal veil, separating the earthly from the divine, and not the veil we associate with women’s covering.

So, we can safely pose the question here: at which point did a reading of the Quran start to shift in order to impose such laws on women? We could even go so far as to ask who was the first to read those verses of the Quran and say: “Yes, women must now wear a hijab to veil themselves from the world?” I am by no means attempting to attack Islam; I am by no means attempting to attack any religion. As I have said, religion is not intrinsically bad. However, I have also said when religion does become “bad”.

5.2 Post-Truth and Religious Knowledge

This is the turning point where I could argue that we could start considering religious knowledge as a post-truth concept. If I abide by the Oxford Dictionary definition, the law which states that (referring to the previous example) women must wear a hijab falls into the category of personal feelings and beliefs, rather than facts. The facts that Ali presented were that women can choose how they want to dress (according to the two rules presented), and that the hijab is a veil separating the earthly from the divine, not a face covering.

Ali perfectly captures this once more:

> I hope it's not any surprise to you that this isn't by accident. For the past few decades, the very people who have been given the important task of reading and interpreting the Quran in a variety of different Muslim communities, certain clerics have been inserting a certain meaning into those three verses concerning women. For instance, that verse I told you about earlier: 'O prophet, tell your wives, your daughters and the women of the believers to draw upon themselves their garments. This is better, so that they not be known and molested.' Some clerics, not all, some clerics have added a few words to that, so that in certain translations of the Quran, that verse reads like this: 'O prophet, tell your wives, your daughters and the women of the believers to draw upon themselves their garments; parenthesis – a garment is a veil, that covers the entire head, the face, the neck, the breasts, all the way down to the ankles and all the way to the wrists. Everything on a woman's body is exposed except for one eye, because she must see where she is headed. And the hands must be covered in gloves. /…/ Ext. ext. ext. ext. on and on and on and on, end of parenthesis – so that she not be known and molested.' (Ali, 2017: 12:34)

This is not just concerning Islam, of course, these are merely examples provided because Hashemi and Bagherpour's focus is on an Islamic country. The example of the Salem Witch Trials is another case similar to this. So, to what does all of this lead us? Ali's TED Talk is in accordance with the Oxford Dictionary's definition of post truth; it goes in accordance with Arendt describing opinion, and not truth, as a basis for power; it even goes in accordance with Hashemi and Bagherpour's deliberate problematic relationship with the truth. It can even be thought of, as McIntyre puts it, deliberately ignoring facts, and/or manipulating them to fit a person's own agenda (in this case, the manipulation of the Quran to fit the notion that women should be veiled, or choosing to believe that witches exist).

> Deniers and other ideologues routinely embrace an obscenely high standard of doubt toward facts that they don't want to believe, alongside complete credulity toward any facts that fit with their agenda. /…/ This is not the abandonment of facts, but a corruption of the

> process by which facts are credibly gathered and reliably used to shape one’s beliefs about reality. (McIntyre, 2018: 11).

Simply put, it allows us to start thinking of an old concept, such as religious knowledge, in the terms of a newer concept, such as post-truth.

6 Conclusion

It seems that Hashemi and Bagherpour were on the right track when responding to Soroush’s work as a post-truth problem, even if they did not continue to expand upon it even further. They have successfully laid grounds, however, to what I have attempted to argue throughout this paper: that we could view religious knowledge as a post-truth concept. If it is indeed the way Souroush describes religious knowledge as being the product of human interpretation, then I have given several arguments (especially backed up by Ali’s profound TED Talk) on how interpretation cannot always be considered truth, because it can easily be manipulated to fit with one’s own agenda (of course, that is not to say that interpretation cannot be done correctly, it is just to say that we must pay attention to how it is done). With the help of Oxford Dictionary’s definition of post-truth, McIntyre’s expansion upon that definition, Arendt’s case on opinion and power, and Hashemi and Bagherpour pointing out the deliberate, problematic relationship with the truth, I have argued how these coincide with religious knowledge and how that can sometimes be harmful, especially in a religiously-governed state (even if religion is not intrinsically bad). Indeed, the concept of religious knowledge may be very old, and the concept of post-truth relatively new, but the latter gives us the opportunity to observe the former from a different angle.

Religijsko »védenje« kot post-resnični koncept?

Beseda »post-resnica« se je prvič pojavila, ko je leta 2016 Donald Trump postal predsednik Združenih držav Amerike. Od takrat je beseda prevzela veliko definicij in pomenov, med drugimi »kjer se nekateri čutijo tako pogumne, da poskusijo izkriviti stvarnost, da se ta prilagodi njihovemu mnenju«, »namenoma kompliciran odnos z resnico« in preprosta definicija slovarja Oxford Dictionary: »ki se nanaša ali označuje okoliščine, kjer so objektivna dejstva manj vplivna pri oblikovanju javnega mnenja kot sklicevanje na čustva in osebna mnenja«.

Čeprav je veliko ljudi obravnavalo post-resnico z različnih vidikov, sem se sama osredotočila na povezavo te z religijo, oz. bolj specifično z religijskim »védenjem«, ki je definirano na naslednji način: »religijo *per se* ustvari Bog, religijsko védenje pa ustvari človek. Sveti zakon je ustvarjen s strani Boga, razumevanje tega pa je človeška domena«. Zdi se, da religijsko »védenje« skoraj po definiciji spada v kategorijo post-resnice – prav zato uporabljam ta izraz zelo ohlapno; »védenje« implicira »resnico«, pravzaprav pa je morda situacija prav nasprotna.

Preučujem, kako različne definicije izraza post-resnica vplivajo na naše razumevanje religijskega »védenja«, pri čemer se sklicujem na delo *A Theory of Evolution of Religious Knowledge in a Post-Revolutionary Iran: And a New Frontier for Sociology of Knowledge* (Hashemi in Bagherpour, 2018). Osredotočam se predvsem na islam, ker je to religija, na katero se osredotočata Hashemi in Bagherpour, vendar to ne pomeni, da ugotovitve tega prispevka ne moremo aplicirati na druge religije.

Ključne besede: religija, religijsko védenje, post-resnica, islam, porevolucionarni Iran

Works Cited

Akhavi, S. (1988). “Islam, Politics and Society in the Thought of Ayatullah Khomeini, Ayatullah Taliqani and Ali Shariati.” *Middle Eastern Studies*, *24*(4), pp. 404–431.

Ali, S. (2017). “What does the Quran really say about a Muslim woman's hijab.” *TEDx Talks* (2021, January 25). URL = https://www.youtube.com/watch?v=_J5bDhMP9lQ&ab_channel=TEDxTalks.

Arendt, H. (1977). *Between Past and Future*. New York: Viking Press.

Britannica, T. Editors of Encyclopaedia (2021a). “ʿAli Shariʿati”. *Encyclopedia Britannica* (2021, January 25). URL = https://www.britannica.com/biography/Ali-Shariati.

Britannica, T. Editors of Encyclopaedia (2021b). “Ruhollah Khomeini”. *Encyclopedia Britannica*. (2021, January 25). URL = https://www.britannica.com/biography/Ruhollah-Khomeini.

Carmichael, P. A. (1949). “Limits of Religious Knowledge”. *Philosophy and Phenomenological Research*, *10*(1), pp. 53–64.

Eteraz, A. (2007). “Post-Islamism.” *The Guardian* (2021, January 25). URL = https://www.theguardian.com/commentisfree/2007/oct/31/postislamism.

Hashemi, M., Bagherpour, A. R. (2018). “A Theory of Evolution of Religious Knowledge in a Post-Revolutionary Iran: And a New Frontier for Sociology of Knowledge”. In Stenmark, M., Fuller, S. and Zackariasson, U. (Eds.), *Relativism and Post-Truth in Contemporary Society*. Cham: Palgrave Macmillan, pp. 71–84.

Holy Bible, *The New International Version*. (1979, 1984, 2011). Bible Gateway (2021, January 25). URL = from https://www.biblegateway.com/verse/en/Exodus%2022%3A18.

McIntyre, L. (2018). *Post-truth*. Cambridge, MA: MIT Press.

Mohammad Mahdi Mojahedi. (2016). “‘Is There Toleration in Islam?’ Reframing a Post-Islamist Question in a Post-Secular Context”. *ReOrient*, 2(1), pp. 51–72.

Plantinga, A. (2013). “Arguing God’s Existence?” *Closer To Truth* (2022, March 21). URL = https://www.youtube.com/watch?v=eeX6Lhb0_6A.

Word of the Year 2016. *Oxford Languages* (2022, April 12). URL = https://languages.oup.com/word-of-the-year/2016/.

Teemu Tauriainen

University of Jyväskylä

Simultaneous Immanence and Transcendence of Truth

Debates on whether truth is dependent or independent of human concerns are present throughout the history of Western thought. According to one side, truth depends on mind-independent states of affairs, like facts or the relevant aspects of extensional reality. For others, truth depends on human-bound factors, like what is useful to believe or what science has at any given time proven to be the case. In more recent history, this debate has been labeled as the realist/anti-realist debate. For realists, the truth of chosen truthbearers is based on them adequately connecting with the world. For anti-realists, truth is determined by what one knows or is justified to believe. The conflict between these views has proven persistent because both sides have persuasive arguments in support of their position. I argue that both sides have lessons to be learned from a historically significant yet nowadays largely dismissed distinction between truths immanence and transcendence. Based on this distinction, I proceed to demonstrate how truth can involve both, human-dependent and independent aspects. My contention is that acknowledging such multifaceted nature for truth is necessary for achieving a philosophically plausible and scientifically legitimate account of its nature that can further be used as an explanatory instrument in defining various socially crucial phenomena, like knowledge, trust, and deception.

Keywords: truth, immanence, transcendence, anti-realism, realism.

1 Introduction

Truth’s nature has been a subject of analysis throughout the history of western thought.[1] Despite the long tradition of theorizing, philosophers are far from reaching a consensus on what the details of a satisfactory, let alone an exhaustive definition of truth would be.[2] Interestingly enough, these definitional issues are in plain contrast with the truths intuitive nature and the significance it bears for our lives. We do not think that there is any confusion with the term when teaching our children to speak the truth, when obliging people to tell the truth and nothing but the truth in our courtrooms, and when relying on experts, scientists, and leaders to tell us the truth about a range of important topics. Our commonsense grasp of truth seems more than sufficient for sustaining the practices that rely on there being a satisfactory account of this notion. Past this, one might argue that supposedly we operated with a notion of truth even before our ability to speak about it, let alone try defining it. We all agree that in one way or another, truth describes things as they stand, and because of this, it bears instrumental value for achieving success in our practices in a systematic and reliable manner. Without this concept, one cannot distinguish between fact and fiction, between reality and illusion, delusion, or wishful thinking. In this sense, truth is not only a normative guideline that we can require from other people. True beliefs bear value, and this is why they are oftentimes worth the trouble of aspiring for.[3] True sentences as the premises of valid inference enable prediction making, the subsequent control of our surroundings, and successful navigation in the world. What, then, is one to make of this Janus-faced notion that is intuitive, important, and valuable, yet which we seem to be incompetent to define in a satisfactory manner, to the extent that consensus is lacking on even the most basic features of an adequate account of its nature.[4]

It is the task of this paper to clarify matters by focusing on the question of whether truth is dependent or independent of human concerns. My contention is that work on this topic advances our collective understanding of truths nature. This, in turn, is necessary for reaching a satisfactory account of said notion. Achieving such an account bears great value for numerous reasons. Without a philosophically tenable and scientifically legitimate notion of truth, the gates are open for criticism towards it, which we all are painstakingly familiar with in light of recent history. Unfortunately, we are in middle of what some have labeled as the crisis or even death of truth that manifests in the form of speech about alternative facts and the post-truth era (Kakutani, 2018). No doubt, there is increase in people’s thirst for alternative facts and explanations, and their inclination to feel

[1] Comprehensive historical discussion on the nature of truth can be found in Glanzberg (2018).
[2] My use of “true” is ambiguous between the concept of truth and the property of being true, but I will emphasize this distinction when it makes a difference. For the sake of simplicity, we can assume that the concept is what gives meaning to the property that is attributed by the predicate “is true”.
[3] More on the value and significance of truth for our concrete practices in Tauriainen (2022a).
[4] Truth’s intimate connection to other concepts is evident when realizing that we use it to define various other concepts, like meaning, knowledge, trust, and validity.

contempt for the truth, especially when it threatens their ways of life. But when we accept deception and lies from our leaders, and when their advisors say things like “truth isn’t truth”, perhaps more than ever, we should focus on developing and defending a robust notion of truth to sustain the practices that constitute our contemporary ways of life. Without such account, we are impaired to define various socially important phenomena, like honesty, trust, and deception, further rendering it difficult to hold those accountable who fail to stand up to the criterion of truth.[5] My contention is that a robust notion of truth is not only a theoretical resource or valuable property of beliefs, but that it plays an indispensable role in sustaining healthy democratic practices, such as responsibility, voting, and trust in our leaders, experts, and educators.[6] It is somewhat embarrassing that in light of the aforementioned significance that truth bears for various domains of life, we are far from reaching an account of it that is not vulnerable to obvious criticism. In this sense, the age-old skeptical question expressed by Pilate “what is truth?” still lives on.

In the following section, I briefly discuss the historical debate on whether truth is dependent or independent of human concerns. This section is a setup for the more thorough examination of the metaphysics of truth that will be executed in the latter half of this paper. After this, I turn to discuss the prominent debate between truth realists and anti-realists and argue that while anti-realists have solid arguments in support of their position, avoiding truths partial human-independence, and thus its partial transcendence, is difficult to avoid on both, theoretical and practical grounds. I conclude by arguing that for truth to play the role that it has in our practices, it best be treated as a realist and partially transcendent notion. Finally, some prospects for avoiding truths partial transcendence will be suggested, but whether or not they ultimately succeed is up to debate.

2 Truths Human-dependent or -Independent Nature

Views about truths nature show significant variety in the history of western thought. From the literature, one can find numerous mutually exclusive definitions with no prospect of agreement in sight.[7] For some, truth is a fully objective matter, a representation of how things stand independent of minds, and for others, truth is that which is useful to believe, or which scientists at any given time prove to be the case. One historically significant distinction regarding truths nature that is

[5] My contention is that without a robust notion of truth, healthy democratic practices become endangered, for we cannot hold politicians accountable for lying, or we cannot explain what science in general aims at past pragmatic utility, which is in itself problematic position to uphold. For example, if we hold that pragmatic utility is the sole purpose of science, then justifying the existence of those disciplines that do not provide pragmatic utility, like the more theoretical aspects of mathematics, physics and ethics becomes difficult if not impossible.

[6] As Hannah Arendt has adequately noted in relation to the totalitarian regimes that: “The ideal subject of totalitarian rule is not the convinced Nazi or the convinced Communist, but people for whom the distinction between fact and fiction (i.e., the reality of experience) and the distinction between true and false (i.e., the standards of thought) no longer exist.” (Arendt, 1951: 474).

[7] More on this in Glanzberg (2021).

relevant to this day concerns its mind-dependent or -independent nature. One side to this debate can be found from Aristotle, according to whom the existence of human's grounds the truth that humans exist, but not the other way around. Indeed, while it is intuitive to think that truth in one way or another depends on how things stand independent of minds, it is questionable whether there would be any truths without humans to uphold them. This point was adequately emphasized by Donald Davidson, perhaps the most prominent theorist of truth of the 20th century: "Nothing in the world, no object or event, would be true or false if there were not thinking creatures." (Davidson, 1990: 279). In *Metaphysics* (1908: IV 7, 1011b27), Aristotle proceeds to emphasize the basic realist intuition that truth is a description of how things stand independent of minds.[8] There are truths about matters insofar as there are states of affairs to ground those truths, but someone still has to uphold or assert them. Whether and to what extent truth is dependent or independent of human concerns, such as our knowledge about it, is questionable even in the case of Aristotle. Nonetheless, it is intuitive that there are truths that we don't know, and truths that we don't know that we don't know, as determined by the reality that we partake in but don't have an exhaustive understanding of. Past this, there seems to be even *in principle* unknowable truths, like the exact rate in which the universe is expanding, or towards what it is expanding.[9] In the most radical end of this dependence/independence spectrum lies Chrysippus, for whom there are contingent truths even about future events, despite anyone's ability to know about them.[10] Aristotle disagreed on this matter on grounds of free-will. For there to be contingent truths about the future, humans would not have free will to affect the course of events, casting doubt on our ability to possess conscious agency. However, Chrysippus has contemporary allies. The now prominent deflationary approaches to truth argue that a sentence is true if and only if things are as the sentence says, and no reference to belief, knowledge or justification is required. The sentence "Trump will go golfing tomorrow" is true if and only if Trump will go golfing tomorrow, full stop. On this account, there are contingent truths about the future even in an indeterministic universe where humans have free will. Simply put, each sentence specifies its own conditions for being true, and it is left for reality to decide which of them adequately connect with its relevant aspects. Finally, for Plato, truth's nature derives from the form of the good, and is as such independent of one's ability to reach for it. Human accessible truth is always approximation of the whole truth, which is an ideal of reason. But as an ideal, it is also something worth pursuing, even though it might be in itself unreachable. From this follows that truth in the form of belief bears inherent normative force:

[8] The dependence relation used here can be understood as either compatibility, isomorphism, identicality, determination etc.

[9] Reason being that one can still guess them and possess a true belief by accident even though there is no way to achieve knowledge about these matters.

[10] The downside of this view is that it assumes determinism and thus compromises the notion of free will.

> So that what gives truth to the things known [beliefs] and the power to know to the knower is the form of the good. And though it is the cause of knowledge and truth, it is also an object of knowledge. Both knowledge and truth are beautiful things, but the [form of the] good is other and more beautiful than they. In the visible realm, light and sight are rightly considered sunlike, but it is wrong to think that they are the sun [the form of the good], so here it is right to think of knowledge and truth as goodlike but wrong to think that either of them is the good—for the good is yet more prized. (Plato, 1997: 508e).[11]

Interestingly enough, this view about the normativity of truth introduces partial transcendence to the notion. We aim and aspire for truth, and in this sense, truth is a goal of reasoning. But for truth to play such role, it needs to transcend our current knowledge or lack thereof about it. For truth to be something that we pursue, aim or aspire for, it must have a nature that transcends our current knowledge about it. This feature of truth as something that infinitely exceeds our grasp about it will be discussed further in the last sections of this paper.

Overall, both Plato's and Aristotle's ideas are intuitive. Surely, some truths are such at least partly because of how things stand independent of minds. Past this, it is both intuitive and practical to hold that there are truths we don't know, but which we *want* to know. We want to know whether human efforts can have a meaningful impact on climate change because the truth concerning this matter bears great instrumental value. We also want to know the truth about the origins of the universe simply to satisfy our inherent curiosity about the world. Interestingly enough, this seems to imply that we have *some* access to the truth, even if not in the eternal or ideal sense of the whole Truth. It is true that we are all conscious here and now, and it is true that the earth is an unevenly shaped ellipsoid. However, as history has proven to be the case, because of our limited perspectives to perceiving the truth, we have to accept the Kantian premise that what we at any given time understand to be true is never the whole story. Science, arguably the highest arbiter of truth, is in constant flux and eventually, most if not all the truths that we now uphold turn out to be false. This, again, is indication of truths at least partial transcendence. When a scientific discovery turns out false, we do not say that it *became* false. Rather, we admit our collective mistake in relation to the *truth*.

In what follows, I set to explore the metaphysics of truth by focusing on the prominent debate between truth realists and anti-realists. I focus specifically on the question of what is the nature of this thing that we aim in our inquiries, that we respect as a property of beliefs, and which we require from our leaders and educators, and which seems to be "out there" for us to discover. After this, I proceed to discuss a related, historically significant yet nowadays largely neglected question about

[11] In contrast, one might hold that truth is only *prima facie* or *pro tanto* correct to believe.

whether truth is immanent or transcendent to our potential understanding about it. My contention is that discussion on these topics helps illuminate the limits and requirements of an adequate and satisfactory account of truth.

3 Realism and Anti-realism

The debate between realists and anti-realists regarding truths nature is prominent in contemporary truth-theoretic literature (Shieh, 2018: 433–476). This distinction is closely related to the aforementioned distinction between truths human dependent or independent nature. One way to understand the former distinction is to see it as arising from a more general question about the relation that truth has to ontological concerns. If truth in the form of true sentences or beliefs depends on or is grounded in mind-independent states of affairs, such as the relevant aspects of extensional reality, then it is to some extent independent of what anyone thinks about it. There is either an even or uneven number of planets in the universe, as determined by the number of planets in existence. Similarly, there are approximately 10 nonillion (10 to the 31st power) viruses on earth alone, and we can identify only a fraction of them. Insofar as truth depends on facts, then it doesn’t mind what anyone thinks about it. In contrast, if truth depends on human-bound factors, such as our ability to know or justify a claim, then no reference to mind- or theory-independent states of affairs is required. Prime example of this are necessary truths, like each entity being identical with themselves, or the number 2 being the smallest prime. There is nothing out there in the world, so to speak, to render these statements as either true or false. They are true based on convention, or in virtue of the language-system in which they are formed.[12] Thus, in short, the realism/anti-realism debate regarding truths nature can be understood from the perspective of truths grounding as the question of whether it is grounded on what exists independent of humans, or whether we have the power to dictate the truth.

As noted, for the realists, truth depends in one way or another on reality, i.e., how things stand independent of minds: “Truths are determined not by what we believe, but by the way the world is.” (Armstrong, 2004: i). The reason for the label “truth realism” follows from the baked in commitment of these theories into ontological realism. Insofar as our thoughts and expressions are about the world, it is this world that renders some of our expressions true and other ones false; truth depends on the adequate connection that our thoughts and language have to the relevant aspects of reality, as perceived from the given theoretical standpoint that is. For example, somewhat poetically, we simply have access to a language that includes a class of truth-apt beliefs and expressions, and it is left for reality to pick from this gallery those that adequately connect with its

[12] This is, of course, debatable.

relative aspects. It is up to reality, ultimately, to decide which of our expressions are true and other ones false. This explanation is conveniently in harmony with the way in which scientific inquiry is executed. In the words of W.V.O. Quine, perhaps the most prominent naturalist to date: “Such is scientific method: interrogation of nature in a cosmic true-false test. Man proposes, nature disposes.” (Quine, 1994: 500). Insofar as our theories aim to be about the world as given by a commitment to this or that form of metaphysical realism, then we require them to be true by consisting of true sentences, in the sense that it is precisely these *types* of sentences that provide the intermediary between language and the world.[13] This approach is persuasive, not least because we indeed do require some standard of correctness to answering questions about the nature of reality or how things stand.

There is some variation in the way in which realists understand the relation between our expressions and the world. According to a restrictive approach, the world *determines* the truth of our expressions. Less ambitiously, one could claim that only *compatibility* with our expressions and the world is required. Further, one might be a naïve realist and presume that we have direct access to the world, or perhaps a scientific realist, for whom only indirect access is granted as mediated by the necessary theoretical standpoint that one has to deploy to conceptualize reality in the first place. Nonetheless, as described above, the basic thesis of truth depending in one way or another on existence stands. While realism is perhaps the most prominent approach to defining truth in contemporary literature, it is in no way devoid of problems, as noted by David:

> We cannot step outside our own minds to compare our thoughts with mind-independent reality. Yet, on the realist correspondence view of truth, this is what we would have to do to gain knowledge of the world. We would have to access reality as it is in itself, to determine whether our thoughts correspond to it. Since all our access to the world is mediated by our cognition, this is impossible. Hence, on realism, knowledge of the world would be impossible. Since knowledge of the world is possible, realism must be wrong. (David, 2020: 9.2)

Indeed, correspondence theories of various sorts have traditionally been treated as a prime example of realist theories. These theories are rich with metaphysical implications, not least because many of them make explicit commitment to some form of realism. Similar strategy of tying considerations about truths nature to metaphysical commitments has been utilized by anti-realists.[14] For anti-realists, truths nature is independent of the nature of the world. Dummett – perhaps the most prominent anti-realist to date – argues that truth can be defined through justification (1959). Truth

[13] Indeed, if one wants to dispense with a scientifically legitimate account of truth, then they bear the burden of explaining what the contents of our theories are if not *true* sentences.

[14] Note that the correspondence relation itself does not require a commitment to realism. For example, the sentence “I am hungry” can correspond with my *feeling* with being hungry, which does not map consistently to anything in the real world.

is *just* justifiability, warranted assertability or verifiability. From this, we get an account that rejects the law of the excluded middle and bivalence: we know some sentences to be true, and others to be false, but we also don't know and can't even in principle some things. According to one formulation of this thesis, what one is justified to believe in any instance *is* true. More robustly, one could argue that what our most potent scientific explanation of the world warrants us to believe counts as true. But whether or not we want truth to be so relativized is debatable.

On a more delicate level, the notion of anti-realist truth can be illustrated by presuming that as we don't have any direct access to reality in itself, all the objects we perceive, or the ontology of our science, is more or less theoretical. We speak about rocks and planets but not in any mind-independent or transcendent sense, for the referents of these terms are concepts, abstract representations, or classes of entities associated with them. Arguably, none of these things exist outside the domain of conscious beings. We are not speaking about anything *real* in the robust, realist sense where existence is grounded in mind-independent facets of the external reality. From this anti-realist claim follows that the truth of "rocks are solid" and "planets are round" has nothing to do with any particular rocks or planets in the world. Rather, one is free to argue that the terms "rock" and "planet" designate abstract entities, theoretical constructions, and the predicates "is solid" and "are round" simply have consistent rules of application and do not map anything significant in similarities between the entities *in themselves* that satisfy them. The truth of these sentences depends on the human-constructed language in which they are formed, and the internal balance of the truth values of those sentences that constitute the theory that gives rise and justifies them. No reference to extra-theoretic, actual, or mind-independent states of affairs is required.[15]

Relying on the aforementioned distinction between realist and anti-realist approaches, we can separate and contrast the core commitments of both types of theories in the following way:

T-realism: reality exists independent of minds, and our thoughts are about its aspects.

Commitment to ontological realism.

Law of the excluded middle and bivalence.

Warrant independence.

T-anti-realism: truth depends on human-practices, such as knowledge or justification.

No commitment to ontological realism.

Rejection of the law of the excluded middle and bivalence.

Warrant dependence.

[15] Excluding observational sentences like "this rock is wet" that represent only a small portion of truth-apt speech, and it is questionable whether these sentences are any less language and theory-bound in the first place. Overall, the whole question of "reality" past what is conceivable from *some* framework of understanding is suspicious.

Interestingly enough, both realist and anti-realist views can be subjected to what is in contemporary literature known as the *scope problem*. This problem follows form the *either-or* way of thinking about truths grounding. Think about the following true sentences: “killing innocent people for pleasure is wrong” and “earth is an unevenly shaped ellipsoid”. As happens, realists have difficulty in accounting for the truth of the first sentence, for there is no fact in the world that the sentence can correspond with, and the anti-realists have difficulty in accounting for the truth of the second sentence, insofar as they don’t commit to realism. Because both theories limit their scope regarding truths grounding, they face issues with accounting for the truth of the full range of truth-apt discourse. Contrary to this, one can hold that depending on the truth-apt content, its truth can be grounded in either human-dependent or -independent factors. This approach has been labeled as the now prominent pluralist approach to truth: "Wright argues that a pluralistic perspective on truth can be put to work in developing a nuanced understanding of realism and anti-realism about various philosophical domains." (Assay, 2016: 177) For example, the pluralist can argue that there is one general way for truthbearers to be true, yet depending on the *subject matter* of the truth-apt content, its truth can be grounded in different things, i.e., properties. Some sentences are true because they correspond and others because they cohere. While all truths are true in some unified sense, the grounds of truth are many. In this sense, truth becomes relativized on the realist/anti-realist spectrum between categories of speech, i.e., *domains of discourse*. This is only intuitive, for surely there are truths about mind-independent states of affairs, like stones and chairs, and truths about ethics and law, such as what one is permitted or prohibited in doing.[16]

Despite the intuitive appeal of the general pluralist thesis, this approach has failed to attract widespread popularity. One reason being the inability of the current pluralist models to resolve certain technical issues and thus convince scholars of the philosophical tenability of this position (Tauriainen, 2021). However, the skepticism that truth pluralism faces in current debates is only justified, for considering the long tradition of theorizing about truth, one has to be suspicious towards any “novel” approach to defining this notion. As the story goes about Strawson’s introduction of his performative analysis of truth to his colleagues as a new kind of truth theory, one of them allegedly replied: “Come on now, which of the old ones is it?” (Strawson, 1974: 23). Thus, it remains to be seen whether truth pluralism is able to attract widespread popularity by offering a philosophically tenable account of truth that satisfies both realist and anti-realist intuitions. In any event, pursuing such account is, to my contention, a worthy effort.

[16] More on pluralist theories of truth in Wyatt & Lynch (2016), Edwards (2018), Pedersen (2020), Ferrari (2021).

Interestingly enough, one can find a closely related debate on the human-dependence or independence, and realism or anti-realism, from the historical literature on truths nature. Surprisingly, the historically significant distinction between truths immanence and transcendence has largely been neglected in contemporary debates. As proves to be the case, there are lessons to be learned by all, realists, anti-realists, and pluralists from this debate.

4 Truths Immanence and Transcendence

Two prominent commentators on truths immanence and transcendence are Kant and Quine.[17] In Kant's usage, immanence is contrasted with transcendence through cognition and its target.[18] The conditions for the possibility of cognition are transcendent, yet the cognition itself is a precondition for the domain of immanent thoughts, and experiences. In Quine's usage, all concepts, such as knowledge, reality and truth are immanent, or contained in the theory that gives rise to them: "Truth is immanent, and there is no higher. We must speak from within a theory, albeit any of the various" (Quine, 1981). In this latter explanation, all concepts are essentially human inventions, theoretical constructions, and as such bound to the framework in which they are *contained.*[19] Thus, in both Kant's and Quine's usage, immanence is contrasted with the transcendent, which in one way or another describes that which exceeds our cognition or our theories, which are human-bound domains. Again, insofar as humans are creatures with inherently limited capabilities of understand the endlessly complex reality that we partake in, the question of immanence and transcendence is constantly present. We perceive entities, but not in themselves, and we understand their nature, but not completely. This can be illustrated with an example. We know there to be not only waves, but different types of waves. There are small and large waves that vary in their frequency and speed, and in scientific discourse, we acknowledge wind waves, wind surges, and tsunamis. In transcendental terms, however, there are no waves *per se*, let alone different types of them, for there are only masses of water molecules with no boundaries in between that act upon the forces imposed on them. But in some sense, there seems to be truth to both sides of the story. It is a necessary precondition of human understanding that one must deploy *some* immanent perspective to make sense of the world, even though the world itself transcends our ability to know about it in its totality.[20] Simply put, to think and speak at all is to do so from some perspective, better or worse.

Indeed, while the notion of there being a transcendent, human-independent reality that at least sometimes dictates or guides the truth and falsity of our claims is sensible, the idea that truth in the

[17] I will bypass the topic of whether and to what extent these terms have been used in distinct ways, as indicated by Sher: "The word "immanence" has been used in the philosophical literature in a variety of ways." (Sher, 2016: 163).

[18] Immanence doesn't necessarily mean "containment in the world" (medieval use) but "containment in our understanding about the world" (modern use).

[19] Thorough analysis of the immanent/transcendent distinction in both Quine and Kant can be found in Sher (2016).

[20] Here I am assuming that there indeed are transcendent truths. Quine, for example, is critical towards this view: "what reality is really like /…/ is self-stultifying" (Quine, 1992: 405).

immanent, inherently limited sense would be something of secondary interest is problematic for various reasons. First, it is not clear to what extent we have access to the transcendent in the first place, and if we do, then it is reasonable to assume that reality in its near infinite complexity is indeed *too* complex for us to grasp past what is given in our limited human representations in the immanent domain.[21] For example, think about the utility of maps, which oftentimes do not even aim at being fully objective representations of the thing they describe. More often than not, maps are crude simplifications of the geometrical structure of their target that lack dimension. Nonetheless, they are extremely useful instruments for achieving success in our practices. For example, an objective representation of the New York subway would contain too much information for it to be helpful in guiding navigation. For our benefit, we can simplify the transcendent as given by our perceptions about it in our models.[22] To claim that these models wouldn't be true in any meaningful sense is peculiar, to say the least. Indeed, if reality is transcendent in the sense that we cannot grasp it past the mediated access granted by our representation about it, then there seems to be no way for us to know about the transcendent in the first place. Reason being that even the term "transcendent" is a theoretical term, human construction there to help us make sense of the world, and thus a member of the immanent domain. In other words, the term "transcendent reality" is a concept and as such a proxy for something in the world, or se we assume. There simply is no way for us to evaluate the Truth or full objectivity of our claims past the perspective that lays the foundation for this evaluation in the first place.

Of course, we would like to think that truth is as objective as it can be, but asking for it to be fully objective seems to be too much. It is extremely useful for us to hold that some things *are* true, even if not in the fully objective sense. Again, to quote Quine: "there is no extra-theoretic truth, no higher truth than the truth we are claiming or aspiring to as we continue to tinker with our system of the world from within. (Quine, 1975b: 327). We accept our fallible, theory bound notion of truth and deploy it in our theorizing, despite the fact that all the truths we now uphold might turn out false in the long run.[23] Truth, just like all other concepts, is there to aid our aspirations for knowledge and understanding, and if this means that it fails to be fully objective, then so much worse for the notion of full objectivity, were there such thing in the first place.

[21] Remember the aforementioned example about waves. We cannot understand the complexity of masses of water molecules, so we organize reality into bits and pieces to make sense of it with our limited capabilities.

[22] Of course, immanent models of the world can vary in their complexity.

[23] One can perceive two foundational commitments in this discussion: anti-transcendentalism and fallibilism. There is no transcendent, extra-immanent perspective to the world, and immanent perspectives are always limited and correctible to some extent.

However, there is a way to save our intuitions about truths partial transcendence without violating the above-mentioned claim of truths near throughout immanence.[24] In the following and final section, I proceed to argue that truth is after all a partially transcendent notion, and that this partial transcendence is an indispensable feature of a philosophically tenable and scientifically legitimate understanding truth. Thus, instead of being either mind-dependent or independent, or immanent or transcendent, truth indeed turns out to encompass both aspects.

5 Between the Immanent and Transcendent

I find the claim of truths simultaneous immanence and transcendence in the aforementioned sense appealing for various reasons. Let's start with a perverse example with computers, gods, and humans. Insofar as computers don't usually have any sensors for perceiving the world, truth for them is a wholly immanent matter. They are closed systems that operate with some internal logic with no regard for what rests beyond. Of course, there are potential inputs from humans, and the operating logic must come from somewhere. But when the foundation is there, the computer can run as long as it wants in solitude. On the other hand, for an omnipotent being such as God, truth would be arguably immanent in his understanding, but its target would be what we conceive as the domain of the transcendent.[25] In God's understanding, the immanent and transcendent would become one. But humans are neither computers nor gods. Rather, we are beings that in many ways lie in between. We operate with some internal logic, no matter its origins, yet we also have access to a fraction of the transcendent through our perceptions. My firm contention is that it is in this subtle interplay between limited cognitive agents and the infinitely complex reality that lays the foundation for the phenomena of truth to emerge. Simply put, both the immanent perspective and the transcendent domain of existence is required to explain truths nature. Regarding metaphysical commitments, for truth to be even partially transcendent, a commitment to realism in one way or another seeps in. It doesn't matter whether the reality assumed is the domain of extensional entities, or whether it is the domain of entities that some mad professor has fabricated. In both cases, we have an interplay between the immanent perspective or minds, and the transcendent reality in all its complexity. But are there any practical reasons for committing to truth partial transcendence in the aforementioned sense?

Let's assume that there are two mutually incompatible theories or understandings about some aspect of reality. If truth were fully immanent to the theory in which it is formed and to which it belongs,

[24] In this sense, the truth accessible to us humans is neither solely immanent nor transcendent, but an interplay between both. In this sense, the truth pluralist is right in trying to accommodate both realist and anti-realist intuition under one notion.

[25] I understand the limits of this analogy, but I think it is nonetheless illuminating.

then there would be no reason to seek resolving these sorts of conflicts. First, it is not clear what these theories would even argue about, for truth would be relative to these theories separately. But in our concrete practices, we seek to resolve theoretical conflicts, for this is what scientific development for a large part explicitly means. The question that emerges is that in relation to *what* are we disagreeing on, or in relation to what are we trying to resolve conflicts in epistemically serious contexts? Indeed, insofar as our theories are *about* something, this thing that they are about both guides our aspirations towards it, and when reached, helps solve disagreements about its nature. But for our theories to be about something, this something needs to rest *beyond* our theories, or they would simply be closed systems about themselves. If this were the case, then we could simply fabricate the truth in the context of any given theory, and it would not be clear in relation to what competing theories would disagree on. Rather, as seems to be the case with our actual scientific practices, we seek to resolve theoretical conflicts by figuring out the *truth*. Indeed, if and when science encompasses both theoretical branches like mathematics and logic, and branches that study empirically verifiable events, then for all theories there is a class of sentences that have their truths values effected by transcendent factors. In formal sciences, these sentences connect with the formal system and its properties, which we are aiming to discover in our theorizing about them. In natural sciences, these sentences are the ones that directly connect with empirically verifiable events. In both cases, our sentences connect with something that transcends our current theories about it.

Indeed, it is because of those sentences that directly connect with the transcendent, we are able to make predictions and evaluate the success of *all* theories. We want to possess true sentences to have them as premises of valid inferences that preserve their truth to the conclusion, and we want to discover true sentences to simply satisfy our curiosity about the world. If one theory is more accurate in predicting actual states of affairs as given by our perceptions about them, then we treat it as more successful than others. Same applies, to some extent, to the explanatory power of different theories. These claims are analogous to the way in which less formal common-sense reasoning is executed. Think about the phenomena of intersubjectivity. When we assert that the weather is nice today, people tend to either agree or disagree on the truth of this statement based on their own perceptions about *actual states of affairs*. It is hardly arguable that the source of these perceptions lies in the domain of the transcendent.

Thus, even though the claim of truth throughout immanence holds in the sense argued for in the last two sections, the shadow of transcendence is always present. The transcendent is what we assume to exist behind our proxies about it, and through these proxies we have evidence for its existence. This couldn't be more evident that in the phenomena of scientific progress. We might hold as firmly as possible that this or that claim about some aspect of reality is true, but when scientific development shows it to be false, we never say that it *turned out* or *became* false. Rather, we simply admit that we were wrong all along, and correct our beliefs in relation to the *truth* as given by the transcendent. What is the exhaustive nature of this truth that we are near infinitely wrong in relation to? There is good reason to be skeptical towards any answer to this question that exceeds the immanent perspective.

Hkratna imanenca in transcendence resnice

Razprave o tem, ali je resnica odvisna ali neodvisna od človeških interesov, je vseprisotna v zgodovini Zahodne misli. Za nekatere je resnica odvisna od človeku neodvisnega dejanskega stanja, kot so dejstva ali relevantni vidiki razsežne realnosti. Za druge je resnica odvisna od človeških dejavnikov, kot so kaj je koristno verjeti ali kaj je v nekem obdobju znanost dokazala kot resnično. V bolj nedavni zgodovini to razpravo imenujemo razprava med realisti in antirealisti. Za realiste je resnica izbranih nosilcev resničnostne vrednosti osnovana na ustrezni relaciji teh s svetom. Za antirealiste je resnica določena glede na to kaj nekdo vé oziroma v kar je nekdo upravičen verjeti. Konflikt med realizmom in antirealizmom se je izkazal za trdovratnega, saj imata obe strani prepričljive argumente, ki podpirajo njune pozicije. Zagovarjal bom trditev, da bi obema stranema koristilo zgodovinsko pomembno, vendar danes bolj kot ne pozabljeno razlikovanje med imanenco in transcendenco resnic. V skladu s tem razlikovanjem pokažem, kako lahko resnica vključuje tako od človeka odvisne kot tudi od človeka neodvisne aspekte. Moja trditev je, da je upoštevanje takšne večplastne narave resnice nujno za doseganje filozofsko verjetnega in znanstveno legitimnega opisa narave resnice, ki je lahko uporabljeno tudi kot razlagalni inštrument pri definiranju različnih družbeno pomembnih pojavov, kot so védenje, zaupanje in prevara.

Ključne besede: resnica, imanenca, transcendenca, antirealizem, realizem.

Works Cited

Arendt, H. (1951). *The Origins of Totalitarianism*. Berlin: Schocken Books.

Aristotle. (1908). *Metaphysics*. English translation by W.D. Ross. *The Works of Aristotle Translated into English Vol. 8*. Oxford: Clarendon Press.

Armstrong, L. (2004). *Truth and Truthmakers*. Oxford: Cambridge University Press.

Assay, J. (2016). "Putting Pluralism in its Place". *Philosophy and Phenomenological Research*, pp. 175–191.

David, M. (2020). "The Correspondence Theory of Truth". In Zalta, E. N., *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition). URL = <https://plato.stanford.edu/archives/win2020/entries/truth-correspondence/>.

Donald, D. (1990). "The Structure and Content of Truth". *The Journal of Philosophy*, 87, pp. 279–328.

Dummett, M. (1959). "Truth". *Proceedings of the Aristotelian Society*, 59(1), pp. 141–62.

Edwards, D. (2018). *The Metaphysics of Truth*. Oxford: Oxford University Press.

Ferrari, F. (2021). *Truth and Norms: Normative Alethic Pluralism and Evaluative Disagreements*. Lexington Books.

Glanzberg, M. (2018). *The Oxford Handbook of Truth*. Oxford: Oxford University Press.

Glanzberg, M. (2021). *Truth*. In Zalta, E. N., *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition). URL = <https://plato.stanford.edu/archives/sum2021/entries/truth/>.

Kakutani, M. (2018). *The Death of Truth: Notes on Falsehood in the Age of Trump*. Tim Duggan Books.

Pedersen, N. (2020). "Moderate Truth Pluralism and the Structure of Doxastic Normativity". *American Philosophical Quarterly* 57(4), pp. 355–376.

Plato. (1997). *Complete Works. Cooper* J. M., & Hutchinson. D. S. (Eds.). Indianapolis, Ind.: Hackett Publishing.

Quine, W. V. (1992). "Structure and nature". *Journal of Philosophy*, 89(1), pp. 5–9.

Quine, W.V.O. (1981). *Theories and Things*. Cambridge, Mass.: Harvard University Press.

Sher. (2016). *Epistemic Friction: An Essay on Knowledge, Truth, and Logic*. Oxford: Oxford University Press.

Shieh, S. (2018). “Truth, Objectivity, and Realism”. In Glanzberg, Michael (Ed.), *Oxford Handbook to Truth*. Oxford: Oxford University Press, pp. 433–476.

Strawson, P. F. (1974). *Freedom and Resentment, and Other Essays*. London, England: Routledge.

Tauriainen, T. (2021a). “No Safe Haven for Truth Pluralist”. *Acta Philosophica Fennica* 97:183–205.

Tauriainen, T. (2022b). “On Peterson’s Truth”. In Sandra Woien (Ed.), *Jordan Peterson: Critical Responses*. Carus Books: Chicago, pp. 251–264. Forthcoming.

Wyatt, J. & Lynch, M. (2016). “From one to many: recent work on truth”. *American Philosophical Quarterly*, 53(4), pp. 323–340.

Dominika Palka

University of Warsaw

Reasons-Responsiveness: One More Look at the Agent-Based and Mechanism-Based Accounts

According to Fischer and Ravizza, having alternative possibilities of action is not a necessary condition of moral responsibility. The authors are convinced by the Frankfurt examples and motivated to hold this position due to the unresolved issue of compatibility of determinism and the possibility to act otherwise. Instead, they propose that what we need for holding people morally responsible is that they possess a kind of control called "guidance control"; and this is dependent on the mechanism which leads to action's being moderately reason-responsive. The notion of moderate reasons-responsiveness (MRR) is explicated as follows: a type of mechanism M is MRR if and only if (1) in many possible worlds where it operates and S has a sufficient reason to do otherwise, S recognizes that reason as sufficient; (2) S's disposition to recognize different reasons forms an intelligible pattern; (3) in at least one of those worlds, S acts on the recognized sufficient reason. After the authors present their account in their book *Responsibility and Control. A Theory of Moral Responsibility* (1998), a lot of discussion and criticism emerged. One of the main points of criticism was the vagueness of the notion of mechanism and the lack of identity criteria for different types of mechanisms. This led some philosophers, Carl Ginet and Michael McKenna among them, to adopt an agent-based reasons-responsiveness theory. This modification faces other problems. First, it has been suggested that it is vulnerable to Frankfurt-like examples, since the agent with whom the counterfactual intervener is connected is not reasons-responsive, although she is morally responsible. This charge is rejected both by Ginet and McKenna. However, the latter admits that there are possible examples, concerned with "intervener" who is internal to the agent, that may threat the agent-based theory. One of his undeveloped examples, constituting the second problem, is the situation in which the mentally ill agent acts while his sickness is not active, but it would activate and prevent him from acting otherwise in alternative scenarios. In this paper, I present examples based on McKenna's suggestion and argue that they pose a much more serious problem for agent-based account than McKenna would like to admit. I also present a possible counterexample to both agent-based and mechanisms-based account concerning a recovered addict who refuses to take a drug, yet seems to be morally responsible (in this case – praiseworthy) for her resistance and yet is not reasons-responsive in the right way.

Keywords: incompatibilism, semi-compatibilism, reasons-responsiveness, moral responsibility

1 Why Reasons-Responsiveness Theory?

In 1998 John Martin Fischer and Mark Ravizza proposed, in their insightful and heavily discussed book *Responsibility and Control. A Theory of Moral Responsibility,* the new approach to assign moral responsibility to acting agents. They developed this particular view because of two motivations, which can be referred to as positive and negative incentives. First – and this is the negative incentive – the authors acknowledge that there is a potential threat from physical determinism to freedom understood as a possibility to act otherwise. Although neither of them claim to be fully convinced by incompatibilist arguments, they admit that a version of consequence argument proposed by Peter van Inwagen (van Inwagen, 1975: 188–194) is strong enough to urge us to look for accounts of responsibility that would allow us to hold people morally responsible even if both determinism and incompatibilism appear to be true.

The second, positive incentive, is the appeal of Frankfurt examples (Frankfurt, 1969: 835–836) concerning the possibility of being morally responsible for acting in a particular way in the situations where one has no alternative possibilities of action. Those are the cases of the following structure:

1. An agent S is driven to a decision D by her values, beliefs, desires /…/; and then she performs the action A as a consequence of that decision;
2. There exists a factor (originally: another agent) that would intervene if S was about to make any decision other that D in such a way that A would anyway make decision D and consequently do A;
3. It is intuitively plausible that S is morally responsible for A.

Convinced by these type of examples, Fischer and Ravizza try to build the theory that accounts for the responsibility of the agent in situations that fall under the presented scheme.

2 The theory – some important features

The most important notion of Fischer and Ravizza's theory and other theories that are based on it yet differ in some details is the notion of reasons-responsiveness. In the most general manner, this notion can be introduced as follows:

> 'An entity E is reasons – responsive if and only if in an appropriate range of scenarios E recognizes reasons for acting differently than it acted and it modifies its behaviour in accordance with those reasons.'

Simple as that, but there is some nuance to be noticed here: to say that something is reasons-responsive in the specified sense is not to say that ‘it could have modified its behaviour (in the actual scenario); although it is possible to either argue that this is a real meaning of a phrase or to make a revisionist twist and claim that it should mean exactly that.

It is useful to inquire into the intuitions that underlie the view that possessing a property at which such a notion points at least partially grants or grounds moral responsibility – or is a necessary condition of it, dependent on a version of the theory. In a wide sense it focuses on the rationality of the entity whose responsibility is being assessed. In the case of an agent, this demand for rationality stands for the fact that it is generally agreed that agents who we do not consider as sane in a relevant sense should not be held responsible. This sanity can be roughly described as consisting of such traits as:

- having beliefs that are at least partially grounded in reality;
- performing thought and decision processes that are at least partially understandable;
- displaying at least a partial consistency between one’s beliefs, desires, and actions.

What exactly stands for the “entity” in the presented definitional scheme? There are two options, concerning which there are huge disagreements among philosophers who adopt reasons-responsiveness theory. Fischer and Ravizza insist on treating mechanisms that lead to action as subjects of reasons-responsivity. It is so because, so they claim, if we demand that the agent herself is a subject of reasons-responsivity, the account falls prey to Frankfurt-like examples. Recall that, in Frankfurt examples, there were no scenarios in which the agent acted otherwise: in every scenario in which she was about to decide to act differently, the intervener intervened. In a sense, it was something *about the agent* that there has been an intervener associated with her. On the other hand, it was not something in any way about the particular mechanism that led to action. Conversely, the key feature of this mechanism – the actual scenario’s mechanism - is that it is the agent’s own mechanism that operates without any external influence. Holding the mechanism fixed with this feature while considering different scenarios allows us to explain why, when we consider some particular situation, even with an assumed counterfactual intervener, we can focus on what the agent would have done *if no intervener were to be present* when assessing her moral responsibility.

However, the mechanism-based approached faces a serious problem: how to individuate between mechanisms operating in different scenarios? How – and if – can we (not too vaguely) specify what is the mechanism that operates in the actual scenario? Fischer and Ravizza themselves admit that they are not in a position to provide any good specification of this notion and that they are unable to provide strict criteria for the individuation of mechanisms; neither was there anyone who would develop their account in a way that provides these missing elements. Alternatively, agent-based variations of the theory were developed to avoid the problem posed by mechanisms.

3 Into details

Since agent-based approaches are not completely independent types of reasons-responsiveness theory, but rather its modifications based on a common core that is provided by Fischer and Ravizza's analysis of the notion of reasons-responsivity, I will now use some space to recall their detailed explication of this notion. Until now, I was using the words "reasons-responsiveness" without further qualification. From now on, I will be talking about *moderate* reasons-responsiveness (MRR). Reasons-responsiveness comes in degrees, and MRR stands for exactly how responsive the agent must be to be possibly regarded as a proper subject of moral responsibility.The best definition of MRR, extracted carefully from Fischer and Ravizza's book, is provided by Carl Ginet in the following way:

> Moderate reasons-responsiveness for a positive action
>
> The kind of mechanism that issued in S's doing A was moderately reasons-responsive if and only if:
>
> (1) in many different possible worlds where the same kind of mechanism operates in S and S has a different sufficient reason to do otherwise, including some where the reason is a moral one, S recognizes that she has a sufficient reason to do otherwise;
>
> and
>
> (2) S's dispositions to recognize various different reasons (including some moral ones) for doing otherwise form a pattern that is rationally intelligible (given S's beliefs and basic values)
>
> and
>
> (3) in one of these possible worlds S recognizes the sufficient reason to do otherwise and chooses to act, and acts, otherwise, for that reason. (Ginet, 2005: 232–233)

Fischer and Ravizza understand actions as bodily movements. They must therefore provide the extension of their account to the responsibility for consequences of actions, where consequences are taken to be states of affairs. Thus, the consequence of my action A, say, raising my hand, is a state of affairs: *that my hand is raised;* and also some further consequences, for example: *"that I* *voted for obligatory vaccinations"* or any other states of affairs that results from my bodily movement.

The formal account of the consequences of actions consists in two parts: (1) the requirement that the mechanism that led to a bodily movement is reasons-responsive, in a way that was defined in the preceding definition; and (2) the requirement that the process leading to a consequence of an action is properly sensitive to the bodily movement. Formally:

> If S were to move his body in way B* [different than what she actually did at T] at T, all other triggering events (apart from B*) which do not actually occur between T and T+i or which actually occur and bring about C simultaneously or subsequently to T+i were not to occur, and a P-type process were to occur, then C would not occur. (Fischer and Ravizza, 1988: 120–121)

Equipped with these definitions, we can proceed to the main part of the paper: presentation and criticism of the alternative, agent-based account proposed by Michael McKenna.

4 Michael McKenna's agent-based approach

According to Michael McKenna and some other philosophers who discussed the reasons-responsiveness account, e.g., Carl Ginet, what should be reasons-responsive is the agent herself, and not any mysterious mechanism. For McKenna, the first condition of moderate reasons-responsivity – the receptivity condition – is defined in the same way as by Fischer and Ravizza for the mechanisms, but the second condition – the reactivity condition – is modified so that it can be allowed to be reactive without being able to react otherwise.

How can we make sense of the idea of reactivity without being able to react otherwise? The proposal owes its plausibility to the recognition that, in the Frankfurt examples, there is an important difference between scenarios in which an agent recognizes sufficient reasons to act otherwise and *would act on them if there were no intervener* and those in which his recognition of reasons does not even motivate him to (try to) act on them. The difference is that in some scenarios, the intervener must intervene; and because of this, in those scenarios, the agent is not the source of her actions; she does not cause her action in a proper way. McKenna justifies his interpretation of theory by referring to the type of conditional used by new-dispositionalists:

> (RR1): If S were to become aware of a reason R (to act otherwise than X) and if S retained during the relevant duration of time intrinsic agential properties P1-Pn, and if S were not interfered with in a way that would impede the casual efficacy of those properties, then S would not do X.

Although McKenna admits that this conditional may not account for the possibility to do otherwise, he claims that it still says something important about the agent herself: specifically, it says that the agent, just as she is, is reasons-reactive even if she cannot react differently.

Now I will present and advance the problem for agent-based theory that is mentioned but not developed by McKenna in his 2011's "Reasons-Responsiveness, Agents and Mechanisms". (McKenna, 2013: 170–176)

5 The problem of intrinsic undermining-possibility factors

Let there be – instead of some counterfactual external intervener – some property of the agent herself that does not play a part in the actual scenario yet that would prevent her from acting otherwise if she tried or wanted to do so. Specifically, consider the following scenario:

> Anna has just become a student of philosophy and she is deliberating whether she should go to the first exam on her faculty. She is well-prepared and she should pass it without problems; more than that: she knows that if she does not go, she will fail the exam and thus lose her scholarship, which would in turn cause her mother, who is already taking financial care of five of her other children, a great distress. But she is just not in the mood and she decides to watch the movie instead of going to the exam. However, Anna suffers from a serious phobia, which she (nor anyone else) has discovered yet: she is extremely terrified of taking any exams at the University. In any scenario in which she would be just about to make a decision to go, she would become extremely frightened and, in the end, decide to stay home anyway.

It seems that, in the actual scenario, Anna is responsible not only for her decision not to go to the exam, but also for not going there, and therefore for causing her mother's distress. It can be shown, that McKenna's and Fisher's versions of reasons-responsiveness theory yield different results here, and that the result given by mechanism-based theory seems to be the proper one, in contrast to the result which we obtain if we choose the agent-based theory instead. McKenna's account fails in two ways. First, the justification in terms of new-dispositionalists conditional is no longer available, since the conditional is just false:

(a) If we assume that causing her mother distress is a sufficient reason for Anna not to make the decision she makes, it fails just because of the facts obtaining in the actual world, those facts being:

> (a.1) Anna's recognizing a sufficient reason for not making the decision not to go to the exam and Anna's retaining all of the relevant agential properties during the relevant period and
>
> (a.2.) Anna's making the decision not to go to the exam.

Since (a.1) makes the antecedent of the material conditional obtaining in the actual world true, and (a.2) makes the consequent of this conditional false, the whole material conditional is true. And since according to the most probable account of counterfactual conditionals, provided by D. Lewis, if a material conditional is false in world w, the associated counterfactual is also false in w, we can conclude that the new dispositionalist conditional is false.

(b) If we assume that Anna is just not such a decent person and causing her mother's great distress does not count as a sufficient reason for her to refrain from making the decision that causes the distress, we have to consider different possible worlds. Assume that Anna would recognize it as a sufficient reason if her mom were to have a heart attack as a result of her action. We then obtain a following conditional:

> If (b.1) Anna was to become aware that her mom would have a heart attack if she decides not to go to the exam, and (b.2) she retained relevant agential properties P1-Pn during that time and (b.3) if she was not interfered in a way that would impede the casual efficacy of those properties, then (C) Anna would decide to go to the exam.

This conditional is, again, false. If the antecedent (b.2) were true, Anna would also retain the property of having a terrible phobia of taking exams at the University. This property would activate when Anna would be just about to decide to go to the exam and prevent her from making that decision, therefore causing (C) to be false. The truth of (b.3) is not impeded at all, it is just because of the causal efficacy of some of the properties she has that she fails to make the decision.

Also, Anna is not reasons-responsive in a sense specified by McKenna, because she fails to be reasons-reactive even on the weakened reading of reactivity. Let R-Ry be the range of reasons which Anna would recognize as sufficient, and R1-Rn the range of reasons that would motivate Anna to (try to) do otherwise. Anna would react to those *if she hadn't suffered from* phobia; in contrast, she wouldn't react to the reasons from the range Rn-Ry regardless of whether she suffers from phobia or not.

Unfortunately, this time the difference is not enough, at least if we want to claim, as McKenna did, that the key feature of the difference is that, in some of the scenarios, the agent is the ultimate cause of her actions. Both in the scenario in which the reasons from the range R1-Rn are present and in those where Anna is confronted with reasons from Rn-Ry range, it is her – and only her – who or what is the cause of her actions.

On the other hand, according to Fischer and Ravizza's mechanism-based theory, Anna is responsible for both her decision and her mother's distress. Since responsibility for the action itself follows on their account from the responsibility of the action's consequences, I will only show how this account grants Anna's responsibility for her mother's distress. Keeping in mind the pertinent problem of mechanisms individuation, let me vaguely specify the mechanism playing this role in Anna's making the decision not to go to the exam and following this decision as "more or less rational deliberation concerning going to the exam". Now, the first condition for Anna's responsibility is fulfilled: there certainly are possible scenarios in which this (type of) mechanism operates and leads Anna to different action than staying at home watching TV. Also, the second condition is satisfied. There are other bodily movements, constituting Anna making the decision to go to the exam, which, other things being equal, would lead to a situation in which Anna's mother would not be distressed.

Fisher's result – that Anna is responsible for her decision, her following the decision, and her mother's distress – is certainly the more intuitive one. However, without going into any details, McKenna points at two possible ways of defending his theory. I will now consider his proposals and give some arguments against treating them as good ways of proceeding with establishing the plausible account on moral responsibility.

6 (Im)possible solutions to the problem of intrinsic undermining-possibility factors

1.

One option that McKenna finds at least worth considering is to walk away a bit further from the common intuitions and simply accept the conclusion that his account yields in problematic cases. But this proposal should not be pursued. Those are not only the reasons which have to do with the incentive to stay as close to intuitions as possible that witness against it; there are normative reasons as well, and they seem to be even more important.

Cases suggested by McKenna are much more down-to-earth than those usually employed in Frankfurt-style examples, and they find application in stories that can be easily encountered in real life. The agents in question suffer psychological impairments that do not allow them to perform some particular actions; although they can freely perform some actions which fully constitute not-performing that actions. If we accept McKenna's result in case of Anna, we will have to accept that in every situation of this type the mentally impaired agent is not responsible for at least those consequences of her actions that follow from the omission that her actual-scenario action constitutes. Also, we should accept that if an agent performs some positive act – where she performs it on the basis of her factual desires, beliefs and so on – such that she would perform that act anyway, in any differing circumstances, she is not responsible for the act. If we do not simply adopt incompatibilist intuitions, according to which an agent is never responsible for the act which she could not have avoided in the first place, there seems to be no convincing reason why we should adopt this conclusion in the case of agents suffering from some hidden phobias or diseases. To the contrary, it seems that we just should not do that. Being a proper agent of moral assessment is usually strongly connected with personhood, and if we want to claim that in the types of situations discussed – among which we should count inactive phobias, psychosis in regress and some other mental issues that are temporarily not active – we are moving dangerously close to excluding people suffering from mental illnesses from the sphere of moral agents, and thus, plausibly, from subjects that we would call “persons”.

However, it is worth noticing that if we wanted to change our minds and accept that we should not hold Anna ultimately responsible, then it would constitute an independent reason against Fischer’s version of the theory.

2.

The other undeveloped solution which McKenna points at, is to somehow exclude properties constituting psychoses, phobias, and other mental impairments of the relevant kind from those which constitute the agent herself. Although, in contrast with the first proposed solution, this one seems more reasonable from a wider perspective, which deals with the issue of who should be regarded as a moral agent, it does face a technical difficulty: it simply looks like a step back towards some variation on mechanism-based theory. The key problem is that we cannot simply specify some range of conditions, say ‘suffering from phobia or psychosis’, which we would always exclude from the relevant agential properties when assessing agent’s responsibility for the performed action.

We cannot do that because in a lot of scenarios it would yield a seriously undesired result: that the person is reasons-responsive and thus, at least in most cases, responsible when the psychotic episode *is* present.

Consider Jane, who steals a bike belonging to a tourist who carelessly left it by the store without any protection. Jane took the bike because she was strongly convinced that it is a gift for her from

Saint Mother Mary: Jane is a very religious person with schizophrenic tendencies; she was praying relentlessly over the last few weeks for a new bike, and when she entered into the psychotic episode she just came to believe that the first bike she saw is the fulfilment of her prayers.

But now, if we insist on the exclusion mentioned earlier, Jane seems to be perfectly reasons-responsive in performing her action. There are certainly some reasons that Jane would both recognize and act on them in alternative scenarios in which she retains all her relevant properties; her psychosis, which played the key role in the actual scenario, would not be counted among them. Even if we do not want to treat reasons-responsivity as a sufficient condition of responsibility, and thus we may find a way to explain not holding Jane responsible for her action, it simply does not seem right – having in mind all that had been said so fair about reasons-responsiveness – to say that she was reasons-responsive while stealing a bike.

On the other hand, if we try to specify the conditions in order to include some crucial features of the actual scenario, for instance: 'suffering for phobia, which did not work while the agent was performing the action', how far do we really move away from the mechanism-based account?

Faced with the abovementioned difficulties, I conclude that internal non-active undermining-responsibility factors are a serious threat for agent-based theories even if they can be convincingly defended against Frankfurt-type examples with external intervener.

7 Diana, the addict – the problem for both approaches?

Finally, I would like to discuss the example that may at the first sight seem to pose a problem for both mechanisms-based and agent-based versions of reasons-responsiveness theory.
Consider the scenario described below:

> Diana, the recovered drug addict from a quite pathological family faces serious pressure to take a drug during Christmas family meeting. However, although it requires a lot of effort and some suffering on her side, she keeps refusing it and her abstinence is left intact. But Diana is really devoted to her new, sober lifestyle. In fact, she is so devoted that neither the mechanism leading to her refusal nor she herself as an agent is reasons-responsive when it comes to this particular topic. Anyway, it seems proper to think that Diana is responsible – in this case, praiseworthy – for keeping her abstinence.

In order to be able to take intuitions concerning Diana's responsibility seriously, we have to say in more "down-to-earth" terms what it would really mean for our imagined picture of Diana's act if either the act was not the effect of a reasons-responsive mechanism or if Diana herself were not a reasons-responsive agent.

According to Fischer-Ravizza theory, there are few ways in which the mechanism can fail to be reasons-responsive. I will now consider all these option in the context of the discussed example. First, and again despite of the prominent problem of mechanism-individuation, the mechanism in question has to be at least vaguely specified. What I propose here to consider, very much in the spirit of Fischer's own candidates for relevant mechanisms, is the mechanism of rational deliberation. Some details will be added in the following discussion of the available options.

(1) There are **no** different possible worlds in which the same kind of mechanism operates in S and S has a different sufficient reason to do otherwise and S recognizes this reason.

This option is not really appealing in its most simple formulation; however, we can try to defend it by proposing quite complicated scenarios. In its most probable version, it emerges if we assume either that staying sober is the highest point in Diana's priority list or, even better, that *never to decide to take a drug* is such a goal. In the first case, in order for us to have option (1) as undermining reasons-responsiveness, we need Diana to be insensitive to any reasons, including such reasons as 'if you decide not to take a drug at t1, you will take the same dose of a drug at t2, t3 and again at t4; while if you decide to take a drug at t1, it will be for the last time and you will then remain sober for the rest of your life'. However odd the scenario in which Diana is faced with such a complex reason is, it has to be taken into account, and if Diana is supposed not to recognize this reason and any other reason of this structure as sufficient to take a drug at t1, then her lack of recognition stays in direct contradiction with what is her ultimate goal. And in such a situation it seems very implausible to maintain that the mechanism leading to her decision is any kind of "rational" deliberation. And even if we are eager to weaken the specification of the mechanism to "just deliberation" (whatever that is supposed to mean), the kind of deliberation that she performs here seems to be far too irrational to be able to grant moral responsibility; in fact, it seems to simply undermine it. The possible way out here is to suppose Diana's most important goal is to never violate the following rule: *never decide to take a drug on your own.* Now it is disputable whether we can postulate the scenarios in which Diana is faced with a reason of the form: 'if you don't (on your own) decide to take a drug at t1, you will make (on your own) more such decisions at t1, t2, t3 etc. In what sense are these latter decisions Diana's own if she is faced with such a reason in advance? To defend this way of lack of Diana's reasons-responsiveness in the discussed scenario, we must be ready to answer this question.

(2) There are some different possible worlds in which the same kind of mechanism operates in Diana and she has a different sufficient reason to do otherwise, but there is no understandable pattern of her recognition of those reasons.

This is the option that I do not find appealing at all, and I mention it just to provide the full spectrum of options that can be considered. If the pattern formed by Diana's recognitions is not intelligible at all (especially if we remember that it has to be intelligible on the assumption of the agent's values and reasons and not from the point of view of the observer who does not have access to those), then the agent acts in a way which is too irrational to consider her sane in any plausible sense; in fact she seems to act just randomly. And yet, in the present discussion, I want to grant the agent at least some degree of rationality in her acting: I want the mechanism not to be reasons-responsive, but I don't want the agent to be a total lunatic. Such a picture of her is provided by the two options that follow.

(3) There are some different possible worlds in which the same kind of mechanism operates in S and S has a different sufficient reason to do otherwise and S recognizes this reason, and her recognitions form an intelligible pattern, but there are no moral reasons among those which S would recognize.

This option can be easily construed from already considered options. Assume that Diana recognizes, in a coherent manner, all the reasons of the type discussed in option (1). However, she is completely blind to moral reasons that we would think should make her violate her strong resolution. We have to therefore accept that Diana would not recognize, for example, the reason: '(only) if you don't take the drug, the nuclear bomb will be dropped on some village and all those innocent villagers will be killed.' as sufficient to take a drug. Or to be even more direct: '(only) if you don't take the drug, almost all of people in the Earth will die in a terrible suffering within 5 hours from you making your decision.' It seems that these are the reasons that should be taken into account when one makes some decision concerning her own actions. However, it is not obvious at all that inability to recognize such reasons makes one not morally responsible for acting in particular circumstances when no such reasons are present.

(4) Diana recognizes sufficient reasons, including some moral ones, in different possible worlds in which the same kind of mechanism operates in her, but there is no possible world in which the same mechanism operates and Diana acts for such a reason. In other words, Diana is moderately reasons-receptive, but she is not even weakly reasons-reactive.

Such a possibility was analysed and rejected by Fischer (Fischer and Ravizza, 1998: 89–90) during the discussion of quite a similar type of example. Fischer asks what would happen if the individual had made a decision to never act on any sufficient reason concerning a particular action, and he answers to himself that we just should not keep that decision fixed, i.e., we should not regard it as a part of mechanism that led to action. The question here is whether we *can* – coherently and without begging the question – claim that, in our scenario, it is at least possible to keep the decision fixed and still hold Diana morally responsible. Also, it should be done in such a manner that the only clear principle postulated by Fischer and Ravizza for mechanisms – the principle that the description of the mechanism does not entail a particular action by itself – is not violated.

If we are ready to accept any of those options as possibly true of the agent who we still think is morally responsible for her action, we are provided with a counterexample to Fischer and Ravizza's mechanism-based theory. What about the agent-based account? Recall the conditional supporting it. Filled with the details of the discussed scenario, it says:

> If Diana were to become aware of a reason R (to take the drug) and if Diana retained during the relevant duration of time intrinsic agential properties P_1-P_n, and if Diana were not interfered with in a way that would impede the casual efficacy of those properties, then Diana would take a drug.

Does the falsity of this conditional imply that Diana is not responsible for not taking the drug in the actual scenario? I see no good reason for such inference.

Furthermore, according to McKenna, if Diana is not reasons-responsive with respect to action X, it is either true that she does not recognize any sufficient reasons for not performing action X, in which case we can apply considerations from the preceding paragraphs, or she is not reactive. Following McKenna's approach, not being reactive means that there are no scenarios in which Diana would react otherwise if there were no counterfactual interveners or any other intervention than can be perceived as external to her, preventing her from reacting. But nonexistence of such scenarios is exactly as probable as lack of reactivity in Fischer's sense. It seems that, in this type of example, agent-based and mechanisms-based theories stand or fall together.

8 Conclusion

Two variations on reasons-responsiveness theory: the mechanisms-based account and the agent-based account were compared on the grounds that abstract from the most commonly discussed problem of mechanism-individuation. The problems that emerge for agent-based account when we analyse particular scenarios were pointed at and developed, in favour of mechanisms-based approach.

The possible counterexamples to both accounts were presented and it was clarified what follows from accepting such type of counterexamples. It is left for the reader to decide whether it is worth to pursue it further or if it should be abandoned when vague intuitions supporting it are analysed and specified.

Odzivnost na razloge: Še en pregled vidika teorij osnovanih na akterju in teorij osnovanih na mehanizmu

Fischer in Ravizza trdita, da alternativne možnosti ravnanja niso nujni pogoj za moralno odgovornost. Avtorja prepričajo Frankfurt primeri in sta tako pripravljena zagovarjati stališče zaradi nerazrešenih problemov kompatibilnosti determinizma in možnosti drugačnega ravnanja. Namesto tega predlagata, da za pripisovanje moralne odgovornosti ljudem ti potrebujejo »nadzorno usmerjanje« (angl. *guidance control*), ki je odvisno od mehanizmov, ki vodijo do tega, da je dejanje zmerno odzivno na razloge. Koncept zmerne odzivnosti na razloge (*MRR – moderate reasons-responsiveness*) je razložen na naslednji načni: vrsta mehanizma M je MRR če in samo če (1) v več možnih svetovih, kjer je mehanizem dejaven in ima S zadostni razlog, da ravna drugače, S prepozna takšen razlog kot zadostni; (2) S-jeva dispozicija, da prepozna različne razloge tvori razumljiv vzorec; (3) v vsaj enem izmed teh svetov S ravna v skladu s prepoznanim zadostnim razlogom. Ko sta avtorja predstavila svoj vidik v knjigi *Responsibility and Control. A Theory of Moral Responsibility* (1998), se je pojavilo veliko razprav in kritik. Ena izmed glavnih poant kritikov je bila nejasnost koncepta mehanizma in pomanjkanje kriterijev identitete za različne vrste mehanizmov. To je vodilo nekatere filozofe, med drugimi Carl Ginet in Michael McKenna, da so prevzeli teorijo odzivnosti na razloge osnovano na akterju. Ta sprememba je naletela na druge probleme. Prvo je bilo izpostavljeno, da jo ovržejo primeri vrste Frankfurt, ker akter, s katerim je protidejstveni posrednik povezan, ni odziven na razloge, čeprav je moralno odgovoren. S tem se Ginet in McKenna ne strinjata. Vendar pa kasneje priznata, da obstajajo možni primeri, povezani z akterju internim »posrednikom«, ki predstavljajo grožnjo na akterju osnovani teoriji. Eden izmed nerazvitih primerov, ki predstavlja takšen problem, je situacija, kjer duševno bolan akter ravna, ko njegova bolezen ni aktivna, vendar bi se ta aktivirala in preprečila drugačno ravnanje v alternativnih scenarijih. V tem članku predstavim primere, osnovane na predlogih McKenna in trdim, da predstavljajo veliko večji problem za teorijo osnovano na agentih kot McKenna prizna. Prav tako predstavim možni protiprimer tako teoriji osnovani na agentu kot teoriji osnovani na mehanizmu, ki zadeva okrevano odvisnico, ki zavrača droge, vendar se zdi, da je moralno odgovorna (v tem primeru hvalevredna) za njeno upiranje, čeprav ni odzivna na razloge na pravi način.

Ključne besede: inkompatibilizem, polkompatibilizem, odzivnost na razloge, moralna odgovornost.

Works Cited

Fischer, J. M., Ravizza, M. (1998). *Responsibility and Control. A Theory of Moral Responsibility.* Oxford: Cambridge University Press.

Frankfurt, H. (1969). “Alternate possibilities and moral responsibility”. *The Journal of Philosophy*, 66(23), pp. 829-839.

Ginet, C. (2006). “Working on Fischer and Ravizza’s Account on Moral Responsibility”. *The Journal of Ethics*, 10, pp. 229–253.

McKenna, M. (2013). “Reasons-Responsiveness, Agents, and Mechanisms”. In Shoemaker, D., *Oxford Studies in Agency and Responsibility, Volume* 1. Oxford: Oxford University Press.

van Inwagen, P. (1976). “The Incompatibility of Free Will and Determinism”. *Philosophical Studies*, 27, pp. 185–199.

Martin Justin

University of Ljubljana

Is System 2 uniquely human?

The paper examines whether it can be said that older authors like Aristotle anticipated the claim that System 2 is uniquely human, made by some dual-system theorists. Anticipation is understood as implying a hierarchical relation between a prior, naïve theory, and a later, more elaborated one. Thus, it is examined whether the claim that System 2 is uniquely human made by dual-system theorists can be understood as a mature and empirically well corroborated articulation of the intuition that humans have unique cognitive capacities, or rather a vague repetition of that intuition. The first part of the paper briefly presents Aristotle's theory of the soul and modern dual-system theories respectively. In the second part, dual-system theories are examined in more detail using three criteria. It is concluded that dual-system theories do not present a mature articulation the intuition about human uniqueness. Thus, it is not the case that authors like Aristotle anticipated these dual-system claims.

Keywords: Aristotle, De Anima, dual-system theories, two minds hypothesis, philosophy of science.

1 Introduction

In the last four decades, different fields of psychological research saw a significant surge in theories that took notice of an apparent duality of human cognitive processes. Referred to with an umbrella term dual-process theories, they hold that many cognitive tasks can be solved by two distinct types of processes. Type 1 processes are fast, autonomous, and unconscious, while Type 2 ones are slow, deliberate, and conscious (Evans, 2008; Frankish, 2010; Frankish and Evans, 2009; Evans and Stanovich, 2013; Pennycook, 2018). Later, starting in the 1990s, dual-system theories that attributed "two types of process to two separate reasoning systems, System 1 and System 2" (Frankish, 2010: 914) started to emerge. Some researchers even suggested that there are, in fact, two separate minds in our brain, one evolutionarily ancient and shared with animals, and another evolutionarily recent and uniquely human (Evans, 2009: 2014).

However, as dual-process and dual-system theorists also recognize themselves, theories that try to account for duality or some other partitions of human cognition are not new. In their paper, titled "The duality of mind: An historical perspective," Frankish and Evans (2009) present a broad if hurried review of different historical traditions – spanning from Plato's division of the soul to Freud's theory of the unconscious – and their possible similarities with modern dual-process theories. In passing, they mention that "many philosophers have held that humans exhibit a qualitatively different kind of mentality from other animals" (Frankish and Evans, 2009: 3). They mention "Aristotle, Aquinas, and Descartes" who "anticipated" the claim of some dual-system theories that humans have a unique cognitive architecture, a qualitatively distinct System 2 or a "new mind," that separates us from other animals (Frankish and Evans, 2009: 3). A similar claim can be found in Frankish's (2010) paper: "Several authors have proposed that humans exhibit a qualitatively different kind of mentality from other animals, anticipating the modern claim that there is a uniquely human reasoning system" (Frankish, 2010: 915).

In this paper, I will examine the claim that some older authors *anticipated* modern dual-system theories in more detail. If we take the relation of anticipation to imply a hierarchy between a naïve theory and a more developed theory, then it must be show that the claim that System 2 is uniquely human made by dual-system theorists is a mature and scientifically well corroborated articulation of a much older intuition that human and animal cognitions are qualitatively distinct. Three criteria will be used to evaluate this. First, dual-system theories must show that there is strong empirical evidence that humans and animals indeed have different cognitive abilities. Second, they must show that these differences correspond to System 1 and System 2. Third, while accounting for these differences, dual-system theories must stay coherent. I will show that dual-system theories do not meet these criteria. Therefore, I will conclude that modern dual-system theories, rather than providing a mature articulation, just repeat this old intuition about differences between humans and other animals. Given that, claims that authors like Aristotle "anticipated" (Frankish and Evans, 2009; Frankish, 2010) modern dual-process theory will be shown to be unwarranted.

2 Human and animal cognition in Aristotle and dual-system theories

In this chapter, I will briefly present both Aristotle's and dual-system theorists' respective ideas about the differences between human and animal cognition. My goal here is not to present the theories in detail, but to point out some possible parallels between them. I will first sketch Aristotle's theory of the soul and then turn to the contemporary theories.

Aristotle's theory of the soul

Here, I will present Aristotle's theory of the soul in the most general outlines. It is based on his doctrine of hylomorphism which states that all things are formed from two substances, matter (potentiality) and form (actuality). Like all other things than, an organism is unity of an appropriate matter (body) and a form (soul). Or, as Aristotle writes: "it is necessary, then, that the soul is a substance as the form of a natural body which has life in potentiality" (2016: 22). In other words, soul is a form which, together with a body that can be animated, generates a living organism that grows, is nourished, and eventually decays. Being alive distinguishes bodies that are enformed by a soul, i.e., ensouled, from those that are enformed by other forms; that is, the object that have a soul are living (Aristotle, 2016: 24).

Further, Aristotle claims that "living is spoken of in several ways" (Aristotle 2016: 24); besides nourishment and growth, living organisms can also have other faculties: reason, perception, locomotion, imagination, desire. He then makes a connection between these different capacities of organisms and different faculties of the soul, saying that "among the capacities of the soul, all belong to some, to others some of them belong, and to still others only one belongs" (2016: 27). If all living beings have a capacity for growth and nutrition, only some organisms have others. Aristotle holds that the capability to perceive forms (i.e., perception) separates animals from plants, while reason belong only "to humans and to anything else there may be of this or of a more elevated sort" (2016: 27). Given this, we can conclude that the faculties of the soul that Aristotle established function also as a principle of differentiation between animal and non-animal species.

Dual-system theories

Now, I will turn to the presentation of dual-system theories. At the face value, they differ strikingly from Aristotle in everything from methodology and metaphysical commitments to the kinds of empirical evidence they use. Nevertheless, like Aristotle, several dual-system accounts, especially the ones promoted by Jonathan St. B. T. Evans, hold that some of our cognitive abilities are shared with animals, while others are uniquely human (Evans, 2003; 2009; 2014). Although some dual-system theories are completely void of such claims (e.g., Sloman (1996)), the claims were already present in some early dual-system accounts (e.g., Reber's (1996) influential book about implicit learning) and gained traction in 2003 with the publication of Evans' review paper in the journal *Trends in Cognitive Sciences*.

In his 2003 paper, Evans presents a broad account of the two systems, which includes hypotheses about their architecture and evolutionary age. He characterizes System 1 as "universal cognition shared between humans and animals," a "set of sub-systems that operate with some autonomy" that is evolutionarily older, responsible for associative learning, instinctive behavior, and other processes that are "rapid, parallel and automatic" (Evans 2003: 454). System 2, in contrast, is characterized as "slow and sequential in nature" since it makes use of the "central working system" and is responsible for deductive reasoning, hypothetical thinking, explicit learning etc. (Evans, 2003: 454). It is evolutionarily more recent, uniquely human and it can inhibit System 1 responses (Evans, 2003: 454).

Later, several problems were found with this picture (cf. Evans 2009) which prompted a move away from using this theoretically richer concept of a system and towards talking about different kinds of processes. Nevertheless, researchers did not completely abandon the concept of cognitive systems. Namely, Evans (2008; 2014) started using a new distinction between the "old mind" and the "new mind" or what he calls the "two minds" hypothesis. He defines the mind as "as a high-level cognitive system capable of representing the external world and acting upon it in order to serve the goals of the organism" (Evans, 2009: 35). He proposes that the human brain contains two such minds. One that is capable of "associative and procedural learning", is evolutionarily old and shared with animals (Evans, 2014: 131). And the other that is responsible for "controlled attention" and is evolutionarily more recent (Evans, 2014: 131).

Although Evens now argues that "it was an error in earlier forms of dual system theories to describe System 2 as unique to human beings" (Evans, 2014: 131), he still seems to hold that it is the new mind that separates humans from other nonhuman animals. I will present this argument in more detail below; for now, it is only important to notice that, despite some additional qualifications, the claim that System 2 or the new mind is uniquely human still plays an important role in dual-system theories.

3 Evaluating the claim about anticipation

As was show above, both Aristotle and dual-system theorists argue that we humans have some special cognitive architecture that separate us from other animals. But does this mean that Aristotle and other authors anticipated the dual-system theorists' claim that humans have a unique cognitive architecture? To answer this question, we must first understand what kind of relation between theories is implied by *anticipation*.

In essence, anticipation implies a hierarchical relation between a prior, naïve theory, and a later, more developed theory. First, a theory expresses a plausible intuition, but it does not articulate it fully or in a right way; then, another theory comes and manages to articulate it in a much more satisfying way. Or, to put it concretely, Aristotle's theory of the soul only hinted at the idea that humans have a unique cognitive system; later this idea was fully articulated and corroborated by dual-system theorists.[1]

For the claim about anticipation to hold than, it must be shown that dual-system theories can be understood [1] as a mature and empirically well corroborated articulation of the intuition about human uniqueness, rather than [2] a vague repetition of that intuition. The burden of proof seems to lie on [1], since a theory, to be understood as maturely articulated, needs sufficient evidence for all claims that follow from it. Thus, I will place the claims made by dual-system theorists under more scrutiny and try to establish that [1] rather than [2] holds by directly examining them. Specifically, I ask two things: (1) are dual-system theorists' claims about uniqueness of human cognitive capabilities grounded in evidence? And (2) are they a part of a coherent and well though-out theory that can predict and accommodate novel evidence?

To establish this, I will use three criteria. First, dual-system theories must show that there is strong empirical evidence that humans and animals indeed have different cognitive abilities. Second, they must show that these differences correspond to System 1 and System 2. Third, while accounting for these differences, dual-system theories must stay coherent. Furthermore, they should be able to account for new empirical evidence about human vis-à-vis animal cognition, without significant

[1] One way of arguing that Aristotle could not anticipate dual-system theories is to say that scientific theories from different paradigms (e.g., cognitive science versus Aristotelian biology) cannot be compared, since they are simply different. This is the famous "incommensurability thesis" (see Oberheim and Hoyningen-Huene, 2018). I will not be making this argument. Rather, I will argue that dual-system theories are lacking as an articulation of the idea that humans have a unique cognitive architecture and thus are not necessarily better than Aristotle's theory of the soul.

changes made to them.[2] Together, these criteria are sufficient to prove that dual-system theories are a mature articulation, while each of them is necessary.

First and Second Criteria

The first criterion is trivially satisfied. The computer on which I am writing this paper, and most of the other artifacts I interact with day-to-day can serve as evidence that humans have developed in a radically different way than other animal species and that we probably indeed have unique cognitive abilities. But the problem is in establishing how exactly are we different from other animals. Or, in other words, is the first criterion satisfied in a way that the second criterion is also satisfied?

Satisfying the second criterion proves to be a harder challenge for dual-system theories. One problem is that dual-system theorists themselves present scarce evidence to support the claim that System 1 is shared with animals while System 2 is uniquely human.[3] Evans presents some archeological evidence that *homo sapiens sapiens* developed qualitative differently than other animal species (Evans, 2003: 457). For example, he quotes “the qualitative change in the archaeological record c. 50,000 years ago when there was sudden evidence of representational art, religious imagery and rapid adaptations in the design of tools and artefacts” (Evans, 2003: 457). But this evidence is not that convincing; it might suggest that something indeed had happened to the human brain, but it does not support the claim that a separate reasoning system developed. In other words, it only satisfies the first criterion, but not the second.

In addition, Reber, who is often cited as making one of the earliest claims about the evolutionary age of the two systems of reasoning (Frankish and Evans, 2009; Frankish, 2010), introduces his claim that “consciousness is a late arrival on the evolutionary scene” (Reber, 1996: 86) as an axiom. This may not be problematic for his argumentation, but it does not help satisfy the second criterion. Furthermore, there is evidence against the claim that the distinction between System 1 and System 2 can function as a principle of differentiation between humans and animals. It was shown that something resembling the System 1/System 2 dichotomy could also be found in rats, which can, in new circumstances, inhibit established behavior patterns (Toates 2004). Experiments with rats also showed that they are capable of causal reasoning that cannot be explained by associative learning

[2] This third criterion is admittedly a very strong one. I do not hold that any psychological theory (or any theory in the special sciences) should generally satisfy it. However, it is necessary, if we want to claim that a theory is a conceptually mature articulation of an idea that another theory anticipated.

[3] They present both psychological and neurological evidence that indeed points to the existence of two different reasoning systems in humans (cf. Evans, 2003: 455-56). What they lack is evidence which would show that one system is shared with nonhuman animals while the other is uniquely human.

(Blaisdell et al., 2006). If it is System 1 that is responsible for associative learning and System 2 that can inhibit autonomous System 1 responses (Evans, 2003) then at least a basis for both System 1 and System 2 also exist in rats.[4]

Does the two minds hypothesis satisfy the second criterion?

Citing similar evidence, Evans (2009; 2014) himself recognized that the claim that System 2 is uniquely human is too strong. As already mentioned above, rather than claiming that System 2 is only found in humans, he now promotes the idea that it is "uniquely developed in humans" (Evans, 2014: 131) or that what is a "relatively small cognitive facility in animals became magnified greatly in humans"(Evans, 2009: 39).[5]

Nevertheless, he still seems to hold that his two minds hypothesis can conceptualize a qualitative difference between humans and other animal species. For example, he states that our seemingly less instrumentally rational behavior[6] is often a consequence of "the fact that we have a new mind. An animal lacking such higher-order concepts and representations will naturally follow the path to the immediate goals" (Evans, 2014: 140). In the face of the extraordinary human development in the last few millennia, he writes: "Not only is the difference in human achievement from any other species staggering in scale but the thought processes which permit these developments seem qualitatively different" (Evans, 2014: 139). In the conclusion of the paper, he suggests that the old mind is responsible for "a form of instrumental rationality which we share with other animals" (Evans, 2014: 143), implying that the "new mind" is uniquely human.

We can thus ask whether the old/new mind distinction satisfies the second criterion of conceptualizing the differences between human and animal cognitive capacities. To achieve this, Evan should be able to make the old mind/new mind distinction independently from the Type 1/Type 2 distinction which is commonly used to sperate System 1 and System 2.[7] He indeed seems

[4] Arguably, this is also not a strong evidence *against* dual-system theorists claim about human and animal cognition. Nevertheless, they themselves (cf. Evans' 2009 and 2014 papers) accepted it as decisive, which suggest that the strong claim about System 2 being uniquely human was not well corroborated in the first place.

[5] This distinction seems to be based on the fact that although nonhuman animals have capabilities which suggest the presence of System 2, these capacities are quite limited in comparison to human ones. But admittedly, this is a very thin distinction. This becomes especially apparent when Evans restates the uniquely human thesis through the distinction between the new and the old mind, as I will show below. I thank the anonymous referee for pointing out that this distinction might not be obvious for the reader.

[6] He gives an example of a "sunk cost" effect. "If I have expensive tickets bought months ago for an opera tonight but (a) am feeling tired and (b) discover that my favorite football team is playing live on television, then at this point I might prefer to stay home rather than go the opera. But I go to the opera anyway as I have bought the tickets" (Evans, 2014: 139). This type of behavior was traditionally considered as irrational since the money spent for the opera ticked is "lost" in both cases. But, as he points out "phenomena like sunk costs are a much more complex issue because humans have higher-order goals, values, and self-perceptions all of which can affect this kind of decision making" (Evans, 2014: 140).

[7] One possibility of this kind of characterization that I ignored throughout this argument is that some parts of the brain evolved more recently (i.e., the new brain) while others are evolutionarily older (i.e., the old brain). While this may be

to argue that these are two different distinctions: he states that it is not simply true that the "Type 2 processing [is] in the new mind or System 2, and Type 1 processing [is] in the old mind or System 2" (Evans, 2014: 132). The example he gives is modular cognition (e.g., the visual system). It involves processes that "are fast, automatic, mandatory, and parallel, hence fitting the criteria for Type 1 processing" (Evans, 2014: 132). Nevertheless, some "modules play a critical foundation for the new mind" (Evans, 2014: 132); therefore, modules as such cannot simply be categorized under the old mind.

But instead of providing an independent characterization of the old/mind distinction, Evans goes on to state that "the key difference is that the old mind operates through automated Type 1 systems that have mandatory outputs, whereas Type 2 systems are in some sense volitional: the new mind is capable of forming plans and carrying out intentions under controlled attention" (Evans, 2014: 133). Reading this statement as coherent with former emphasis that "most of the processing in the new mind is also Type 1 in the sense of being autonomous, rapid, and preconscious" (Evans, 2014: 132), it seems to mean that the old mind operates *only* through automated Type 1 systems, while the new mind can *also* operate through volitional Type 2 systems.

This characterization of the relation between the old/new mind distinction and Type 1/Type 2 distinction might not be strictly problematic, but it also cannot support the claim that the new mind is only found in humans. If nonhuman animals really are capable of some rudimentary Type 2 processes, as we have seen above (cf. Blaisdell et al., 2006), and these processes are characteristic of the new mind, then nonhuman animals too must possess the new mind, albeit in some rudimentary form.

Therefore, the two minds hypothesis does not provide sufficient ground to suppose a qualitative difference between human and animal cognition. According to Evans' definition of the new and the old mind (Evans, 2014: 133) the distinction between cognitive capacities of humans and other animals is either quantitative (i.e., both have the new mind, but it is developed to different degrees) or it cannot be captured fully by the distinctions he makes. Thus, his theory does not satisfy the second criterion. Rather, it seems to accept the claim that humans have qualitatively different cognitive abilities without providing a substantial evidential ground for it.

true and indeed a big part of telling a story of human uniqueness (or similarity with other animals), it moves us to a different level of analysis altogether: we are seeking distinctively human cognitive capabilities, not physiological features. Furthermore, even if some uniquely human capabilities, e.g., language or hypothetical thinking, can be mapped onto those evolutionarily more recent parts of the brain, an independent distinction between the new and the old brain, which would be consistent with this evidence, must be made. Otherwise, the old and the new brain function only as a shorthand for different parts of the brain and their corresponding capabilities.

Third criterion

Having shown that dual-system theories fail to satisfy the second criterion, namely that System 1/System 2 and old/new mind dichotomies cannot account for differences in cognitive capabilities between humans and other animals, I can now move to the third criterion. This one states that besides providing empirical evidence (which they do not), dual-system theories must also be coherent. This is not a necessary step in my argumentation. Dual-system theories should fulfill all three criteria; showing that they do not meet the second criteria is already enough to conclude that the theories cannot be understood as a mature articulation of the intuition about human cognitive uniqueness.

Nevertheless, I want to point out the general disarray on the field of dual-process studies: even the most fundamental conceptions and distinctions made by the researchers in the field are a subject of continuous debate. Different dual-system and dual-process theories are attacked both by researchers who subscribe to different views of human cognition (Keren and Schul, 2009; Kruglanski and Gigerenzer, 2011; Osman, 2004) and by researchers who themselves commit to dual-system or dual-process view but think specific theories are lacking in some way (Evans, 2009; Evans and Stanovich, 2013).

Take for example Evans' (2009) discussion of symmetry between Type 1 and 2 processes and Systems 1 and 2 (40), favored by some earlier theories, for example Sloman (1996). Sloman (1996) argued that humans can provide responses to problems using two different types of processes, associative and rule-based (9-11). For example, we can categorize different animals based on similarities among them (e.g., we see different animals that all share a similar feature, wings, and then group them together, as birds), or we can apply rules to categorize them (e.g., “if it has wings, then it is a bird”). From this, he concluded that two separate systems exist, one responsible for associative and other for rule-based responses (Sloman, 1996: 11), thus establishing a symmetry between Type 1 and 2 processes and Systems 1 and 2.

But Evans pointed out that such clear distinction between System 1, responsible for one type of processes, and System 2, responsible for another type of processes, is hardly consistent when considered in detail. For him, the problem mainly lies in using working memory access as a defining distinction between System 1 and System 2 (Evans, 2009: 37–39). The idea is that Type 1 processes, e.g., heuristics for solving reasoning tasks, memory retrieval, or word and face recognition, do not have access to working memory and are therefore fast and automatic. In contrast, Type 2 processes, e.g., deductive reasoning, explicit learning, or hypothetical thinking, need working memory and are therefore slow, require high effort, and have limited capacity.

The problem is that working memory cannot be used for distinguishing Systems 1 and 2 in the same way. First, working memory needs content, which is mostly provided by the fast, Type 1 processes. If we say that System 1 is responsible for them, while System 2 has access to working memory, a significant overlap between the two systems is established. This can be solved by saying that System 2 needs access to working memory "among other resources" (Evans, 2009: 38). However, that leaves us with two, equally unsatisfying options. We can either reduce System 2 to working memory alone or say that there is not only one System 2, but maybe a set of Type 2 systems defined as Type 1 units plus working memory (Evans, 2009: 38). Given this, he argues for introducing a new type of processes (Evans, 2009: 42) and discourages the use of the terms System 1 and System 2 (Evans, 2009: 50).

Further, this production of new theoretical concepts has not slowed in recent years. Beside Evans' (2009) introduction of the new type of processes, other authors from the field are introducing concepts like an additional system (Stanovich, 2009), a "conflict monitor" (Pennycook, 2018)) or presenting new theories altogether ("fuzzy trace theory" (Reyna et al., 2017)). This indicates that dual-system and dual-process theories, in their present state of development, are not a coherent theoretical framework that could easily incorporate new evidence. This means that we cannot safely say that they satisfy the third criterion.

4 Conclusion

Given that dual-system theories fail to satisfy both the second and the third criterion, it can be concluded that they cannot be taken as a conceptually mature articulation of the difference between human and animal cognitions. Rather, the claim that System 2 (or the new brain) is uniquely human, while System 1 (or the old brain) is shared with animals, should be understood as a preconceived notion that is neither sufficiently supported by evidence nor theoretically coherent. Therefore, it is not the case that authors like Aristotle "anticipated" (Frankish, 2010; Frankish in Evans, 2009) this claim made by some dual-system theorists. Instead, dual-system theorists try to provide a new conceptualization of a much older intuition but are not much more successful than authors like Aristotle. In other words, the claim made by dual-system theorists' that older authors anticipated their arguments about a uniquely human reasoning system is therefore of not much relevance since these modern claims are still not coherent and well corroborated.

Acknowledgments

The author would like to thank Borut Trpin for his invaluable comments, suggestions, and encouragement.

Je Sistem 2 izključno človeški?

Članek raziskuje trditev nekaterih teoretikov dvojih sistemov [*dual-system theories*], da so klasični avtorji, na primer Aristotel, anticipirali njihovo tezo o obstoju dodatnega, unikatno človeškega kognitivnega sistema. Anticipiranje je razumljeno kot hierarhično razmerje med zgodno, naivno teorijo in novejšo teorijo, ki je bolje razdelana. Članek se tako vpraša, ali lahko trditev teoretikov dvojih sistemov, da je Sistem 2 izključno človeški, razumemo kot empirično podprto in teoretično koherentno artikulacijo starejše intuicije o človeški kognitivni unikatnosti ali pa gre zgolj za nič bolje podprto ponovitev te intuicije. V prvem delu članka so tako najprej na kratko predstavljene Aristotelova teorija duše in sodobne teorije dvojnih sistemov. Drugi del članka pa s pomočjo treh kriterijev natančneje razišče, ali teorije dvojnih sistemov lahko razumemo kot podprto artikulacijo intuicije o človeški unikatnosti. Članek se zaključi z ugotovitvijo, da to ne velja, torej posledično tudi ne velja, da so klasični avtorji (kot Aristotel) predvideli trditve sodobni teoretikov dvojnih sistemov.

Ključne besede: Aristotel, O duši, teorije dvojnih sistemov, hipoteza dvojnega uma, filozofija znanosti.

Works Cited

Aristotle. (2016). *Aristotle: De Anima*. Edited by Christopher Shields. Oxford: Clarendon Press.

Blaisdell, A. P., Kosuke Sawa, Kenneth J. Leising, and Michael R. Waldmann. (2006). "Causal Reasoning in Rats." *Science*, 311 (5763), pp. 1020–22. https://doi.org/10.1126/science.1121872.

Evans, Jonathan St. B. T. (2003). "In Two Minds: Dual-Process Accounts of Reasoning." *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2003.08.012.

Evans, Jonathan St. B. T. (2008). "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology*, 59(1), pp. 255–78. https://doi.org/10.1146/annurev.psych.59.103006.093629.

Evans, Jonathan St. B. T. (2009). "How Many Dual-Process Theories Do We Need? One, Two, or Many?" In *In Two Minds: Dual Processes and Beyond*, pp. 33–54. https://doi.org/10.1093/acprof:oso/9780199230167.003.0002.

Evans, Jonathan St. B. T. (2014). "Two Minds Rationality." *Thinking &* Reasoning, 20(2), pp. 129–46. https://doi.org/10.1080/13546783.2013.845605.

Evans, Jonathan St. B. T., and Keith E. Stanovich. (2013). "Dual-Process Theories of Higher Cognition." *Perspectives on Psychological Science*, 8(3), pp. 223–41. https://doi.org/10.1177/1745691612460685.

Frankish, Keith. (2010). "Dual-Process and Dual-System Theories of Reasoning." *Philosophy Compass*, 5(10), pp. 914–26. https://doi.org/10.1111/j.1747-9991.2010.00330.x.

Frankish, Keith, and Jonathan St. B. T. Evans. (2009). "The Duality of Mind: An Historical Perspective." In *In Two Minds: Dual Processes and Beyond*, pp. 1–30. https://doi.org/10.1093/acprof:oso/9780199230167.003.0001.

Keren, Gideon, and Yaacov Schul. (2009). "Two Is Not Always Better Than One." *Perspectives on Psychological Science*, 4(6), pp. 533–50. https://doi.org/10.1111/j.1745-6924.2009.01164.x.

Kruglanski, Arie W., and Gerd Gigerenzer. (2011). "Intuitive and Deliberate Judgments Are Based on Common Principles." *Psychological Review*, 118(1), pp. 97–109. https://doi.org/10.1037/a0020762.

Neys, Wim De, ed. (2017). *Dual Process Theory 2.0*. Routledge. https://doi.org/10.4324/9781315204550.

Oberheim, Eric, and Paul Hoyningen-Huene. (2018). “The Incommensurability of Scientific Theories.” In Zalta, E. N., *The Stanford Encyclopedia of Philosophy*. URL = https://plato.stanford.edu/entries/incommensurability/#Aca.

Osman, Magda. (2004). “An Evaluation of Dual-Process Theories of Reasoning.” *Psychonomic Bulletin & Review*, 11(6), pp. 988–1010. https://doi.org/10.3758/BF03196730.

Pennycook, Gordon. (2018). “A Perspective on the Theoretical Foundation of Dual Process Models.” In *Dual Process Theory 2.0*, 5–27. https://doi.org/10.4324/9781315204550-2.

Reber, Arthur S. (1996). *Implicit Learning and Tacit Knowledge*. *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195106589.001.0001.

Reyna, Valerie F., Shahin Rahimi-Golkhandan, David M. N. Garavito, and Rebecca K. Helm. (2017). “The Fuzzy-Trace Dual Process Model.” In *Dual Process Theory 2.0*, pp. 82–99. Routledge. https://doi.org/10.4324/9781315204550-6.

Sloman, Steven A. (1996). “The Empirical Case for Two Systems of Reasoning.” *Psychological Bulletin*, 119(1), pp. 3–22. https://doi.org/10.1037/0033-2909.119.1.3.

Stanovich, Keith E. (2009). “Distinguishing the Reflective, Algorithmic, and Autonomous Minds: Is It Time for a Tri-Process Theory?”. In *In Two Minds: Dual Processes and Beyond*, pp. 55–88. https://doi.org/10.1093/acprof:oso/9780199230167.003.0003.

Toates, Frederick. (2004). “‘In Two Minds’ – Consideration of Evolutionary Precursors Permits a More Integrative Theory”. *Trends in Cognitive Sciences*, 8(2), 57. https://doi.org/10.1016/j.tics.2003.12.005.

ANαliZA, časopis za kritično misel, številka 1, letnik XXV., 2021

Izdajatelj in založnik: Društvo za analitično filozofijo in filozofijo znanosti,
Aškerčeva 2, 1000 Ljubljana
http://www.drustvo-daf.si/analiza/
e-naslov: analiza@drustvo-daf.si

Naročila sprejemamo na zgornji naslov.
Na leto izideta dva zvezka Analize, ki sta prosto dostopna na spletni strani Društva za analitično filozofijo in filozofijo znanosti.

Glavni in odgovorni urednik: Božidar Kante

Tehnična urednika številke: Smiljana Gartner in Tadej Todorović

Oblikovanje: Tadej Todorović

Naklada: elektronska verzija

Izid te revije je sofinancirala
JAVNA AGENCIJA ZA RAZISKOVALNO DEJAVNOST REPUBLIKE SLOVENIJE

Revija je indeksirana v THE PHILOSOPHER’S INDEX.

ISSN - 2712-4916

Navodila avtorjem

Prispevke oddajte po elektronski pošti (analiza@drustvo-daf.si). Priložiti je treba povzetek (v slovenščini in v angleščini), ki povzema glavne poudarke dela. Povzetku v angleškem jeziku je treba dati tudi angleški naslov. Povzetek ne sme presegati 150 besed. Na koncu povzetka navedite do 5 ključnih besed (deskriptorjev). Na primer, Ključne besede: filozofija jezika, metafora, analogija; Keywords: philosophy of language, metaphor, analogy.

Prispevki naj ne presegajo obsega ene in pol avtorske pole (45.000 znakov s presledki). Uporabite urejevalnik besedil Word, standardno obliko pisave brez dodatnih slogovnih določil. Napisani naj bodo z dvojnim razmikom med vrsticami; za literaturo, opombe in povzetek pa uporabite enojni razmik. Prispevki naj bodo notranje razčlenjeni, torej razdeljeni na razdelke, in opremljeni – če je mogoče – z mednaslovi. V besedilu dosledno uporabljajte dvojne narekovaje (na primer: pri navajanju člankov, citiranju besed ali stavkov in kadar navajate strokovne in posebne izraze). Izjema so citati znotraj citatov. Pri teh uporabite enojne narekovaje. Naslove knjig, periodike in tuje besede je treba pisati ležeče. Primeri: Kritika praktičnega uma, Anthropos, a priori.

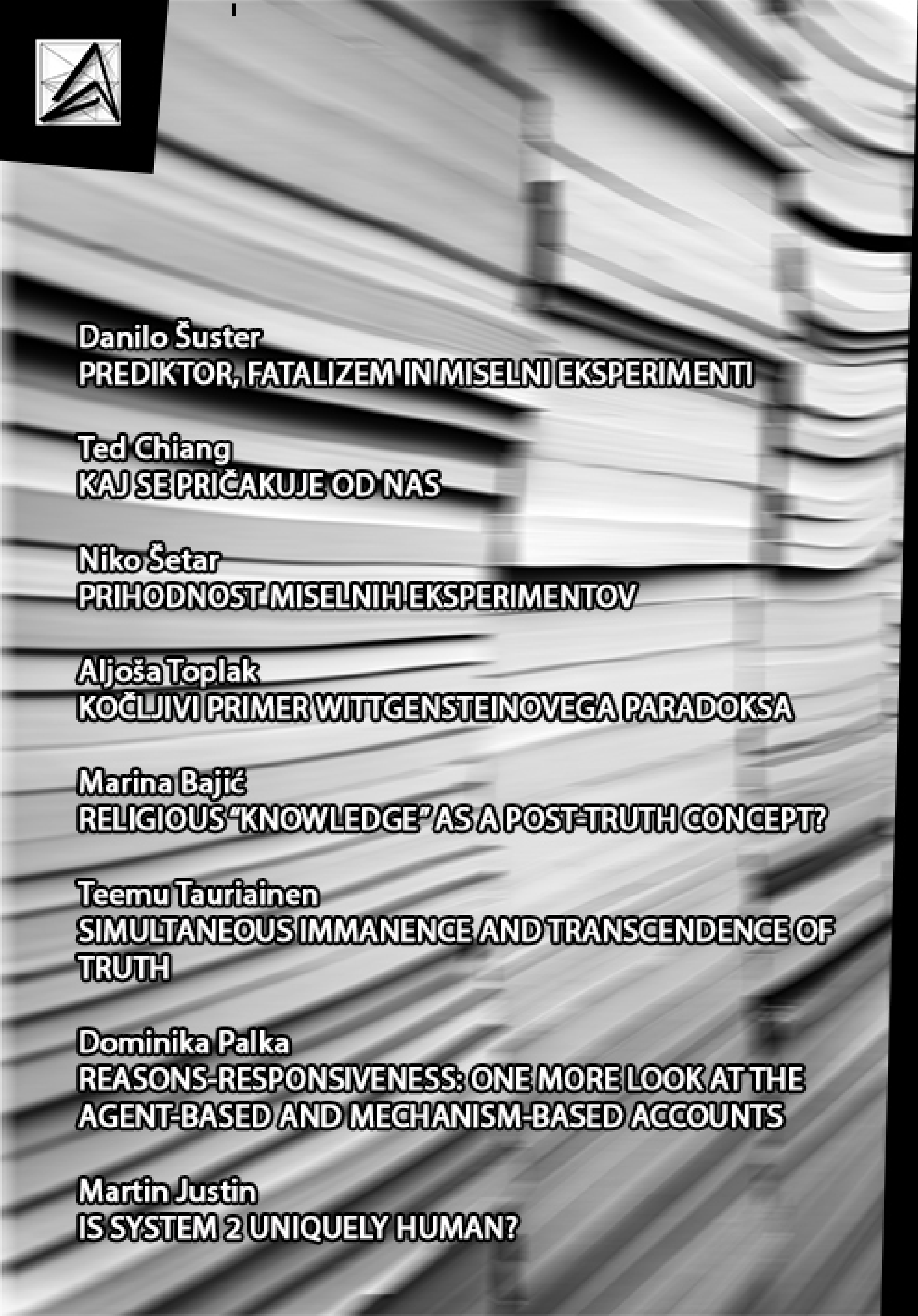
Za citiranje in navajanje virov se uporablja 7. verzija APA standardov.

Pod črto navajajte samo opombe, ki naj ne bodo predolge. V besedilu je treba opombe označiti z dvignjenimi indeksi1. Ne uporabljajte opomb za navedbo reference.

Zaradi anonimnega recenzijskega postopka (“double-blind peer-review”) zapišite svoje ime in kontaktne podatke v ločenem dokumentu in poskrbite, da iz samega prispevka ni mogoče razbrati vaše identitete.

Sprejemamo tudi ocene knjig (do 12.000 znakov s presledki).

Uredništvo ne sprejema prispevkov, ki so bili že objavljeni ali istočasno poslani v objavo drugam. Z objavo se morajo strinjati vsi avtorji. Vse moralne avtorske pravice pripadajo avtorju, vse materialne avtorske pravice za članek pa avtor brezplačno prenese na izdajatelja. Avtor s tem dovoljuje tudi objavo na spletu.



1/21