

Društvo za analitično filozofijo in filozofijo znanosti

časopis za kritično misel

ANALIZA

02

2016 letnik 20.

AN_αliZA, časopis za kritično misel, številka 1, leto XX., 2016

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 izidejo trije zvezki Analize (dve enojni in ena dvojna številka).
Cena zvezka v prosti prodaji je 6,00 EUR, za naročnike 4,00 EUR oz. letna
naročnina 12,00 EUR (z vključenim DDV).

Izdajateljski svet:

Bojan Borstner, David Owens, Matjaž Potrč, Marko Uršič, Matthias Varga von
Kibéd, Elizabeth Valentine, Bojan Žalec

Glavna in odgovorna urednica: Maja Malec

Uredniški odbor:

Smiljana Gartner, Božidar Kante, Vojko Strahovnik, Toma Strle, Borut Trpin,
Sebastjan Vörös

Lektor: Borut Cerkovnik

Oblikovanje: K.reg@

Tisk: Demat, d.o.o., Stegne 3, 1000 Ljubljana

Naklada: 200 izvodov

Izid te revije je sofinancirala

JAVNA AGENCIJA ZA KNJIGO REPUBLIKE SLOVENIJE.

Revija Analiza je indeksirana v THE PHILOSOPHER'S INDEX.

ISSN 1408-2969

Kazalo

Moralna teorija

Vojko Strahovnik Razsežnosti moralne teorije	5
---	---

Nezdružljivost svobode in determinizma

Tim Koprivnik Zakaj nismo svobodni (v smislu moralne odgovornosti) in kaj iz tega (ne) sledi.....	29
---	----

Filozofija umetnosti

Lea Lampret Filozofija umetniškega sloga: kaj nam ima o umetniškem slogu povedati Kendall Walton?	51
---	----

Filozofija duha

Tadej Todorović Večvrstna realizacija in Turingov test: preslikava	63
Alen Lipuš Pojasnjevanje pojasnjevalne vrzeli.....	75

Filozofski razmislek o superinteligenci

Blaž Cotman Izziv »superintelligence« – komentar k Bostromovi knjigi <i>Superintelligence: Paths, Dangers, Strategies</i> (2014)	91
--	----

Filozofija ekonomije

Brin Blatnik Problem racionalnega agensa	107
---	-----

Recenzija

Borut Trpin L. A. Paul: Transformative Experience.....	119
---	-----

Vojko Strahovnik

Univerza v Ljubljani

Razsežnosti moralne teorije¹

V prispevku načrtamo temeljne razsežnosti moralne teorije. Na začetku predstavimo razliko med moralnimi vprašanji prvega reda in vprašanji drugega reda ter slednje povežemo z moralno teorijo in metaetiko. V nadaljevanju razlikujemo med petimi osrednjimi razsežnostmi moralne teorije, tj. moralno ontologijo, moralno semantiko, moralno epistemologijo, moralno psihologijo in moralno fenomenologijo, ter predstavimo osnovna stališča v okviru teh razsežnosti. Članek sklenemo z razmisleki o najboljši vstopni točki v metaetično razpravo ter nekaterih novih razvojnih težnjah na tem področju moralne filozofije.

Ključne besede: moralna teorija, metaetika, moralni (i)realizem, kognitivizem, nekognitivizem, moralna motivacija, moralna fenomenologija

Uvod

Tako etika kot metaetika sta področji filozofije, ki se ukvarjata z moralnostjo, pri čemer se metaetika prvenstveno ukvarja z naravo moralnosti, etika pa z njeno vsebino. Če se v okviru slednje na primer sprašujemo, katere stvari so vrednote, katera dejanja so naše dolžnosti ali kakšna motivacija naj nas vodi pri naših dejanjih, potem si metaetika zastavlja vprašanje o samih izhodiščih teh prvotno zastavljenih vprašanj. Za predmet si izbere značilnosti naše moralne misli in jezika in sprašuje, npr. kaj sploh pomeni, da je nekaj vredno, ali pa ali sploh obstajajo od nas neodvisni, objektivni odgovori na vprašanje glede tega, kaj moramo storiti, kako lahko dosežemo moralno vednost oz. upravičimo naše moralne sodbe, kako so naši moralni pogledi in naravnosti povezane z delovanjem ipd. V tem smislu lahko rečemo, da se torej etika ukvarja z etičnimi vprašanji prvega reda, z vsebinskimi vprašanji, ki prevladujejo tudi v naši vsakodnevni moralni misli in praksi, medtem ko metaetika zastavlja vprašanja drugega reda, tj. vprašanja o teh vprašanjih in njihovih predpostavkah. Podobno razlikovanje lahko ubesedimo tudi preko razlike med formalnimi in vsebinskimi vprašanji. Meja med obema vrstama vpra-

¹ Članek je prirejena oblika prvega poglavja knjige *Moralna teorija*, ki bo do konca leta 2016 izšla pri založbi Aristej. Knjiga vključuje dvanajst poglavij, ki sistematično obravnavajo posamezna stališča v moralni teoriji, tudi z vidika, kakor so si ta sledila v njenem razvoju.

šanj ni povsem jasna oz. določna; v vmesnem, sivem področju najdemo vprašanja povezana z moralno odgovornostjo, svobodno voljo in motivacijo za moralno delovanje. Pa tudi sicer med obema področjema ne moremo potegniti stroge meje, kajti tudi naša vsakdanja moralna misel in jezik pogosto presežeta zgolj ukvarjanje z vsebinskimi vprašanji ter predpostavljata, pogosto zgolj implicitno, metaetična stališča.

Tudi skozi večino zgodovine same moralne filozofije ti dve vrsti vprašanj nista bili strogo ločeni kot posebni področji moralne filozofije ter prepoznani kot takšni, saj v teorijah posameznih avtorjev nahajamo tako odgovore na vsebinska vprašanja etike o tem, kako moramo živeti, kaj vrednotiti in po čem stremeti, kot tudi razlago različnih vidikov narave moralnosti, na primer o izvorih moralne vednosti, povezanosti moralnih sodb z dejanji ali načinu obstoja moralnih lastnosti. Pomemben prelom nastopi z izidom knjige G. E. Moora *Principia Ethica* (1903), v kateri Moore jasno oddeli obe vrsti vprašanj. Kot izhodišče postavi, da si moramo najprej sploh odgovoriti na osnovna metaetična vprašanja, torej vprašanja drugega reda, da bi se sploh lahko pričeli ukvarjati tudi z vsebinskimi vprašanji in odgovori nanje, posebej glede tega, katere stvari so neodnosno oz. notranje vredne in katera dejanja so moralno pravilna. Metaetične teme imajo zanj tako prednost – ne v smislu pomembnosti, temveč metodološko – pred vsebinskimi temami. V tem smislu Moore razume metaetiko kot predprostor etike; kot orodje, s katerim postavimo temelje razpravi v moralni filozofiji (Darwall, 2006: 20–23). S tem pa je nastala vrzel med obema področjema, ki se je po Mooru še povečevala, hkrati pa se je slabila brv med obema področjema, in sicer v smeri, da je metaetika pričela prevladovati nad celotnim poljem moralne filozofije, vsebinska vprašanja pa so bila iz nje potisnjena. Mnogi filozofi so v Moorovem filozofskem programu pojmovne analize temeljnih etičnih pojmov ter stavkov, v katerih ti nastopajo, videli prvenstveno ali celo edino nalogo etike. Prevladovala je analiza pomena in vloge moralnih izjav, kakor se te pojavljajo v običajnem govoru, ter še posebej pojmovna analiza ključnih moralnih izrazov, ki v njih nastopajo. Vsebinska oz. normativna etika je bila izrinjena izven polja moralne filozofije.

V zadnjih desetletjih je v moralni filozofiji prisotnih vse več pozivov v smeri preseganja te vrzeli in ponovne tesnejše združitve področji metaetike in vsebinske etike, da bi tako lahko dospeli do trdne, prepričljive in ubranljive filozofske etike. To je pomembno, predvsem upoštevajoč vidik, da nam takšen vključujoč pristop omogoča, da oblikujemo normativne ideale in obveznosti, ki jih bomo na takšni podlagi lahko sprejeli tako zaradi ustreznega upravičenja njihove vsebine z normativnimi razlogi kakor tudi utemeljili z ustreznim teoretičnim razumevanjem te vsebine ter možnosti doseganja moralne vednosti v zvezi z njimi (Darwall, 2006: 25).

Metaetiko lahko določneje opredelimo tudi v okviru delitve moralne filozofije na tri podpodročja, (i) metaetiko, (ii) normativno etiko in (iii) aplikativno etiko. Me-

taetiko razumemo kot to, ki si zastavlja vprašanja o naravi moralnosti, npr. o objektivnosti moralnih sodb, utemeljenih v nanašanju na domnevno od nas neodvisno moralno stvarnost, o pomenu temeljnih moralnih izrazov ter o vlogi moralnih sodb oz. o logiki moralnega jezika in misli ali pa izkustvenih vidikih moralnega presojanja in delovanja. Normativna etika se ukvarja s splošnimi vsebinskimi vprašanji, na primer katere so temeljne vrednote; kaj je moralno pravilno in kaj napačno ravnanje oz. kaj dela naša dejanja moralno pravilna ali napačna; katere vrline naj gojimo oz. spodbujamo v skupnosti, da bodo posamezniki v njej lahko dosegali dobro življenje; kateri motivi naj nas vodijo pri delovanju; kako vzpostaviti pravično družbo in podobno. Pri tem normativna etika seveda ne ostaja zgolj na ravni določanja in naštevanja vrednot, vrst dejanj ali kreposti, ampak si hkrati zastavlja tudi vprašanja, zakaj so ravno izbrana dejanja moralno pravilna, kaj utemeljuje ta njihov moralni status in podobno. Odgovori na ta splošna vsebinska vprašanja potem predstavljajo osnovo tudi za njihov prenos oz. uporabo na bolj specifičnih področjih človekovega življenja oz. delovanja, kar vzpostavlja tretjo raven uporabne (tudi uporabljene) oz. aplikativne etike. Ta področja v tem smislu oblikujejo odgovore na bolj posebna etična vprašanja in izzive, s katerimi smo soočeni, na primer o moralnem statusu splava v različnih stopnjah razvoja zarodka, ali o dopustnosti smrtne kazni za najhujše zločine, o pravičnih socialnih in zdravstvenih politikah, o našem odnosu do drugih čutečih bitij, posebej z vidika upoštevanja njihovega trpljenja, o izzivih biotehnologije, o uničevanju in varovanju naravnega okolja ipd.

Poleg izrazov metaetika in normativna etika se pogosto uporablja tudi izraz moralna teorija (kot nasprotje praktične etike; tako govorimo na eni strani o teoretični etiki in na drugi o praktični etiki). Ta se uporablja raznovrstno, nemalokrat tudi kot sinonim za metaetiko, a v literaturi najdemo tudi rabe, kjer ta zamenjuje normativno etiko. Najbolj smiselno je med obema poljema razlikovati na način, da izraz moralna teorija zajema sklop metaetičnih stališč in odgovore na nekatera temeljna strukturna normativna vprašanja, npr. vprašanja po vrednostnem monizmu ali pluralizmu, deontološki ali posledicistični utemeljenosti moralnosti ipd. To zadnje bi lahko imenovali tudi normativna struktura določenega etiškega stališča. In v tem smislu lahko namesto delitve moralne filozofije na metaetiko, normativno etiko in uporabno etiko, dobimo le dvojico moralna teorija in praktična etika.

Področja metaetike in vrste metaetičnih vprašanj

Temeljna vprašanja o naravi moralnosti lahko razdelimo v več vsebinskih sklopov. *Moralno metafiziko* zanima predvsem vprašanje obstoja moralnih lastnosti ter narave moralnih dejstev. S pomenom moralnih izrazov in logiko moralnih sodb in jezika se ukvarja *moralno pomenoslovje* oz. *moralna semantika*. Osrednji predmet *moralne spoznavne teorije* je možnost moralnega spoznanja ter upravi-

čenja moralnih sodb. *Moralno psihologijo* zanimajo psihološki vidiki moralnih sodb, posebej odnos med našimi moralnimi sodbami ter delovanjem oz. vprašanje moralne motivacije. Zadnje področje pa je področje *moralne fenomenologije* oz. *moralnega izkustvoslovja*, ki raziskuje kvalitativne vidike našega moralnega izkustva ter posledice za moralno teorijo, ki jih lahko iz njih izpeljemo. To so seveda samo jedrna področja metaetike, ki bi jim lahko pridružili še kakšno. Naslednja tabela podrobneje izpostavi najbolj značilne teme znotraj teh petih področij metaetike.

METAETIKA	metafizika	Kakšna je narava oz. metafizičen temelj moralnih sodb? Ali obstajajo moralne lastnosti in ali obstajajo moralna dejstva? Če obstajajo, na kakšen način obstajajo in kakšne vrste so? Kako so moralne lastnosti povezane ali odvisne od naravnih lastnosti? Ali lahko moralne lastnosti zvedemo na naravne lastnosti oz. ali lahko moralna dejstva zvedemo na naravna dejstva? Kakšna je ta morebitna zvedba?
	semantika	Kakšna je pomenska vloga moralnih sodb oz. moralnega govora? Je ta vloga opisna ali pa imajo moralne sodbe vlogo izražanja zavezanosti, norm oz. predpisov (volitivna ali priporočilna vloga, predpisovalna oz. predpisna vloga) ali naše čustvene naravnosti do predmetov teh sodb (izrazna oz. ekspresivna vloga)? Lahko moralnim sodbam pripišemo resničnostno vrednost in kaj so resničnostni pogoji teh sodb? Kakšen je pomen izrazov oz. pojmov »dobro«, »slabo«, »moralno pravilno«, »moralno dopustno«, »moralno napačno« itd.? Kako so moralni pojmi povezani med seboj? Na kakšen način (lahko) te pojme uporabimo v naših sodbah in sklepanjih? Kako se razlikujejo od ne-moralnih pojmov? Kako so t.i. tanki moralni pojmi (tj. pojmi, ki imajo zgolj normativno ali vrednostno razsežnost, npr. moralno pravilno, dopustno, nedopustno, dobro, slabo) povezani z debelimi moralnimi pojmi (tj. pojmi, ki imajo poleg normativne ali vrednostne razsežnosti

		tudi bolj določno opisno vsebino, npr. pravično, iskreno, pogumno, ljubeznivo, kruto)?
	epistemologija	Ali nam je moralna vednost sploh dosegljiva? Če da, na kakšen način? In če ne, zakaj je temu tako? Če obstajajo moralna dejstva, na kakšen način so nam spoznavno dosegljiva? Ali lahko in na kakšen način lahko upravičimo svoje moralne sodbe? Kakšna je logika moralnega mišljenja in vrednotenja?
	psihologija	Kaj so oz. izražajo moralne sodbe? Ali so oz. izražajo prepričanja ali druga, nekognitivna duševna stanja ali pa morda celo kakšno posebno duševno stanje, ki je sestav prej omenjenih? Kaj je psihološki izvor ter sodb? Kako so naše moralne sodbe povezane z našo motivacijo in našim delovanjem? Ali obstaja med obojim kakšna nujna povezava v smislu, da iskrena moralna sodba pomeni tudi, da bomo vedno motivirani ravnati v skladu z njo? Če temu ni tako, kaj je potem lahko izvor naše moralne motivacije, ki bi pojasnil povezanost med moralnostjo in njeno praktično pomembnostjo?
	fenomenologija	Kakšna je narava moralnega izkustva in kako se le-to razlikuje od drugih vrst izkustva? Kaj nam lahko kvalitativna izkustva, ki so povezana z moralnim presojanjem in delovanjem, razkrijejo o sami naravi moralnosti? Kaj vse zaobjemajo ta izkustva? Kakšen pomen jim lahko pripišemo oz. kakšna je moč fenomenoloških argumentov v metaetiki v primerjavi z metafizičnimi, semantičnimi in epistemološkimi argumenti?

Glede na ta temeljna področja in zastavke razprave se je oblikovalo tudi več možnih odgovorov nanje. V nadaljevanju so poglavitna stališča kratko orisana, ker je to pomembno za razumevanje nadaljnje razprave.

Metafizika

Na polju metafizičnih vprašanj v moralni teoriji se je v zgodovini razprave zarisalo predvsem razlikovanje med moralnim realizmom na eni strani in moralnim anti-realizmom na drugi (pri slednjem srečamo tudi poimenovanji moralni irealizem in moralni nihilizem). Nekatera sodobna stališča sicer vse bolj brišejo mejo med obema. Moralni realizem je v svojem jedru stališče, da obstajajo moralne lastnosti, ki so bodisi lahko objektivne v smislu, da obstaja od našega jezika in misli neodvisna moralna stvarnost, bodisi subjektivne v smislu, da je obstoj teh lastnosti sestavno odvisen od moralnega jezika in misli (posameznikova želja ali okus, delovanje racionalne volje, odločitev presojevalca v nekem izhodiščnem ali posebnem prednostnem položaju), kot je to na primer stališče, da je dobro to, kar bi si želeli želeli.

Ta metafizična vprašanja so tesno povezana tudi z vprašanji pomena in vednosti, saj je tudi sicer metafizični realizem pogosto prepleten s semantičnimi tezami in ima pomembne posledice za naš pogled na vednost. Posamične opredelitve moralnega realizma so zelo raznovrstne ter so lahko podane v okvirih dejanskega obstoja, neodvisnosti od jezika/misli, objektivnosti, resničnosti moralnih sodb in podobno. Tako lahko srečamo precej celostne in določne opredelitve, ki moralni realizem razumejo kot sklop naslednjih petih trditev. Moralne lastnosti in moralna dejstva obstajajo. Te lastnosti in ta dejstva obstajajo objektivno (neodvisno od našega jezika in misli, sistema pojmov, prepričanj ali teorij o njih). Moralni govor je dejstveni govor (v smislu, da imajo moralne izjave/sodbe vlogo opisovanja moralne stvarnosti). Moralne izjave so bodisi resnične bodisi neresnične glede na to, ali se ujemajo z objektivnimi moralnimi dejstvi. Nekateri (trdilni) moralne izjave so (ujemalno) resnične (Timmons, 1999: 72). Russ Shafer-Landau moralni realizem opredeli kot stališče, da obstajajo moralne resnice, ki so neodvisne od katere koli izbrane perspektive oz. gledišča, kar tudi pomeni, da moralna merila, ki določajo moralna dejstva, niso resnična vsled njihove potrditve iz kakršnega koli dejanskega ali hipotetičnega gledišča (2005: 15). Spet drugi opredeljujejo moralni realizem precej bolj minimalno, in sicer kot spoj izhodiščne trditve, da so moralne izjave (vzete dobesedno) dobesedno resnične ali neresnične, in nadaljnje trditve, da so vsaj nekatere teh moralnih izjav (dobesedno) resnične (Sayre-McCord: 1988).

Ne glede na te razlike, pa je jedro moralnega realizma običajno tesno povezano s pogledom na to, kakšna je narava in kaj je vloga moralnih sodb, npr. da je moralni govor dejstveni govor, ki opisuje moralno stvarnost in je (ne)resničen, glede na

to, ali se s to stvarnostjo (ne) ujema. Glede na to, kakšna je narava moralnih lastnosti, lahko nadalje razlikujemo med moralnim naturalizmom in moralnim nenaturalizmom. Naturalistični moralni realizem moralne lastnosti in moralna dejstva razume bodisi kot posebno vrsto naravnih lastnosti in dejstev bodisi kot zvedljive na naravne lastnosti in dejstva (tj. lastnosti in dejstva, ki so predmet ali del naravoslovnih znanosti in psihologije, npr. presežek ugodja nad bolečino, zadovoljitev interesov, preživetje vrste ipd.). Nenaturalistični moralni realizem pa moralne lastnosti in dejstva razume kot posebne, *sui generis* lastnosti in dejstva, ki niso identične z ali zvedljive na naravne lastnosti in dejstva. Nasprotje moralnega realizma je moralni anti-realizem, ki zavrača od nas neodvisen obstoj moralnih lastnosti.

Semantika

Poglavitna ločnica glede vprašanj moralne semantike oz. pomenske interpretacije moralnih sodb poteka med moralnim kognitivizmom in moralnim nekognitivizmom. Vzemimo preprosto moralno sodbo »Trpinčenje čuteče živali zgolj zaradi našega lastnega užitka je moralno nedopustno.« Kakšno duševno stanje izraža ta sodba in kakšna je njena primarna pomenska vloga? Moralni kognitivizem trdi, da ta in druge podobne moralne sodbe prvenstveno so oz. izražajo prepričanja (kognitivna stanja). Moralne sodbe kot prepričanja so podobno kot vsa preostala prepričanja lahko resnična ali neresnična, njihova semantična vloga oz. funkcija pa je opisna, saj merijo na moralno stvarnost, ki jo v sodbi poskušamo zajeti. Glede zgornje sodbe bi tako lahko dejali, da izraža naše prepričanje, s katerim takšni vrsti dejanjem pripisujemo moralno napačnost, ter je posledično lahko resnično ali neresnično. Nadalje nekateri avtorji razlikujejo med močnim in šibkim kognitivizmom (Miller, 2003: 4–6). Močni kognitivizem predstavlja pravkar opisano stališče, da lahko moralne sodbe ovrednotimo glede na njihovo resničnostno vrednost, ki je posledica našega kognitivnega oz. spoznavnega dostopa do dejstev v njihovi podlagi, ki jih naredijo resnična ali neresnična. Za razliko od takšnega stališča, šibki kognitivizem sprejema le prvo trditev, hkrati pa zanika, da bi bile moralne sodbe ovrednotene kot resnične ali neresnične na temelju neke naše kognitivne zmožnosti zaznave ali dostopa do moralne stvarnosti (dejstev, stanj stvari), ampak njihovo resničnost ali neresničnost razume kot posledico tega, da naše najboljše, najbolj trdne in utemeljene moralne sodbe določajo obseg moralnih predikatov ter jih glede na te lahko tudi ovrednotimo.

Nasprotno nekognitivizem zagovarja pogled, da moralne sodbe niso in ne izražajo prepričanj, ampak druga, ne-prepričanja, ne-spoznavna oz. nekognitivna stanja, kot so na primer naša čustva, želje, naravnosti, odobravanje, sprejemanje izbrane splošne moralne norme, zapovedi ali pa kombinacijo le-teh. Kot takšne moralne sodbe tudi ne morejo biti (v najbolj neposrednem smislu) resnične ali neresnične, saj je to slednje zgolj lastnost pristnih prepričanj, ki so opisna. Za nekog-

nitiviste moralne sodbe nimajo vloge opisovanja moralnih lastnosti predmetov teh sodb oz. nimajo opisne vloge, kar pomeni, da njihova primarna vloga ni izjavljati nečesa o svetu in za kar bi lahko rekli, da je resnično ali neresnično. V tem smislu nekognitivisti razumejo moralne sodbe kot psevdo-sodbe, saj s tem poudarijo, da se na prvi pogled res morda zdi, kot da bi takšne sodbe izražale prepričanja, saj so podobne našim ostalim prepričanjem in tudi nastopajo kot takšne v našem moralnem mišljenju, vendar je ta njihova površinska deklarativna oz. opisna gramatična oblika zavajajoča. V tem smislu nekognitivizem tudi moralne izraze razume kot psevdo-izraze, ki nič ne pridajo k sami vsebini stavka, ampak izražajo našo naravnost do te vsebine.

Glede na to, kakšna je narava te naravnosti, lahko razlikujemo med različnimi vrstami nekognitivizma. Njegova najzgodnejša oblika, *emotivizem* (Ayer, Stevenson) trdi, da so moralne sodbe izraz naših čustev, torej gre pri njih primarno za izražanje čustvenih naravnosti do predmetov teh sodb, takšna raba moralnega jezika pa v medosebnem prostoru služi tudi za vplivanje na čustva in naravnosti drugih. *Preskriptivizem* oz. *predpisni pristop* (Carnap, Hare) zagovarja pogled, da so moralne sodbe prvenstveno oblika predpisov in prepovedi, ter da tako na primer naš zgornji primer moralne sodbe pomeni našo zavezanost temu, da je trpinčenje živali nedopustno, ter oznanitev drugim, da tega ne odobravamo. *Ekspresivizem norm* (Gibbard) je stališče, da moralne sodbe kot normativne sodbe izražajo delovalčevo sprejemanje teh norm ali sistema pravil. V ozadju moralnih vprašanj so vprašanja glede racionalnosti določenih čustev, kot na primer občutka krivde, jeze ali zamere, ki jih ima delovalec, ki izvrši moralno nepravilno dejanje, ali ki jih občuti presojevalec, ki dejanje presoja. *Kvazi-realizem* (Blackburn) je različica ekspresivizma (tudi izrazna teorija), ki poskuša v prvi vrsti razložiti oz. prilagoditi to, zakaj se nam zdi moralni govor dejstveni govor, ki je lahko resničen ali neresničen, in nadalje, zakaj smo upravičeni do tega, da ga kot takšnega tudi jemljemo, četudi strogo vzeto moralne sodbe ne morejo biti niti resnične niti neresnične. Moralne sodbe so glede na številne vidike podobne običajnim prepričanjem, čeprav so v svoji osnovi ne-prepričajska stanja. Moralne sodbe lahko razumemo kot neke vrste projekcijo oz. odraz naših naravnosti v svet, ki jih nadalje obravnavamo, kot da bi odražale objektivna dejstva v tem svetu. Ob tem ni potrebe, da bi poskušali zaobiti to iluzijo, kajti ravno ta ustvarja resno in pretehtano razmišljanje, ki naj bi spremljalo oblikovanje naših moralnih naravnosti.

Ta izhodišča delitve na kognitivizem in nekognitivizem so v zadnjih desetletjih postala precej manj jasna. Eden najpomembnejših premikov v razpravi je razlikovanje med dvema različnima tezama, vsebovanima v stališču kognitivizma. Prva je pomenoslovna teza, da imajo moralne sodbe opisno vlogo oz. izražajo propozicije ter so lahko resnične ali neresnične, druga pa je psihološka teza, da te sodbe tipično so oz. izražajo prepričanja. Glede na to, da se tezi pokažeta kot medsebojno neodvisni, vznikneta na tem temelju dve novi metaetični stališči, od katerih vsako sprejema le eno izmed teh tez, drugo pa zavrača. *Neopisni kognitivizem*

(tudi *kognitivistični ekspresivizem*) tako sprejema psihološki vidik tradicionalnega kognitivizma (tudi glede na izkustvene vidike, ki spremljajo moralne sodbe ter so podobni izkustvenim vidikom običajnih prepričanj), zavrača pa, da bi imele moralne sodbe opisno vlogo. *Moralni fikcionalizem* pa sprejema njihovo opisno vlogo, a zavrača, da bi do te opisne vsebine zavzemali pristno prepričanjsko naravnost, ter jih razume bolj v smislu umislekov ali kvazi-zatrditev.

Moralna epistemologija

Izhodiščno vprašanje moralne epistemologije je, če in kako je možna moralna vednost. Glede na odgovor na to vprašanje razlikujemo med moralnim anti-skepticizmom in moralnim skepticizmom. Moralni anti-skeptik zagovarja, da je moralna vednost dosegljiva, ter lahko nadalje ponudi raznolike razlage te dosegljivosti, ki so podobne tudi siceršnjim spoznavno-teoretskim stališčem, ki se opirajo med drugim na temeljnost, zanesljivost ali tesno povezanost in skladnost naših prepričanj kot vir njihovega upravičenja. Podobno so raznoliki tudi razlogi za sprejemanje moralnega skepticizma, ki je prvenstveno stališče, ki zavrača možnost ali dosegljivost moralne vednosti, čeprav ga širše povezujemo tudi z zavračanjem možnosti moralnih resnic, obstoja moralnih dejstev in moralne stvarnosti ali pa dobrih razlogov, zakaj delovati moralno.

Po eni strani se nam zdi, da moralni skepticizem glede moralne vednosti ne more biti resnično stališče, saj bi malokdo zanikal, da vemo oz. smo lahko upravičeno prepričani v vsaj nekatere izmed naših moralnih presoj, kot npr. da je moralno nedopustno trpinčiti otroka zgolj zaradi naše zabave, da je genocid zločin, ali pa da je suženjstvo moralno nedopustno. V tem smislu ga lahko označimo za stališče, ki očitno nasprotuje našemu vsakdanjemu razumevanju moralne misli in prakse. Vendar nas premislek o izvorih ter načinu upravičenja teh sodb hitro sooči z zahtevnimi izzivi. Če kot moralni realisti sprejemamo obstoj moralne stvarnosti, potem se nam z vidika moralnega spoznavoslovja seveda postavi vprašanje, kako nam je ta stvarnost dosegljiva, tudi glede na razumevanje te moralne stvarnosti (npr. če so moralne lastnosti vrsta naravnih lastnosti ali vsaj nanje zvedljive, potem lahko verjetno oblikujemo pogled na naš spoznavni dostop do njih, podoben našemu siceršnjemu spoznavnemu dostopu do sveta; če so moralne lastnosti *sui generis*, nenaravne lastnosti, potem moramo pojasniti, kako so nam kot take spoznavno dostopne).

Kot rečeno obstaja več spoznavnih teorij, na katere se glede tega lahko opremo. Moralni intuicionizem tako predstavlja vrsto spoznavnega fundacionalizma ter zagovarja, da so naša temeljna moralna prepričanja upravičena preko razumskega moralnega uvida oz. moralne intuicije. Moralna vednost glede na ta pogled vključuje posebno vrsto spoznanja, in sicer neposredni uvid, razumevanje, dojem, moralno (za)vest oz. moralno intuicijo kot vrsto spoznavne zmožnosti, ki nam omo-

goča doseganje moralne vednosti. Posamezne teorije in avtorji pojem intuicije razumejo različno, vendar osnovna ideja ostaja ta, da je vsaj del naše (najbolj temeljne) moralne vednosti vezan na moralno intuicijo kot spoznavni dostop do moralne resnice.

Spet druge moralne teorije razumejo moralno upravičenje bolj v okvirih skladnosti in ujemanja med našimi moralnimi prepričanji ter so blizu splošnemu spoznavnemu koherentizmu. Glede na to stališče je izbrano prepričanje upravičeno, v kolikor se ujema in sklada z drugimi prepričanji ter je torej v upravičenjskem smislu z njimi povezano. Koherentisti pogosto uporabijo primero splava, ki plava na površini vode in predstavlja varno in trdno površino, ne zaradi določenega posameznega debela ali deske (posamezno prepričanje), ampak zaradi celote teh in načina, kako trdno in pogosto so povezane skupaj (mnoštvo prepričanj in povezave med njimi). Podobno lahko razumemo upravičenje glede naših moralnih prepričanj kot izhajajoče iz povezane celote teh prepričanj, ki je medsebojno skladna in kjer se prepričanja (na različnih ravneh splošnosti) podpirajo med sabo.

Če sprejmemo opredelitev in razumevanje vednosti kot resničnega upravičenega prepričanja, potem uvidimo, da ima lahko na drugi strani moralni skepticizem v širšem smislu prav tako več izvorov. Prvi izmed njih je nekognitivizem, saj z zavračanjem pogleda na moralne sodbe kot kognitivna oz. spoznavna stanja spodmaknemo možnost, da bi jih neposredno razumeli kot resnična upravičena prepričanja. Drugi vir za skepticizem je zavračanje obstoja moralne stvarnosti na način, da hkrati ne dopuščamo možnosti, da bi bile moralne sodbe kadar koli resnične. Tretji poglavitni izvir moralnega skepticizma pa je stališče (vzporedno splošnemu spoznavnemu skepticizmu), da so sicer moralne sodbe prepričanja, ki so lahko resnična ali neresnična, vendar glede nobenega izmed njih nikoli ne moremo biti upravičeno prepričani. Na te in podobne vidike prepletenosti med metafizičnimi, semantičnimi in epistemološkimi vprašanji nakazujejo tudi nekatera poimenovanja metaetičnih stališč. J. L. Mackie (1977: 15) tako svojo teorijo, katere bistvo predstavlja moralni anti-realizem, poimenuje moralni skepticizem (Strahovnik, 2008), Brink (1989: 18 op. 6) pa moralni kognitivizem tesno povezuje z moralni anti-skepticizmom. Osrednje razumevanje moralnega skepticizma je povezano z zadnjim omenjenim vidikom, torej z vidikom upravičenja.

Eden izmed jedrnih premislekov, ki takšen moralni skepticizem podpirajo, je tako povezan z nazaj nadaljujočim se nizom upravičenja, saj se zdi, da je smiselno sprejeti, da so naša moralna prepričanja upravičena sklepalno (izpeljano) glede na naša druga moralna prepričanja (npr. moje prepričanje, da bi bilo moralno napačno prelomiti mojo obljubo, ki sem jo dal sosеду, bi lahko upravičevalo moje prepričanje, da je moralno napačno ne biti zvest dani besedi, in še nadalje prepričanje, da to pomeni kršitev človekovega dostojanstva). Toda postavi se vprašanje, kaj pa potem upravičuje ta druga prepričanja, iz katerih izpeljujemo upravičenje prvih. Tako lahko posplošimo, da naše izbrano moralno prepričanje kot vednost

upravičuje neko nadaljnje, morda bolj temeljno prepričanje. Postavljeni smo pred problem spoznavnega nazadovanja, kajti če sledimo verigi teh prepričanj, ki upravičujejo tiste za njimi, se lahko utemeljeno vprašamo, kjer se to upravičenje sploh konča in ali lahko najdemo nekakšen drug način upravičenja te izhodiščne točke, kajti glede na predpostavko temeljnosti, le-te ne moremo upravičiti z nadaljnjim nazadovanjem po tej verigi. Moralni skeptik trdi, da nikoli ne moremo priti do takšnih neodvisno upravičenih prepričanj, torej tudi nima nobenega smisla trditi, da bi lahko iz njih izpeljali prepričanja, ki bodo prav tako upravičena. Za razliko od skeptičnega stališča smo lahko zgoraj videli, da intuicionizem in koherentizem trdita, da se ta niz upravičenja bodisi konča pri določenih temeljnih moralnih prepričanjih ali pa v skladnem sklopu prepričanj, ki podpirajo druga drugo.

Drugi vidik, na katerega moralni skeptiki pogosto opozarjajo, je obstoj razmeroma razširjenega moralnega nestrinjanja glede moralnih vprašanj, ki se zdi nerazrešljivo in neodpravljivo ter torej nakazuje v smer, da nimamo opravka z moralno vednostjo, temveč zgolj z nekakšnim množtvom mnenj in pogledov. Vendar, kot smo nakazali že zgoraj, je moralni skepticizem stališče, ki izenačuje moralni status vseh moralnih prepričanj (tako manj kot bolj spornih) ne glede na njihovo vsebino, in zdi se, da ga torej ne moremo vzpostaviti zgolj z opozorilom na dejstvo, da se glede nekaterih moralnih vprašanj, kot npr. moralna dopustnost evtanazije, splava ali prehranjevanja z mesom živali, ki jim način proizvodnje hrane povzroča precejšnje trpljenje, ne moremo poenotiti.

Vsak tak teoretični moralni skepticizem pa nima nujno tudi posledic za samo moralno prasko (teoretični oz. epistemološki moralni skepticizem zadeva obstoj dobrih razlogov za prepričanje, medtem ko gre pri praktičnem skepticizmu za razloge za delovanje). Nekdo, ki je moralni skeptik, je lahko do svojih moralnih prepričanj oz. sodb prav tako trdno zavezan ter jim v svojem delovanju sledi, kakor ostali. Moralni skepticizem ne pomeni nujno odsotnosti vsake moralne motivacije za delovanje v skladu z moralnimi prepričanji niti tega, da ta prepričanja ne bi ponujala razlogov za delovanje.

Moralna psihologija

Polje moralne psihologije zaseda predvsem vprašanje moralne motivacije, to je, kako so naše moralne presoje povezane s tem, da smo tudi motivirani delovati v skladu z njimi oz. nam dajejo razloge, da v skladu z njimi ravnamo. V tem smislu razlikujemo med moralnim internalizmom in moralnim eksternalizmom, pri čemer prvi pritruje tej notranji povezanosti sodb z motivacijo, slednji pa jo zanika. Hkrati se moralna psihologija deloma prekriva z razpravo o moralni semantiki, in sicer v tistem njenem delu, kjer teče razprava o tem, kakšne vrste duševnih stanj so oz. izražajo moralne sodbe, ter nadalje, kakšne so značilnosti teh dušev-

nih stanj. V tem uvodnem pregledu nas bo ožje zanimalo zgolj vprašanje moralne motivacije.

Glede na naše vsakdanje moralno izkustvo lahko izhajamo iz pripoznanja, da nas naše moralne sodbe vedno (ali pa vsaj zelo pogosto) motivirajo, da v skladu z njimi tudi ravnamo oz. – v primeru nasprotja med različnimi motivacijami – da bi ravnali v skladu z njimi ob odsotnosti drugih premislekov. Lahko tudi rečemo, da imajo naše iskrene moralne sodbe nad nami razumno avtoriteto v smislu, da nam dajejo dobre razloge, da jim v svojem delovanju sledimo, tudi če to delovanje ni nujno v našem lastnem interesu oz. tudi če pred tem v sebi ne prepoznamo neke določene želje, da bi tako ravnali. Gotovo bi vsaj podvomili v iskrenost moralnih prepričanj nekoga, ki bi s svojimi moralnimi izjavami na primer zatrjeval, da je določena vrsta dejanja moralna nepravilna, hkrati pa dosledno ravnal nasprotno temu.

Moralni internalizem zagovarja bistveno oz. notranjo povezanost moralnosti oz. moralnih sodb in motivacije za delovanje. Brink (1989: 42) ga opredeli kot sklop naslednjih treh trditev:

- (i) moralni premisleki (sodbe, izjave, ...) nujno motivirajo oz. priskrbijo razloge za dejanje;
- (ii) ker zgornja povezava izhaja iz samega *pojma* moralnosti, je teza o motivacijski moči ali razumni avtoriteti moralnih premislekov apriorna;
- (iii) ker zgornja povezava izhaja iz samega *pojma moralnosti*, razumna avtoriteta in motivacijska moč moralnih premislekov nista odvisni od vsebine moralnosti, dejstev o delovalcu (njegovih želja, interesov, ...) ali vsebinskih podmen teorije racionalnosti.

Internalizmu nasprotno stališče je moralni eksternalizem, ki zanika zgornje postavke in trdi, da sta motivacijska moč in razumnost moralnosti odvisni od dejavnikov, ki so moralnosti oz. moralnim sodbam zunanji. Povezava med sodbami in motivacijo je zgolj naključna, spoznavna *a posteriori* in je nadalje lahko odvisna od same vsebine moralnih sodb ali dejstev o delovalcu (njegovih želja, nagnjenj, interesov, stremljenj, splošnejših zavezanosti k moralnosti ipd.). Če kot izhodiščni premislek vzamemo antično vprašanje »Čemu biti moralen?« in ga na primer razumemo kot to, ki vprašuje »Čemu naj bi storil to, za kar sodim, da je moralno pravilno ali dobro?«, potem je za motivacijskega internalista to vprašanje absurdno, nesmiselno, ali pa vsaj nerazumno. Kdor ga bi zastavil, očitno ne bi poznal uporabljenih pojmov oz. ne bi vedel, po čem vprašanje sploh sprašuje. Za motivacijskega eksternalista pa to isto vprašanje ni nesmiselno, temveč kvečjemu nemoralno; kdor ga zastavlja je kvečjemu moralno nezrel, nerazvit, neobčutljiv ali celo sprijen, ne pa tudi nujno nerazumen.

Ta grobi zaris obeh nasprotnih stališč lahko nadalje dopolnimo in pretanimo glede na to, kaj natančno imamo v mislih z moralnostjo, kaj imamo v mislih z motivaci-

jo in glede na obseg in moč razmerja med obema. Glede prve točke lahko razlikujemo med sodbenim internalizmom (tudi prepričanski internalizem) in med internalizmom obstoja (tudi motivacijski internalizem v ožjem smislu ali moralni internalizem; Shafer-Landau, 2005: 142–3). Prvi na prvo stran obravnavanega razmerja, tj. na stran moralnosti, postavlja moralne sodbe/izjave oz. moralna prepričanja, drugi pa moralne obveze oz. dolžnosti same. Eno je, če trdimo, da je delovalec, ki *sodi* ali je *prepričan*, da je neko dejanje moralno pravilno, vsled tega tudi motiviran oz. ima razlog, da ravna v skladu s tem; drugo pa, da če *ima* delovalec moralno obvezo ali dolžnost storiti neko dejanje, potem je nujno, da ima tudi motivacijo oz. razlog, da v skladu z njo ravna (in to neodvisno od tega, ali ima o tej moralni obvezi neko prepričanje; od tod tudi poimenovanje internalizem obstoja, ker sam obstoj moralne obveze predstavlja razlog za dejanje). Slednje stališče lahko označimo tudi z izrazom internalizem delovalca (glede na sam pojem moralnosti moralne obveze motivirajo delovalca ali priskrbijo delovalcu razlog za dejanje), prvega pa internalizem presojevalca (moralna prepričanja ali moralne sodbe motivirajo presojevalca oz. delovalca ali priskrbijo presojevalcu razlog za dejanje). Ob tem lahko ugotovimo, da je internalizem delovalca oz. obstoja objektivni v smislu, da motivacijo ali razloge naveže na moralne obveze, neodvisno od tega, ali jih nekdo prepozna kot take. Internalizem presojevalca je to, kar sem zgoraj imenoval sodbeni internalizem in je subjektiven v smislu, da motivacijo ali razloge naveže na presojevalca in njegove sodbe ali prepričanja, neodvisno od tega, ali so ta prepričanja resnična oz. sodbe upravičene (Brink, 1989: 40–1).

Glede drugega vprašanja lahko na eni strani govorimo o motivaciji za delovanje in na drugi o razlogih za dejanje. Internalizem glede motivacije je stališče, da je nujno: če neka oseba iskreno sodi, da je dejanje moralno pravilno, potem bo tudi motivirana, da ravna v skladu s to sodbo. (Podobno tudi glede internalizma obstoja: če ima neka oseba moralno obvezo, potem bo tudi motivirana, da jo izpolni.) Internalizem glede razlogov pa trdi, da je nujno: če neka oseba iskreno sodi, da je dejanje moralno pravilno, potem ima tudi razlog, da ravna v skladu s to sodbo. (To spet lahko preslikamo v okvire internalizma obstoja, kar bi torej pomenilo, da obstoj moralne obveze pomeni, da ima oseba razlog, da ravna v skladu z njo.)

Tretja razsežnost v razpravi med internalizmom in eksternalizmom pa je moč (in s tem tudi obseg) razmerja med moralnimi sodbami in motivacijo. Šibki (sodbeni) internalizem zagovarja stališče, da so moralne sodbe notranje povezane z motivacijo za delovanje, vendar ni nujno, da bo ta motivacija nujno prevladujoča in bo močnejša od drugih silnic v naboru naših motivacij za delovanje, zato tudi ni nujno, da bomo v delovanju vedno sledili našim, sicer iskrenim moralnim sodbam. Motivacija, ki je povezana z moralno sodbo, je torej lahko povožena s strani drugih premislekov. Močnejša, pa tudi precej manj sprejemljiva različica internalizma pa zagovarja, da moralna motivacija, izhajajoča iz naših moralnih sodb, vedno prevlada na drugimi motivacijami in željami. Znotraj te razsežnosti naj omenimo še to, da internalizem pogosto poskuša omejiti obseg delovalcev, pri katerih

so moralne sodbe nujno povezane z motivacijo, na način, da iz tega obsega izključi osebe, ki niso povsem praktično razumne oz. ne zadostijo naboru določenih pogojev. S tem na primer izključimo osebe, ki so v globoki depresiji in ki navkljub temu, da ohranijo jedro svojih vrednostnih oz. moralnih sodb in prepričanj, pogosto nimajo nobene motivacije, da bi lahko ravnale v skladu z njimi. Tudi glede tega vidika lahko torej govorimo o močnejši in šibkejši različici motivacijskega internalizma, pri čemer prvi ne pozna pravkar omenjene omejitve.

Moralna fenomenologija

Moralna fenomenologija oz. moralno izkustvoslovje v prvi vrsti zadeva kvalitativne, izkustvene vidike moralnosti, torej subjektivna, kvalitativna izkustva, odzive in stanja, ki spremljajo moralno presojanje, odločanje in delovanje, na primer ob tem, ko presodimo, da je določeno dejanje naša moralna dolžnost, se ujamemo v težko razrešljivo moralno dilemo ali pa moralno obžalujemo naše moralno napako ravnanje. Kot takšna se moralna fenomenologija šele v zadnjih desetletjih vzpostavlja kot posebno polje znotraj metaetike, poprej pa so podobni fenomenološki premisleki nastopali kot del splošnejših premislekov o naravi moralnosti, posebej tistih, ki so bili vezani na skladnosti posameznih moralnih teorij (ali še ožje, njihovih posameznih vidikov) z običajnimi prepričanji o morali. Razloga za to predhodno odsotnost moralne fenomenologije sta dva. Prvi vzrok je Moorovo zavračanje naturalističnega razumevanja etike, kar je posledično pomenilo, da nam niti naravoslovne znanosti niti psihologija ne morejo razkriti narave moralnosti, ampak nas kvečjemu lahko zavedejo v zmoto glede le-te. In ker je fenomenološki opis moralnega izkustva vezan (tudi) na opisno moralno psihologijo, bi to že lahko pomenilo prvi korak na poti do zagrešitve naturalistične zmote. Drugi vzrok pa je izrazita prevlada pomembnosti moralne semantike (pomen moralnih izjav in njihova logika) v metaetični razpravi, kar zopet pomeni, da so psihološki, izkustveni vidiki ostajali v ozadju.

V filozofiji sam izraz fenomenologija povezujemo s kvalitativnimi vidiki našega izkustva, s »kako je biti« (ang. *what it's like*) vidiki, vendar pa nekateri zajamejo njeno področje tudi širše, torej da poleg »kako-je-biti« konkretne moralne izkušnje, le-ta zajema tudi nekatere značilnosti slovnice in logike morale oz. moralnega jezika ter določene kritične prakse, kot je npr. udeleženosť v moralnem nestrinjanju (Horgan in Timmons, 2008). Glede fenomenoloških vidikov moralnega izkustva si lahko na primer zastavimo naslednja vprašanja. Kako se nam kaže moralna vrednost in moralni status ob tem, ko moralno presojamo? Se nam ta kaže kot objektivna, kot od nas neodvisna, kot prisotna tam zunaj v svetu samem ali pa ju vidimo kot vznikajočo nad našimi odzivi na sicer ne-moralni svet okoli nas? Kako občutimo moralno obvezo, za katero sodimo, da jo imamo in da se moramo nanjo odzvati? Od kod prihaja avtoriteta moralnih sodb in kaj jo izkustveno tvori? (Potrč in Strahovnik, 2014) Na kaj merimo oz. kaj iščemo, ko moralno presoja-

mo? Od kod izvira motivacija za moralno delovanje in kakšna je njena narava? Gre torej za t.i. »gole moralne občutke«, tj. pred-teoretične občutke in odzive, ki se v delovalcu oz. presojevalcu vzbudijo v moralni situaciji, ter ki so mu introspektivno oz. notranjeuvidno dostopni.

Kirchin v zvezi z njimi na naslednji način izpostavi vzporednice z običajno, nemoralno fenomenologijo. Ko gledamo predmete, lahko preprosto, nereflektirano občutimo in mislimo, da njihove barve res pripadajo njim na intrinzičen način (kot je to z njihovo dolžino ali obliko) in da njihov obstoj sploh nima opravka z nami. Tukaj gre za določen *goli* občutek, ki ga imamo in ki še ni preoblikovan z nobeno teorijo barve in barvne zaznave. Seveda lahko spremenimo svoje poglede, ko o barvni zaznavi bolje razmislimo. Podobno imam takšne gole občutke v mislih, ko govorim o moralni fenomenologiji. Ko opazujemo ali razmišljamo o moralnih dejanjih in situacijah, imamo ob tem določene pred-teoretične občutke in odzive, ki so nevplivani in nepreoblikovani glede na kakšno metaetično teorijo, ki bi jo izbrali. Npr., lahko imamo občutek, da dejanje očitno in dejansko ima moralno vrednost, ki jo ima – prav tako kot ima lastnosti trajanja ali druge lastnosti –, in da so ljudje preprosto v zmoti, če je ne opazijo. Podobno lahko občutimo, da obstaja neke vrste avtoriteta moralnosti, kateri smo mi sami in vsi drugi zavezani (Kirchin, 2003: 242–3).

Ti kvalitativni izkustveni vidiki so lahko povezani z raznolikimi vrstami vsakdanjega moralnega izkustva, od takšnih, ki predhodijo samemu oblikovanju moralne sodbe, do moralnih sodb o vrednosti ali dolžnostih prvega ali drugega reda, ki jih oblikujemo spontano in intuitivno ali pa po tehtnem premisleku.

Za moralno teorijo so ti kvalitativni izkustveni vidiki moralnosti pomembni, ker nastopajo v t.i. fenomenoloških argumentih. Fenomenološki argument je argument, ki izhaja iz fenomenološkega opisa izbranih vidikov moralnega izkustva ter iz njih izpelje posledice za metaetično oz. teoretsko razumevanje moralnosti. Izhodiščna premisa takšnega argumenta je torej opis značilnosti omenjenih pred-teoretičnih občutij, pri čemer ji dodamo bodisi močnejšo premiso, da je narava moralnosti takšna, kakršna se nam kaže neposredno v izkustvu, ali pa izbrano različico nekoliko šibkejše premise, npr. da mora biti moralna teorija sposobna prilagoditi naravo omenjenih izkustev, tj. jih razložiti ali nakazati izvore njihove zmo-te. Ustrezna metaetična teorija ne sme preveč iznakaziti narave oz. izravnati bogatosti moralne fenomenologije. Kot pravi Kirchin:

Argumenti, ki se pomaknejo od pred-teoretičnih občutij do metaetičnih sklepov, so ti, ki jim pravim fenomenološki argumenti. /.../ [G]lavna ideja v njih ozadju je, da je gola narava moralnih izkustev (ang. *the raw character of one's ethical experience*) odsevana v metaetičnem stališču na način, da imamo dobre in zdrave razloge za misel, da je s stvarmi dejansko tako, kot zatrjuje metaetično stališče. Takšna argumentacija ima malo opraviti z drugimi, prepoznavno metafizičnimi, epistemološkimi, ... argumenti (Kirchin, 2003: 243).

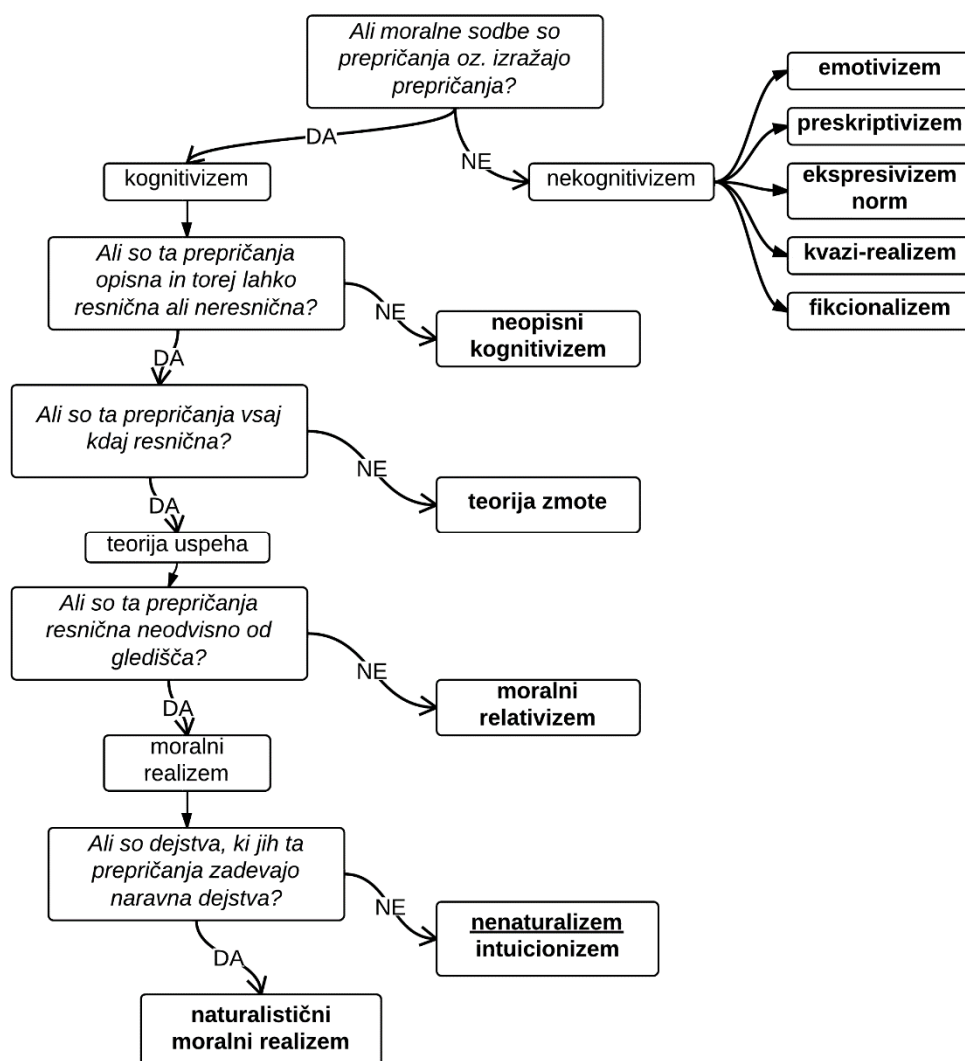
Vendar mora takšen fenomenološki argument izpolnjevati določene pogoje. Pri njem mora iti res za značilno fenomenološki argument, ki mora biti različen in karseda neodvisen od siceršnjih metafizičnih, semantičnih in epistemoloških premislekov in argumentov, ki so vezani na moralno teorijo. Opirati se mora na introspekcijo oz. notranji vpogled in na preko njega spoznane vidike (lastnega) moralnega izkustva; ti vidiki nam morajo biti notranje dostopni. Fenomenološki argument tudi predpostavlja, da ljudje široko delijo skupno moralno fenomenologijo in da nam to zagotavlja zadostne pred-teoretične temelje, na katere se potem lahko sklicujemo (prim. Horgan in Timmons 2005; 2008).

Fenomenološki argumenti niso omejeni zgolj na posledice za moralno metafiziko, ampak jih lahko podamo tudi za ali proti (ne)kognitivistični interpretaciji moralnih sodb, glede premislekov o moralni motivaciji ter tudi glede same normativne strukture moralnih teorij. Ob tem pa na polju metaetike ni strinjanja glede moči oz. pomena takšnih fenomenoloških argumentov. Na eni strani jih lahko razumemo kot zajemajoče sicer zanimiv opis moralnega izkustva in kot prispevek k opisni psihologiji, toda kot povsem nepomembne za metaetično razpravo. Drugič, moralna fenomenologija je lahko pomembna za razpravo o metaetiki, a z omejenim pomenom, ki nikakor ne dosega vloge, ki jo sicer imajo metafizični, semantični in epistemološki argumenti. Zakaj bi sploh sprejeli sklep, da je neko stališče zmotno, če ne zmore ustrezno zajeti kvalitativnih vidikov našega izkustva? Moralna fenomenologija je kvečjemu dobra izhodiščna točka naših metaetičnih razmišljanj, a komaj kaj več od tega. Težava je tudi v tem, kako najti ustrezen fenomenološki opis nekega izbranega vidika moralnega izkustva, ki bo v zadostni meri nevtralen, da ne bo že vnaprej izključeval določenih metaetičnih stališč – to bi pomenilo vložiti preveč v moralno fenomenologijo samo (Kirchin 2003: 244, 251–2, 257). Tretjič, do moralne fenomenologije lahko zavzamemo tudi precej bolj naklonjen odnos, po katerem ji namenimo mesto ob siceršnjih metafizičnih, semantičnih in spoznavno-teoretskih osiščih metaetičnih razprav. Gotovo je res, da zgolj s fenomenologijo ne bomo mogli dokončno ubraniti ali zavrniti nekega metaetičnega stališča, a enako drži tudi za ostale vidike razprave. Pri moralni teoriji gre za ocenitev njenih primerjalnih prednosti in slabosti in znotraj te primerjave moramo dati fenomenologiji njeno lastno mesto (Strahovnik 2007).

Vstopne točke v metaetično razpravo

Glede na zgoraj predstavljenih pet področij metaetike ter različna stališča na vsakem izmed njih si lahko zastavimo vprašanje, kaj je najboljša vstopna točka v metaetično razpravo. Tukaj ne moremo najti enotnega odgovora, saj si pristopi različnih avtorjev nasprotujejo. Najbolj pogost pristop je ta, da pričnemo z razlikovanjem posameznih metaetičnih stališč glede na moralno semantiko, torej glede na razumevanje pomena in vloge moralnih sodb. Na ta način vsa stališča razdelimo v dve skupini, na kognitivistična in nekognitivistična stališča, ter jih potem

nadalje podrobneje delimo glede na preostale vidike (slika spodaj; Miller 2003). Drugi avtorji v izhodišče kot temeljno postavljajo razliko med moralnim realizmom in antirealizmom (McNaughton, 1988), spet tretji med naturalizmom in nenaturalizmom (Moore, 2000; Ross, 1939; Timmons, 1999), med objektivizmom in subjektivizmom, med moralnih relativizmom in nerelativizmom ali pa med moralnih anti-skepticizmom in skepticizmom.



Ostala razumevanja metaetike in povezanost z drugimi področji

Poleg omenjenih obstaja še nekaj posebnih, bolj specifičnih problemskih sklopov, ki se jih lotevamo znotraj moralne teorije, kot so na primer vprašanje obsega etike, vprašanje statusa etike kot filozofske discipline, obravnavanje značilnosti posameznih moralnih teorij, vprašanje metod etike ter vloge moralnih in zdravo-razumskih intuicij.

Pod okrilje metaetičnega raziskovanja gotovo sodi medsebojno primerjanje različnih moralnih teorij ter raziskovanje njihovih značilnosti in teoretskih zavezanoosti. Kakšni so lahko kriteriji za primerjavo oz. izbiro moralnih teorij? Prednost moralnih teorij *qua* teorij je vsekakor izkazovanje določenih teoretskih vrlin oz. odlik, kot so konsistentnost, razlagalna moč, varčnost, določnost in stabilnost pri rabi na primerih, izogibanje *ad hoc* predpostavkam, ohranitev obstoječih prepričanj, skladnost z moralnim izkustvom ipd. Konsistentnost moralne teorije pomeni, da v njej ne sme biti neodpravljljivih napetosti med njenimi posameznimi metaetičnimi vidiki in da lahko načela v njeni normativni strukturi skupaj z ustreznimi dejstvenimi informacijami vodijo do konsistentnih ovrednotenj dejanj, oseb in ostalih predmetov vrednotenja. Varčnost pomeni, da moralna teorija ni zavezana odvečnim teoretskim postavkam. Razlagalna moč glede na metaetične vidike pomeni, da lahko moralna teorija v čim večji meri pojasni, zakaj so izbrani metaetični pogledi pravilni in na kakšen način ti pojasnjujejo tudi naša običajna moralna izkustva. Glede na normativni del moralne teorije pa ima razlagalna moč tista teorija, katere načela pojasnjujejo bolj specifične in usmerjene moralne sodbe in nam pomagajo razumeti, zakaj imajo predmeti vrednotenja določen moralni status. Določnost in stabilnost pri uporabi pomenita, da morajo biti moralna načela takšna, da vodijo do določenih moralnih sodb glede na širok obseg primerov. Prav tako mora moralna teorija ohranjati smiselnost in intuitivno privlačnost običajnih prepričanj o naravi moralnosti in običajnih intuicij o posameznih primerih moralnih situacij ali pa njih zmotnost ustrezno pojasniti (Timmons, 2002: 12–17, 271–2). Moralna teorija je trdnejša glede na to, v kolikšni meri nam pomaga pri odgovarjanju na moralna vprašanja, pri katerih v odgovore nismo povsem prepričani ali pa o njih vlada nestrinjanje. Ko imamo pred sabo izbrano metaetično teorijo, jo lahko tako primerjamo z drugimi glede na vse te izbrane razsežnosti ter ocenimo, kako dobro se glede njih izkaže v primerjavi z drugimi teorijami.

Poleg naštetih se kot vrlina oz. prednost zgolj metaetičnih stališč pogosto navaja tudi združljivost z vsemi (ali vsaj s čim večjim številom) vrstami normativnih etičnih teorij. V prednosti naj bi tako bilo tisto metaetično stališče, ki je lahko teoretsko ozadje čim večjega števila raznolikih normativnih stališč, npr. deontološke etike, teleološko utemeljene etike, etike vrlin, ipd. Vendar je to slednje razumevanje po drugi strani že v napetosti z na začetku izpostavljeno povezanostjo metaetičnih in normativnih vidikov moralnih teorij. Nadalje, sprejemljiva moralna teorija se mora skladati z globoko ukoreninjenimi predpostavkami običajnega mo-

ralnega govora in moralne prakse. Sprejemljiva moralna teorija se mora skladati s sprejemljivimi in splošno sprejetimi stališči in izhodišči drugih relevantnih področij raziskovanja (Timmons, 1999: 12).

Pri prvem vidiku gre za to, da se mora izbrana moralna teorija na čim ustrežnejši način skladati z zdravo-razumskimi intuicijami o moralnosti oz. jih pojasniti na način, da jih bodisi v čim večji meri ohranja bodisi (če so te zmotne) pojasni, kakšni so izvori njihove globoko zakoreninjene zmote in zakaj le-ta v običajni misli vztraja. Pogosto izpostavljene predpostavke vsakdanjega in zdravo-razumskega pogleda na moralnost so med drugim te, da je moralni govor opisni, dejstveni govor; da nam gre pri moralnem nestrinjanju za iskanje pravih odgovorov na moralna vprašanja in ne le za govorjenje eden mimo drugega; da je možna moralna zmota; da lahko imamo moralno vednost ipd. Pri drugem vidiku gre za ujemanje s predpostavkami splošnejših oz. širših pristopov v filozofiji, skladnost s preostalinimi filozofskimi disciplinami in tudi drugimi področji (znanstvenega) raziskovanja. Kajti, kot je izpostavil že Mackie, »[m]oralna načela in moralne teorije ne stojijo same zase: zadevajo in so same zadevane od prepričanj in predpostavk drugih področij, še posebej psihologije, metafizike in religije« (1977: 203).

Zato je metaetika tesno povezana z drugimi filozofskimi področji (poleg pomenoslovja, metafizike in spoznavne teorije), posebej s filozofijo jezika, filozofijo kognitivne znanosti, logiko, fenomenologijo in estetikom, pa tudi z drugimi znanostmi, kot so (socialna) psihologija, nevroznanost (posebej nevroetika), antropologija ali evolucijska biologija.

Prihodnost metaetike

V zadnjih desetletjih razvoja metaetike lahko razberemo nekaj pomembnih teženj, ki vsekakor lahko napovedujejo tudi njen razvoj v prihodnje. Prva takšna značilnost, ki jo lahko izpostavimo, je preseganje prvenstveno semantičnih in metafizičnih razprav glede pomena moralnih izjav in ontoloških postavk moralnosti ter vse večji poudarek na vprašanjih moralne psihologije, spoznavne teorije in moralne fenomenologije. Ta slednja vprašanja se tudi vse bolj tesno povezujejo z znanostmi (psihologija, nevroznanost, sociologija, antropologija) ter ostalimi področji filozofije (filozofija duha, eksperimentalna filozofija, nevroetika), kar pomeni tudi razširjanje polja metaetičnih vprašanj in pogledov na področja, kot so moralna presoja, odločanje, kognitivni izvor moralnih sodb in teorija upravičenja moralnih sodb. Pa tudi stara, semantična in metafizična vprašanja so lahko na tej podlagi, predvsem glede moralne fenomenologije, zastavljena in odgovorjena v novi luči. Hkrati je s to prepletenostjo vprašanj narasel širši filozofski pomen metaetike.

Vse to je odprlo pot tudi novim stališčem, predvsem pa tudi drugačni zastavitvi samih izhodiščnih vprašanj. Tako gre druga zaznana težnja razvoja v smeri vznika-

nja novih, križanih oz. hibridnih metaetičnih stališč, ki se vse bolj posvečajo iskanju vmesnega prostora med tradicionalnimi, večkrat povsem nasprotnimi in nezdružljivimi metaetičnimi stališči, za katera se je še do nedavnega zdelo, da jih ne moremo preseči (npr. razlike med kognitivizmom in nekognitivizmom). Tri stališča, ki izrazito stremijo k preseganju teh mej, so kvazi-realizem, moralni fikcionalizem in kognitivistični ekspresivizem (Strahovnik 2008). V njihovem okviru so tradicionalna nasprotja zabrisana ter vse bolj se poudarja celovit pogled na naravno moralnost, ki bo filozofsko utemeljen, združljiv z ugotovitvami znanosti ter tudi blizu našemu vsakdanjemu izkustvu moralnosti (Horgan in Timmons, 2006: 14–15). Po drugi strani pa gre razvoj tudi v smeri vse večjega povezovanja strogo metaetičnih vprašanj z vsebinskimi oz. normativnimi vprašanji. To lahko vidimo, če pogledamo na primer razpravo o moralnem partikularizmu oz. vlogi moralnih splošnosti v naši moralni misli in delovanju, kjer težko povsem razločimo metaetične vidike od ostalih. Podobno je v nekaterih razpravah o sodobnih različicah etike vrlin, kjer najdemo sorodno prepletenost normativnih vidikov in predvsem vidikov moralne psihologije. Kot zadnji vidik razvoja pa lahko izpostavimo vse večjo vpetost metaetike v širše filozofske projekte in stališča, posebej polji meta-normativnosti ter spoznavne teorije.

The Dimensions of Moral Theory

The paper investigates the basic dimensions of moral theory. In the first part, we present the distinction between the moral questions of the first order and the second order and connect the latter with moral theory and metaethics. Then we distinguish between the five core dimensions of moral theory, namely moral ontology, moral semantics, moral epistemology, moral psychology, and moral phenomenology. Additionally, we present the basic positions within the framework of these dimensions. The article concludes with reflections on the best point of entry into the metaethical debate and on some new development trends in this field of moral philosophy.

Keywords: moral theory, metaethics, moral (i)realism, cognitivism, non-cognitivism, moral motivation, moral phenomenology

Literatura

- Brink, O. D. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge, MA: Cambridge University Press.
- Darwall, S. (2006). »How Should Ethics Relate to (the Rest of) Philosophy? Moore's Legacy«. V Horgan, T. in Timmons, M. (ur.), *Metaethics After Moore*, Oxford: Clarendon Press, str. 17–37.
- Horgan, T. in Timmons, M. (2005). »Moral Phenomenology and Moral Theory«. *Philosophical Issues*, 15(1), str. 56–77.
- Horgan, T. in Timmons, M. (ur.) (2006). *Metaethics after Moore*. Oxford: Clarendon Press.
- Horgan, T. in Timmons, M. (2008). »Prolegomena to a Future Phenomenology of Morals«. *Phenomenology and the Cognitive Sciences*, 7(1), str. 115–131.
- Kirchin, S. (2003). »Ethical Phenomenology and Metaethics«. *Ethical Theory and Moral Practice*, 6(3), str. 241–64.
- Mackie, J. L. (1977). *Ethics – Inventing Right and Wrong*. New York: Penguin. [V slovenščino je prevedeno prvo poglavje »Subjektivnost vrednot«, (prev. V. Strahovnik). V *Analiza*, 2008, 12(1/2), str. 69–92.]
- McNaughton, D. (1988). *Moral Vision: An Introduction to Ethics*. Oxford: Blackwell.
- Miller, A. (2003). *An Introduction to Contemporary Metaethics*. Cambridge: Polity Press.
- Moore, G. E. (2000) [1903]. *Principia Ethica* (prev. F. Klampfer). Ljubljana: Študentska založba. (Izvirno besedilo dostopno na: <http://fair-use.org/g-e-moore/principia-ethica>).
- Potrč, M. in Strahovnik, V. (2014). »Phenomenological Objectivity and Moral Theory«. *Synthesis philosophica*, 29(1), str. 159–173.
- Ross, W. D. (1939). *Foundations of Ethics*. Oxford: Clarendon Press.
- Sayre-McCord, G. (1988). »The Many Moral Realisms«. V Sayre-McCord, G. (ur.), *Essays on Moral Realism*, Itacha & London: Cornell University Press, str. 1–23.
- Shafer-Landau, R. (2005). *Moral Realism: A Defence*. Oxford: Oxford University Press.
- Strahovnik, V. (2007). »Moralna fenomenologija in fenomenološki argumenti«. *Analiza*, 11(4), str. 73–95.

Strahovnik, V. (2008). »Teorija zmote: moralni kognitivizem in irealizem«. *Analiza*, 12(1/2), str. 49–67.

Timmons, M. (1999). *Morality without Foundations*. New York in Oxford: Oxford University Press.

Timmons, M. (2002). *Moral Theory: An Introduction*. Lanham: Rowman and Littlefield Publishers.

Tim Koprivnik

Filozofska fakulteta Univerze v Mariboru

Zakaj nismo svobodni (moralno odgovorni) in kaj iz tega (ne) sledi

Sledeči članek predstavlja in kritično ocenjuje Pereboomov članek »Determinizem *al dentek*«, v katerem omenjeni dokazuje predvsem dve stvari: prvič, da je *trdi inkompatibilizem* – zaradi »argumenta iz 4 primerov« – najbolj prepričljivo teoretsko stališče o vprašanju svobode, determinizma in njune kompatibilnosti; in drugič, da ugovori zoper trdi inkompatibilizem, češ da onemogoča doseganje pomembnih stvari v življenju, spodbuja nemoralnost in neosebne odnose in/ali da je praktično nemogoč, niso prepričljivi.

Ključne besede: svoboda, moralna odgovornost, (in)determinizem, (in)kompatibilizem, Pereboom

0 Uvod

V pričujočem članku nameravamo predstaviti in kritično oceniti Pereboomov članek »Determinizem *al dentek*«. Njegov članek lahko v grobem razdelimo na dva dela. V 1. delu (v izvirniku je to od I. do IV. razdelka) Pereboom najprej s pomočjo t.i. »argumenta iz 4 primerov«¹ dokazuje, da so v zmoti *kompatibilisti*,² ki trdijo, da je determinizem resničen in da imamo svobodo, potrebno za moralno odgovornost, saj determinizem *je* kompatibilen s svobodo, potrebno za moralno odgovornost.³ Nato pa s pomočjo treh kratkih argumentov dokazuje, da so v zmoti

¹ Morda bi bilo bolj smiselno, če bi ga imenovali »argument iz manipulacije« ali »argument iz pomanjkanja nadzora«, saj dá takšno poimenovanje jasno vedeti, da je bistven problem, ki ga morajo nasprotniki argumenta rešiti, manipulacija oz. pomanjkanje nadzora (ki je posledica determiniranja/manipulacije).

² Lahko jih imenujemo tudi »mehki deterministi«.

³ Bistvenega pomena je, da se zavedamo, da tisti, ki zagovarjajo, da ljudje imamo moralno odgovornost, razumejo in uporabljajo izraz »moralno« v izključno *retributivnem* (*povračilnem*) smislu razumevanja morale, kar pomeni, da je njihova argumentacija grajanja, kaznovanja in hvaljenja – kar so glavni vidiki uporabe morale – sledeča: nekoga je potrebno kaznovati/grajati/hvaliti zato, ker si to *zasluži* oz. ker mora za svoje preteklo dejanja dobiti *povračilo*. Takšno moralo lahko zato imenujemo tudi »v-preteklost-usmerjena morala«. Kot bomo videli, bo Pereboom dokazoval, da obstaja tudi drugačna vrsta morale, in sicer »v-prihodnost-usmerjena«, ki pa *ne* zahteva (predpostavlja), da so oz. morajo biti ljudje moralno odgovorni.

tudi *libertarci*, ki menijo, da determinizem ni resničen, ampak je resničen nek indeterminizem (v nas samih in/ali v svetu)⁴, in da imamo svobodo, potrebno za moralno odgovornost, saj je indeterminizem kompatibilen s svobodo, potrebno za moralno odgovornost. V 2. delu (v izvirniku je to od V. do VII. razdelka) pa Pereboom odgovarja na ugovore zoper stališče, ki ga zagovarja sam, tj. *trdi inkompatibilizem*,⁵ ki trdi, da svoboda, ki jo potrebujemo za moralno odgovornost, ni kompatibilna z determinizmom (niti z indeterminizmom). Ugovori trdemu inkompatibilizmu dokazujejo, da ta onemogoča doseganje pomembnih stvari v življenju, spodbuja nemoralnost in uničuje medosebne odnose in/ali je praktično nemogoč.

1 Zakaj nismo svobodni (moralno odgovorni)

1.1 Nekaj uvodnih besed o »argumentu iz 4 primerov« in predstavitev kompatibilističnih pojmovanj svobode (v smislu, ki ga zahteva moralna odgovornost)

Pereboom prične svoj članek s podatkom, da je v zahodni filozofiji največ kompatibilistov, nekaj je libertarcev, najmanj pa trdih inkompatibilistov. Ker kakor Pereboom tudi sami menimo, da je trdi inkompatibilizem z vidika logične argumentacije (pa tudi empiričnih spoznanj) daleč najbolj prepričljivo teoretsko stališče o vprašanju svobode, determinizma in njune kompatibilnosti, ugovori zoper njega pa za lase privlečeni, in ker menimo, da se nasprotniki trdega inkompatibilizma ne zavedajo prave moči Pereboomovega *argumenta iz 4 primerov*, se nam zdi edino prav, da slednjega predstavimo, dodatno obrazložimo in dopolnimo s svojimi mislimi tudi sami. Preden pa to storimo, bomo še predstavili, kako kompatibilisti ta problematičen pojem »svobode« (v smislu, ki ga zahteva moralna odgovornost) sploh razumejo.

Najbolj razširjeno kompatibilistično pojmovanje svobode je tisto Davida Huma, po katerem se za svobodno dejanje, kakršno zahteva moralna odgovornost, šteje dejanje, ki ga posameznik pač *želi izvesti*. Recimo, če mož želi umoriti svojo ženo in jo potem tudi zares umori, je to storil svobodno (moralno odgovorno), zaradi česar ga je potrebno retributivno grajati in/ali kaznovati (npr. poslati v zapor).

⁴ Obstaja več vrst libertarcev, ki zagovarjajo različne tipe indeterminiranosti. O njih bo več govora v nadaljevanju članka.

⁵ Pereboom sicer v omenjenem članku svoje stališče imenuje »trdi determinizem«. To je zavajajoče, saj Pereboom pušča odprto možnost, ali je resničen determinizem ali indeterminizem. Še več, po njegovem mnenju sploh ni pomembno, ali so naša dejanja posledica determinističnih ali indeterminističnih vzročnih procesov, saj nimamo *nadzora* nad nobenimi. Pomanjkanje nadzora nad temi vzročnimi procesi pa je po njegovem mnenju – ki ga bo upravičil z argumentom iz 4 primerov – dovolj velik razlog za odvzem svobode, ki jo potrebujemo za moralno odgovornost (Pereboom, 2007: 185). Zato je njegovo stališče bolje imenovati »trdi inkompatibilizem« (kar v svojih kasnejših letih spozna tudi Pereboom sam).

Drugo pojmovanje svobode (v smislu, ki ga zahteva moralna odgovornost) zagovarja Harry Frankfurt, ki trdi, da lahko o svobodi v smislu moralne odgovornosti govorimo le takrat, *ko se delujoče želje skladajo z željami drugega reda*. Povedano drugače, svobodna oz. moralno odgovorna oseba je tista oseba, ki je (1) storila, kar je želela, in ki (2) si je želela tega, kar si je želela.⁶ Če torej mož želi umoriti svojo ženo, si obenem želi imeti željo umoriti svojo ženo in če ženo potem tudi zares umori, je svoboden (moralno odgovoren), zaradi česar ga je potrebno retributivno grajati in/ali kaznovati (npr. poslati v zapor).

Tretje pojmovanje svobode pa prihaja od Bernarda Gerta in Timothyja Duggana, ki trdita, da je svoboda, kakršno zahteva moralna odgovornost, *sposobnost hotenja oz. dovzetnost za razloge*. Po njunem je torej dejanje svobodno (v smislu, ki ga zahteva moralna odgovornost) takrat, *ko je rezultat »razumnega preudarka razlogov*, ki so v neki situaciji pomembni«, ki pa »mora biti takšen, da bi dejavnik vsaj v nekaterih alternativnih okoliščinah – tam, kjer obstajajo zadostni razlogi za drugačno ravnanje od dejanskega – bil za te razloge dovzeten in bi na osnovi istega mehanizma preudarjanja, ki je dejansko vodil do ravnanja, zares ravnal drugače« (Pereboom, 2007: 173). To na konkretnem primeru pomeni tole: mož je svoboden (moralno odgovoren), če si po tem, ko izve, da danes z ženo prejmeta obisk, premisli in se namesto tega, da ženo ubije danes, raje odloči, da jo bo umoril jutri (ko bo obiska že konec). Če pa bi mož kljub zavedanju, da danes z ženo prejmeta obisk, in kljub pretehtanju različnih drugih možnosti ostal pri prvotni odločitvi, da ženo umori danes, ne bi bil svoboden (moralno odgovoren).

Menimo, da ima Ted Honderich povsem prav, ko pravi, da kompatibilisti v vseh svojih pojmovanjih svobode (v smislu, ki ga zahteva moralna odgovornost) vedno znova pozabljajo na dejstvo, da moralna odgovornost *predpostavlja*, da je delujoči *izvor* dejanja, in da je že trmasto vztrajanje pri takšnem neustreznem pojmu moralne odgovornosti nekaj, kar šteje proti njim (Honderich, 1988). A prav ima tudi Pereboom, ki opozarja, da četudi je Honderichev očitek pravilen, na takšen način kompatibilistov nikdar ne bomo prepričali, da se motijo. Zato je sam mnenja, da je potreben močnejši argument, in sicer argument, ki ne bo zagovarjal drugačnih pojmovanj svobode (v smislu, ki ga zahteva moralna odgovornost) od pojmovanj kompatibilistov, a bo argumentativno še vedno pripeljal do sklepa, ki s kompatibilizmom ne bo združljiv (Pereboom, 2007: 174). Takšen argument naj bi bil po Pereboomu *argument iz 4 primerov*. Zdaj je napočil čas, da si ga ogledamo in pokažemo, zakaj ovrže kompatibilizem.

⁶ S Frankfurtovimi besedami: svobodna oz. moralno odgovorna oseba je oseba, ki je »storila, kar je hotela, da je to storila zato, ker je tako hotela in da je bila volja, ki jo je motivirala, njena volja, saj je bila to volja, ki jo je oseba hotela« (Frankfurt, 2007: 130).

1.2 Zavrnitev kompatibilizma – argument iz 4 primerov v akciji

Situacija, v kateri so zajeti vsi trije kompatibilistični pomeni svobode (ki ga zahteva moralna odgovornost), je takšna:

Denimo, da gospod Novak zaradi neke osebne koristi ubije gospo Pavšič. Vzrok umora so tako bile *želje, ki so v resnici njegove*, njegova želja, da bi ubil gospo Pavšič pa *se sklada z njegovimi željami drugega reda*. Poleg tega se njegove želje spreminjajo in nekatere tudi nastanejo zaradi razumnega premisleka relevantnih razlogov, njegov proces preudarjanja je *dovzeten za razloge*. Na primer, če bi vedel, da bodo posledice njegovega zločina zanj veliko hujše, kot je dejansko verjetno, gospe Pavšič ne bi ubil. (Pereboom, 2007: 174)⁷

Na vprašanje, »Če privzamemo, da je determinizem resničen, ali je smiselno trditi, da je gospod Novak odgovoren za svoje dejanje?«, bomo seveda dobili dva različna intuitivna odgovora – inkompatibilističnega in kompatibilističnega. Prvi se glasi takole:

Po determinističnem nazoru so tako želje prvega in drugega reda, kakor tudi za razloge dovzetni procesi, ki vodijo gospoda Novaka v zločin, *neogibni glede na svoje vzroke, ti vzroki pa so neogibni glede na svoje vzroke*. V oceni moralne odgovornosti za storjeni umor se vračamo nazaj po *deterministični verigi vzrokov*, ki so privedli do želja in razmišljanja gospoda Novaka, dokler ne pridemo do *dejavnikov, nad katerimi Novak nima nadzora* – do vzročnih dejavnikov, ki jih ni mogel ustvariti, spremeniti ali preprečiti. /.../ [In] za dejanja, ki so učinek determinističnega vzročnega procesa, ki vodi nazaj do dejavnikov, nad katerimi nima nadzora, dejavnik *ni* moralno odgovoren. (Pereboom, 2007: 174)

Drugi pa takole:

[M]oralna odgovornost *ne izgine*, ko se deterministični vzročni proces odvija v času nazaj. Gospod Novak je moralno odgovoren, *červno* veriga zadostnih vzrokov za zločin, ki ga je storil, sega daleč onkraj njegovega nadzora, celo v čas, ko še ni živel. Za umor je odgovoren, ker so vzrok dejanja *njegove želje* prvega reda, ker se te želje *skladajo z njegovimi željami drugega reda* in ker so vse nastale v kontekstu njegove *razumne ocene razlogov*. (Pereboom, 2007: 174)

Seveda se sedaj pojavi vprašanje, kako ugotoviti, kateri odgovor je ustrezen? Pereboomova metoda je izredno bistra: predstavimo 4 primere (med katerimi so eni zgolj hipotetični, drugi pa realni), ki (1) vključujejo primere determiniranosti z vzročnimi procesi, ki segajo nazaj do dejavnikov, nad katerimi gospod Novak – kateri služi kot prototip vseh ljudi – *nima* nadzora (ker je determini-

⁷ Poudarki v tem in vseh nadaljnjih citiranjih so naši.

ran/zmanipuliran), (2) vanje vključimo vsa tri kompatibilistična pojmovanja svobode, potrebna za moralno odgovornost, in (3) pogledamo, kakšen intuitiven⁸ odgovor dobimo sedaj. In ker uvidimo, da je intuicija velike večine izmed nas vsaj v enem izmed teh 4 primerov inkompatibilistična (gospod Novak *ni* svoboden v smislu, ki ga zahteva moralna odgovornost), in ker ne najdemo nobenega dobrega razloga, zakaj bi bili vsi ti primeri determiniranosti *načelno različni*, (4) ugotovimo, da so kompatibilisti v zmoti, ko mislijo, da svoboda, ki se zahteva za moralno odgovornost, je kompatibilna z determinizmom. Takšna je torej struktura *argumenta iz 4 primerov*. Zdaj pa k predstavitvi 1. primera:

Gospod Novak je običajno človeško bitje, le da *so ga ustvarili nevroznanstveniki*, ki lahko z njim *neposredno manipulirajo* s pomočjo tehnologije podobne radiu. Vzemimo, da ti nevroznanstveniki neposredno manipulirajo Novaka tako, da se ta loti procesa razmišljanja, ki *ustvarja in spreminja njegove želje*, ter da se njegova delujoča želja prvega reda, ubiti gospo Pavšič, *sklada* z njegovimi željami drugega reda. Nevroznanstveniki manipulirajo z njim tako, da med drugim pritisnejo na vrsto gumbov tik preden prične razmišljati o svojem položaju, s čimer povzročijo, da je *proces njegovega razmišljanja racionalno egoističen*. Njegov proces razmišljanja je *dovzeten za razloge*, saj bi bile njegove izbire drugačne v situacijah, kjer bi bili egoistični razlogi drugačni. Gospod Novak pa ne razmišlja in ne deluje proti svojemu značaju, kajti nevroznanstveniki običajno z njim manipulirajo tako, da je racionalno egoističen. (Pereboom, 2007: 175)

Kot vidimo, dejanje gospoda Novaka izpolni vsa tri kompatibilistična pojmovanja svobode, ki so potrebna za moralno odgovornost – tako Humovo, Frankfurtovo, kot tudi Gertovo in Dugganovo. Toda intuitivno bi velika večina ljudi dejala, da gospod Novak *ni* moralno odgovoren –, determiniran je namreč z dejanji nevroznanstvenikov, nad katerimi *nima* nadzora.

Kompatibilisti se seveda domisljijo načina, kako temu primeru ugovarjati. John Martin Fischer recimo temu primeru ugovarja tako, da pravi, da »je v primeru *neposredne* manipulacije možganov verjetno, da miselni proces, ki vodi do dejanja, *ni* dovzeten za razloge; medtem ko dejstvo, da je miselni proces vzročno determiniran, samo po sebi ne vpliva na to, ali je ta proces dovzeten za razloge ali ne« (1987, v Pereboom, 2007: 175). Kot njegov ugovor povzame Pereboom, gre za idejo, da je Frankfurtovo vrsto svobode sicer mogoče sprožiti na nevro-fiziološki način, da pa to *ne* velja za procese, dovzetne za razloge (2007: 175).

⁸ Morda bi morali tukaj izraz »intuitiven« nadomestiti z »razumsko intuitiven«, saj »intuicijo« načeloma razumemo kot *razumsko nepodprto* prepričanje. Intuicija pa po preučitvi teh 4 primerov ni razumsko nepodprta, saj smo jo dopolnili z vsaj dvema razumsko podprtima spoznanjema, namreč prvim, kaj sploh pomeni ne imeti nadzora, in drugim, v čem (če sploh čem) je razlika med načini determiniranostmi z dejavniki, nad katerimi nimaš nadzora. To omenjamo zato, ker nekateri kritiki argumentu iz 4 primerov *neupravičeno* očitajo, da se opira na intuicije ljudi, katerim ni moč zaupati, saj so razumsko nepodprte.

Fischerjev ugovor ne drži. Ne poda namreč niti enega empiričnega razloga, zakaj to ne bi bilo mogoče. Da se njemu kaj takšnega ne *zdi* verjetno, pač ni noben ugovor. A še več, če nismo nekakšni dualisti, ki verjamejo, da (nekateri) miselni procesi niso nevrofiziološki in da potekajo na neki višji, ne-nevrofiziološki ravni,⁹ potem bi se nam moralo zdeti neverjetno ravno nasprotno kot to, kar se *zdi* neverjetno Fischerju – neverjetno bi se nam moralo zdeti, da nekaterih miselnih procesov *ne* bi bilo mogoče zmanipulirati. Kakor pravi tudi Pereboom: »Vse dotlej, dokler ta proces zahteva le *fizikalno* izvedljive sposobnosti, ga lahko sproži dovolj dobro opremljena ekipa znanstvenikov« (2007: 175).

Drugi možen ugovor 1. primeru bi se lahko glasil, da je proces, ki vodi do dejanja, v prvem primeru resda dovzeten za razloge, vendar pa ga sproži neposredna manipulacija *tik pred samim dejanjem*, zaradi česar je primer zelo podoben *pranju možganov* (Pereboom, 2007: 175). Morda pa bi kdo tudi ugovarjal, da je dovzetnost za razloge v tem primeru »*preveč umetna*, saj bi lahko nevroznanstveniki z drugačno vrsto nadzora dosegli, da ne bi bil dovzeten za razloge« (str. 176).

Da sta obe možnosti zgornjega ugovora prav smešni oz. da se jima je zelo preprosto izogniti, lahko dokažemo kar z 2. primerom, ki gre takole:

Gospod Novak je običajno človeško bitje, le da so ga ustvarili nevroznanstveniki. Ti ga sicer ne morejo neposredno nadzorovati, vendar so ga *sprogramirali* tako, da je njegovo vedênje *racionalno egoistično*. To pomeni, da je v vsaki situaciji, podobni tej, v kateri se je zdaj znašel, vzročno determiniran, da sproži proces, ki je *dovzeten za razloge* in *poseduje takšne želje prvega in drugega reda*, ki vodijo do umora gospe Pavšič. (Pereboom, 2007: 176)

V tem primeru gre torej za manipuliranje oz. programiranje¹⁰, ki je bilo *udejanje-no precej časa nazaj*, s čimer se izognemo ugovoru iz *neposrednosti* manipuliranja. Sedaj pa se vprašamo: kakšna *načelna* razlika je med 1. in 2. primerom? V čem je razlika z vidika determiniranosti, če si bil zmanipuliran oz. sprogramiran 2 sekundi nazaj ali pa 30 let nazaj? Razlike seveda ni. Časovni interval med programiranjem in dejanjem je popolnoma *nepomemben* (Pereboom, 2007: 176). Če torej sprejmemo, da gospod Novak ni svoboden (v smislu, ki ga zahteva moralna odgovornost) v 1. primeru – kar velika večina bi sprejela –, potem ni svoboden (v smislu, ki ga zahteva moralna odgovornost) niti v 2. primeru. Kaj pa v 3. primeru? Ta gre takole:

⁹ Da ne bi smeli biti dualisti, pa dokazujejo nevroznanstvene ugotovitve druga za drugo.

¹⁰ Razlika med »manipulacijo« in »programiranostjo« je zgolj *pojmovna (jezikovna)* – za isto stvar imamo oz. uporabljamo dva različna pojma. Zmanipulirati nekaj namreč ne pomeni nič drugega kakor (pre)programirati (oz., s še drugimi besedami, (pre)oblikovati, (pre)določiti nekaj). In, obratno, (pre)programirati nekaj ne pomeni prav nič drugega kakor zmanipulirati nekaj.

Gospod Novak je običajno človeško bitje, le da ga je *stroga vzgoja* dôma in njegove skupnosti *determinirala*, da je *racionalno egoističen*. Vzgoja se je pričela v njegovi zgodnji dobi, preveč zgodaj, da bi lahko bil sposoben preprečiti ali spremeniti prakse, ki so determinirale njegov značaj. Gospod Novak je vzročno determiniran, da sproži ustrezen, *za razloge dovzeten* proces ter da *poseduje takšno organizacijo želja prvega in drugega reda*, ki vodijo do umora gospe Pavšič. (Pereboom, 2007: 176)

Ta primer je človeški psihologiji še najbolj domač. In tukaj je intuicija kompatibilistov takšna, da gospod Novak *je* svoboden (v smislu, potrebnem za moralno odgovornost). Vendar pa morajo kompatibilisti to svojo intuicijo razumno *upravičiti*. In upravičijo jo lahko le tako, da pokažejo na neko *načelno* razliko, ki obstaja med 2. in 3. primerom (ali med 1. in 3. primerom). Kaj hitro pa nam postane jasno, da takšne načelne razlike ni mogoče najti (tudi kompatibilisti še niso predlagali nobene). To pa zato, ker gre pri vseh treh primerih za *načelno* povsem *enako* determiniranost: gospod Novak je determiniran z dejavniki, nad katerimi *nima nadzora*. Kakor ni pomembno, kdaj si bil determiniran, tudi ni pomembno, kako, kje, zakaj si bil determiniran. Determiniranost je determiniranost, pa če je posledica (1) manipuliranja nevroznanstvenikov iz prejšnjega tedna, ker so si zaželeli malce eksperimentiranja na ljudeh, (2) naravne selekcije pred nekaj sto tisoči ali milijoni let, ker se evoluciji fučka zate, (3) vzgoje (kulture) pred 10 leti, ker so starši in širša družba želi oblikovati konformne in ubogljive posameznike, ali (4) česarkoli že drugega. Kakorkoli pogledamo, nadzora nad tem nimamo. Smo takšni, kakršni smo, ker nas je nekaj determiniralo (zmanipuliralo) na nek način. Zaključiti je treba, da gospod Novak *ni* svoboden (v smislu, ki ga zahteva moralna odgovornost) niti v 3. primeru. Kompatibilistom tako ostane le še ena možnost, in sicer v 4. primeru, ki gre takole:

*Fizikalni determinizem*¹¹ *je resničen*. Gospod Novak je *racionalno egoističen*, drugače pa je običajno človeško bitje, vzgojeno v normalnih okoliščinah. Umor gospe Pavšič je rezultat tega, da gospod Novak sam sproži proces presoje, ki je *dovzeten za razloge*, ter da *poseduje ustrezno organizacijo želja prvega in drugega reda*. (Pereboom, 2007: 177)

A kaj, ko je tudi 4. primer *načelno enak* prejšnjim trem – enako je posledica determinističnih vzročnih procesov, ki segajo nazaj do dejavnikov, nad katerimi gospod Novak *nima nadzora* (2007: 177).

¹¹ Fizikalni determinizem je stališče, ki trdi, da bi, če bi imeli popoln opis fizikalnih stanj v svetu v kateremkoli trenutku in popolno poznavanje fizikalnih zakonov, lahko napovedali prav vsako prihodnje fizikalno stanje stvari (Ginet, 1990: 92). A treba se je zavedati, da fizikalni determinizem *ne* predpostavlja, da so vsa mentalna stanja in procesi identični ali realizirani v fizikalnih procesih. Tako da reči, da je naša volja naravno vezana na nevrnalne procese, pomeni zgolj reči, da je to pač naravni zakon, da bosta naša volja in nevrnalni procesi vezani (Ginet, 1990: 92).

Edina razlika med tem primerom in prejšnjimi tremi je ta, da tukaj vzročna determiniranost (zmanipuliranost) zločina gospoda Novaka ni posledica delovanja drugih *ljudi* (ali nevroznanstvenikov ali staršev ali širše skupnosti), ampak *zakonov narave*. To bi sicer morda kdo lahko razumel kot dokaz *načelne* razlike med primeri determiniranosti in zato kot ugovor argumentu iz 4 primerov. Vendar pa to seveda ni *načelna*, temveč zgolj *pojmovna* razlika. Ljudi in zakone narave lahko zaobjamemo pod širši skupni pojem »nekaj (karkoli že), nad čimer nimamo nadzora« in s tem pojmovna razlika izgine. Ali, kakor to poanto izrazi Pereboom: »Toda, če bi prve tri primere popravili tako, da bi bila determiniranost posledica *spontano delujočega in brezumnega stroja*, to ne bi oslabilo intuicije, da gospod Novak ni moralno odgovoren« (2007: 177).

Pereboomov argument iz 4 primerov je tako zaključen: (1) zamislili smo si 4 primere, ki vključujejo primere determiniranosti z vzročnimi procesi, ki segajo nazaj do dejavnikov, nad katerimi gospod Novak – kateri služi kot prototip *vseh ljudi – nima* nadzora (ker je determiniran/zmanipuliran), (2) v njih vključili vsa tri kompatibilistična pojmovanja svobode, potrebna za moralno odgovornost, (3) videli, da je naša intuicija vsaj pri enem izmed teh 4 primerov inkompatibilistična, ter (4) dokazali (oz. vsaj ne-ne-dokazali), da so vsi ti primeri determiniranosti *načelno enaki*, iz česar pa končno sledi, da je kompatibilistična intuicija – svoboda, ki jo potrebujemo za moralno odgovornost, je kompatibilna z determinizmom – preprosto *zmotna oz. neprepričljiva*.

Vendar pa kompatibilistom ostane na voljo še ena možnost ugovarjanja (za le-to se lahko zahvalijo Harryju Frankfurtu). Trdijo lahko, da je oseba moralno odgovorna za svoja dejanja *tudi, če ni mogla ravnati drugače*. To potem pomeni, da Frankfurt in njegovi somišljeniki zagovarjajo, da gospod Novak, četudi ni svoboden, *je* moralno odgovoren v prav vseh 4 primerih. V svoj prav nas Frankfurt poskuša prepričati z naslednjim hipotetičnim primerom:

Gospa Scarlet resno razmišlja, ali naj umori polkovnika Mustarda. Istočasno si tudi nevroznanstvenik profesor Plum želi polkovnikove smrti. Skrbi ga, da si gospa Scarlet ne bi premislila in pustila Mustarda pri življenju. Zato je v njene možgane vsadil napravo, ki bi, če bi gospa Scarlet zaradi razlogov, ki nasprotujejo umoru, preveč omahovala, povzročila, da bi ga vseeno ubila. Toda gospa Scarlet se odloči za umor in ga brez omahovanja izvrši (1969, v Pereboom, 2007: 179).

Ni kaj, treba je priznati, da je omenjen primer izredno premeten. Kakor nas opozarja tudi Pereboom, izrecno namreč ne omenja prav nobenega determinizma (2007: 180). Nepozornega bralca lahko tako kaj hitro prepriča. Vendar pa je premetenost obenem tudi vse, kar ta primer premore. Povsem jasno je namreč, da vsak kompatibilističen poskus argumentiranja *mora* zajeti tezo determinizma. Ravno za tezo determinizma morajo namreč kompatibilisti dokazati, da je kompa-

tibilna s svobodo, potrebno za moralno odgovornost, oz. – če pojma svobode sploh ne uporabljamo – da je kompatibilna z moralno odgovornostjo.

Nenazadnje pa je lepo rekel že Thomas Nagel: »Zdi se *neracionalno* lastiti si zasluge ali kriviti se za stvari, nad katerimi oseba nima nobenega nadzora, ali za njihov vpliv na rezultate, nad katerimi ima oseba le delni nadzor« (2007: 228).

Sami torej ne vidimo drugega zaključka kakor tega, da je stališče kompatibilizma neprepričljivo in da ga ne bi smeli jemati resno, dokler ne ponudi korektne, dobro argumentirane rešitve za argument iz 4 primerov.

Sedaj pa je čas, da pogledamo, kako Pereboom opravi še z libertarci.

1.3 Zavrnitev libertarizma

Kot smo že omenili, libertarci trdijo, da determinizem sploh ne obstaja, temveč da obstaja nek indeterminizem (v nas samih in/ali v svetu) in da *imamo* svobodo (v smislu, potrebnem za moralno odgovornost), saj svoboda, potrebna za moralno odgovornost, in indeterminizem *sta* kompatibilna. Omenili smo tudi, da obstaja več vrst libertarcev. Prva vrsta libertarcev trdi, da so naša dejanja svobodna, ker jih (deloma) tvorijo *indeterministični* naravni dogodki. Lukrecij se je denimo skliceval na *naključni* odklon padajočih atomov (Lucretius, 1959). Danes se libertarci te vrste sklicujejo na spoznanja iz kvantne mehanike, ki naj bi dokazovala, da »[t]i odkloni in dejanja, ki so na njih utemeljena, so *naključni*, vsaj v pomenu, da *niso določeni z nobenim predhodnim stanjem fizikalnega sveta*« (Pereboom, 2007: 180). Vendar pa, kakor izpostavlja tudi Pereboom, četudi v svetu indeterminiranost zares obstaja, ta »sama po sebi [nikakor] ne more razložiti svobode, kot jo zahteva moralna odgovornost« (prav tam). Povedano na kratko, naključnost ti še ne prinese svobode, potrebne za moralno odgovornost, saj z naključnostjo ne dobimo prav nič več *nadzora* (str. 181).

Drugi libertarci (zaradi tega) raje zagovarjajo *alternativno teorijo vzročnosti-dejavnika*. Ta pravi, da svobodnega dejanja ne moremo (enostavno) razložiti z naključnimi dogodki, temveč z *dejavniki, ki so sposobni zavestnega povzročanja svojih dejanj*. Pereboom takšno stališče ovrže tako, da izpostavi, da ideja, da dejavnik (človek) poseduje neko posebno, »ne-humovsko vzročno moč zavestnega povzročanja dejanj, pri čemer povzročanje teh dejanj *ni* proizvedeno s procesi zunaj njegovega nadzora«, pač ni prepričljiva, saj (zaenkrat) nimamo prav nobenega (empiričnega) razloga za verjetje v to, »da takšne bitnosti zares obstajajo« (2007: 181). Ravno nasprotno: zdi se, da imamo dandanes zaradi spoznanj iz kognitivne znanosti mnogo razlogov za verjetje v to, da takšne bitnosti *ne* obstajajo.

Obstaja pa še tretja vrsta libertarcev, namreč tista »kantovska«. Immanuel Kant je namreč trdil, da nam že *hipotetična možnost*, »da ne-determinirani ne-fizikalni dejavniki vselej svobodno izberejo *prav tista* potencialna dejanja, katerih fizikalne sestavine so [potem tudi dejansko] vzročno determinirane«, daje vse, kar potrebu-

jemo za *racionalno* vero v indeterministično svobodo (2011, v Pereboom, 2007: 182). Pereboom tej vrsti libertarcev odgovarja podobno kot prejšnji, namreč da »je sovpadanje, ki ga tako stališče implicira, *preveč neverjetno*, da bi vanj lahko verjeli«¹² (Pereboom, 2007: 182). To, kar trdi Kant, je zares nerazumno in preveč za lase privlečeno. Od kdaj pa namreč zgolj *hipotetična* možnost daje *racionalen* razlog za sprejem nekega stališča? Kolikor je nam znano, poskušamo kot razumni misleci težiti k temu, da sprejemamo tista prepričanja, ki so *čim manj* hipotetična in *čim bolj* empirično in logično prepričljiva, ne pa nasprotno. To, kar omenja Kant, je pravzaprav definicija za *iracionalno*, ne pa »racionalno vero« v indeterministično svobodo.

A obstaja še en problem z neverjetnim sovpadanjem, namreč ta, da so po današnji fiziki »*predhodne verjetnosti* fizikalnih sestavin človeških dejanj ... *fiksne*«, pa naj bo »determinizem resničen ali pa naj velja kvantna nedoločenost« (Pereboom, 2007: 184). Če torej v fiziki (in posledično v nevrofiziologiji in psihologiji) ne bo prišlo do novih zares revolucionarnih spoznanj, pri katerih se bo dopuščalo, da fizikalni dogodki *sploh nimajo* nikakršnih determiniranih predhodnih verjetnosti, problem neverjetnega sovpadanja ne bo rešljiv (str. 184–185).

Tako lahko zaključimo, da na temeljno inkompatibilistično intuicijo nimajo razumnega protiargumenta ne kompatibilisti ne libertarci. Preprosto je inkompatibilistična intuicija toliko logično in empirično podprta (upravičena), da trdega inkompatibilizma ne bo mogoče ovreči po takšni poti, da bi ga zamenjali s kakšnim boljšim alternativnim.

A obstaja še druga pot, kako (poskusiti) ovreči trdi inkompatibilizem, namreč tako, da ga obtožimo, da (1) onemogoča doseganje pomembnih stvari v življenju, (2) spodbuja nemoralnost in neosebne odnose med ljudmi in/ali da (3) je vsaj praktično nemogoč (in tako teoretsko iracionalen). Vsi misleci se namreč prikrito zavedamo, da ne obstaja lažja, čeprav nekorektna, pot za ovržbo nekega teoretskega stališča kakor ta, da ga prikažemo v neprivlačni luči. Kot rečeno, Pereboom si je v 2. delu svojega članka zadal nalogo, da skuša takšne ugovore čim bolje ovreči. Čas je, da pogledamo, kako dobro mu je to uspelo (in kako mu lahko pri tej nalogi pomagamo še drugi).

¹² Zato ta problem mnogi imenujejo »problem neverjetnega sovpadanja«.

2 Kaj (ne) sledi iz tega, da nismo svobodni (moralno odgovorni)

2.1 Trdi inkompatibilizem in ugovor iz nezmožnosti doseganja pomembnih stvari v življenju

Prvi ugovor zoper trdi inkompatibilizem izpostavlja, da »v primeru njegove resničnosti *ne bi imeli nobenega razloga, da bi si za karkoli prizadevali*. Ne bi se trudili izboljšati naših življenj ali družbe, ker naše presoje in izbire *ničesar ne bi spremenile*« (Pereboom, 2007: 185).

Takšen ugovor je zgrešen. Takšen ugovor bi namreč bil utemeljen le, če bi šlo za *fatalizem*, ne pa za *determinizem*. Med enim in drugim namreč obstaja ogromna razlika. Če bi živeli v svetu, v katerem bi bili enako pametni, če bi se učili ali ne, enako zdravi, če bi jedli zdravo ali nezdravo, enako cenjeni in priljubljeni, če bi se do drugih ljudi obnašali prijazno ali nesramno, potem bi zares držalo, da ne bi imeli nobenega razloga, da bi si za karkoli prizadevali. Vendar *ne* živimo v takšnem fatalističnem svetu. V svetu, v katerem živimo, so posledice zelo drugačne, če se učimo ali ne, če jemo zdravo ali nezdravo, če smo do drugih ljudi prijazni ali nesramni. V našem (delno) determinističnem svetu torej naše presoje in izbire še kako spreminjajo potek življenja. Kot pravi Pereboom, »tisto, kar se zgodi, je še vedno lahko povzročeno z *našimi* presojami in izbirami« (2007: 186). Zato je tudi še kako pomembno, kakšne presoje in izbire imamo – *četudi* le-te niso svobodne (vsaj ne v smislu, ki ga zahteva moralna odgovornost). Prav tako lahko naše presoje in izbire – vsaj do neke mere in z različnimi sredstvi, pri čemer so nekatera učinkovitejša od drugih – spreminjamo. Ne sicer svobodno (v smislu, ki ga zahteva moralna odgovornost), a vendarle. Nič od tega determinističen svet ne onemogoča.

Drugi ugovor zoper trdi inkompatibilizem izpostavlja, da imamo ljudje močan občutek, da imamo moč izbire, tj. moč, da lahko stvari naredimo drugače, kakor smo jih dejansko naredili (Pereboom, 2007: 186). Povedano na kratko, *zdi* se nam, da smo svobodni.

Jasno je, kaj bo na to odgovoril trdi inkompatibilist: drži, vsi imamo (občasno) takšen občutek svobode, vendar pa je takšen občutek *iluzija* (Pereboom, 2007: 185).¹³ Mnogi bi dejali, da je takšen odgovor zgolj »izogibanje *pravemu* odgovoru« (ta ugovor lahko pogosto slišimo). Vendar pa to ne drži, saj je omenjen odgovor podprt s precejšnjo zbirko spoznanj iz nevrozanosti, kognitivne psihologije in vedenjske ekonomije, ki vse znova in znova ugotavljajo, kako zelo in glede koliko stvari nas naša introspekcija (naš intuitivni občutek) goljufa – nas zavaja, nam laže.

¹³ To trdijo tudi mnogi misleci, ki se nimajo za trde inkompatibiliste.

A Kant je podal še en razlog, zakaj se tega, četudi lažnega, občutka svobode ne bi smeli znebiti. Sam je namreč menil, da je občutek, da imamo več možnih izbir oz. da je vzročno možna več kot ena izbira dejanja, sploh *predpogoj* za naše presojanje (Kant, 2008, v Pereboom, 2007: 186).

Trdi inkompatibilist ima na ta ugovor dva možna odgovora. Prvi priznava, da med presojo v trenutku izbire dejansko napačno ali neupravičeno, torej *iracionalno*, *teoretsko* predpostavljamo, da imamo na voljo več kot en način ravnanja, a obenem izpostavlja, da pa je privzem te kognitivne drže kljub temu legitimen, saj je racionalen na praktični ravni, torej je *praktično racionalen* – zaradi te kognitivne drže namreč uspemo priti do neke odločitve. Povedano drugače, po tem stališču gre za nek neizogiben konflikt med teoretsko racionalnostjo (sprejeti resnico, da nimamo več možnih izbir) in praktično racionalnostjo (nadaljevati s presojanjem), ki pa ga je edino smiselno odpraviti tako, da sprejmemo praktično racionalnost in, posledično, teoretsko iracionalnost (Pereboom, 2007: 186).

Sami menimo, da je že ta odgovor sprejemljiv. To pa zato, ker s priznanjem takšne »teoretske iracionalnosti« trdega inkompatibilizma samega po sebi ne ogrozimo v prav nobenem smislu. Zakaj ne, bomo poskusili pojasniti v nadaljevanju.

Za Perebooma pa je teoretska iracionalnost vseeno preveč »moteča«, zaradi česar podaja še drugi odgovor, ki pa »ne zahteva prekoračitve zakonov teoretične racionalnosti«, saj lahko *zanika* nujnost, da moramo v trenutku presoje predpostaviti, da je vzročno mogoča več kakor ena izbira, in namesto tega trdi, da se mora predpostaviti zgolj to, da v trenutku presoje *izida* le-te *ne poznamo že vnaprej* (Pereboom, 2007: 187).

Odgovor je povsem korekten. Kako lahko namreč zares vemo, da je ravno občutek večih možnih izbir predpogoj za naša presojanja, ne pa morda občutek vnaprejšnjega nepoznavanja izida presoja? Še več, porodi se nam lahko vprašanje, zakaj bi sploh morali pred presojo karkoli predpostavljati? Empiričnih razlogov za kaj takega (še) nimamo. Zato bi lahko rekli *à la* Wittgenstein, da je ta Kantov ugovor zgolj *jezikoven* (in zato psevdo), ne pa realen problem.

2.2 Trdi inkompatibilizem in ugovor iz nemoralnosti

Tretji ugovor zoper trdi inkompatibilizem izpostavlja, da slednji spodkopava »[z]elo pomembn[o] značilnost našega vsakdanjega pojmovanja moralnosti«, namreč »prepričanje, da si oseba *zasluži* čast in hvalo, kadar namenoma izvrši moralno zgledno dejanje, oziroma *grajo*, ko namenoma izvrši moralno neupravičeno dejanje«, in da, če ne druga, je »vsaj bolje misliti in ravnati, *kakor da si* [oseba] zasluži [čast, hvalo in grajo], saj je takšno razmišljanje in ravnanje sijajna *metoda za promoviranje* moralne reforme in izobrazbe« (Pereboom, 2007: 187).

Ta ugovor je izredno zanimiv, saj četudi trdemu inkompatibilizmu ugovarja, da onemogoča promoviranje moralne reforme in izobrazbe, nekako prikrito priznava,

da je kompatibilizem tisti, ki je teoretsko iracionalen (je neprepričljiv oz. dokazano napačen), ne pa trdi inkompatibilizem. To je zanimivo zato, ker kompatibilisti preko prejšnjega Kantovega ugovora in ugovora iz psihološke teze trdim inkompatibilistom teoretsko iracionalnost štejejo kot problem, sami pa jo na tem mestu kar pometejo pod preprogo, češ da ni bistvena. O intelektualni poštenosti takšnega početja presodite sami.

A ugovarjanje, češ »da je mišljenje in ravnanje, kot da si ljudje včasih zaslužijo grajo, pogosto *nujno*, če hočemo učinkovito promovirati moralno reformo in izobrazbo«, je že v osnovi neprepričljivo, saj je možno tudi drugačno mišljenje in ravnanje za promoviranje moralne reforme in izobrazbe, ki se namesto na grajo sklicuje na moralna *svarila* in *spodbude*. V tem smislu bi torej moralno reformo in izobrazbo promovirali tako, da bi recimo prestopniku »razložili, da je to, kar je napravil, napačno in ga nato *spodbudili*, naj v prihodnje podobna dejanja opusti«. Po tem mišljenju tako »prestopnika ni potrebno imeti za moralno odgovornega za njegova dejanja, temveč raje *dovzetnega za moralno opominjanje in spodbujanje*« (Pereboom, 2007: 188).

A kritik že ima pripravljen naslednji ugovor: »Kaj pa tisti, ki redno in namenoma ravna napak in zavrača pošteno delovanje? Ali ni res, da mu trdi [inkompatibilist] ne more veliko povedati? Trdi [inkompatibilist] lahko *samo* opominja, zagovornik moralne odgovornosti pa lahko tudi graja [in kaznuje]« (Pereboom, 2007: 188).

Na tem mestu Pereboom predstavi dva možna odgovora. Prvega recimo zagovarja Ted Honderich, ki pravi, da bi od v-preteklost-usmerjenega grajanja in kaznovanja zares morali odstopiti, da pa to še ne pomeni, da se nad nemoralnimi ravnanji ljudi ne bi smeli *zgražati*. Kot pravi on sam: »Nič me ne ovira, da se zgražam nad željami in nameni nezvestega moža, ki vnaprej sluti razvezo, oziroma, pomembneje, da se zgražam nad njim, možem, katerega osebnost in značaj sta skladna s temi željami in nameni ter jih podpirata« (Honderich, 1988, v Pereboom, 2007: 188).

Odgovor neko smiselno poanto ima, saj nakazuje, da vendarle obstaja neka razlika med grajanjem *osebe*, ker si pač to zasluži, in zgražanjem nad *dejanjem* osebe z namenom *spodbujanja*, da v prihodnje podobna dejanja opusti (Pereboom, 2007: 188). Recimo, nad možem, ki pretepa svojo ženo, se lahko zgražamo zato, da bi ga pripravili do tega, da bi s tem svojim ravnanjem v prihodnosti prenehal. Nad njegovim ravnanjem se torej ne zgražamo zato, ker si on to zasluži oz. ker je pravično, da se nad njim zgražamo, temveč zato, da bi njegova dejanja – in ja, posledično mogoče tako tudi njega samega – *preoblikovali na bolje*, tj. tako, da v prihodnosti žene ne bi več pretepal. Je pa res, kar nas opozarja tudi Pereboom, da podatkov o tem, ali bi z zgražanjem, za razliko od zgolj skrbi in žalosti, lahko kako »prispevali k dobrem«, (še) nimamo, zaradi česar nek dvom o tem vendarle obstaja (prav tam: 190).

Kaj pa drugi odgovor? Ta pa priznava, da trdi inkompatibilist, za razliko od kompatibilista, moralnih prestopnikov¹⁴ zares ne more grajati, vendar pa obenem poudarja, da mora kompatibilist še zmeraj *empirično* dokazati, da ima grajanje kakršnokoli »praktično prednost« napram zgolj opominjanju (Pereboom, 2007: 188). In ker tega kompatibilist še ni storil, njegov ugovor ni prepričljiv.

A pravzaprav je ugovor slab že zato, ker *napačno* trdi, da lahko trdi inkompatibilist *zgolj* opominja. Trdi inkompatibilist namreč lahko *tudi* kaznuje (le argumentacijo kaznovanja zastavi drugače kot kompatibilist). In kako lahko kaznuje? Trdi inkompatibilist se ravna po dveh v-prihodnost-usmerjenih moralnih načelih: (1) *načelu posameznikovega preoblikovanja na bolje* – pri katerem se vpraša, »Kako prestopnika najučinkoviteje *preoblikovati* na bolje, torej tako, da ne bo več (tako hud) prestopnik?«¹⁵ – in (2) *načelu družbene varnosti* – pri katerem se vpraša, »Kako pred prestopnikom obvarovati ostale in/ali kako ostalim preprečiti, da tudi oni ne bi postali prestopniki?«¹⁶. Teži k temu, da bi zadovoljil obe načeli, saj bi s tem prišli do *dobrih posledic* tako za prestopnika samega kot tudi za druge (za družbo). Ker pa živi v nepopolnem, omejenem svetu, se mora velikokrat sprijazniti že z zadovoljtvijo enega izmed obeh načel.

A kritik nam že postavi novo vprašanje: kaj pa je s hvalo – ali ima kakšen smisel, da nekoga hvalimo za dejanja, čeprav je »determiniran, da iz pravih razlogov pravilno ravna«? (Wolf, 1990, v Pereboom, 2007: 190)

Odgovor je seveda ta, da ga ima. Hvale namreč ne rabimo nujno utemeljiti na zasluznosti, ampak jo lahko utemeljimo tudi na dobrih posledicah, ki jih le-ta prinese za seboj – hvala namreč ljudi pripravi do tega, da so ponosni na svoje ravnanje (in tako zadovoljni sami s sabo), prav tako pa jih tudi motivira, da v prihodnosti ravna jo dobro (Pereboom, 2007: 191). Torej, četudi trdi inkompatibilisti hvale ne moremo teoretsko upravičiti iz v-preteklost-usmerjenega vidika, pa jo lahko iz v-prihodnost-usmerjenega vidika.

¹⁴ »Moralni prestopnik« je posameznik, ki s svojim ravnanjem škoduje drugim ljudem.

¹⁵ Tukaj se je seveda potrebno zavedati, da vsi prestopniki nimajo enako dovzetnih možganov za »moralni kompas« – enim ga bo lažje prisvojiti, drugim težje; eni ga bodo prisvojili že z opozorilom, denarno kaznijo, z družbeno koristnim delom, z rehabilitacijami ipd., drugi pa zgolj, če jim bomo direktno preoblikovali njihov genetski material ali njihove možgane. O tem glej Eagleman (2011: 151–192).

¹⁶ In zakaj je to načelo moralno korektno? Ker če je moralno, da recimo nosilce hudih prenosljivih bolezni postavimo v karanteno samo zato, da pred njimi zaščitimo družbo (ostale ljudi), je potemtakem enako moralno tudi to, da v imenu zaščite družbe izoliramo nevarne zločince. In če je moralno, da v karanteno postavimo zločince, je moralno tudi to, da ljudem vnaprej povemo, da jih bomo izolirali, če bodo zagrešili zločin (Schoeman, 1979: 30; Pereboom, 2001: 174). Bi pa si bilo smiselno prizadevati, da bi ta »karantena« oz., z drugimi besedami, »zapor« bil razumljen drugače, kakor ga razumemo danes. Obstaja namreč ogromno študij, ki dokazujejo, da pošiljanje ljudi v zapor, kakor ga razumemo danes, zadeve tako za zapornike kot za druge le še poslabša. O tem glej Leach (2014), Greenstreet (b.d.), Rodham (2012), Newitz (2015), Dewar Gleissner (2011) itd.

Da tako zaključimo: čeprav drži, da trdi inkompatibilist raje opominja, kakor graja in kaznuje, in čeprav drži, da ne more grajati, kaznovati in hvaliti iz *v-preteklost-usmerjenega* vidika, pa še to ne pomeni, da ne more kaznovati in hvaliti (pa se tudi zgražati, v kolikor ima to praktično korist) iz kakšnega drugega, namreč iz *v-prihodnost-usmerjenega* vidika. Edina stvar, ki jo torej mora trdi inkompatibilist storiti, je le to, da eno argumentacijo grajanja, kaznovanja in hvaljenja (grajam/kaznujem/hvalim te, *ker si zaslužiš*) zamenja za drugo (zgražam/ kaznujem/hvalim *zaradi dobrih posledic*, ki jih le-to prinese za seboj). A ne samo, da je to *možno* storiti, temveč je to tudi edino *smiselno* storiti. V-prihodnost-usmerjena morala je namreč praktično *učinkovitejša* od v-preteklost-usmerjene, saj se pri slednji argumentacija moralnosti nanaša izključno na povračila za pretekla dejanja (za posledice pri vsem tem pa ji je popolnoma vseeno), medtem ko se pri prvi nanaša izključno na dobre posledice, skratka na učinkovitost.

2.3 Trdi inkompatibilizem, ugovor iz psihološke teze in ugovor iz neosebnih odnosov

Končno nam ostaneta še ugovora, ki poskušata dokazati, da četudi v-prihodnost-usmerjena morala je teoretsko smiselna in upravičena morala, pa nekateri povratni odzivi, ki jih spodbuja in implicira, zmeraj *niso praktično mogoči* (ne moremo zmeraj ravnati tako, kakor nam teoretsko veleva trdi inkompatibilizem) ali pa so vsaj v praksi *nevarni* (*uničujejo medosebne in spodbujajo neosebne odnose*). Začnimo z ugovorom, da povratni odzivi, ki jih takšna morala spodbuja, niso praktično mogoči.

Ta ugovor, ki ga nekateri imenujejo »psihološka teza« in katerega sta najbolj zavzeto zagovarjala David Hume (2008) in Peter Frederick Strawson (2007), pravi, da »splošno prepričanje v determinizem ali v katerokoli podobno abstraktno stališče *na povratne odzive ne more vplivati*«, zaradi česar bi bil vsak poskus spreminjanja ali izkoreninjenja odzivov zaradi determinističnega prepričanja ... jalov« (Pereboom, 2007: 193).

Pereboom temu ugovoru odgovarja s primerom o Robertu Harrisu,¹⁷ ki je leta 1978 v Kaliforniji na krut način umoril dva mladoletna fanta. Seveda je naš takojšnji povratni odziv po tem prebranem stavku gnus in misel, kako brezsrčen in zloben je ta možakar. A ko beremo dalje in dodatno spoznamo, kako so Harrisa v otroštvu zlorabljali (kar ga je definitivno pomembno determiniralo), se naš povratni odziv kaj hitro ublaži (Watson, 1987). Še več, možakar se nam zelo verjetno tudi zasmili (Pereboom, 2007: 194).¹⁸ To pa pomeni, da teoretsko prepričanje v trdi inkompatibilizem *do neke mere lahko* povzroči preoblikovanje naših pre-

¹⁷ Kot prvi ga je sicer izpostavil že Gary Watson (1987).

¹⁸ Presodite pri sebi, če ste tudi vi že kdaj ob dodatnih informacijah – ki v takšnih primerih niso nič drugega, kakor informacije o determiniranosti ravnanja nekoga – spremenili svoj prvotni povratni odziv. Mislimo, da bo vaš odgovor pritrdilen.

vladujočih povratnih odzivov (prevladujoči povratni odzivi so jeza, zamera, odpuščanje, hvaležnost) v t.i. »vzporedne odzive« (zavračanje jeze, graje in zamer, izogibanje hvaležnosti).

Se pa je treba zavedati, da je ta odgovor le delno prepričljiv – kar se zaveda tudi Pereboom –, saj povratnih odzivov vendarle ni mogoče *zmeraj* preoblikovati in da je tako naša vrojena »psihologija« vsaj v nekaterih situacijah zares *premočna*. Recimo, »[č]e bi kdo okrutno umoril vašo družino, bi se vam, psihološko gledano, bržčas zdelo nemogoče kadarkoli izkoreniniti občutje intenzivne jeze in besa do morilca« (Pereboom, 2007: 194).

Vendar pa Pereboom kljub zgornjemu priznanju vztraja pri misli, da je ugovor z vzporednimi odzivi še vedno dovolj dober, da lahko ovrže ugovor iz psihološke teze (2007: 194). Še več, na neki točki celo pravi, da posameznik z vzporednimi odzivi, ki bi »zavračal jezo, grajo in zamero, če bi doživel krivico ..., se za povzročeno krivico sploh ne bi zmenil ...[in] se hvaležnosti izogibal«, realno *je* mogoč (str. 198).

Tu gre Pereboom po našem mnenju predaleč. Takšnega posameznika si je namreč izredno težko predstavljati. V kolikor želi biti Pereboom prepričljiv, ga nam bo moral pokazati. Predvidevamo pa, da bi na podlagi samorefleksije večina trdih inkompatibilistov pošteno priznala, da niso takšni posamezniki. Tako se nam zdi, da je imel Hume s slavnim rekom »razum je suženj strasti« vseeno bolj prav, kakor mu je pripravljen priznati Pereboom.

Zakaj Pereboom *precenjuje* takšno zmožnost preoblikovanja prevladujočih povratnih odzivov, je seveda preprosto za razumeti: zato, ker ne želi sprejeti niti kančka teoretske iracionalnosti. Po njegovem mnenju (in mnenju nekaterih kompatibilistov) namreč sprejem tega, da ljudje ne moremo *zmeraj* ravnati tako, kakor nam teoretsko velewa trdi inkompatibilizem (kar pomeni vsaj delni sprejem psihološke teze), vodi trde inkompatibiliste v *teoretsko iracionalnost*, kar pa je problem (Pereboom, 2007: 196).

S tem njegovim mnenjem (in mnenjem kompatibilistov) se ne strinjamo, saj nikakor ne razumemo, zakaj bi naj (občasna) *praktična* nezmožnost preoblikovanja prevladujočih povratnih odzivov v vzporedne odzive sploh trdi inkompatibilizem lahko *teoretsko* ovrгла. Sami menimo, da je (občasna) nezmožnost preoblikovanja prevladujočih povratnih odzivov v vzporedne odzive za trdi inkompatibilizem prav toliko problematična kot je zanj problematična psihološko vrojena človeška intuicija, da se ljudje odločamo svobodno, skratka *nična*! Trditi, da lahko naše intuitivno, vrojeno *mišljenje* in *ravnanje* povesta, katero stališče je *teoretsko* pravilno, se zdi prav smešno! Povedano drugače, to, da ne moremo (*zmeraj*) misliti in ravnati tako, kakor nam velewa neko teoretsko stališče, stališča samega po sebi še *ne* ovrže. Če bi namreč to trdili, potem bi morali tudi trditi, da je napačnih vsaj 80 % vseh znanstvenih spoznanj – približno toliko jih je namreč neintuitivnih. Torej, četudi nam naša psihologija v praksi ne dopušča, da bi se dojeli takšne, kakršni

smo, namreč determinirani, temveč nam raje daje iluzorni občutek, da smo svobodni, in četudi nam naša psihologija ne dopušča, da bi se v praksi zmeraj odzvali tako, kakor nam velewa trdi inkompatibilizem, to trdega inkompatibilizma samega po sebi še ne ovrže.

A ne samo, da ne razumemo tega, zakaj bi bila teoretska iracionalnost problem za trdi inkompatibilizem, temveč ne razumemo niti tega, zakaj bi sploh morali trditi, da gre za kakršnokoli teoretsko iracionalnost. Ali namreč ni mogoče povratnega odziva, recimo, jeze zelo preprosto teoretsko *racionalizirati* tudi iz vidika v-prihodnost-usmerjene morale? Da pojasnimo na primeru: če naš prijatelj zmeraj zamuja, ga z našim jeznim odzivom lahko pripravimo do tega, da se bo naučil, da nas njegovo zamujanje zelo moti in, tudi, da se bo v prihodnje naučil (ali se vsaj zelo trudil), da ne bo več zamujal. Skratka, z jeznim odzivom ga lahko preoblikujemo na bolje – tj. tako, da ne bo več zamujal. In nobenega razloga ne vidimo, zakaj si iste v-prihodnost-usmerjene argumentacije ni mogoče (vsaj v nekaterih situacijah) zamisliti tudi za druge prevladujoče povratne odzive, kot so hvaležnost, zamera in odpuščanje.

Zaradi tega pa po našem mnenju tudi P. F. Strawsonov ugovor, ki pravi, da bi bili prepuščeni nevarnemu »neosebni odnosu«¹⁹ med ljudmi, če bi sprejeli trdi inkompatibilizem in zavrgli klasične povratne odzive (kar so po njegovem zamera, jeza, odpuščanje, hvaležnost in zrela ljubezen), ne more biti prepričljiv. Kot smo namreč dokazovali, klasičnih povratnih odzivov sploh ni potrebno (zmeraj) zavrniti.

A obstajata še druga dva odgovora na ta zgornji ugovor P. F. Strawsona. Prvega zagovarjata Honderich in Pereboom, ki izpostavljata, da ima Strawson sicer prav, ko pravi, da bi takšna neosebnost odnosov uničila medosebne odnose, da pa se moti, ko misli, da bi bili takšni neosebni – tj. objektivni, brezkrvni in upraviteljski – odnosi do drugih posledica sprejetja trdega inkompatibilizma oz. trdega determinizma²⁰ (Honderich, 1988: 532–533; Pereboom, 2007: 195).

Takšen odgovor je povsem korekten, saj P. F. Strawson dejansko *ne* uspe podati prav nobenega *empiričnega* in/ali *teoretskega* dokaza v prid temu, da nas sprejem trdega inkompatibilizma vodi v takšen neoseben pogled na druge ljudi, kakor pra-

¹⁹ Po P. F. Strawsonu »[p]rivzeti neosebni odnos ... pomeni videti drugo človeško bitje kot stvar, nemara stvar socialne politike; kot predmet, ki v najširšem smislu zahteva nek postopek. Se pravi kot nekaj, kar je na vsak način potrebno, morda celo preventivno, upoštevati, presoditi ali potrebuje vodstvo, obravnavo, zdravljenje ali vzgojo. Morda gre za nekaj, česar se je potrebno izogniti, čeprav to ni ravno značilno za primere neosebnega odnosa. Neosebni odnos je lahko čustveno obarvan na različne načine, ne pa na vse možne načine. Vključuje lahko odpor ali strah, sočutje in celo ljubezen, vendar ne vsakršne ljubezni. Nikakor pa ne more vključevati niza povratnih odzivov in čustev, ki izhajajo iz običajnega sodelovanja oziroma udeležbe v medčloveških vezeh. Tako ne more vključevati zamere, hvaležnosti, odpuščanja, jeze ali tiste vrste ljubezni, za katero dve odrasli osebi pravita, da jo občutita druga do druge« (P. F. Strawson, 2007: 98).

²⁰ Honderich se namreč šteje za trdega determinista.

vi sam. Njegov strah pred neosebniimi odnosi je tako (vsaj zaenkrat) neupravičen in nerazumen.

Drugi možen odgovor pa podaja Galen Strawson (sin P. F. Strawsona), ki pa sicer sprejema očitek, da trdi inkompatibilizem vodi do neosebnih odnosov, *ne* sprejema pa, da je to za odnose in moralo kakorkoli nevarno in slabo. Kot dokaz za svoje stališče podaja budistične menihe, ki domnevno »ne gledajo nase in na druge kot na samodoločujoče izvore dejanj in privzamejo do sebe in drugih objektivni odnos«, kar pa ne pomeni, da se odpovedo sočutju in usmiljenju« (G. Strawson, 2007: 221).

Tudi ta odgovor se zdi enako korekten kot zgornji. In to je pravzaprav vse, kar lahko tukaj še rečemo.

Sklep 2. dela je ta, da je prav vsak ugovor zoper trdi inkompatibilizem *neprepričljiv*, saj ali zgreši na celi črti ali ni empirično in logično podprt ali pa preprosto temelji na nekem iracionalnem strahu. Tako menimo, da bodo morali kritiki trdega inkompatibilizma narediti nekaj bistveno več, da bodo prepričljivo ugovarjali trdemu inkompatibilizmu.

3 Sklep

V pričujočem članku smo predstavili in kritično ocenili Pereboomov članek z naslovom »Determinizem *al dente*«. Kar se tiče 1. dela članka, smo prišli do zaključka, da je Pereboomova argumentacija pravzaprav brezhibna, saj je, kot prvo, podal korekten argument, ki lepo kaže, da kompatibilistična pojmovanja svobode (v smislu, ki ga zahteva moralna odgovornost) *niso* kompatibilna z determinizmom – nobena od njih namreč ne uspe zadostiti kriteriju *nadzora* –, iz česar pa sledi, da je kompatibilizem neprepričljivo stališče; kot drugo pa predstavil še tri krajše argumentacije, s katerimi je pokazal, da so neprepričljiva tudi libertaristična stališča, saj so ali nezadostna (naključnost še ne vodi do svobode) ali pa preveč za lase privlečena (neverjetna sovpadanja in neverjetne nadnaravne vzročne moči dejavnikov), da bi jih lahko jemali resno.

Kako pa je z njegovim 2. delom, kjer zavrača ugovore, da trdi inkompatibilizem onemogoča doseganje pomembnih stvari v življenju, spodbuja nemoralnost in neosebne odnose in/ali da je vsaj praktično nemogoč (in tako teoretsko iracionalen)? Kar se tiče ugovora iz nezmožnosti doseganja pomembnih stvari v življenju, kot je recimo prizadevanje za neke cilje, Pereboom lepo pokaže, da je ta ugovor zgrešen na celi črti, saj omenjenega trdi inkompatibilizem ne ogroža prav v nobenem smislu. Kar se tiče ugovora iz nemoralnosti, prav tako lepo pokaže, da ne drži, saj lahko trdi inkompatibilist namesto *v-preteklost-usmerjene* morale zagovarja *v-prihodnost-usmerjeno* moralo, ki pa, za razliko od prve, s trdim inkompatibilizmom ni nezdržljiva. Da je neprepričljiv tudi ugovor iz neosebnih odnosov, po-

kaže tako, da nas opozori, da je takšen ugovor zgolj neupravičen strah, ki nima nobene empirične in teoretske podlage. Edini dve črni piki lahko najdemo v njegovem odgovoru na ugovor iz psihološke teze. To pa zato, ker, kot prvo, preceňuje zmožnost vzporednih odzivov (zmeraj jih vendarle nismo možni) in, kot drugo in bistveno, ker ne uspe uvideti, da psihološka teza za trdi inkompatibilizem sploh ni pravi problem.

Kako naj torej zaključimo? Ali je Pereboomu njegov osnovni cilj – dokazati, da »je trdi [inkompatibilizem] stališče, ki bi ga najlažje sprejeli« (Pereboom, 2007: 199) – uspel? Naš odgovor je pozitiven.

Why we are not free (morally responsible) and what (does not) follow(s) from this

This article presents and critically evaluates Pereboom's article "Determinism *al Dente*" in which he mainly argues two things: firstly, *hard incompatibilism* is – because of "4-case argument" – the most convincing theoretical stance on freedom, determinism and their compatibilism; and secondly, objections against hard incompatibilism, namely that it prevents us from achieving important things in life, promotes immorality and impersonal relationships, and/or is practically impossible, are not convincing.

Keywords: freedom, moral responsibility, (in)determinism, (in)compatibilism, Pereboom

Literatura

Dewar Gleissner, J. (28. 2. 2011). »Some Reasons Why Incarceration Does Not Work Very Well«. *Corrections.com*. Dostopno na:

<http://www.corrections.com/articles/27870-some-reasons-why-incarceration-does-not-work-very-well>

Eagleman, D. (2011). *Incognito*. Edinburgh: Canongate Books.

Frankfurt, H. G. (2007). »Svoboda volje in pojem svobode«. V Šuster, D. (ur.), *O svobodni volji: od Leibniza do Libeta*, Maribor: Aristej, str. 117–131.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press. Dostopno na:

https://books.google.si/books?id=mRns0NYdns0C&pg=PA92&redir_esc=y#v=onepage&q&f=false

Greenstreet, S. (b.d.). »Prison Doesn't Work«. *Philosophy Now*. Dostopno na:

https://philosophynow.org/issues/102/Prison_Doesnt_Work

- Hume, D. (2008). *An Enquiry Concerning Human Understanding*, Section 8, Part 2. Dostopno na: <http://www.earlymoderntexts.com/assets/pdfs/hume1748.pdf>
- Honderich, T. (1988). *A Theory of Determinism*. Oxford: Oxford University Press.
- Leach, A. (24. 6. 2014). »Prison doesn't work 50% of the time, so why do we keep sending people there?«. *Mirror*. Dostopno na: <http://www.mirror.co.uk/authors/amp3d-from-mirror/prison-doesnt-work-50-time-3748230>
- Lucretius, T. C. (1959). *De rerum natura = O naravi sveta*. Ljubljana: Slovenska matica.
- Nagel, T. (2007). »Moralna sreča«. V Šuster, D. (ur.), *O svobodni volji: od Leibniza do Libeta*, Maribor: Aristej, str. 225–237.
- Newitz, A. (23. 7. 2015). »New Evidence That Sending People to Prison Makes Crime Worse«. *Gizmodo*. Dostopno na: <http://gizmodo.com/new-evidence-that-sending-people-to-prison-makes-crime-1719621868>
- Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2007). »Determinizem *al dente*«. V Šuster, D. (ur.), *O svobodni volji: od Leibniza do Libeta*, Maribor: Aristej, str. 171–199.
- Rodham, T. (29. 8. 2012). »Cruel and inhumane: Why prison doesn't work and what to do about it«. *Religion and Ethics*. Dostopno na: <http://www.abc.net.au/religion/articles/2012/08/29/3578660.htm>
- Schoeman, F. D. (1979). »On Incapacitating the Dangerous«. *American Philosophical Quarterly*, 16(1), str. 27–35.
- Strawson, G. (2007). »Nemožnost moralne odgovornosti«. V Šuster, D. (ur.), *O svobodni volji: od Leibniza do Libeta*, Maribor: Aristej, str. 203–220.
- Strawson, P. F. (2007). »Svoboda in zamera«. V Šuster, D. (ur.), *O svobodni volji: od Leibniza do Libeta*, Maribor: Aristej, str. 91–113.
- Watson, G. (1987). »Responsibility and the Limits of Evil«. V Schoeman, F. (ur.), *Character, and the Emotions*, Cambridge: Cambridge University, str. 256–287.

Lea Lampret

Filozofska fakulteta Univerze v Mariboru

Filozofija umetniškega sloga: kaj nam ima o umetniškem slogu povedati Kendall Walton?

Članek posega na področje filozofije umetnosti in v ospredje postavlja razprave o teorijah sloga. Walton trdi, da je umetniški proces nosilec umetniškega sloga. Za določanje sloga umetnine je torej ključno samo dejanje ustvarjanja in ne umetniško delo kot končni produkt. Članek zastavlja vprašanje, ali je za slogovno opredelitev res dovolj samo vpogled v proces nastajanja umetniškega dela ali pa so za pripis ključne vizualne lastnosti umetnine.

Ključne besede: filozofija umetnosti, teorija sloga, Kendall Walton, umetniški slog, performans

Kendall Walton svojo obravnavo sloga umetniških del izpelje iz vprašanja, ali so umetnostni slogi lastnosti (umetniškega) objekta ali lastnosti dejanj (ustvarjalnega procesa umetnika). Zanima ga, na osnovi česa predmetom sploh pripisujemo slogovne značilnosti. V obravnavo želi zajeti čim širši spekter umetnosti, torej tradicionalne likovne zvrsti (slikarstvo, kiparstvo, risbo, grafiko), arhitekturo in performans; kar pa je tudi odlika njegove teorije. Pred obravnavo Waltonovega teksta, »Style and the Products and Processes of Art« (2008), je najprej potrebno prikazati splošni oris obravnave sloga, tako s področja filozofije kot umetnosti.

Preden so vprašanja o slogu prešla v domeno filozofije umetnosti, so bila dolgo časa izrecna skrb zgodovinarjev in umetnostnih teoretikov. Izhodiščno točko za vključevanje razprav o slogu v filozofijo najdemo v začetku 19. stoletja, ko je pojmovanje sloga dobilo nove razsežnosti. Do konca 18. stoletja v razpravah o umetnosti namreč ne najdemo povezav med slogovnimi značilnostmi in pomenom umetniškega dela, pač pa so do tega časa o slogu govorili le kot o nekakšnem stranskem elementu umetnosti. Pred letom 1800 pravzaprav ni bilo teoretskega razumevanja, da obstajala možnost izbire (individualnega) sloga, pač pa so umetniška dela preprosto sovpadala s sočasnimi slogovnimi lastnostmi, ki sta jih je narekovala doba in okolje. Šele v 19. stoletju je slog postal ključni element v razu-

mevanju umetniškega dela in je zavzel mesto, ki so ga v prejšnjih stoletjih zavzemali tradicionalni elementi umetnosti: antični vzori, naturalizem in racionalizem (Carroll, 2002: 129). V tem času je slog postal ključni vidik umetniškega ustvarjanja, ki ga je mogoče prilagajati individualnim potrebam brez spreminjanja vsebine umetniškega dela. Slog, kot spreminjajoči element umetnikovega ustvarjanja, pa je kmalu postal glavni nosilec pomena umetniškega dela (Carroll, 2002: 129; Dileo, 1995). Gre za kompleksen zbir opisov, sestavljenih iz kombinacije pojavljajočih se lastnosti in značilnosti v umetnosti določene osebe, dobe in/ali kraja (Munro, 1946, 129). V nadaljevanju bomo izpostavili ključne prispevke k razvoju teorije sloga v zgodovini in predstavili prispevek sodobne filozofije umetnosti, skozi besede Kendalla Waltona.¹

Mnogi filozofi so se pri svojih teorijah o slogu nanašali na umetnostne teoretike, predvsem umetnostne zgodovinarje, in v njihovih definicijah iskali gradivo za nadaljnji razmislek. A umetnostna teorija razlikuje med različnimi vrstami slogov: individualni slog, periodični/historični slog, nacionalni/krajevni slog (Menaše, 1971: 2054–2055). Za vsakega izmed naštetih vrst umetniških slogov veljajo drugačne lastnosti in vsak je definiran na drugačen način. Umetnostni zgodovinar Johann J. Winckelmann je prvi uporabil koncept periodičnega sloga, ki ga je utemeljil na podlagi vizualnih lastnosti umetniških del. S pomočjo lastnega modela slogovne analize je na podlagi predhodno razvrščenih umetniških del, določenemu zgodovinskemu obdobju pripisal tudi nesignirana dela in dela brez opredelitve avtorstva (Winckelmann, 2008: 204–244). Iz Winkelmanovega modela je izhajal tudi Heinrich Wölfflin, ki je periodični slog natančneje definiral. Svojo teorijo je zasnoval na slogovni analizi, ki izhaja iz primerjave posameznih umetniških del (Wölfflin, 1917: 6; Munsterberg, 2008-2009a). Na Wölfflinovo teorijo se je v veliki meri navezoval tudi Izidor Cankar, ki je pri slogovni teoriji izhajal iz dveh pozicij: naturalizma in idealizma (Cankar, 1995: 49–50). Veliko pozornosti je v razpravah o slogu doživel tudi koncept individualnega umetnikovega sloga. Razlikovanje med slogovnimi značilnostmi umetnikov je pravzaprav staro kot umetnost sama, na zakonitostih osebnega sloga pa so oblikovane številne sodobne metode slogovne analize umetniških del (Munsterberg, 2008-2009b).

¹ Kendall L. Walton (1939) je upokojen profesor Univerze v Michiganu, kjer je poučeval filozofijo, umetnost in oblikovanje, je pripadnik analitične veje filozofije. Velja za enega najbolj uglednih sodobnih filozofov umetnosti zahodne filozofske tradicije. Njegovi članki o osrednjih vprašanih estetike so podlaga za nastanek vidnejših razprav na področju filozofije umetnosti. V veliki meri se osredotoča na teoretična vprašanja o umetnosti, prepoznaven pa je tudi na področju filozofije duha, metafizike in filozofije jezika. Med drugim je pisal o naravi leposlovja in ontologiji fiktivnih subjektov, zanimal se je za slikovno reprezentacijo, fotografijo, estetske vrednosti, metafore, domišljijo in glasbeno izraznost. Njegova najvidnejša dela so: *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (1990), *Marvelous Images: On Values and the Arts* (2008) in *In Other Shoes: Music, Metaphor, Empathy, Existence* (2015) (Povzeto po Churchill, 2011). V sklopu revije *Analiza* je leta 2007, 11(1-2) Božidar Kante objavil prevod Waltonovega spisa »Kategorije umetnosti«.

Morda naj na tem mestu opozorim še na terminologijo. Poimenovanje »osebni slog« se v starejši literaturi pogosto pojavlja pod nazivom »manira«. Izraz je v *Slovarju slovenskega knjižnega jezika* opredeljen kot »publ.: pogosto se ponavlja-joč način izražanja, zlasti v umetnosti, navada: slikarske manire – način izražanja« (Likar, 2000). Beseda »manira« sicer izhaja iz italijanske besede »*maniera*« in v likovni umetnosti pomeni poseben način izražanja umetnika oz. osebni stil umetnika (Bernik, 1985: 162; Likar, 2000).

Tudi pogledov na umetnostni slog s filozofske perspektive je veliko – pravzaprav vsaka filozofska tradicija prispeva drugačne poglede na teorijo sloga, strinjanja med posameznimi filozofi pa tako rekoč ni. Problem oblikovanja definicije sloga med filozofi se izkaže v razumevanju njegovih ključnih lastnosti in iskanju njegovega izvora, ki ga pripisujemo umetniškim delom. Nekateri slog tako razumejo kot moralni oz. družbeni problem, saj zanje predstavlja neko neestetsko prvino umetniškega dela. Ta sicer omogoča pogled na umetnikovo individualno pojmovanje lepega, v katero nam je omogočen le posredni pogled (preko umetnikove izraznosti v konkretnem). Drugi slogovne značilnosti umetniških del pripisujejo izbiri tehnike, v kateri je neko delo izvedeno (Mueller, 1941: 105–106; Lah, 2015: 215).

Walton izhaja iz predpostavke, da je za razumevanje umetniškega sloga potrebno pogledati v sam proces nastajanja umetnine. Za določanje njenega sloga je torej ključno samo dejanje ustvarjanja in ne umetniško delo kot končni produkt. Pa vendar s tem ne trdi, da sama umetnina ni pomembna; ravno nasprotno. Oba vidika sta med seboj tesno povezana. Svojo izbiro utemelji na podlagi hipoteze, da vsak umetniški produkt predpostavlja dejanje umetnostne kreacije. Ni umetnine brez ustvarjanja in ni ustvarjanja brez človeka. Zanj je pomembno, da o slogu govorimo le v povezavi s tistimi deli, ki jih je ustvarila človeška roka (artefakti); delom, ki pa so bila ustvarjena v procesu masovne produkcije (npr. obleka, pohištvo,...), pa lahko pripisujemo le estetsko, ne pa tudi slogovne vrednosti (torej sodimo o njih le kot o lepih ali grdih) (Walton, 2008: 222). Na primer, baročnemu stolu lahko pripišemo slogovne značilnosti, saj ga je zasnoval in oblikoval človek – izkušen obrtnik, ki je materialu podal novo obliko in namembnost. Na drugi strani pa slogovnih lastnosti ne moremo pripisati slavnemu stolu Rex-u, saj gre za predmet masovne produkcije; nekaj, kar je človek sicer zasnoval, a je materialna izvedba rezultat industrijskih postopkov. Prav tako slogovnih značilnosti ne moremo pripisovati naravnim pojavom (Walton, 2008: 222). Če na primer vidimo sončni vzhod na nebu, ne moremo govoriti o mehkih linijah, nežni konturi ali močnem kontrastiranju kot o lastnostih naravnega. Te lastnosti lahko naravi pripišemo le na podlagi analogije z umetniškim delom.

Umetniška dela torej posedujejo določene lastnosti. Kadar opisujemo umetniško delo, uporabljamo številne predikate, ki se z umetnino povezujejo in služijo za opisovanje umetnikove dejavnosti v procesu ustvarjanja. Zbir predikatov pred-

stavlja lastnosti, ki jih je umetnik v aktivnem procesu ustvarjanja vtisnil v umetnino. Te lastnosti so v umetniškem delu ponotranjene in zato predstavljajo osnovo za izgradnjo slogovnega vidika umetnine. Gre za lastnosti (npr. sentimentalnost, ekspresivnost, morbidnost,...); ki pa še ne pomenijo sloga samega, pač pa so sestavine za izgradnjo slogovne analize umetniškega dela (Walton, 2008: 223).

Walton izpostavlja pomembnost umetniškega dejanja tudi s t.i. trodelnim modelom, ki ga poimenuje »model čevljarja«. Gre za miselni eksperiment, ki v ospredje postavlja razmerje med čevljarjem, čevljem (kot končnim produktom) in kupcem oz. uporabnikom čevljev. Bistvo čevljarjevega dela je, kako dobro se bo končni produkt, torej čevlji, prilegal kupcu (udobje, velikost, estetska pričakovanja, namen uporabe). Kakor hitro se kupec odloči kupiti čevlje, postane razmišljanje o procesu njihove izdelave zanj irelevantno, in to celo takrat, kadar gre za čevlje, ki so artefakt (Walton, 2008: 223). Walton je ta primer apliciral na področje umetnosti.

Če torej predstavljeni model uporabimo za umetnost, potem sledimo razmerju med umetnikom, umetnino in (kritičnim) gledalcem. Vsakdanji gledalec se namreč ne bo kaj prida ukvarjal z umetnikovim ustvarjalnim procesom. Še več, morda celo z umetnino, kaj več kakor v individualno estetskem smislu, ne. A Walton od poznavalca in kritičnega gledalca umetnosti zahteva, da se pri razmišljanju o umetniškem delu ne ustavlja zgolj pri umetnini kot objektu, pač pa mora preudariti tudi sam proces ustvarjanja. Ta je namreč povod za obstoj umetniškega dela, v smislu končnega produkta (Walton, 2008: 223). Če poenostavimo, gledalec, ki želi razmisliti o slogu umetniškega dela, ki ga opazuje, pri slogovni analizi ne bo uspešen, če bo opazoval samo končni produkt, pač pa bo za uspešno identifikacijo sloga moral pogledati v samo dejanje nastanka umetnine in pri tem upoštevati vse dejavnike (individualne, periodične in krajevne vrvine).

Ustavimo se, na primer, pri umetnosti Jacksona Pollocka. Kako bi lahko njegovo umetnost razumeli, če ne skozi sam ustvarjalni proces? Svoje slikarstvo je Pollock dojemal kot nekaj nezavednega, pa vendar je procesu ustvarjanja pripisoval veliko pomembnost, kar nam lepo pokaže njegova izjava:

Ko sem v svojem slikanju, se ne zavedam, kaj počnem. Šele po neke vrste »spoznavnem« obdobju sprevidim, kaj sem pravzaprav počel. Nič se ne bojim spreminjanj, pa tega, da bi pokvaril podobo itd., ker ima slika svoje lastno življenje. Trudim se, da bi presevalo skozi to. Šele kadar izgubim stik s sliko, je končni izid godlja. Sicer pa gre za eno samo harmonijo, za lahkoten, neobremenjen daj-dam, in slika se posreči. (Lyton, 1994: 231)

Iz citata je razvidno, da je za umetnost pomembno, da se zavedamo obeh vidikov, predvsem pa, da je ustvarjalni proces nekaj, kar veliko pove o umetniku kot osebnosti.

Tovrsten princip pripisovanja slogovnih značilnosti pa se ne obnese samo pri tradicionalnih likovnih zvrsteh, pač pa lahko model uporabimo tudi za utemeljevanje performativne umetnosti. Zakaj sploh lahko performanse, ki so se množično pojavili v 20. stoletju, vključujemo v razprave o umetnosti? Če bi na njih gledali iz vidika končnega produkta, v njih ne bi našli nič povednega in umetniškega. Na performativno umetnost moramo tako gledati iz perspektive umetnikove aktivnosti, tj. dejanja, skozi katerega se umetnik izraža. Pri tem nas ne smejo zanimati posamezni gibi, ki jih umetnik opravi, pač pa je treba gledati na dejanje kot celoto (Walton, 2008: 224–225). V tem smislu se performans povezuje tudi z gledališko umetnostjo. Šele ko pogledamo na celoto prizorov, na vse elemente, ki so bili predstavljeni na odru, lahko zajamemo bistvo predstave. Tako je tudi s performansom. Če vam nekdo pokaže zgolj fotografske posnetke performansa z določenim naslovom – npr. fotografije performansa Gašperja Kunšiča z naslovom *Life is Amazing* – iz fotografij ne moremo opredeliti niti sporočila dela, niti slogovnih značilnosti performansa, saj si nismo ogledali celotnega dogajanja, ki ga je v procesu izvajal umetnik. Šele ko bi si performans ogledali v živo, bi po Waltonu lahko govorili o umetnikovem sporočilu in slogovnih značilnostih umetnikovega dejanja.



Gašper Kunšič, Life is Amazing, 2015.

Kendall Walton nas opozarja, da je zmeda pri pripisu sloga še toliko večja, ker umetniki gledalcem pogosto ne sugerirajo, na kaj naj se ob ogledu njihovih umetniških del osredotočajo. Walton umetnikom očita, da pogosto o svojih delih govo-

rijo kot o nekakšnih mojstrskih delih in s tem interesno področje gledalca usmerjajo stran od ustvarjalnega momenta (Walton, 2008: 225). Le redko umetniki v ospredje postavijo sam proces kreiranja, čeprav ni mogoče, da bi kakršnokoli umetniško delo nastalo neodvisno od umetniškega procesa. Dober primer utemeljevanja umetniškega sloga skozi dejanja ustvarjanja je tudi opredeljevanje *ready-made* umetnosti, kjer lahko opazujemo še en zanimiv pojav: simbolika in ekspresija ustvarjanja predmetov je drugačna kakor simbolika in ekspresija razstavljanja predmetov.



Jure Markota, Nakopičenje jeze II, 2014.

Walton torej slogovne lastnosti umetniškega dela išče v dejanju umetnikovega ustvarjanja. Slog razume kot način, kako se preko umetniškega dela izraža ustvarjalni proces, torej umetnikova dejanja, ki jih je izvrševal med kreacijo umetnine (Walton, 2008: 228). A slikarji in kiparji pri svojih delih pogosto tudi spretno zakrivajo proces, v katerem so delo izvedli. Sledi, ki jih je za sabo pustil ustvarjalni proces, pozorni opazovalci sicer lahko zaznajo (npr. grundiranje, poteze čopičev, sledi kiparske obdelave materiala...), a so prepogosto spregledane (Walton, 2008: 231–232). V tem oziru lahko slog povezujemo ne samo z ustvarjalnim procesom, pač pa tudi z izbiro tehnike, v kateri je delo izvedeno.

Slog je Walton povezal tudi z estetskimi lastnostmi umetniških del, saj so le-te odvisne od ustvarjalnega procesa. S tem pa se slogovna teorija iz področja filozofije umetnosti razširja tudi na estetsko področje (Walton, 2008: 233).

Njegovo pojmovanje sloga lahko povežemo z opredelitvijo sloga pri umetnostnem zgodovinarju Ernstu Gombrichu. Zanj je slog vsak razlikovalen in s tem prepoznaven način izvedbe umetniškega dejanja oz. način ustvarjanja in izvedbe artefakta (Gombrich, 2009: 129). Torej tudi on za osnovo slogovne analize vzame umetniško dejanje, ki je vzrok za nastanek umetniškega dela. V umetnini so lastnosti, ki jih je umetnik preko ustvarjalnega in intencionalnega procesa vtisnil v delo, zato ustvarjalnega dejanja in končnega umetniškega dela ne moremo obravnavati ločeno.

James Ackerman, ameriški umetnostni zgodovinar in teoretik umetnosti, nasprotuje Waltonovi teoriji in slog umetniškega dela pojmuje kot zbir atributov umetnostnega objekta (umetnine). Trdi, da so umetniška dela sama na sebi vir primarnih podatkov, ki vplivajo na našo sodbo o tem delu. Ko naredimo zbir teh primarnih podatkov, na njihovi podlagi oblikujemo določene lastnosti umetniškega dela, ki morajo biti stabilne, pa vendar do določene mere fleksibilne za interpretacijo. Te lastnosti pa predstavljajo slog. Zanj je slog (tako individualni kakor periodični) neizogibno zgodovinski pripomoček, saj nam slogovna analiza odpre poti do številnih spoznanj tako o umetnosti sami, kakor tudi o kulturi in času, v katerem je bila ustvarjena (Ackerman, 1962: 227).

Med obe skrajnosti lahko umestimo Meyerja Schapira, umetnostnega zgodovinarja, ki pravi, da je za uspešno slogovno analizo potrebno upoštevati oba vidika: tako vidik umetniškega dejanja kakor vidik končnega dela. Slog zanj predstavlja kompleksno povezavo med formo, likovnimi elementi in umetnostnimi izrazi, ki jih najdemo v umetnosti posameznika ali skupine. Izraz slog pa se nanaša tudi na celotno dejavnost posameznika (ali skupine), in sicer takrat, ko govorimo o »življenjskem slogu« (Schapiro, 1998: 143).

Menim, da je Waltonova težnja po poenotenju metode slogovne analize, s katero bi lahko zajeli vse umetniške zvrsti (likovno umetnost, gledališko umetnost, glasbo in literaturo), nekoliko pretirana. Strinjam se, da je proces umetniškega dejanja osnova za obstoj umetniškega dela, pa naj bo to slika, kip, drama, opera ali sonet. Brez ustvarjalnega procesa umetnosti seveda ne bi bilo. A vsaka izmed umetnostnih zvrsti ima lastno vrsto izražanja in s tem posebne lastnosti. To, kar gledalcu omogoča slogovno opredelitev slikarstva, ni enako temu, kar bralcu omogoča slogovno opredelitev poezije. Čeprav pogosto več umetnostnim zvrstem pripisujemo enak slogovni termin, menim, da ne gre za isti slog. Kriterij, po katerem krajinam Marka Pernharta pripisujemo romantični slog, ni enak kriteriju, po katerem poezijo Franceta Prešerna umeščamo v romantično pesništvo. Kakor hitro ne gre za enak kriterij, pa mislim, da ne moremo govoriti o istem slogu.

Če se osredotočim zgolj na likovno umetnost, lahko izpostavim, da je slog v večji meri značilnost končnega produkta, kakor pa procesa, v katerem nastaja. Ob tem ne trdim, da sta ustvarjalni proces in okoliščine nastanka umetnine za njene slogovne značilnosti nerelevantne, pač pa zgolj to, da je slog v večji meri odvisen od vizualnih lastnosti umetnine. Morda se zdi, da je Walton slog enačil s t.i. umetnostno tradicijo, tj. z načinom, kako je neka umetnina ustvarjena. Če je nek slikar, na primer, slikal v barbizonski tradiciji, to še ni zadosten pogoj za to, da lahko njegovo delo označimo za impresionistično. Pridevnik impresionistično lahko pripišemo samo vizualnemu efektu umetniškega dela, ki je sicer res rezultat umetnikovega ustvarjalnega postopka, a samo dejanje ustvarjanja nima impresionističnih lastnosti. Pravzaprav si ne morem niti zamisliti impresionističnega dejanja. Je pa res, da bi moral pozoren gledalec ob ogledu umetniškega dela razmisliti tudi o postopku njegovega nastanka, saj lahko poznavanje umetnikovega ustvarjalnega procesa veliko pripomore k razumevanju njegove umetnosti. Tako umetnikovo življenje in delo kot tudi odraz časa in kraja nastanka umetniškega dela gotovo vplivata na slogovne značilnosti umetnosti, vendar menim, da je slogovna opredelitev veliko bolj odvisna od vizualnih lastnosti umetniškega dela.

Ob Waltonovi teoriji se lahko sprašujem tudi o slogovni opredelitvi performativne umetnosti. Ali performativna umetnost od gledalca sploh pričakuje, da sodi o slogu performansa? Performans je v celoti odvisen od enkratnega dogajanja, ki ga umetnik izvede pred občinstvom v določenem prostoru in času. Čeprav ima performans veliko sporočilno moč menim, da mu ne moremo pripisovati slogovnih značilnosti, vsaj ne tistih likovno-vizualnega značaja.

Ne glede na drobne pomisleke naj na koncu poudarim, da ima Waltonova teorija izjemen doprinos v razpravah o teorij umetniškega sloga. Njegov pristop je opozoril, da moramo pomemben vidik slogovne analize umetniških del iskati tudi v procesu nastajanja, oz. ustvarjanja umetniškega dela. Razprave o slogu ostajajo aktualne, tako v filozofskih kakor likovnoteoretskih krogih, za seboj pa puščajo mnogo odprtih vprašanj in prostora za nadaljnjo interpretacijo.

Philosophy of Artistic Style – The Contribution of Kendall Walton

This article deals with the philosophy of art, in general, and with the theories of style in particular. Walton claims that the bearer of artistic style is the creative process. In determining the style of an artwork, the key role is played only by the process of creation, and almost no role is assigned to the final product, the artwork itself. We raised the question whether a mere insight into the genesis of artwork is really sufficient, or whether its visual characteristics are indeed the key element for ascribing stylistic characteristics.

Keywords: philosophy of art, theory of style, Kendall Walton, artistic style, performance

Literatura

- Ackerman, J. S. (1962). »A Theory of Style«. *The Journal of Aesthetics and Art Criticism*, 20(3), str. 227–273.
- Bernik, S. (1985). *Likovna umetnost*. Ljubljana: Cankarjeva založba.
- Cankar, I. (1995). *Uvod v umevanje likovne umetnosti: Sistematika stila*. Ljubljana: Karantanija.
- Carroll, N. (2002). *Philosophy of Art: A contemporary introduction*. London, New York: Routledge.
- Churchill, S. J. (2011). »Regents Communication«. [9. 10. 2016]
<http://www.regents.umich.edu/meetings/12-11/2011-12-VI-Walton.pdf>.
- Dileo, J. R. (1995). »The Question of Style in Philosophy and the Arts« (recenzija). [9. 10. 2016]
http://muse.jhu.edu/journals/philosophy_and_literature/summary/v021/21.1br_van_eck.html.
- Gombrich, E. (2009). »Style«. V Preziosi, D., *The Art of Art History: A Critical Anthology* (2. izd.), Oxford: Oxford University Press, str. 129–140.
- Kunšič, G. (2015). *Life is Amazing*, performans.
<http://cargocollective.com/gasperkunsic>
- Lah, N. (2015). »How Style Became Famous and Irrelevant at the Same Time«. *Ars & Humanitas*, 2, str. 215–230.
- Likar, V. (ur.) (2000). »manira (1)«. V *Slovar slovenskega knjižnega jezika*. [9. 10. 2016] http://bos.zrc-sazu.si/cgi/a03.exe?name=sskj_testa&expression=ge=manira.
- Lytton, N. (1994). *Zgodba moderne umetnosti: Pregledi likovne umetnosti 20. stoletja*. Ljubljana: Cankarjeva založba.
- Markota, J. (2014). *Nakopičenje jeze II*, instalacija.
<http://www.juremarkota.com/>
- Menaše, L. (1971). *Evropski umetnostnozgodovinski leksikon*. Ljubljana: Mladinska knjiga.
- Mueller, G. E. (1941). »Style«. *The Journal of Aesthetics and Art Criticism*, 1(2/3), str. 105–122.
- Munro, T. (1946). »A Method of Stylistic Analysis«. *The Journal of Aesthetics and Art Criticism*, 5(2), str. 128–158.
- Munsterberg, M. (2008-2009a). »Period Style«. [9. 10. 2016]
<http://writingaboutart.org/pages/periodstyle.html>.

Munsterberg, M. (2008-2009b). »Personal Style«. [9. 10. 2016]
<http://writingaboutart.org/pages/personalstyle.html>.

Schapiro, M. (1998). »Style«. V Preziosi, D., *The Art of Art History: A Critical Anthology* (1. izd.), Oxford: Oxford University Press, str. 143–149.

Walton, K. (2008). »Style and the Products and Processes of Art«. V Walton, K., *Marvelous Images: On Values And The Arts*, Oxford: Oxford University Press, str. 221–248.

Winckelmann, J. J. (2008). »Misli o umetnosti«. *Nova revija*, 27(309/311), str. 204–244.

Wölfflin, H. (1917). *Kunstgeschichtliche Grundbegriffe: Das Problem der Stilentwicklung in der neueren Kunst*. München: Hugo Bruckmann, Verlag. [9. 10. 2016]
<https://archive.org/stream/kunstgeschichtli00wluoft#page/n7/mode/2up>

Tadej Todorović

Filozofska fakulteta Univerze v Mariboru

Večvrstna realizacija in Turingov test: preslikava

Članek v prvem delu predstavi večvrstno realizacijo in Turingov test. Drugi del je namenjen dokazu, da je način sklepanja pri Turingovem testu in večvrstni realizaciji identičen. V tretjem delu predstavi enega izmed ugovorov Turingovemu testu in njegovo rešitev s pomočjo t.i. *oblikovnega* kriterija. V četrtem delu nato to rešitev preslikamo na večvrstno realizacijo, kar nas pripelje do sklepa, ki je popolnoma nasproten sklepu večvrstne realizacije. To nas vodi do ugotovitve, da je edini razlog za sprejem večvrstne realizacije njena metafizična zamisljivost, kar pa, kot pokažemo, ni dovolj. V petem delu predstavimo možno strategijo zagovarjanja večvrstne realizacije in težave, ki se lahko pri tem pojavijo.

Ključne besede: večvrstna realizacija, mentalna stanja, Turingov test, ugovor iz črne škatle, takšnosti

0 Uvod

V članku bomo v prvem delu najprej predstavili večvrstno realizacijo, kot si jo je zamislil Hilary Putnam v letih 1960–1970, in pa Turingov test, kot ga je predstavil Alan Turing. Poskušali bomo ugotoviti, na podlagi česa sklepamo v primeru večvrstne realizacije in na podlagi česa v primeru Turingovega testa. Ugotovili bomo, da tako Turingov test kot tudi večvrstna realizacija gradita svoji osnovni premisi na isti predpostavki, ki jo bomo imenovali behavioristična predpostavka. V tretjem delu članka bomo izpostavili enega izmed ugovorov Turingovemu testu, ki to predpostavko postavi pred velik vprašaj – ugovor iz črne škatle. S pomočjo *oblikovnega* kriterija bomo nato nadgradili Turingov test. To nadgradnjo bomo zatem preslikali na tezo večvrstne realizacije. Ugotovili bomo, da z dodano nadgradnjo pridemo do sklepa, ki je nasproten tezi večvrstne realizacije, in da je edini razlog za sprejetje večvrstne realizacije njena metafizična zamisljivost. V zadnjem delu članka bomo poskusili ponuditi strategijo zagovarjanja večvrstne realizacije in opozorili na težave, ki se lahko pri tem pojavijo. Poskusili bomo tudi prikazati jasen primer večvrstne realizacije v relaciji s človeškimi možgani, ki

ne bo zgolj analogija, njegov princip bo predstavljal morda premočen, vendar zadosten kriterij za oblikovanje večvrstnih realizacij v problemu duha in telesa.

I Večvrstna realizacija in Turingov test

Večvrstna realizacija je teza, ki jo je Hilary Putnam uvedel v filozofijo duha v seriji člankov v letih 1960–1970.¹ Trdi, da so mentalna stanja (npr. bolečina) lahko realizirana preko različnih fizičnih sistemov. (gl. Bickle, 2016: 1) Če uporabimo primer bolečine: tako mi (ljudje) kot druga živa bitja (hobotnice, psi, mačke) doživljamo subjektivno izkustvo bolečine. Vendar se način nastajanja bolečine v ljudeh, mačkah, psih in hobotnicah (zaradi različne zgradbe možganov) razlikuje, kar vodi v to, da lahko različni fizični sistemi (fizični sistem možganov ljudi, psov, mačk in hobotnic) realizirajo isto mentalno stanje – tj. bolečino. Iz tega sledi, da mentalnih stanj ne moremo enačiti s fizičnimi stanji, ki mentalna stanja realizirajo. Večvrstna realizacija je tako eden izmed glavnih argumentov proti redukcionizmu kot teoriji duha in telesa,² ki trdi, da je mentalno stanje identično fizičnemu stanju, ki ga realizira.

V deduktivni obliki lahko argument iz večvrstne realizacije zapišemo v naslednji obliki:

- 1) Vse mentalne pojave lahko večvrstno realiziramo z različnimi fizičnimi sistemi. (teza večvrstne realizacije)
- 2) Če lahko mentalne pojave večvrstno realiziramo z različnimi fizičnimi sistemi, potem le-ti ne morejo biti identični s specifičnim fizičnim sistemom.
- 3) ($\therefore$) Noben mentalni pojav ni identičen specifičnemu fizičnemu sistemu. (Bickle, 2016: 1)

Pod drobnogled bomo vzeli prvo premiso argumenta – tezo večvrstne realizacije, »resničnost katere so kritiki³ v zadnjih letih poskušali ovreči« (Bickle, 2016, 2.4). Preko analogije med večvrstno realizacijo in Turingovim testom bomo poskušali tezo večvrstne realizacije ovreči ali vsaj postaviti pod vprašaj. Da lahko to naredimo, si najprej oglejmo, kaj je Turingov test.

Turingov test je preizkus, ki si ga je leta 1950⁴ zamislil Alan Turing. Namenjen je temu, da presodi, ali stroj misli, to pa naredi s pomočjo igre oponašanja. V igri nastopajo trije akterji – dva človeka in stroj. Stroj in prvi človek sta zaprta vsak v svojo sobo. Obe sobi sta komunikacijsko povezani z nadzorno sobo (danes lahko

¹ Recimo v člankih »Psychological Predicates« (Putnam, 1967) in »The Nature of Mental States« (Putnam, 1975).

² Za več o redukcionizmu glej Bregant, 2004: 61.

³ Zangwill, 2002; Shapiro, 2000; Couch, 2004.

⁴ Gl. Turing, 1950.

to zelo elegantno izvedemo s pomočjo lokalne internetne povezave), v kateri je drugi človek. Le-ta nato preko te povezave z obema komunicira – ne vidi v notranjost sobe in ne sliši stroja in prvega človeka. Tako torej drugi človek ne ve, v kateri sobi je stroj in v kateri je prvi človek. Drugi človek (imenujmo ga spraševalec) nato zastavlja kakršnakoli vprašanja obeh akterjema v sobah (enostavna, površinska, zapletena, globoka) z namenom, da bi ugotovil, v kateri sobi se nahaja stroj in v kateri človek. Naloga stroja je, da pripelje drugega človeka (spraševalca) do zmotne identifikacije (torej, da spraševalec identificira stroj kot človeka, in obratno). To bo storil najlažje, če bo spraševalcu podajal odgovore, ki so nerazločljivi od odgovorov človeka – stroj bo torej v odgovorih oponašal človeka. Če se spraševalec v 50% primerov zmoti ali če se ne more odločiti, v kateri sobi je stroj in v kateri človek, potem moramo priznati, da je stroj zavesten – da torej misli (gl. Bregant, 2010: 64–65).

V prvem delu smo predstavili večvrstno realizacijo in Turingov test, v drugem delu pa si bomo ogledali, na kakšen način sklepamo, ko govorimo o Turingovem testu in večvrstni realizaciji.

II Primerjava načina sklepanja pri Turingovem testu in pri večvrstni realizaciji

Ne bo nas zanimal torej Turingov test sam,⁵ temveč samo predpostavka, na kateri Turingov test sloni. Imenovali jo bomo behavioristična predpostavka, ki bo trdila, da: če se stroj obnaša enako kot človek oziroma če med obnašanjem človeka in stroja ni razlike, potem moramo zavoljo konsistence priznati, da stroj misli. Ali kot to povzame Ayer: »Če vem, da se objekt v vseh pogledih obnaša tako, kot se po definiciji obnaša zavestno bitje, potem vem, da ima v resnici zavest« (Ayer, 1978: 178).

Zapisano v prečiščeni obliki: *če se stroj obnaša enako (nerazločljivo) kot zavestno bitje, potem ima stroj zavest*. Zdi pa se, da je ta predpostavka zelo podobna predpostavki teze večvrstne realizacije, ki pravi, da lahko vse mentalne pojave večvrstno realiziramo z različnimi fizičnimi sistemi. Zakaj? Da namreč lahko smiselno trdimo, da tako kot človek, tudi pes in mačka doživljata bolečino, moramo sprejeti behavioristično predpostavko, in sicer *da če se nek organizem obnaša na enak (nerazločljiv) način, kot se obnašamo mi v stanju bolečine, potem ta organizem prav tako doživlja bolečino*. Imenujmo prvo predpostavko »predpostavka Turingovega testa« in drugo »predpostavka teze večvrstne realizacije« in obe zapišimo:

Predpostavka Turingovega testa: če se stroj obnaša enako (nerazločljivo) kot človek v stanju zavesti, potem ima stroj zavest.

⁵ Za več o samem testu in ugovorih nanj glej Bregant, 2010.

Predpostavka teze večvrstne realizacije: če se nek organizem obnaša enako (nerazločljivo) kot človek v stanju bolečine, potem je organizem v stanju bolečine.

Vidimo lahko, da obe predpostavki sledita istemu pravilu. Če se namreč neka stvar obnaša enako (nerazločljivo) kot človek v stanju x, potem je ta stvar v stanju x. Logično sta torej enaki, še več, rečemo lahko, da je to v bistvu ena in ista predpostavka – imenujmo jo *behavioristična predpostavka*, le stanje x je različno (v prvem primeru je stanje x bolečina, v drugem pa zavedanje/mišljenje).

Predpostavka Turingovega testa in predpostavka večvrstne realizacije bi tako, če želimo biti konsistentni, morali biti podvrženi enako strogi kritiki v filozofiji, saj smo ugotovili, da sta v bistvu ena in ista predpostavka. Pa vendar ni tako. Resničnost Turingovega testa je bila v 20. stoletju večkrat postavljena pod vprašaj, že v času nastajanja samega testa⁶ kot tudi kasneje⁷ – kot enega najbolj znanih lahko omenimo ugovor ameriškega filozofa Johna Searla, *argument kitajske sobe*.⁸ Nasprotno pa v resničnost večvrstne realizacije do leta 2000 ni dvomil skoraj nihče. Kot ugotavlja Shapiro: »Čeprav se je nekaj filozofov vprašalo, ali je večvrstna realizacija res v nasprotju z redukcionizmom, /.../ še, kolikor vem, nihče ni resno podvomil v resničnost same teze večvrstne realizacije« (Shapiro, 2000: 635–636).

Zato bomo sedaj poskusili najti ugovor Turingovemu testu, ki bi bil po analogiji lahko tudi ugovor večvrstni realizaciji.

III Ugovor iz črne škatle in nadgradnja Turingovega testa

Eden izmed ugovorov Turingovemu testu se imenuje ugovor iz črne škatle. Turingov test temelji samo na obnašanju stroja, zanemari pa, kaj se dogaja v samem stroju – torej njegovo zgradbo. Ko sočloveku pripisujemo mentalna stanja (zavest), vemo, kako delujejo možgani, kateri elektrokemični procesi so odgovorni za katero mentalno stanje ipd. Nasprotno nam je delovanje stroja neznano, zato stroju zaradi obnašanja samega ne moremo kar pripisati mišljenja/zavesti (Bregant, 2010: 67).

Oglejmo si podrobneje, kaj želi ta ugovor sporočiti. Prva interpretacija je naslednja: zaradi tega, ker ne poznamo zgradbe stroja, mu samo zaradi enakega obnašanja ne moremo pripisati zavesti. To se zdi absurdno, saj bi mi tudi danes, čeprav ne poznamo vseh elektrokemičnih procesov v možganih, pripisali zavest sočloveku, saj ne želimo pasti v solipsizem. Z drugimi besedami, velika večina ljudi ne ve, kako možgani delujejo, pa vendar soljudem pripisujejo zavest – torej poznavanje delovanja stroja ne more biti pogoj za pripisovanje zavesti. Bolj se zdi verjetna druga interpretacija: zaradi tega, ker ne vemo, ali je zgradba stroja enaka kot

⁶ Gl. Bregant, 2012.

⁷ Za vse sodobne ugovore Turingovemu testu gl. Oppy in Dowe, 2016.

⁸ Za več o argumentu kitajske sobe gl. Searle, 1981.

zgradba naših možganov, stroju ne pripišemo zavesti samo zaradi obnašanja. Torej ljudje v nasprotju s tem, kar je predpostavil Turing, ne pripišemo zavesti sočloveku samo zaradi obnašanja, temveč tudi zaradi podobne zgradbe možganov. Šele ta dva pogoja v konjunkciji tvorita zadosten pogoj, da pripišemo sočloveku zavest. Ali, zapisano v logični obliki: če se x obnaša kot jaz (zavestno bitje) in ima podobno strukturo možganov kot jaz, potem sledi, da je tudi x zavestno bitje.

Zdi se, da je treba ugovor iz črne škatle interpretirati v drugem smislu. To podkrepi tudi protiugovor temu ugovoru (torej poskus reševanja Turingovega testa), ki ga je postavil Copeland: »Oplemeniten Turingov test bi torej moral vsebovati dva elementa: prvič, izvirno posnemovalno igro, s pomočjo katere bomo ugotovili, ali je izpolnjen *behavioristični kriterij*⁹ (tj. da so odgovori¹⁰ računalnika nerazločljivi od odgovorov človeka), in drugič, morali bi pregledati sam računalnik,¹¹ kar bo razjasnilo, ali je izpolnjen tako imenovan *oblikovni kriterij*« (Copeland, 1993: 50–51). Seveda je vprašanje, kaj ta oblikovni kriterij je. Copeland predlaga dve možnosti – oglejmo si obe in njune probleme. Prva verzija *oblikovnega* kriterija pravi, da mora stroj »delati stvari, ki jih dela, na način, ki je približno enak načinu delovanja človeških možganov« (Copeland, 1993: 51). Z drugimi besedami – fizično (elektrokemično) delovanje stroja mora biti zelo podobno fizičnemu delovanju možganov. Ugovor, na katerega takšna verzija *oblikovnega* kriterija nalleti, je, kot pravi Copeland sam, da »... je premočan. Izgleda, da je neupravičeno antropocentrično zahtevati, da računalnik deluje na podoben način kot človek« (Copeland, 1993: 51). Pa je to res neupravičeno? »Izviren Turingov test je brez sramu antropocentričen, saj zahteva, da računalnik zmore proizvesti obnašanje, nerazločljivo od človeškega« (Copeland, 1993: 51).

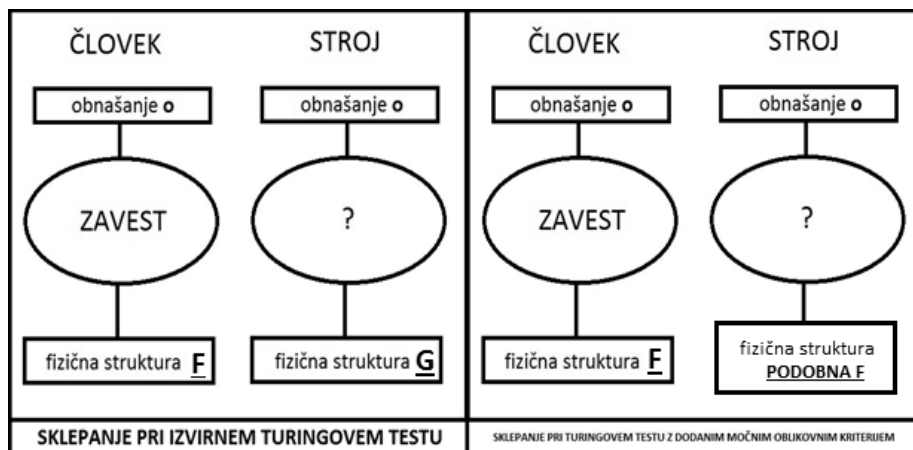
Druga (šibka) verzija *oblikovnega* kriterija pa pravi, da mora biti računalnik modularen. To pomeni, da mora obstajati možnost, da program »zapakiramo« in vključimo v drug program in tako ustvarimo program večje kompleksnosti. Na primer, mora biti možno, da program prenesemo na program, ki nadzira motorične in čutne sisteme robota, rezultat tega pa mora biti robot, ki ima možnost govoriti o tem, kaj dela (Copeland, 1993: 51). Žal ni najbolj jasno, katero verzijo *oblikovnega* kriterija sprejeti. »Mislim, da iz diskusije sledi, da ni nobena od teh verzij očitno dobra ali slaba, niti ni nobena očitno boljša kot druga« (Copeland, 1993: 52). Ne glede na to bomo poskušali podati razlog za sprejetje močnejše verzije *oblikovnega* kriterija.

⁹ Avtor tukaj uporabi besedno zvezo *output criterion*, vendar bomo zaradi jasnosti in že uvedene terminologije uporabljali besedno zvezo *behavioristični kriterij*. Enak prevod uporabi tudi Bregant (2010: 67).

¹⁰ Ponovno avtor uporablja besedno zvezo *conversational output*, kar prevajamo kot odgovori računalnika/človeka – mislim, da s tem obdržimo semantičen pomen besedne zveze.

¹¹ Avtor uporabi besedo »program«, vendar mislim, da avtor misli na pregledovanje delovanja računalnika in programa.

Ko sklepamo samo na podlagi behavioristične predpostavke (če se *x* obnaša kot jaz, ki sem zavestno bitje, potem sledi, da je tudi *x* zavesten), sklepamo na podlagi analogije. Sklepamo namreč tako: jaz (*y*) imam lastnost *z* (zavest) in lastnost *o* (specifično obnašanje). Stroj *x* (ki »zmaga« na Turingovem testu) ima lastnost *o* (specifično obnašanje, ki je nerazločljivo od mojega). Zato ima stroj *x* tudi lastnost *z* (zavest). Ampak sklepanje po analogiji je induktivno, torej ne-nujno, analogija pa je dobra takrat, ko »so podobnosti med primerjanima pojavoma ali dogodkoma precej številčnejše od razlik« (Klampfer, Bregant, Gartner, 2008: 78). V dotičnem primeru imamo samo eno razliko in eno podobnost, kar je za dobro sklepanje po analogiji premalo. Če ponazorimo na primeru: 85% temna čokolada in 99% temna čokolada sta enake barve (obe temno rjavi), pa to še ne pomeni, da če je nekomu dobra 85% temna čokolada, mu bo dobra tudi 99% temna čokolada. Pri sklepanju po analogiji je zato bolj »varno«, če obstaja več skupnih lastnosti, na podlagi katerih sklepamo na neko tretjo lastnost (v primeru Turingovega testa je to zavest). Razlika, ki jo želimo prikazati, je ponazorjena na Shemi 1. Na levi strani vidimo sklepanje pri izvirnem Turingovem testu, kjer sklepamo *zgolj* na podlagi obnašanja – zdi se, da je takšno sklepanje prešibko (ugovor iz črne škatle in argument kitajske sobe). Na desni strani pa sklepamo (z dodanim oblikovnim kriterijem) na podlagi dveh skupnih lastnosti, kar je zagotovo močnejše sklepanje, ki naredi naš argument močnejši.



Shema 1

Kot pravi Copeland (za močni *oblikovni* kriterij), se sicer zdi, da je ta kriterij premočan (antropocentričen). Nekdo lahko namreč ugovarja z naslednjim miselnim eksperimentom (imenovali ga bomo miselni eksperiment iz zamisljivosti): lahko si zamislimo bitje, ki bi imelo bistveno drugačno fizično strukturo in delovanje možganov, pa bi vendarle imelo zavest. Vendar tudi če je to zamisljivo, je naloga tistega, ki to trdi, da dokaže, da takšno bitje obstaja. Prav tako z logičnega vidika ne moremo očitati nekomu, ki se s tem ne bi strinjal in bi rekel: »sprej-

mem, da ima to bitje enako obnašanje in drugačno fizično stanje, ampak morda gre ravno zaradi tega, ker ima bitje drugačno fizično stanje od mojega, na mentalni ravni za nek popolnoma drugačen pojav, ki pa si deli z zavestjo skupno lastnost – obnašanje.« »Ne sprejemem, da sta dva sistema, ki sta funkcionalno izomorfna, realizaciji iste vrste samo zaradi tega, ker sta funkcionalno izomorfna« (Shapiro, 2000: 638). Samo dejstvo, da se neko bitje obnaša enako kot mi, torej še ni zadostni razlog za to, da trdimo, da smo oboji tudi v istem mentalnem stanju (tj. da smo zavestni).

Tukaj zagovornika miselnega eksperimenta iz zamisljivosti postavimo v nemogoč položaj, saj se zdi, da več ne obstaja način, na katerega bi lahko dokazal, da ima tudi takšno bitje zavest – samo obnašanje ni več zadosten pogoj za priznavanje zavesti, in ko imamo dva sistema z različnima fizičnima stanjema, ne moremo sploh več govoriti o tem, ali gre za identičen mentalni pojav. Edini način, na katerega lahko ugovarja, je, da še naprej trdi, da si je takšno bitje mogoče zamisliti – v zamišljanju bitja, ki bi imelo zavest in drugačen fizični sistem, ni nič protislovnega. Takšno bitje je torej logično oz. metafizično zamisljivo, ni pa dokazano, da dejansko obstaja v našem svetu (vendar tudi ni dokazano, da ne obstaja). In zaradi metafizične možnosti takšnega bitja mora biti vsak pogoj za pripis zavesti stroju zgolj zadosten, ne pa tudi nujen. Torej, če se stroj obnaša kot človek in če je v podobnem fizičnem stanju (to pomeni, da »možgani« stroja delujejo na podoben način kot naši), potem nujno velja, da ta stroj misli. Ne velja pa, da stroj misli samo takrat, ko izpolni ta dva pogoja (torej obnašanje in *oblikovni* kriterij), saj si je metafizično možno zamisliti bitje, ki bi imelo zavest in bi se obnašalo enako kot človek, pa bi vendar bilo v različnem fizičnem stanju (možgani takšnega bitja bi delovali na popolnoma drugačen način kot pri človeku). V formalni obliki se zato pogoj za pripis zavesti stroju glasi takole: *če se stroj x obnaša kot človek in ima fizično strukturo procesiranja podobno človeškemu možganom, potem ta stroj misli.*

Prišli smo torej do konjunkcije dveh kriterijev (obnašanja in *oblikovnega* kriterija), ki skupaj tvorita zadosten, nesporen pogoj za priznavanje zavesti stroju, hkrati pa smo pustili odprto možnost, da ima lahko neko bitje zavest tudi brez izpolnitve teh dveh kriterijev, pri čemer pa je breme dokazovanja obstoja takšnega bitja na njegovih zagovornikih.

IV Večvrstna realizacija in Turingov test: preslikava

V drugem delu članka smo ugotovili, da večvrstna realizacija in Turingov test temeljita na isti (behavioristični) predpostavki. V tretjem delu pa smo pokazali, da je behavioristična predpostavka za pripisovanje zavesti pri Turingovem testu pomanjkljiva. Da bi to pomanjkljivost odpravili, smo zato obnašanju dodali *oblikovni* kriterij. Ta dva kriterija v konjunkciji tvorita kakovosten zadosten pogoj za priznavanje zavesti. Če je bila behavioristična predpostavka pomanjkljiva pri Turingovem testu, potem je takšno sklepanje pomanjkljivo tudi v primeru večvrstne

realizacije, saj smo ugotovili, da tudi argument iz večvrstne realizacije uporablja isto behavioristično predpostavko. Zato bomo preslikali rešitev, ki smo jo uporabili pri Turingovem testu, na tezo večvrstne realizacije. Z drugimi besedami, pri Turingovem testu smo to pomanjkljivost odpravili tako, da smo dodali *oblikovni* kriterij. Večvrstna realizacija uporablja enako (behavioristično) predpostavko kot Turingov test, zato ima enake pomanjkljivosti. Torej bomo tudi tezi večvrstne realizacije dodali *oblikovni* kriterij.

Izvirna predpostavka večvrstne realizacije se je glasila takole: če se nek organizem obnaša enako (nerazločljivo) kot človek v stanju bolečine, potem je ta organizem v stanju bolečine. V tretjem delu članka smo ugotovili, da obnašanje samo ni dovolj dober kriterij za pripisovanje identičnega mentalnega stanja, zato je potrebno dodati *oblikovni* kriterij. Tako dobimo naslednji zadostni pogoj za pripisovanje bolečine nekemu organizmu: če se nek organizem obnaša enako (nerazločljivo) kot človek v stanju bolečine *IN ima podobno fizično strukturo kot človek (oblikovni kriterij)*, potem je TA organizem v stanju bolečine. Če sedaj to modificirano predpostavko zapišemo v abstraktni obliki, dobimo naslednje: *če se nek sistem obnaša enako (nerazločljivo) kot človek v stanju x IN ima podoben način delovanja »možganov« kot človek, potem je ta sistem v stanju x*. Če to velja za Turingov test, potem mora veljati tudi za večvrstno realizacijo, ker je predpostavka pri argumentu iz večvrstne realizacije identična tisti pri Turingovem testu.

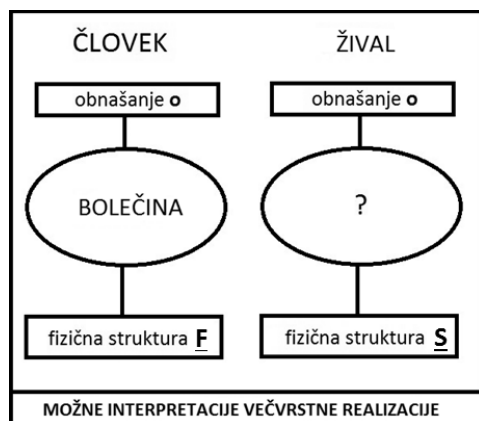
To pa je v popolnem nasprotju s tezo večvrstne realizacije, ki trdi, da lahko vse mentalne pojave večvrstno realiziramo z različnimi fizičnimi sistemi. Zagovornik teze večvrstne realizacije trdi (v logični obliki) dva zadostna pogoja za pripisovanje istega mentalnega stanja (bolečine). Prvič trdi, da: če se nek organizem obnaša enako (nerazločljivo) kot človek v stanju bolečine *in NIMA podobnega sistema delovanja možganov kot človek*, potem je ta organizem v stanju bolečine. Hkrati pa mora zagovornik večvrstne realizacije seveda sprejeti tudi naš izboljšan pogojnik (*če se nek organizem obnaša enako (nerazločljivo) kot človek v stanju bolečine IN ima podoben sistem delovanja možganov kot človek, potem je ta organizem v stanju bolečine*), saj priznava, da tudi soljudje izkušajo bolečino. Zagovornik večvrstne realizacije mora tako popolnoma zanemariti sistem delovanja (fizično stanje) organizma, saj se mora strinjati z obema pogojnikoma in s tem, da če ima organizem lastnost O (obnašanje) in lastnost F (človeku podobna fizična struktura), potem je ta organizem v stanju B (bolečina), in s tem, da če ima organizem lastnost O in nima lastnosti F, potem je ta organizem v stanju B (bolečina). V logičnem zapisu to izgleda takole: $((O \wedge F) \vee (O \wedge \neg F) \rightarrow B)$

Vendar leži razlog za priznavanje $(O \wedge F) \rightarrow B$ v sklepanju na podobnost (analogija), ki je zaradi dveh skupnih lastnosti kakovostno in empirično podkrepljeno. Medtem ko je sklepanje $(O \wedge \neg F) \rightarrow B$, če priznamo, da samo obnašanje ni dovolj za pripisovanje bolečine – kar zaradi ugovora iz črne škatle in ugovora kitajske sobe skorajda moramo –, utemeljeno zgolj z razlogom, da je to metafizično za-

misljivo. »Resnico teze večvrstne realizacije je potrebno oceniti na podlagi empiričnih dokazov« (Shapiro, 2000: 637). Sama metafizična zamisljivost torej ne more biti dovolj za sklepanje na obstoj večvrstne realizacije, saj je teza večvrstne realizacije empirična teza. Če si lahko nekaj metafizično zamislimo, to še ne pomeni, da kaj takšnega v dejanskem svetu tudi obstaja. Kot pravi Shapiro: »Morda pa je samo eno fizično možno elektrokemično stanje, ki lahko realizira bolečino« (Shapiro, 2000: 641). V vsakem primeru je očitno, da je breme dokazovanja – *onus probandi* – kot smo zapisali za Turingov testi, tudi v tem primeru večvrstne realizacije na plečih zagovornikov le-te.

V Kako dokazati večvrstno realizacijo

Kako naj torej zagovornik večvrstne realizacije dokaže svojo tezo, če obnašanje samo ni dovolj dober kriterij, v primeru *oblikovnega* kriterija pa sploh nismo več prepričani, ali res gre za večvrstno realizacijo? Sama metafizična zamisljivost prav tako ne more služiti kot dokaz za obstoj večvrstne realizacije, saj je večvrstna realizacija empirična teza. Prej smo omenili, da je zaradi nezmožnosti vpogleda v izkustvo nekega drugega organizma skoraj nemogoče prepričljivo dokazati večvrstno realizacijo (dvomimo lahko že v enakost izkustva sočloveka), razen seveda preko analogije s samo eno skupno lastnostjo, kar pa se zdi premalo. Kaj torej zagovorniku večvrstne realizacije preostane? Ali kot je zapisal Nagel: »Vedeti hočem, kako je *netopirju* biti netopir. In če si to poskušam predstavljati, me omejujejo možnosti mojega duha, in te možnosti so za nalogo neprimerne. Ne morem je izvršiti niti ob zamišljanju dodatkov k mojemu sedanjemu izkustvu niti ob zamišljanju segmentov, ki mu jih postopoma odvzema, niti ob zamišljanju nekakšne kombinacije dodatkov, odvzemanj in sprememb« (Nagel, 1974/1990: 402).



Shema 2

Tudi če dokažemo, da možgani živali delujejo na popolnoma drugačen način, bo nasprotnik večvrstne realizacije vedno imel možnost, da zanika, da gre za enako subjektivno kvalitativno izkustvo (takšnost). Če dokažemo, da se mehanizmi de-

lovanja možganov živali in ljudi razlikujejo, imamo še vedno na voljo dve interpretaciji (glej Shemo 2). Prvič: gre za večvrstno realizacijo bolečine, saj obstaja (po argumentu iz podobnosti) ena podobnost – enako obnašanje. Drugič: gre za različno izkustvo (žival ne čuti bolečine, temveč nekaj drugega), *ravno zaradi tega*, ker se fizična struktura razlikuje. »Če se realizirajoči vrsti resnično razlikujeta v njihovih vzročno relevantnih lastnostih, potem se zdi, da gre za različni vrsti. Toda če sta različni, potem nista ista vrsta, in tako to ni primer, ko je ena sama vrsta večkrat realizirana« (Shapiro, 2000: 647).

Kako torej rešiti večvrstno realizacijo? Odgovor se morda skriva v plastičnosti možganov. »... možgani so presenetljivo plastični. Preprosto povedano, videti je, da so lahko identična mentalna stanja realizirana v različnih možganskih (fizičnih) stanjih v različnih ljudeh ali pa v različnih možganskih stanjih v isti osebi« (Shapiro, 2000: 642). En način dokazovanja bi torej lahko ležal v nevrofiziologiji. Če bi dokazali, da lahko človek izkusi bolečino, ko je v dveh različnih fizičnih stanjih, ki bi se razlikovali v »vzročno relevantnih lastnostih – tj. v lastnostih, ki pripomorejo k zmožnosti, namenu, cilju in drugim stvarem, ki opredelijo, da je T¹² takšne vrste« (Shapiro, 2000: 646), potem bi lahko rekli, da je bolečino možno večvrstno realizirati. Problem je seveda določiti vzročno relevantne lastnosti – kdaj se mehanizma na elektrokemični ravni razlikujeta do te mere, da lahko temu rečemo večvrstna realizacija, in pa kako takšno mero postaviti.

Shapiro v svojem članku *Multiple Realizations* poskuša prikazati vzročno relevantne lastnosti na primeru različnih odpiralcev buteljk, z različnimi mehanizmi (glej Shapiro, 2000: 643–644). To Shapiro šteje za jasen primer večvrstne realizacije, še vedno pa nimamo kriterija oziroma vsaj paradigmatičnega primera v samem problemu duha in telesa, ki bi nam omogočal jasno razločevanje in ocenjevanje različnih fizičnih sistemov, in preko katerega bi lahko ugotovili, ali gre za primer večvrstne realizacije ali ne – vprašanje je torej, kaj bi lahko navedli kot primer jasne večvrstne realizacije v samem problemu duha in telesa.

Odgovor, ki ga bomo podali, se nanaša na osnovno predpostavko večvrstne realizacije (če se neka stvar obnaša enako (nerazločljivo) kot človek v stanju x, potem je ta stvar v stanju x) in služi zgolj postavljanju kriterija različnosti (vzročno relevantnih lastnosti), ki bi ga morali vzpostaviti (znotraj človeških možganov), da bi lahko zagotovo rekli, da je nek pojav večvrstno realiziran. Če bi torej v človeških možganih našli dva mehanizma, preko katerih bi lahko subjekt izkusil bolečino, in če bi se ta dva mehanizma razlikovala tako, kot se razlikujeta delovanje človeških možganov in delovanje računalnika, potem bi lahko zagotovo rekli, da je takšna bolečina večvrstno realizirana. Zelo verjetno je, da je ta pogoj premočan (je zadosten pogoj, ni pa nujen) – takšne razlike v delovanju možganov najverjetneje ne bomo našli –, vendar nam predstavi jasen primer razlike v dveh mehanizmih,

¹² T je črka, ki jo avtor uporabi za neko mentalno stanje – recimo bolečino.

kjer kot eden izmed mehanizmov nastopajo človeški možgani, ne pa, kot je v navadi filozofov, analogije z odpirachi, mačkami in mišelvami.

Zagovorniki večvrstne realizacije morajo torej na podlagi nevrofizioloških dejstev poiskati različna fizična stanja, ki realizirajo identična mentalna stanja, ugotoviti, kaj neko fizično stanje naredi za bistveno različno od drugega, in postaviti kriterije, na podlagi katerih bomo lahko presojali, ali obstajata dve fizični stanji, ki isto mentalno stanje realizirata na bistveno različen način. Povedati morajo torej, kaj so vzročno relevantne lastnosti, ki naredijo dve fizični stanji različni.

VI Sklep

V članku smo pokazali, da se večvrstna realizacija in Turingov test pri postavljanju lastnih premis naslanjata na isto predpostavko. S pomočjo ugovora iz črne škatle smo prikazali, da je ta (behavioristična) predpostavka pomanjkljiva in je kot takšne ne moremo sprejeti. Zato smo predpostavko v primeru Turingovega testa nadgradili s pomočjo *oblikovnega* kriterija. Nadgradnjo smo nato preslikali na večvrstno realizacijo, kar nas je vodilo v sklep, nasproten sklepu večvrstne realizacije. Ugotovili smo, da je edini razlog za sprejetje večvrstne realizacije njena metafizična zamisljivost, kar pa ni dovolj, saj je teza večvrstne realizacije empirična teza. Zato smo poskušali izdelati načrt, kako bi lahko na kakovosten način dokazali obstoj večvrstne realizacije: s pomočjo nevrofiziologije se poišče dve različni fizični stanji v človeku, obe korelirani z izkustvom istega mentalnega stanja (npr. bolečine). Če bi odkrili dve takšni stanji in prikazali bistvene razlike med njunima delovanjema, bi s tem dokazali tudi večvrstno realizacijo.

Multiple Realizability and Turing's Test: The Transference

The first part of the article introduces the following two concepts: multiple realizability and Turing's test. The aim of the second part is to prove that both Turing's test and multiple realizability use an identical line of reasoning. In the third part we present one of the objections against Turing's test, which shows that the reasoning in part two is fallacious, and show how it can be solved by adding the so-called *design* criterion. This solution is then transferred to the multiple realizability thesis in part four, bringing us to a conclusion opposite to that of multiple realizability. This then leads us to the conclusion that the metaphysical conceivability of multiple realizability is the only reason for accepting it, which, as we show, is insufficient. In the fifth part we present a possible defence strategy of multiple realizability and the problems that may arise.

Keywords: multiple realizability, mental states, Turing's test, black box objection, qualia

Literatura

- Ayer, A. (1978). *Language, Truth and Logic*. Harmondsworth: Penguin.
- Bickle, J. (2016). »Multiple Realizability«. V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (pomlad 2016). Dostopno na: <http://plato.stanford.edu/archives/spr2016/entries/multiple-realizability/>
- Bregant, J. (2004). *Misel kot vzrok: ali so mentalna stanja vzročno učinkovita?* Maribor: Pedagoška fakulteta Maribor.
- Bregant, J. (2010). »Ali lahko stroj misli?«. *Analiza*, 14(4), str. 55–72.
- Bregant, J. (2012). »Turing se je motil.« *Analiza*, 16(1/2), str. 5–31.
- Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell Publishers Ltd.
- Couch, M. (2004). »Discussion: A Defense of Bechtel and Mundale«. *Philosophy of Science*, 71, str. 198–204.
- Klampfer, F., Bregant, J. in Gartner, S. (2008). *Telovadnica za možgane: uvod v kritično mišljenje*. Maribor: Svarog.
- Nagel, T. (1974/1990). »Kako je biti netopir?«. V Hofstadter, D. in Dennett, D. (ur.), *Oko duha*, Ljubljana: Mladinska knjiga, str. 399–410.
- Oppy, G., in Dowe, D. (2016). »The Turing Test«. V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (pomlad 2016). Dostopno na: <http://plato.stanford.edu/archives/spr2016/entries/turing-test/>
- Putnam, H. (1967). »Psychological Predicates«. V Capitan, W. H. in Merrill, D. D. (ur.), *Art, Mind, and Religion*, Pittsburgh: University of Pittsburgh Press, str. 37–48.
- Putnam, H. (1975). »The Nature of Mental States«. V Putnam, H., *Mind, Language and Reality: Philosophical Papers, Vol. 2*, Cambridge: Cambridge University Press, str. 429–440.
- Searle, J. (1981). »Minds, Brains, and Programs«. *Behavioral and Brain Sciences*, 3(3), str. 417–457.
- Shapiro, L. (2000). »Multiple Realizations«. *The Journal of Philosophy*, 97(12), str. 635–654.
- Turing, A. (1950). »Computing Machinery and Intelligence«. *Mind*, 59(236), str. 433–460.
- Zangwill, N. (1992). »Variable Reduction Not Proven«. *Philosophical Quarterly*, 42, str. 214–218.

Alen Lipuš

Filozofska fakulteta Univerze v Mariboru

Pojasnjevanje pojasnjevalne vrzeli

Glavni namen besedila je predstavitev osrednjega problema zavesti, tj. pojasnjevalne vrzeli, in prikaz alternativnega pristopa k njenem pojasnjevanju. Pojasnjevalna vrzel se nanaša na občutek, da nekaj izpuščamo, kadar pojasnjujemo mentalna stanja z znanstveno zgodbo elektro-kemičnih signalov. Ne glede na dovršenost znanstvene zgodbe, občutka »kako je biti« v tem in tem mentalnem stanju z njo ni mogoče zaobjeti, zato je ta znanstvena zgodba nekaj izpušča. Predstavili smo tradicionalno pojmovanje pojasnjevalne vrzeli in pokazali, da ima pomanjkljivosti, katere pa naša strategija, ki razume pojasnjevalno vrzel kot rezultat intuitivnega okvirja, nima. Videli smo, da lahko z intuicijo različnosti bolje pojasnimo obstoj pojasnjevalne vrzeli in s tem obranimo fizikalizem. Nato smo ilustrirali t.i. intuitivno vrzel in predstavili razloge za obstoj intuicije različnosti ter podali pojasnila za njen obstoj, ki so združljiva s fizikalizmom.

Ključne besede: pojasnjevalna vrzel, identiteta duha in telesa, intuicija različnosti, *a priori*, izpeljivost.

0 Uvod

V tem besedilu bomo pokazali, kako je možno problematiko duha in telesa¹ razumeti kot posledico našega intuitivnega okvirja. Kadar se soočamo s sodobno filozofijo, nam je njena snov načeloma predložena skozi probleme. Za slednje se zdi, da stojijo sami zase, da osvetljujejo dejanske vrzeli v svetu. Okoli njih se razvije široka polemika, katere dialektična situacija je tako zapletena, da bralec komaj sledi znova in znova revidiranim stališčem. Zato morda ni prostora in ne časa, da bi se spraševali o sami naravi problemov v filozofiji, o tem, kaj jih legitimizira in zakaj sploh so problemi. Ravno to vprašanje zaseda osrednje mesto v tem besedilu. Iz praktičnih razlogov se bomo omejili zgolj na problem duha in telesa, formuliran kot pojasnjevalna vrzel.

¹ Načeloma govorimo o problemu vzročnosti in problemu zavesti. Temljita na težavah usklajevanja določenih pojavov s stališčem, da je svet v ontološkem smislu fizičen in da je svet vzročno zaprt. Filozofsko delo pri problemu vzročnosti zajema iskanje načina, kako pokazati, da imajo naše misli in namere dejanske učinke na fizičen svet. Delo pri problemu zavesti pa zajema iskanje načina, kako zapreti pojasnjevalno vrzel (Campbell, 2009: 34).

Pojasnjevalna vrzel se nanaša na občutek, da nekaj izpuščamo, kadar pojasnjujemo mentalna stanja z znanstveno zgodbo elektro-kemičnih signalov. Kako je lahko naše specifično, privatno doživljanje sveta nekaj, kar je moč pojasniti z nevrofiziološkim dogajanjem v telesu. Jaz namreč nisem telo, jaz imam telo. Ne glede na dovršenost znanstvene zgodbe, občutka »kako je biti« v tem in tem mentalnem stanju z njo ni mogoče zaobjeti, torej ta znanstvena zgodba nekaj izpušča.² Ilustrirali bomo klasično, osnovno polemiko v filozofiji duha in predstavili problem pojasnjevalne vrzeli. Pokazali bomo, zakaj je splošno sprejeto stališče o izvoru pojasnjevalne vrzeli zmotno in predlagali alternativno pojasnilo. To pojasnilo temelji na misli, da so težave pri sprejetju fizikalizma v filozofiji duha posledica našega intuitivnega (modalnega) aparata in ne nezmožnosti apriornega izpeljevanja mentalnih resnic iz znanstvenih opisov, kot so prepričani nekateri filozofi. Na podlagi tega razvijamo to strategijo naprej tako, da najprej argumentiramo sam obstoj intuicije različnosti duha in telesa, in nato pokažemo, kako je možno to intuicijo uskladiti s fizikalizmom. Predstavimo tri možne načine, kako lahko fizikalisti pojasnijo obstoj intuicije različnosti znotraj fizikalizma in s tem slednjega ubranijo pred številnimi, v klasični polemiki, močnimi napadi.

1 Kaj je pojasnjevalna vrzel?

Če bi želeli popolni opis človeškega ustroja in bi zato uporabili vse znanstveno znanje, bi vseeno imeli občutek, da takšen opis nekaj izpusti. Znanost nam lahko pove vse o živčnih poteh in nevronske aktivnosti, kadar smo v stanju bolečine, toda, zdi se, da nam ne more povedati, kako je biti v tem stanju, in vendar se nam to zadnje vprašanje zdi bistveno za mentalna stanja, kot je bolečina. Zdi se nam, da znanstveni opis bolečine nekaj izpusti, tako kot se Nagelu (1990) zdi, da nikdar ne bomo vedeli, kako je biti netopir. Poanta te linije mišljenja je, da je za mentalna stanja, se pravi stanja, za katera se lahko smiselno vprašamo, kako je biti v teh stanjih, bistvena neka komponenta, ki pa jo tretjeosebni opisi ne zajemajo.

Razširjeno je stališče filozofov, da je za to pojasnjevalno vrzel odgovorna naša nesposobnost apriornega izpeljevanja mentalnih stanj iz fizičnih (Papineau, 2011: 6). Ti filozofi torej zagovarjajo, da če bi lahko *a priori* izpeljevali mentalna stanja iz fizičnih stanj, potem pojasnjevalna vrzel ne bi obstajala in s tem se nam tudi velik del trenutnega problema duha in telesa ne bi zdel problematičen. Prav tako iz tega stališča sledi, da obstajajo druge identitete, ki so *a priori* izpeljive. Navajajo se primeri, kot so voda in H₂O ter toplota in gibanje molekul. Če bi namreč vedeli vse, kar je možno vedeti o H₂O, potem bi lahko, ko bi prvič zagledali reko,

² Takšen psihološki odziv je nekaj kar je prisotno pri večini ljudi, čeprav pa je dobro vprašanje, ali je takšna psihološka situacija prisotna pri vseh kulturah in zgodovinskih obdobjih. Ali gre res za strukturno lastnost kognicije ali pa za socio-lingvistično in historično pogojenost, je sicer pomembno empirično vprašanje, ki pa za našo obravnavo pojasnjevalne vrzeli nima pretirane teže.

dejali, da je to voda. Takšen korak pri scenariju z mentalnimi stanji ni možen. Lahko vemo vse, kar je moč vedeti o fizični zgodbi mentalnih stanj, toda to nam ne omogoča, da mentalna stanja tudi identificiramo. Poleg popolne fizične zgodbe potrebujemo še prvoosebno izpoved, da gre za neko mentalno stanje, ki ga lahko nato povežemo kot »identičnega« s fizičnim dogajanjem ob tistem času.

Razliko v teh dveh scenarijih najbolje povzamemo s pomočjo t.i. opisnih in neopisnih pojmov (Loar, 1990: 89). Naravne vrste, kot je voda, so opisni pojmi, o njih razmišljamo na opisni način. Voda je tekoča, prozorna, brez okusa, brez vonja, teče v rekah, se razširi, ko je zmrznjena. Glede na naše znanje o H₂O lahko pregledamo seznam tekočin in ugotovimo, katera ustreza našemu pojmu vode. Mentalna stanja pa niso opisni pojmi, do njih imamo neposreden dostop. Ne rabimo vedeti vzročne vloge bolečine, da bi jo doživeli. Ne rabimo vedeti, kakšne so nevrofiziološke značilnosti vizualnega izkustva, da vidimo rdeč paradižnik. Za tovrstne informacije je dovolj, da imamo primerno izkustvo. Lahko bi rekli, da sta izkustvena in fizična plat mentalnih dogodkov ločena oziroma izolirana druga od druge, in to nam onemogoča apriorno izpeljivost:³

- i. Bolečina = vzdraženost C-vlaken
- ii. Voda = H₂O

Razlog, zakaj se nam zdi, da pri i) nekaj izpuščamo, je ta, da identiteta i) ni izpeljiva iz fizičnih dejstev, medtem ko identiteta ii) je. Skratka, o mentalnih vsebinah razmišljamo na način, ki nam ne omogoča apriorne izpeljave mentalnih dejstev iz fizičnih dejstev, in to je glede na konvencionalno zgodbo razlog za pojasnjevalno vrzel (Papineau, 2003: 11).

2 Šibkosti splošno sprejetega pogleda na pojasnjevalno vrzel

Ali je zgornje pojasnilo za obstoj pojasnjevalne vrzeli dobro? So naše različne reakcije na znanstvene in duh-telo identitete res odvisne od različnih zmožnosti apriornega izpeljevanja iz fizičnih dejstev? S temi vprašanji bomo problematizirali tradicionalno zgodbo o obstoju pojasnjevalne vrzeli. Naš namen ni zanikati obstoja le-te, ampak natančno preveriti tradicionalno zgodbo in podati ustrezno razlago za njen obstoj. Ne bomo torej zanikali tega, kar opaža Levine (1983), da imamo ob identitetah duh-telo občutek, da je nekaj izpuščeno. Ta občutek oziroma dozdevanje zaenkrat priznavamo kot pristno mentalno dejstvo. Zanimal nas bo odgovor na vprašanje, *zakaj imamo občutek, da identiteta duha in telesa izpušča nekaj (bistvenega)?* Preden pričnemo z odgovorom na to vprašanje, se bomo

³ Chalmers (1996) in Jackson (1994, 1998) sta sicer v manjšini, ko argumentirata, da iz tega sledi neresničnost materializma. Večina filozofov tako brez zadržkov sprejme situacijo, v kateri so identitetne izjave duh-telo resnične, čeprav ne obstajajo opisni pojmi, ki bi omogočili apriorno izpeljavo mentalnih resnic iz fizičnih.

posvetili kritiki ustaljenega pogleda, ki razume pojasnjevalno vrzel kot rezultat pojmovnega dualizma, ki ovira izpeljavo mentalnih stanj iz fizičnih.

Šibkost I: če ustaljen pogled drži, potem bi morali občutek, da je z identiteto duha in telesa nekaj izpuščeno, pridobiti šele tedaj, ko se zavemo, da mentalne resnice niso izpeljive iz fizičnih. Še več, občutek bi moral nastopiti šele tedaj, ko dejansko sprejmemo identiteto duha in telesa kot resnično. Pravzaprav pa imamo ta občutek ne glede na to, ali smo zavezani identiteti duha in telesa, ter ne glede na to, ali se zavedamo nezmožnosti apriorne izpeljivosti. Preprosto se nam zdi ideja, da so naše mentalne vsebine identične in torej ista stvar kot fizični dogodki v telesu, nesprejemljiva. Poleg tega pa, ali res pričakujemo od vsakega, ki se kdaj čudi naravi duha in se sprašuje, zakaj imajo določeni fizični procesi moč proizvajati mentalne vsebine, določeni pa ne, da je poprej uvidel pogoj smiselnosti apriornih identitet, tj. pogoj apriorne izpeljivosti? To se zdi precej optimističen in radodaren pogled na povprečnega misleca.

Šibkost II: obstaja veliko drugih identitet, poleg identiteta duha in telesa, ki niso izpeljive iz fizičnih dejstev, in ob njih nimamo občutka, da nekaj izpuščajo. Identitete, ki vključujejo lastna imena in niso *a priori* izpeljive, ampak vseeno nimamo občutka, da je kaj izpuščeno, kadar naznamo, da je Cicero Tulij. Prav tako nimamo problema, kadar rečemo, da je danes petek (Papineau, 2011: 8). Pri obeh primerih nimamo opravka z opisnimi pojmi, se pravi, nam je apriorna izpeljivost onemogočena, in hkrati nimamo občutka, da smo nekaj (bistvenega) izpustili.⁴

Šibkost III: Morda bi se morali vprašati, kakšna je razlika med aposteriornimi oziroma empiričnimi identitetami in identitetami, ki so (naj bi bile) *a priori* izpeljive. Nemara imajo slednje večjo pojasnjevalno moč, toda kaj bi naj pojasnjevale. En očitni odgovor je, da tovrstne identitete pojasnjujejo same sebe. Toda identitete so nujne in ne potrebujejo pojasnila, saj ne bi moglo biti drugače, kot je. Nesmiselno se je vprašati, zakaj je Cicero Tulij, s tem namreč sprašujemo, zakaj je Cicero isti samemu sebi. Mogoče je pojasnjevalna moč apriornih identitet v tem, da nam pojasnijo, zakaj smo upravičeno prepričani v njihovo resničnost. Toda upravičeno smo prepričani v razna agregatna stanja vode, čeprav so odkrita *a posteriori* in ne *a priori* iz našega pojma vode. Podobno je z bolečino in vzdraženostjo C-vlaken (Papineau, 2003: 8). Namreč to, da bolečina kot mentalni pojem ne razodeva ključnih fizičnih dogodkov, ne pomeni, da nismo upravičeno prepričani, kadar vzročno verigo bolečine identificiramo *a posteriori*.⁵

⁴ Potrebno je tudi poudariti, da ni povsem jasno, kako je lahko identiteta voda = H₂O *a priori* izpeljiva. Podrobnosti te razprave presegajo meje tega besedila, omenimo zgolj to, da nekateri filozofi opozarjajo na težave tradicionalnega pogleda, da je identiteta voda = H₂O *a priori* izpeljiva. Kljub temu se zdi, da te težave ne ogrozijo smiselnosti identitete, oziroma kljub temu nimamo občutka, da identiteta voda = H₂O nekaj izpušča.

⁵ Nasprotniki fizikalizma radi uporabijo strategijo prosojnosti in od mentalnih pojmov zahtevajo spoznavno prosojnost oziroma da morajo razodeti fizično zgodbo v ozadju.

3 Pojasnjevalna vrzel kot intuitivna vrzel

Glede na to vidimo, da apriorne identitete nimajo nobene prednosti pred aposteriornimi, vidimo, da nimajo večje pojasnjevalne moči. Ideja, h kateri bi radi preusmerili našo pozornost, je, da pojasnjevalna vrzel obstaja zaradi tega, ker se ne zmoremo dobro prepričati v resničnost identitete mentalnega in fizičnega. Občutek, da je s fizično zgodbo nekaj izpuščeno, tako izvira iz nečesa osnovnejšega od nezmožnosti apriornega izpeljevanja. Občutek, da je nekaj izpuščeno, izvira iz nam inherentnega dualističnega načina mišljenja (Papineau, 2003; 2006; 2011). Za možganske procese se v znanostih govori, kot da povzročajo, proizvajajo, generirajo in realizirajo mentalna stanja. Kako pa bi drugače govorili o teh stvareh? Že sama smiselnost tega vprašanja kaže na obstoj nekega intuitivnega odpora do ideje identitete duha in telesa ter osvetljuje vpetost dualističnega načina mišljenja v naravnem, opisnem jeziku. V tovrstni uporabi jezika je moč zaznati implicitno ontološko predpostavko, in sicer, da je zavest nekaj drugega od možganov. Če se v primeru »vode« ne sprašujemo, kako je lahko H₂O voda, zakaj se potem enako ne sprašujemo tudi v primeru »zavesti« oziroma zakaj pa se tukaj sprašujemo, kako je lahko zavest vzdraženost C-vlaken?

Intuicija različnosti, se pravi občutek oz. način mišljenja, da sta duh in telo različna, lahko pojasni, zakaj imamo vtis, da pojasnjevalna vrzel obstaja. Tudi za fizikaliste je namreč na intuitivni ravni težko sprejeti teoretično prepričanje, da so mentalni dogodki, kot je na primer bolečina, fizični procesi. Celo največji fizikalisti imajo na neki intuitivni ravni probleme s fizikalizmom, saj je propozicijo *bolečina = vzdraženost C-vlaken* težko sprejeti v celoti. Lahko jo zagovarjamo teoretično in razumsko, toda neka intuitivna vpetost v dualistični nazor se zrcali ali skozi vsakdanji jezik ali pa skozi sam občutek, da bi lahko bilo drugače (Papineau, 2007: 132). To trdovratno prepričanje o ontološki različnosti duha in telesa imenujemo *intuicija različnosti* (ang. *intuition of distinctness*).

Problem pojasnjevalne vrzeli torej vznikne zaradi teoretičnih (tudi morda družbenih in kognitivnih) okvirov, znotraj katerih razmišljamo o duhu, in ena izmed zamolčanih predpostavk je ravno predpostavka o ontološki različnosti duha in telesa. Razmišljanje znotraj te predpostavke organsko vodi do vprašanj (in jezikovno specifičnih formulacij, ki nadalje še utrjujejo to ne-teoretično prepričanje) o odnosu med duhom in telesom. Zdi se, da noben teoretični uvid ne more odpraviti intuicije različnosti, ker je slednja nekaj notranjega temu, kako sami sebe dojemamo kot »biti bitje v svetu«. Pojasnjevalna vrzel obstaja, ker si preprosto ne moremo pomagati, da ne razmišljamo v dualističnih pojmi, ki jo proizvajajo.

4 Ustreznost pristopa z intuicijo različnosti

Najprej bomo na kratko ilustrirali dialektično situacijo problema duha in telesa in pokazali, na kakšne ovire naletimo pri poskusih reševanja tega problema. Pri tem

bomo pokazali, da je zanikanje spoznavne veljavnosti intuicije različnosti boljša strategija, kot pa zanikanje ostalih dveh intuicij, tj. intuicije vzročnosti in intuicije o vzročni zaprtosti sveta. Začetna izhodiščna točka je dilema, ali je zavest le učinek možganov ali pa je neodvisna (nefizična)⁶ substanca. Zavest lahko opredelimo kot nekaj, kar označuje naše misli, občutke, zaznave, predstave, skratka, izkustva sleherne vrste (Strawson, 2006: 10). Polemiko o naravi zavesti in naravi odnosa med duhom in telesom lahko na grobo opišemo s tremi intuicijami, ki so izražena s tremi stališči (Robinson, 2007: 324).

1. *Intuicija vzročnosti*. Predpostavlja vzročno zvezo med zavestjo in možgani. Intuicija vzročnosti torej pomeni, da so vsaj nekateri mentalni dogodki v vzročni zvezi s fizičnimi (Davidson, 1988: 8). Gre torej za povzročanje fizičnega z mentalnim.
2. *Intuicija o vzročni zaprtosti sveta (popolnost fizike)*. Vsak fizičen pojav ima zadosten fizičen vzrok. Nekateri vidijo v tem problem, saj sprejetje popolnosti fizike implicira, da je zavest fizičen pojav. Posledično bi naj morali pojasniti, kako je možno, da ima fizičen pojav lastnosti, kot so mišljenje, občutki in subjektivnost.⁷
3. *Intuicija različnosti*. Imamo občutek, da so mentalne vsebine nekaj poleg našega telesa, da jih ni možno popolnoma zaobjeti s fizičnim opisom.

Videli bomo zakaj, če zanikamo prvo intuicijo, pristanemo pri *epifenomenizmu*,⁸ in zakaj, če zanikamo drugo intuicijo, pristanemo v interakcionizmu.⁹ Med sodobnimi dualisti skoraj ni mogoče najti koga, ki bi zagovarjal interakcionizem (Robinson, 2007: 324). Očitno se celo dualistom zahteva po popolnosti fizike zdi prevelik filozofski davek. Pristanemo torej v epifenomenalizmu. Toda, če je epifenomenalizem resničen in je naša spoznavna situacija takšna, kot je (zavest se nam zdi uganka), potem zavest¹⁰ oziroma karkoli mentalnega ne more biti vzrok za našo spoznavno situacijo.¹¹ Če pa mentalni svet ne more biti vzrok za našo

⁶ V filozofiji duha razumemo fizikalizem kot stališče, katerega osrednja ideja je ta, da so mentalna stanja v nekem smislu fizična stanja (Bregant, 2004: 34). In obratno, dualizem, razumemo kot stališče, da so mentalna stanja v nekem smislu različna od fizičnih stanj.

⁷ Galen Strawson se s tem ne strinja. Po njegovem to vprašanje vsebuje predpostavko, da imamo o fizični stvari popolno znanje, le tako je namreč možen sklep – na katerega nas to vprašanje napeljuje –, da zavest ne more biti fizična (Strawson, 2006: 11–12).

⁸ Epifenomenalizem je stališče, ki predpostavlja, da mentalni dogodki niso vzročno učinkoviti.

⁹ Interakcionizem je stališče, ki predpostavlja, da sta mentalno in fizično dve različni substanci, med katerima obstaja vzročna zveza.

¹⁰ Tu imamo v mislih samo metafizično naravo mentalne stvari v najsplošnejšem pomenu.

¹¹ Sklepati iz naše spoznavne situacije na podobno metafiziko je v tem primeru sporno, saj epifenomeni po definiciji nimajo vzročne moči.

spoznavno situacijo, potem vpeljava mentalnosti kot epifenomena nima nobene spoznavne vrednosti in pojasnjevalne moči, saj ne more pojasniti naše spoznavne situacije oz. ne more pojasniti pojasnjevalne vrzeli. Torej epifenomenalizem ne more biti resničen in intuicije o vzročnosti ne moremo zanikati (Papineau, 2011: 15). Glede na to je torej najbolje, da izberemo tretjo možnost in tako zanikamo spoznavno veljavnost intuicije, ki nas napeljuje na misel, da sta duh in telo različna v ontološkem smislu. Če namreč zanikamo intuicijo o vzročnosti, nas to vodi v epifenomenalizem,¹² če pa zanikamo intuicijo o popolnosti fizike, pristanemo v interakcionizmu, stališču, ki ni razširjeno niti pri dualistih. Tako pristanemo pri tretji možnosti, da zanikamo¹³ intuicijo različnosti kot spoznavno veljavno. Pokazali bomo, da ima ta strategija veliko pojasnjevalno moč, saj lahko najbolje razloži obstoj pojasnjevalne vrzeli.¹⁴ Glede na njeno pojasnjevalno moč lahko vidimo smiselno strategijo za fizikaliste, da zanikajo spoznavno veljavnost intuicije različnosti in s tem odpravijo velik del problemov fizikalizem.

5 Obstoj intuicije različnosti

Dosledni filozofi se zavedajo, da morajo zanikati možnost filozofskih zombijev (mikrofizični dvojniki brez mentalnih stanj), toda na intuitivni ravni dojemajo zombije kot metafizično možne. In ravno ta raven je tista, ki upraviči obstoj neke vrste intuitivne odpornosti do ideje mentalno-fizične identitete. Če bi namreč filozofi popolnoma sprejeli identiteto duha in telesa, potem se jim zombiji ne bi kazali kot metafizično možni. Na reflektivni ravni se fizikalisti zavedajo, da morajo zanikati modalno smiselnost dualističnih scenarijev, toda sam obstoj polemike okoli zombijev kaže, da se na intuitivni ravni zombiji večini filozofom kažejo kot koherentna ideja. To, da polemika o metafizični možnosti filozofskih zombijih na nek način dokazuje intuitiven odpor (intuicija različnosti) do psihofizične identitete najbolje prikažemo z analognim primerom Cicera in Tulija.

Ali je možno, da Cicero ni Tulij oz. je lahko Cicero prisoten, medtem ko je Tulij odsoten? Zdi se nam, da to ni možno, saj gre vendar za eno in isto osebo, in predlagati, da bi ista oseba hkrati bila prisotna in odsotna je nesmisel. Toda, zakaj je potem primer duh-telo odprt za zombi ugovor? Če so mentalna stanja identična s

¹² Prav tako paralelizem, okazionalizem, prestabilirana harmonija, ampak to so stališča, ki se še redkeje pojavljajo v sodobnih razpravah.

¹³ V tem besedilu se ne bomo ukvarjali z dokazovanjem zmotnosti intuicije različnosti. Naš cilj je predstaviti način pojasnjevanja problemov v filozofiji duha s strategijo intuicije različnosti. Pri tem se lahko vprašamo, ali je intuitivna smiselnost dualizma dokaz zanj. To bi bilo morda res, če bi dualizem bolje pojasnil našo spoznavno situacijo oziroma obstoj intuicije različnosti. Zgoraj pa smo videli, da dualizem ne nudi dobre pojasnitve za intuicijo različnosti. Dualisti s tem še niso pobiti, lahko zagovarjajo tezo, da ravno obstoj intuicije različnosti pomeni dober indic za dualizem (glej Robinson, 2007).

¹⁴ Lahko tudi odgovori na Kripkejev izziv (glej Lipuš, 2015), ki je sicer zelo soroden problemu pojasnjevalne vrzeli. Preko tega lahko posledično pojasni smiselnost dualističnih modalnih scenarijev.

fizičnimi stanji, potem bi morali zombiji izzvati podobne intuicije o nesmislu tega scenarija, kot se to zgodi pri primeru Cicero je Tulij. Zakaj pri duhu in telesu, ob predpostavki neke relacije identitete, enako govorimo o tem, da lahko nekaj obstaja brez samega sebe, da je lahko nekaj prisotno in odsotno hkrati. Zakaj potem takem tega scenarija ne dojemamo kot nekoherentnega? Mogoče zato, ker nismo popolnoma prepričani v fizikalizem oz. fizikalisti niso popolnoma prepričani v fizikalizem? Če bi bili, potem zombi sveta ne bi smeli niti prvotno (intuitivno) dojemati kot koherentnega, tako kot scenarij, ko je Cicero prisoten brez Tulija, dojamemo kot nekoherentnega takoj, ko spoznamo, da Cicero je Tulij. Če bi namreč bili prepričani v identiteto duha in telesa, potem bi se zombiji zdeli enako absurdni kot prisotnost Cicera in odsotnost Tulija (Papineau, 2011: 12).

Še več, primer Cicero je Tulij je sestavljen iz opisnih pojmov:

- i) Cicero = veliki rimski retorik
- ii) Tulij = veliki rimski državnik.

Pri tem primeru imamo opravka z opisnimi pojmi in vidimo, da je primer veliki rimski retorik (opis Cicera) ni veliki rimski državnik (opis Tulija) smiselno, do njega nimamo intuitivnega odpora. Torej bi imeli pojasnilo, če bi se nam identiteta Cicero je Tulij zdela sporna. Toda ta identiteta se nam ne zdi sporna, medtem ko se nam identiteta duh je telo zdi, pa čeprav, kot smo že pokazali, pri njej ne nastopajo opisni pojmi, ki bi nudili neko pojasnilo glede dozdevne napačnosti identitete. Se pravi, da imamo na eni strani opisne pojme in kontingentno referiranje in na drugi strani ne-opisne pojme in neposredno referiranje, in kljub temu dojemamo identitete iz prvega primera kot neoporečne, medtem ko dojemamo identitete iz drugega primera kot vsaj prvotno metafizično napačne. Glede na povedano bi bilo lažje sprejeti, da Cicero ni Tulij, kot pa, da bolečina ni vzdraženost C-vlaken, pravzaprav pa je težje. Torej, kljub danosti opisnih pojmov je reakcija na primer Cicero ni Tulij, da njuno ne-identiteto zavrnemo kot nesmiselno. Toda, celo brez danosti opisnih pojmov v primeru duh je telo, filozofi (fizikalisti) nimajo (prvotnih oz. intuitivnih) težav s primerom zombijev. Iz tega sledi, da filozofi (fizikalisti) preprosto niso, na intuitivni ravni, popolnoma prepričani v fizikalizem (Papineau, 2011:13).

6 Pojasnjevanje intuicije različnosti

Fizikalisti morajo sedaj, če se želijo sklicevati na intuicijo različnosti kot na izvor pojasnjevalne vrzeli, pokazati ne samo, kako je obstoj te intuicije združljiv s fizikalizmom, temveč tudi pojasniti, zakaj ta intuicija sploh obstaja. Zgoraj smo ugotovili, da dualizem ne more nuditi ustreznega pojasnila za njen obstoj, saj dualisti pristanejo v epifenomenalizmu, ki pa ne more biti vzrok za našo mentalno situacijo in je tako ne zmore pojasniti. Ali to zmore fizikalizem? V nadaljevanju bomo

obravnavali tri možne pristope znotraj fizikalizma pri pojasnjevanju intuicije različnosti.

6.1 Nagelova opomba

Eden izmed možnih odgovor na vprašanje, zakaj sploh imamo intuicijo različnosti in posledično pojasnjevalno vrzel, sloni na različnih tipih domišljije, kot jih omenja Nagel v enajsti opombi v članku *Kako je biti netopir (1990)*:¹⁵

Teorija, ki bi razložila, kako je nujno razmerje duh – možgani, bi nam še zmeraj pustila Kripkejev problem razlage, zakaj se navsezadnje prikazuje kot kontingentno. To težavo se mi zdi mogoče premagati na naslednji način: lahko si nekaj zamislimo in si to predstavljamo bodisi zaznavno, sočuteče ali pa simbolno. Ne bom skušal razložiti, kako deluje simbolna domišljija, a del tistega, kar se zgodi v preostalih dveh primerih, je tole: če si nekaj predstavljamo zaznavno, se pripravimo do zavestnega stanja, *ki je podobno stanju v katerem bi bili, če bi to nekaj dejansko zaznavali. Če si nekaj predstavljamo sočutečno, se pripravimo do zavestnega stanja, ki je podobno stvari sami. (Ta metoda je uporabna edinole za zamišljanje mentalnih stanj in dogodkov – lastnih ali tujih.)* Ko si skušamo predstavljati mentalno stanje, ki se pojavlja brez z njim povezanega stanja možganov, si najprej sočuteče predstavljamo pojavnost mentalnega stanja: se pravi, da se preselimo v stanje, ki ga mentalno predstavlja. Ob istem času si skušamo zaznavno predstavljati nenavzoče s tem povezanega fizičnega stanja tako, da se predstavljamo v novo stanje, nepovezano s prvim: tako, ki predstavlja tisto, v katerem bi bili, če bi dosegli nenavzoče fizičnega stanja. Kjer je zamišljanje fizičnega zaznavno in mentalnega sočuteče, se nam zdi, da si lahko zamišljamo vsako izkustvo, ki se pojavlja brez z njim povezanega stanja možganov, in obratno. *Zveza med njima se bo zaradi neodvisnosti različnih tipov imaginacije zdela kontingentna, četudi je nujna.* (Nagel, 1990: 407–408)

Hill (2014) na podlagi te opombe zasnuje način pojasnjevanja vrzeli med duhom in telesom. Obstoj pojasnjevalne vrzeli je motiviran s strani psihičnih refleksov ali intuicij, ki jih imamo, ko razmišljamo o modalnih scenarijih. Problem za fizikalista je ta, da mora trditi, da je zveza med fizičnim in mentalnim nujna, čeprav se nam zdi, da ni. Če namreč lahko pokažemo, da sta propoziciji a) identiteta duha in telesa in b) zanikanje te identitete združljivi,¹⁶ potem smo naredili napredek pri obrambi fizikalizma. Nagel v opombi nakaže način, kako sta propoziciji lahko združljivi, pri tem pa se opira na tri tipe domišljije:

¹⁵ V slovenskem prevodu gre za deseto opombo. Slovenski prevod je na mestih netočen. Z ležečo pisavo so označena mesta, kjer smo slovenski prevod spremenili v skladu z izvirnim besedilom.

¹⁶ Združljivi v ne-logičnem smislu, drugače bi kršili zakon neprotislovja. Poleg tega pa nimamo nobene motivacije zagovarjati obrambo parakonsistentne logike, pa tudi ta naloga presega namene tega besedila. To, da sta propoziciji in njeno zanikanje lahko združljivi, razumemo na spoznavni način. Zanikanje propozicije a) tako razumemo kot proizvod intuitivnega aparata, namreč kot videz oz. privid (iluzijo).

- a) zaznavna domišljija (kadar si zamišljamo možganska stanja);
- b) sočuteča domišljija (kadar si zamišljamo mentalna stanja);
- c) simbolna domišljija.

Najprej Nagel omenja primer breztelesnosti, kjer si zamišljamo mentalno stanje stvari brez spremljajočih fizičnih stanj. Po njegovem v tem primeru združimo dva načina predstavljanja in sicer sočuteči in zaznavni način. Pod sočutečim načinom si zamišljamo situacijo, kjer je prisotno neko mentalno stanje, medtem ko si pod zaznavnim načinom zamišljamo situacijo, kjer je odsotno neko možgansko stanje. Ta dva domišljijška načina nekako »zlepimo« skupaj in tako dobimo neko predstavo. Dejstvo, da se nam zamišljena predstava zdi koherentna, ne pomeni, da je takšno bitje dejansko možno. Po Nagelu moramo to sposobnost koherentnega zamišljanja prisoditi prej omenjenim, neodvisnim tipom domišljanja (Hill, 2014: 120).

Pri zagovarjanju fizikalizma Hill razvija Nagelovo opombo in izpostavi pomembno vprašanje, in sicer: »Kaj pomeni zamišljati si možgansko stanje?« Slednjih namreč ne zaznavamo; kako torej lahko njihovo zamišljanje spada pod zaznavni način domišljanja? Hill razlikuje med posredno in neposredno zaznavo, ali na kratko: možganska stanja si lahko zamišljamo tako, da si zamišljamo, da zaznavamo aparat, ki odčitava možganske procese. V nadaljevanju poda nov način pojasnjevanja intuicij o breztelesnosti in intuicij o zombijih. Tako želi pojasniti vrzel med duhom in telesom:

Rečemo lahko, da so mehanizmi, ki so odgovorni za intuicije o breztelesnosti in zombijih, člani družine F-mehanizmov, ki jih lahko opišemo tako: če je M nek psihološki mehanizem, potem je M član družine F, če M deluje tako, da združuje predstave odsotnosti oziroma prisotnosti zdravorazumskih pojavov (t.i. pojavov, do katerih imamo dostop preko zdravorazumskega zavedanja, kot sta npr. introspekcija in vizualna zaznava) in predstave odsotnosti oziroma prisotnosti teoretičnih pojavov (t.i. pojavov, do katerih imamo dostop zgolj preko teoretske konstrukcije in laboratorijskih pripomočkov). (Hill, 2014: 125)

V skladu s Hillovim razmislekom imajo torej zamišljene propozicije dvoje vsebin, eno teoretično in eno zdravorazumsko. Zdravorazumska komponenta se nanaša na vsebino, ki jo pridobimo ali z introspekcijo ali z zaznavo. Teoretična komponenta pa zadeva vsebino, ki nam je dostopna zgolj preko neke teorije in tehnologije. Za Hilla torej velja, da v primerih, ko si zamišljamo zombije, združujemo zdravorazumsko vsebino s teoretično vsebino in tovrstna zamišljanja za Hilla niso verodostojna. Z razlikovanjem med zdravorazumsko in teoretično vsebino lahko pojasnimo modalne intuicije, na primer glede ločljivosti vode in H_2O , vročine in gibanja molekul ter svetlobe in elektromagnetnega sevanja. Z razlikovanjem med teoretičnimi in zdravorazumskimi komponentami aposteriornih resnic pojasnimo pojasnjevalno vrzel kot rezultat zmotnih (modalnih) intuicij. Iz tega, da

imamo zmotne modalne intuicije na primer glede ločljivosti vode in H₂O, vročine in gibanja molekul ter svetlobe in elektromagnetnega sevanja, lahko nato sklepamo, da imamo podobno zmotne modalne intuicije tudi glede duha in telesa.

6.2 Kultura

Ali je možno, da smo intuitivni dualisti zaradi načina vzgoje? Mogoče je za našo zmožnost zamišljanja dualističnih scenarijev kriva z Descartesovo intelektualno tradicijo obogatena kultura? Zdi se, da je (bil) dualizem v zahodni kulturi lep čas samoumeven. Nekateri filozofi, kot npr. Stephen Yablo, so prepričani, da se bo intuitivna smiselnost dualističnih scenarijev, kot so zombiji, s tehnološkim in znanstvenim napredkom razblinila (Yablo, 2000: 119). Študija Blooma in kolegov (Bloom, et al. 2014: 7) je pokazala, da so posamezniki in posameznice, ki so, ali odrasli v zahodni kulturi ali pa bili in so izpostavljeni oz. imajo stike z zahodno kulturo, načeloma manj nagnjeni k dualističnemu načinu mišljenja, kot pa tisti, ki niso bili nikdar v stiku z zahodno kulturo. To pojasnjujejo z izobraževalnim sistemom, ki temelji na predpostavki fizičnega monizma, kot se ta uporablja v znanostih na sploh (Bloom, 2013: 7). Toda, hkrati Bloom in ekipa zagovarjajo tezo, da je dualizem intuitivna in ne kulturna inklinacija. Glede na njihovo raziskavo ima namreč predpostavka, da je dualizem intuitivna značilnost večjo pojasnjevalno moč.¹⁷ Glede na to študijo se dualistične intuicije razvijajo z minimalnim kulturnim učinkom. Seveda to ne pomeni, da kultura ne oblikuje način, kako mislimo in vidimo svet (Bloom, 2013: 7). Torej, če je dualizem intuitivna značilnost našega uma, potem gre za neko strukturno potezo naše mentalne arhitekture, ki ne more biti popolnoma negirana s kulturo.¹⁸

6.3 Strukturna značilnost

Znani ameriški razvojni psiholog z Univerze Yale Paul Bloom je pred več kot desetimi leti izdal knjigo *Descartes' Baby*, v kateri zagovarja tezo, da je naša naklonjenost dualizmu vpeta v naš kognitivni ustroj. Razvija misel, da je dualizem naravna posledica razvoja človeka, saj imamo dva različna kognitivna sistema, ki sta vsak zase odgovorna za dožemanje fizičnega in mentalnega (Bloom, 2004: 33). Tako naj bi ljudje imeli sistem za branje misli, s katerim pripisujemo mentalna stanja intencionalnim agentom, in sistem za fiziko, s pomočjo katerega razmišljamo o vsakdanjem, čutnem svetu. Glede na vrsto pojava, se aktivira primeren sistem, vendar nikdar oba hkrati, kar je tudi razlog – tako sklepa Bloom –, da se nam duh in telo zdita nezdružljiva (Bloom, 2004: 44).

¹⁷ Z njo so lahko bolje pojasnjevali rezultate in jih tudi bolje napovedovali.

¹⁸ Če gre za strukturno značilnost, potem bi pričakovali, da bo vsaj problem duha in telesa, tako kot ga razumemo danes, oz. intuicija različenosti prisotna skozi vso filozofijo oz. zgodovino.

Papineau predlaga, da smo pod vplivom *antipatetične zmote*, kadar razmišljamo o mentalnih vsebinah. Kadar namreč razmišljamo o mentalnih vsebinah uporabljamo pojavne pojme,¹⁹ z razliko od materialnih pojmov, ki jih uporabljamo, kadar razmišljamo o materialnih dogodkih. Tako se lahko zgodba o bolečini drastično razlikuje, kadar jo pripovedujemo s pomočjo uporabe prvoosebni pojavnih pojmov in pa tretjeosebni materialnih pojmov. Navzlic seznanjenosti o identiteti duha in telesa, se nam zdi, da zaradi različne zgodbe materialni pojmi in pojavni pojmi ne morejo imeti istega referenta. Ta vtis Papineau poimenuje antipatetična zmeta, za katero misli, da je vir intuicije različnosti (Papineau, 2007: 137).

Spet tretji filozofi se sklicujejo na to, da so mentalni pojmi posebni, *epistemološko neprosojni*²⁰ pojmi. To pomeni, da od mentalnih pojmov ne smemo zahtevati, da nam razkrivajo bistvene značilnosti referenta. Tako skušajo obraniti fizikalizem pred nastalo zagato, pri kateri se nam prav fizikalizem zdi napačna doktrina (Papineau, 2006: 8). Na drugi strani pa so filozofi, kot denimo David Chalmers (1996), ki ravno to značilnost mentalnih pojmov jemljejo kot odločljiv argument proti fizikalizmu. Toda v tem primeru lahko vprašanje, ali je epistemološka neprosojnost mentalnih pojmov argument proti fizikalizmu ali pa je vseeno združljiva z njim, postavimo na stran, saj nas zanima vprašanje, ali lahko z idejo epistemološke neprosojnosti pojasnimo nastalo pojasnjevalno vrzel, kadar predpostavljamo fizikalizem. Tako si lahko zastavimo vprašanje: ali smo ljudje intuitivni dualisti zaradi tega, ker mislimo, da bi naj s pomočjo pojavne refleksije spoznali bistveno naravo pojavnih stanj? Če je to res, potem ta pristop slika povprečnega misleca kot tako filozofsko kot introspektivno sofisticiranega, kar pa ne moremo trditi oz. vsaj zdi se, da ljudje v vsakdanjem življenju nismo občutljivi za takšne filozofske podrobnosti (Papineau 2011: 17).

7 Sklep

Prikazali smo osnovno dialektično situacijo v sodobni filozofiji duha in videli, da je izmed treh metafizičnih postavk v filozofiji duha najbolj opustiti idejo o različnosti duha in telesa. Če zanikamo stališče o vzročnosti, nas to vodi v epifeno-menalizem, za katerega smo pokazali, da je nevzdržen. Če pa zanikamo stališče o vzročni zaprtosti sveta oziroma o popolnosti fizike, pa pristanemo v interakcionizmu, s katerim se sicer nismo podrobneje ukvarjali, ker gre za stališče, ki ni priljubljeno niti pri zagovornikih dualizma. Tako smo sklenili, da se je bolje osredotočiti na raziskovanje tretje možnosti, tj. intuicije o različnosti duha in telesa in

¹⁹ Pojavni pojmi so pojmi, ki označujejo kvaliteto značilnost izkustev.

²⁰ Epistemološka prosojnost pomeni, da dobimo znanje o vseh bistvenih značilnosti stvari, ko pridobimo pojem, ki jo označuje. Epistemološka neprosojnost mentalnih pojmov za fizikaliste pomeni, da nam v trenutku, ko dobimo nek mentalni pojem (npr. prvič izkusimo bolečino), ni omogočen dostop do vseh njenih lastnosti, npr. ob občutenju bolečine ne spoznamo informacij o vzdraženosti C-vlaken.

zanikanja njene modalne veljavnosti. Najprej smo predstavili klasično stališče glede pojasnjevalne vrzeli in ga zavrgli. Nato smo vpeljali novo strategijo odpravljanja pojasnjevalne vrzeli, ki temelji na intuiciji različnosti. Na kratko smo se dotaknili vprašanja, ali je intuitivnost dualizma pravzaprav dokaz zanj, in videli, da to ni res, saj dualizem ne zmore zadostno pojasniti te intuicije. Videli smo, da lahko s pomočjo intuicije različnosti najboljše pojasnimo, zakaj pojasnjevalna vrzel obstaja. Toda, če je glavni krivec za pojasnjevalno vrzel intuicija različnosti in če fizikalisti želijo ubrati to pot pri zagovarjanju lastnega stališča, potem morajo podati zadovoljivo pojasnilo za sam obstoj intuicije različnosti. Takšno pojasnilo naj bi intuicijo različnosti uskladilo s fizikalizmom. Tako smo za pojasnjevanje obstoja intuicije različnosti predstavili tri možne pristope, ki so sicer na meji z empirično znanostjo, kjer pa se dogajajo zanimive raziskave. Toda vloga empirije je v tem primeru rezervirana zgolj za izbor pravega pojasnila intuicije različnosti, medtem ko je za fizikalista najbolj pomembno le to, da intuicija različnosti obstaja in, da so na voljo pojasnila zanjo, ki so v skladu s fizikalizmom.

Explaining the Explanatory Gap

The main purpose of this text is to present the main problem of consciousness, the so-called explanatory gap problem, and to provide an alternative explanatory approach. The explanatory gap refers to the feeling that something is left out when we explain our mental states with the scientific story of electro-chemical impulses. No matter how sophisticated the scientific story is, the feeling of "how it is like" to be in such and such mental state cannot be encompassed by it, therefore the story leaves something out. We present the traditional understanding of the explanatory gap and show its shortcomings. Then we turn to our proposed strategy which eschews them by understanding the explanatory gap as the result of an intuitive framework. We saw that with the intuition of distinctness we can devise a better understanding of the explanatory gap and at the same time defend physicalism. We then focus on the illustration of the so-called intuitive gap, present the reasons for the existence of the intuition of distinctness and put forward possible explanations for it, all of which are compatible with physicalism.

Keywords: explanatory gap, mind-body identity, intuition of distinctness, *a priori*, derivability

Literatura

Bloom, P. (2004). *Descartes' Baby*. New York: Basic Books.

Bloom, P., Birch, S., Chudek, M., Henrich, J. in McNamara, R. (2013).

»Developmental and Cross-Cultural Evidence for Intuitive Dualism« Dostopno na:

http://www2.psych.ubc.ca/~henrich/pdfs/ChudekEtAl_IntuitiveDualism_WorkingPaper_June2014.pdf [3. 7. 2016].

Bregant, J. (2004). *Misel kot vzrok*. Maribor: Pedagoška fakulteta.

- Campbell, N. (2009). »Why We Should Lower Our Expectations about the Explanatory Gap«. *Theoria*, 75, str. 34–51.
- Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
- Davidson, D. (1988). *Raziskave o resnici in interpretaciji*. Ljubljana: Studia humanitatis.
- Hill, Ch. (2014). *Meaning, Mind and Knowledge*. Oxford: Oxford University Press.
- Jackson, F. (1982). »Epiphenomenal Qualia«. *Philosophical Quarterly*, 32: 127–136.
- Jackson, F. (1986) »What Mary Didn't Know«. *Journal of Philosophy*, 83: 291–295.
- Levine, J. (1983). »Materialism and Qualia: The Explanatory Gap«. *Pacific Philosophical Quarterly*, 64, str. 354–361.
- Lipuš, A. (2015). »Odgovor na Kripkejev izziv: kako se nam lahko zdi a posteriorna nujnost kontingentno resnična?« *Analiza*, 19: 109–125.
- Loar B. 1990. »Phenomenal States«. *Philosophical Perspectives*, 4, str 81–108. Dostopno na:
<http://www.nyu.edu/gsas/dept/philo/courses/consciousness97/papers/loar.html> [4.7.2016].
- Nagel, T. (1990). »Kako je biti netopir?« V Dennet, D. in Hofstadter R. (ur.), *Oko duha*, Ljubljana: Mladinska knjiga, str. 399–410.
- Papineau, D. (2003). »Confusions about Consciousness«. *Richmond Journal of Philosophy*, 5, str. 6–12.
- Papineau, D. (2007). »Phenomenal and Perceptual Concepts«. V Alter, T. in Walter, S. (ur.), *Phenomenal Concepts and Phenomenal Knowledge*, Oxford: Oxford University Press, str. 117–144.
- Papineau, D. (2011). »What Exactly is the Explanatory Gap?«. *Philosophia*, 39, str. 5–19.
- Robinson, S.W. (2007). »Papineau's Conceptual Dualism and the Distinctness Intuition«. *Synthesis Philosophica*, 44, str. 319–333.
- Strawson, G. (2006). »Realistic Monism: Why Physicalism Entails Panpsychism«. *Journal of Consciousness Studies*, 13, str. 3–31.
- Yablo, S. (2000). »Textbook Kripkeanism and the Open Texture of Concepts«. *Pacific Philosophical Quarterly*, 81, str. 98–122.

Blaž Cotman

Filozofska fakulteta Univerze v Ljubljani

Izziv »superintelligence« – komentar k Bostromovi knjigi *Superintelligence: Paths, Dangers, Strategies* (2014)

Članek orisuje zahteven izziv, ki ga predstavlja možnost »superintelligence«. Sledeč nekaterim ključnim poudarkom švedskega filozofa Nicka Bostroma, ki jih najdemo v njegovi knjigi *Superintelligence: Paths, Dangers, Strategies* (2014), skušam začrtati nekaj relevantnih filozofskih (metafizičnih, epistemoloških in etičnih) vprašanj, s katerimi smo soočeni, ko razmišljamo o umetnih agensih, ki presegajo človeško kognitivno zmogljivost na domala vseh pomembnih področjih. Začnem s kratko obnovo Bostromove »Nedokončane basni o vrabcih« in nadaljujem z opredelitvijo pojmov, kot so inteligenca, eksplozija intelligence in superinteligence. Zatem pojasnim, zakaj velja razlikovati med vprašanjem o tem, ali je neka entiteta inteligentna, in vprašanjem, ali poleg tega poseduje tudi zavedajoča se mentalna stanja, ter naznačim možne tehnološke poti do superintelligence in mnogovrstne negotovosti, ki obdajajo z njo povezana vprašanja. Trdovratnost in nerazrešenost teh vprašanj sugerirata, da preprostih rešitev preprosto ni in da imamo opraviti s perečim problemom, ki zahteva temeljito nadaljnjo filozofsko, znanstveno in nenazadnje tudi širšo družbeno pozornost.

Ključne besede: superinteligence, umetna inteligenca, zavest, Nick Bostrom

Uvod

V pričujočem članku bom predstavil nekaj pomembnih vprašanj in dilem, s katerimi nas v svoji novejši knjigi z naslovom *Superintelligence: Paths, Dangers, Strategies* sooča švedski filozof Nick Bostrom.¹ V njej pretresa kompleksne izzi-

¹ N. Bostrom (roj. 1973), nekakšen »znanstveni prerok« (Bostrom, 2009: 58), je direktor Oxfordskega inštituta za prihodnost človeštva, eden od ustanovnih članov Svetovne transhumanistične zveze, avtor knjig

ve, ki jih predstavlja možnost superinteligence, in na zvečine zgledno poglobljen način premišlja o tem, kako bi se lahko nanje najboljše odzvali (če in ko bi do »vzleta« superinteligence sploh prišlo, seveda, kar pa se zdi avtorju *zelo* verjetno, in sicer v ne tako daljni prihodnosti²). Po Bostromovem prepričanju je to »bržkone najpomembnejši in najstrašljivejši izziv, s katerim se je človeštvo kdajkoli soočilo« (Bostrom, 2014: vii).³

Ker gre za razmeroma obsežno, večplastno in zapleteno delo, ki ga na tem mestu niti slučajno ne morem zadovoljivo, kaj šele izčrpno, predočiti in analizirati, se bom v nadaljevanju omejil na oris tistih problematik, ki se mi zdijo še posebej relevantne in so iz takšnih ali drugačnih razlogov pritegnile mojo pozornost. Začel bom s kratko obnovo Bostromove »Nedokončane basni o vrabcih« (»The Unfinished Fable of the Sparrows«) in nadaljeval z opredelitvijo pojmov, kot so inteligenca, eksplozija inteligence in superinteligence. Zatem bom pojasnil, zakaj velja skrbno razlikovati med vprašanjem o tem, ali je neka entiteta inteligentna, in vprašanjem, ali *poleg tega* poseduje tudi zavedajoča se mentalna stanja, ter naznačil možne tehnološke poti do superinteligence in raznovrstne negotovosti, ki obdajajo z njo povezana vprašanja. V tem kontekstu bom med drugim skušal nakazati nepredvidljivost superinteligentnega vedenja ter izmuzljivost razmerja med moral(n)o(stjo) in inteligenco. Menim, da je razumevanje tega razmerja dokaj pomembno za prihodnost človeške vrste in planeta – zlasti v kolikor sprejmemo argumente v prid možnosti in/ali (precejšnji) verjetnosti eksplozije inteligence v (bolj ali manj bližnji) prihodnosti ter s tem idejo, da se bodo superinteligentni agensi prej ali slej pojavili, pa če to hočemo ali ne.⁴ Vse to se mi zdijo pomembni vidiki naslovne tematike, vredni temeljite nadaljnje filozofske, znanstvene in nenazadnje tudi širše družbene pozornosti.

Anthropic Bias: Observation Selection Effects in Science and Philosophy (2002) in zgoraj omenjene *Superintelligence: Paths, Dangers, Strategies* (2014), člankov, kot so »The Doomsday Argument is Alive and Kicking« (1999), »Are You Living in a Computer Simulation?« (2001), »Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing« (2008), »Cognitive Enhancement: Methods, Ethics, Regulatory Challenges« (2009), »The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents« (2012), »Existential Risk Prevention as Global Priority« (2013) in »Strategic Implications of Openness in AI Development« (2016), ter sourednik zbornikov *Human Enhancement* (2009) in *Global Catastrophic Risks* (2011).

² Bostromu se zdi *verjetno*, da se bo ključen preboj na področju umetne inteligence dogodil »enkrat v tem stoletju« (Bostrom, 2014: vii).

³ Bostrom pravi, da bo šlo za »bržčas zadnji izziv, s katerim se bomo spoprijeli« (Bostrom, 2014: vii). Zdi se namreč, da »bomo dobili eno samo priložnost« (prav tam) – ko bo superinteligence enkrat ustvarjena, bo naša usoda »zapečaten« (prav tam), v dobrem in/ali slabem.

⁴ Za koncizno predstavitev in ovrednotenje nekaterih temeljnih argumentov glej Chalmers, 2010.

Kratek pretres trdovratnega izziva superintelligence

Naj za začetek kar se da na kratko obnovim »Nedokončano basen o vrabcih«, s katero Bostrom prične svojo knjigo. Gre za skrajno sugestivno, poučno pripoved o vrabcih, ki v času napornega spletnja gnezd sanjarijo o sovi, ki bi jim pomagala pri tem duhamornem opravilu. Ker bi jih lahko znatno razbremenila dela, in sicer ne le tako, da bi jim pomagala pri že omenjeni gradnji gnezd, marveč tudi s tem, da bi pazila in skrbela za njihove najmlajše in najstarejše sorojake, oprezala za mačkami ter drugimi nevarnostmi ipd., se odločijo, da čimprej pričnejo z iskanjem sovinega jajca, izvaljeno ptico pa nato udomačijo in ustrezno izurijo. Do takšnega, na prvi pogled neproblematičnega načrta pa je že od samega začetka skeptičen enooki vrabec, ki v tem preprostem planu razpozna potencialno usodno nevarno pomanjkljivost – po jajce se odpravljajo, ne da bi pred tem razrešili problem udomačitve in urjenja sove, ko se bo ta enkrat izvalila iz najdenega jajca (Bostrom, 2014: v).

Četudi se vrabčji poglavar strinja, da je izziv, s katerim jih sooča dvomljivec, pomemben, nazadnje vendarle presodi, da je dovolj težavna že sama prisvojitve sovinega jajca in da je zato bolje, da se najprej lotijo kar te, o »problemu nadzora« oziroma (ne)možnosti(h) udomačitve in urjenja sove pa premlevajo in razpredajo pozneje. Razplet basni ni znan, Bostrom pa svojo knjigo posveti skeptičnemu vrabcu in njegovim somišljenikom, ki niso sodelovali pri »sovoiskateljstvu«, temveč so raje svarili pred nevarnostmi in skušali dognati odgovor na zagonetno vprašanje, kako sovo pravzaprav sploh vzgojiti, udomačiti in izuriti (prav tam). Zdi se mi, da se nekaj podobnega godi tudi z »iskanjem« (»močne«) umetne intelligence – večina proučevalcev hiti, da bi do nje kar najhitreje prispeli, le redki pa konkretnije in temeljiteje razmišljajo o tem, kakšne daljnosežne posledice bi utegnili imeti takšno prispetje in kako bi se veljalo že predhodno pripraviti na dogodek, ki bo, če bo do njega prišlo, po vsej verjetnosti »eden najbolj pomembnih dogodkov v zgodovini planeta« (Chalmers, 2010: 10).⁵

⁵ Na področju filmske umetnosti ne primanjkuje stvaritev, ki so upodobile dileme in probleme, ki jih predstavlja umetna inteligenca. Na tem mestu se bom omejil na omembo enega samega sodobnega primera, ki se mi zdi v okviru razpravljanja o umetni (super)inteligenci posebej relevanten in sugestiv. V *Ex Machini* (2015), najnovejšem filmu angleškega scenarista in režiserja Alexa Garlanda, uspe programerju Nathanu Batemanu (Oscar Isaac) ustvariti unikatno humanoidno robotko Avo (Alicia Vikander). Izkaže se, da premore ta – če sklepamo po njenem zavidljivo kompleksnem vedenju, vključno z njeno zmožnostjo koketiranja, pretvarjanja in nenazadnje premetenega načrtovanja poba – precej globlje in manj predvidljivo »notranje« življenje (ali, če hočete, simulacijo notranjega življenja), kot prvotno domnevata Nathan, ki ji pozneje celo izrecno pripiše zmožnost samozavedanja, ter njegov enotedenski pomočnik Caleb (Domhnall Gleeson). V kolikor bi se dejansko zgodilo kaj takega, kar prikazuje Garlandov film, bi to predstavljalo, če uporabimo Nathanove besede, »največji znanstveni dogodek v zgodovini človeštva«, ki bi lahko privedel do »eksplozije intelligence«, o kateri piše Bostrom: prvokrat bi bil ustvarjen »umetni« subjekt, »popolnoma etični ['umetni'] agens« (Tonkens, 2009), ki bi posedoval zmožnost samozavedanja, čustev, avtonomnega upora in predirljivega mišljenja, ki bi med drugim porajalo vprašanja takšne vrste, kot ga na neki točki zastavi Ava: »Mar ni čudno ustvariti nekaj, kar te sovraži?« Nemalo proučevalcev se strinja, da se je treba na vse pretege izogniti stvaritvi česa takšnega, kar bi nas

Bostrom se v nadaljevanju izogne in izrecno odreče rabi izraza »tehnološka singularnost«, ki je po njegovem mnenju zadobil pretirano »tehnoutopične konotacije« (Bostrom, 2014: 2).⁶ Prisega na »bolj precizno terminologijo« (prav tam), zaradi česar v kontekstih, v katerih bi običajno govorili o singularnosti, razpreda o »eksploziji intelligence« (ang. *intelligence explosion*), o kateri je slaven matematik I. J. Good (1916–2009) pisal že davnega leta 1965 v članku »Speculations Concerning the First Ultraintelligent Machine«. Temeljni pomen omenjene sintagme je koncizno pojasnil znani ameriški proučevalec umetne intelligence Eliezer Yudkowsky, ki je opozoril na to, da pri eksploziji intelligence ne gre za tezo o »splošni ekonomski in tehnološki pohitritvi« (Yudkowsky, 2013: 1), temveč je osrednja ideja v naslednjem: ker je tehnologija proizvod intelligence, bi v primeru stvaritve inteligentnih strojev slejkoprej prišlo do »učinka snežne kepe – prvi inteligentni umi bi lahko ustvarili inteligentnejše ume, inteligentnejši umi pa bi lahko izdelali še inteligentnejše ume« (prav tam) in tako naprej v potencialno neskončnost.

Toda kaj sploh sta inteligenca in superinteligenca? Inteligenco lahko razumemo kot »zmožnost doseganja ciljev v različnih okoljih« (Bostrom, 2014), kot »spособnost doseganja ciljev kljub oviram prek sprejemanja odločitev, ki temeljijo na racionalnih (resnici podrejenih) pravilih« (Pinker, 2009: 74).⁷ Superinteligenca pa

bilo *zmožno* (za)sovražiti in delovati v skladu s tem sentimentom. To pa razpira vprašanje, ali zamejitev zmožnosti, ki bi zadostovala za gotovo izognitev scenariju, ki smo mu priča v *Ex Machini*, ne bi obenem neizogibno pomenila onemogočenja stvaritve *resničnega* agensa in/ali *dejanske* superinteligence. Toda, kakor upravičeno opozarja Eliezer Yudkowsky, se s tem še vedno ne pritaknemo najglobljega in najpomembnejšega problema, ki ima opraviti s precej preprostejšo zmožnostjo (super)inteligentnih strojev – njihovo močjo, da krojijo svet in tako (so)oblikujejo (našo) prihodnost: »UI vas ne sovraži in prav tako ne ljubi, toda narejeni ste iz atomov, ki jih lahko uporabi za kaj drugega« (Yudkowsky, 2008: 333). Pri čemer gre tudi za to, da »[t]renutno zasedamo ogromno prostora na planetu«, in zatorej utegnemo predstavljati »potencialno grožnjo« (Agar, 2016). Več o tem v nadaljevanju.

⁶ Kot opozarja Marko Uršič, je futuristična raba izraza »singularnost« analoška in izvira iz matematike ter fizike: »V matematiki je singularnost točka, v kateri funkcija ni definirana (oz. ni odvedljiva). .../ Singularnost v fiziki je 'črna luknja' v prostoru-času; pri tem pa je treba razlikovati med črno luknjo kot *območjem prostora-časa* za 'horizontom dogodkov' (kar pomeni, da informacije o dogodkih v njej niso dosegljive tistim, ki so/smo zunaj nje) – in samo singularnostjo, ki pa je zgolj teoretsko *hipotetična* točka v črni luknji oziroma znotraj območja prostora-časa za horizontom dogodkov, v kateri fizikalne količine domnevno presežejo vse končne vrednosti« (Uršič, 2015: 452). Tehnološka singularnost po besedah futurista Vernorja S. Vinge predstavlja točko, »v kateri moramo opustiti svoje stare [miselne] modele, saj zavlada nova realnost« (Vinge, v Uršič, 2015: 453), pri čemer je, kakor poudarja Uršič, »mišljena v dvojni analogiji s fiziko: gre za *enkratnost* tega Dogodka, za iztrganost Singularnosti iz zgodovinskega prostora-časa, in za njeno *zastrtost*, saj je (še) onkraj našega izkustvenega 'horizonta dogodkov'« (Uršič, 2015: 454).

⁷ Roman V. Yampolskiy, avtor nedavne knjige *Artificial Superintelligence*, inteligenco označi za »zmožnost najdejanja vzorcev v podatkih« (Yampolskiy, 2016: 98). Lakhmi C. Jain, Anas Quteishat in Chee Peng Lim pa so v članku »Intelligent Machines: An Introduction« opozorili, da so med »glavnimi dejavniki, vključenimi v "inteligenco", zmožnosti avtonomnega učenja in prilagajanja okolju« (Jain, Quteishat in Lim, 2007: 1). – Jasno je, da so vse navedene opredelitve zgolj provizorične in poleg tega ne implicirajo (nujno) druga druge, čeprav so medsebojno vsekakor tesno povezane. Vprašanje, katera definicija je najustreznejša, je zapleteno in ga na tem mestu puščam ob strani (za razčlenbo in zagovor – abstraktne, splošne – opredelitve intelligence, podobne zgornji(m), glej Legg, 2008: 1–22).

je po Bostromu pojmovana kot »razum, ki močno presega človeško kognitivno zmoglost na tako rekoč vseh pomembnih področjih« (Bostrom, 2014: 22).

Čeprav lahko v »vsakem diskontu /.../ kupimo računalnik, ki presega človeško sposobnost računanja, shranjevanja in priklica podatkov, risanja, preverjanja črkovanja, usmerjanja poti in tipografije« (Pinker, 2009: 95; prevod je nekoliko prirejen), se velja zavedati, da takšni računalniki še zdaleč niso superinteligentni, vsaj v Bostromovem smislu ne – njegova oznaka se namreč nanaša *zgolj in samo* na sisteme, ki imajo nadčloveško raven *splošne* intelligence.⁸

Naj na tem mestu izrecno pripomnim, da se zgoraj navedena definicija v ničemer vnaprej ne opredeli glede tega, za kakšno vrsto superintelligence naj bi šlo, temveč preprosto govori o razumu, ki izjemno presega človeško splošno kognitivno zmoglost. Prav tako ne pove ničesar o tem, ali se bo nadčloveško inteligentna entiteta dejansko zavedala svojega razuma, svojih presežnih »spoznavnih sposobnosti« in bo potemtakem svet izkušala prvoosebno, subjektivno, zavestno,⁹ tako da

⁸ Računalniki, s katerimi imamo opraviti dandanes, še vedno ne predstavljajo nečesa, kar bi »močno presevalo človeško kognitivno zmoglost na tako rekoč vseh pomembnih področjih« (Bostrom, 2014: 22). Se več, ne samo, da te zmoglosti ne presegajo – zaenkrat se človeški *splošni* inteligenci niti približajo ne, saj gre za »sisteme, ki imajo ozek nabor kognitivnih zmoglosti« (prav tam: 16). Bostrom se dobro zaveda, da so se »tehnične težave izdelovanja inteligentnih strojev izkazale za večje, kot so predvideli pionirji [UI]« (prav tam: 4), a poudarja, da to ne pove ničesar o tem, »kako velikanske so te težave in kako daleč smo trenutno od tega, da jih presežemo« (prav tam). Poleg tega ne smemo pozabiti, da se »[v]časih /.../ izkaže, da ima problem, ki sprva izgleda brezupno zapleten, presenetljivo preprosto rešitev (červeno je obratno verjetno pogostejše)« (prav tam).

⁹ S filozofom Richardom Brownom delim prepričanje, da gre pri vprašanju, »ali lahko dejansko ustvarimo sintetično ali umetno zavest«, za *odprto empirično* vprašanje, saj »se utegne izkazati, da je za zavest potrebno nekaj izrazito [*distinctly*] biološkega« (Brown, 2012: 212). To pomeni, da po mojem mnenju obstaja *možnost*, da je zavest biološki fenomen v smislu, da je »kavzalno vezana na določeno biokemijo svojih virov, kot so laktacija, fotosinteza ali katerikoli drugi biološki pojavi« (Searle, 1990: 379; prevod je nekoliko prirejen), in da je, denimo, bistveno odvisna od »fizikalnokemijskih lastnosti /.../ možganov« (prav tam), zaradi česar prav tako vnaprej ne izključujem možnosti, da je stvaritev »umetne«, nebiološke zavesti – ali »prenos, 'teleportacija' mentalnega softvera v neki *drugi*, ne nujno biološki hardver« (Uršič, 2015: 319) – nekaj nemogočega: »Nihče ne bo trdil, da lahko proizvajamo sladkor ali mleko tako, da z računalnikom simuliramo formalne sekvence laktacije oziroma fotosinteze, ko pa gre za vprašanje duha, je v takšne čudeže pripravljeno verjeti veliko ljudi, in sicer zaradi globoko zakoreninjenega dualizma; duh, menijo, je stvar formalnih procesov in je za razliko od mleka in sladkorja neodvisen od povsem določenih materialnih vzrokov« (Searle, 1990: 379; prevod je mestoma prirejen). Daniel C. Dennett izpostavlja in zagovarja verjetnost nasprotne možnosti, po kateri obstaja »(po)poln izomorfizem med človeško zavestjo in računalniškim programom/procesom« (Uršič, 1987: 197): »Če je jaz 'zgolj' pripovedno težišče in če lahko vse vidike človeške zavesti pojasnimo 'zgolj' z dejavnostmi virtualnega stroja, udejanjenega v izjemno prilagodljivih povezavah človeških možganov, potem bi ustrezno 'programiran' robot s silicijevimi računalniškimi možgani načeloma lahko bil zavesten in imel svoj jaz. Natančneje, lahko bi obstajal zavesten jaz z robotskim telesom in računalniškimi možgani« (Dennett, 2012: 509). Podobnega mnenja je tudi Douglas R. Hofstadter: »Duhovi obstajajo v možganih in bodo nemara obstajali tudi v programiranih strojih. Če in ko bodo takšni stroji nastali, njihove kavzalne moči ne bodo izvirale iz substanc, iz katerih bodo narejeni, temveč iz njihovega načrta in programov, ki bodo tekli v njih. Ali imajo te kavzalne moči, pa bomo zvedeli tako, da bomo govorili z njimi in skrbno prisluhnili tistemu, kar bodo imeli povedati« (Hofstadter, 1990: 389). A pripomniti velja, da celo iz tega, da je samozavedajoči se stroj nekaj nemogočega, *samo po sebi* (brez dodatnega argumenta) ne bi sledilo, da je nemogoča tudi stvaritev *superinteligentnega* stroja – vsaj ne, če sprejmemo (dozdevno sprejemljivo) tezo, da entitete, ki presegajo človeško (splošno) inteligenco, niso *nujno* tudi zavestne.

bomo lahko – à la Thomas Nagel – govorili in se spraševali o tem, *kako je biti* – »biti v 'prvi osebi'« (Nagel, 1990: 400) – ta entiteta.¹⁰ Bostrom se raje kot na vprašanja o metafiziki duha in/ali zavesti osredotoča na »vzročne antecedense in konsekvence superintelligence« (Bostrom, 2014: 22), s čimer se skuša izmakniti prenekaterim nadvse zagonetnim filozofskim problemom.¹¹ V članku »Why We Need Friendly AI?«, ki ga je napisal skupaj z Lukom Muehlhauserjem, razjasnjuje, čemu ubira takšen pristop: »Čeprav nekateri dvomijo, da bi lahko stroji posedovali mentalne lastnosti, kakršna je zavest, pomanjkanje tovrstnih lastnosti strojem ne preprečuje tega, da bi postali precej zmogljivejši od ljudi, da bi učinkovito krmarili prihodnost pri zasledovanju svojih ciljev« (Muehlhauser in Bostrom, 2014: 41). Uršič je podobno opozoril, da je vprašanje o tem, »ali ima lahko kompjuter (ali mreža ipd.) sploh svojo *lastno voljo*: ali lahko *avtonomno*, brez človeškega 'ukaza' sproži kakšno dejanje« (Uršič, 2015: 479), verjetno celo trdovratnejše in prvotnejše od vprašanja o zavestnosti superintelligence (vsaj z evolucijskega in, ne nazadnje, praktičnega ozira).¹² Tudi (super)inteligentni stroji, ki bi bili »zombiji«, oropani vseh subjektivnih doživetij, bi potemtakem predstavljali potencialno grožnjo naši blaginji.¹³ Kot pojasnjuje Rob Bensinger (2015), je te-

¹⁰ Kot pravi Nagel, ima »organizem [ali entiteta] zavedajoča se mentalna stanja tedaj in le tedaj, če obstaja nekaj takšnega kot *biti* ta organizem [ali entiteta] – nekaj takšnega *za* ta organizem [ali entiteto]« (Nagel, 1990: 400).

¹¹ Mednje – poleg vprašanja razlage intencionalnosti, subjektivno-kvalitativnega značaja izkustva, težkega problema zavestnega izkustva in brezštevilnih drugih izjemno zanimivih in težavnih znanstveno-filozofskih izzivov, ki so tako ali drugače povezani z duhom in zavestjo – sodi tudi vprašanje razmerja med inteligenco in (fenomenalno) zavestjo (ali – širše gledano – mentalnostjo). Naj pristavim, da so nekateri, med njimi zlasti prepričljivo Rosalind Picard, opozorili, da »imamo dokaze, da so čustva dejaven del intelligence, posebej zaznave, racionalnega mišljenja, odločanja, načrtovanja, ustvarjalnosti itn.« (Picard, 1997: 247), kar problem le še dodatno poglobi. Če namreč drži, kar trdi Picard, potem se *zdi*, da vprašanj (super)intelligence preprosto ni mogoče jasno razmejiti od vprašanj (super)čustev, posledično pa tudi ne od vprašanj (super)duševnosti.

¹² Uršič poudarja, da je na vprašanje avtonomnosti »nemara še težje odgovoriti kot na vprašanje, ali ima načeloma lahko kompjuter zavest, sámozavedanje. Kajti tudi če na vprašanje zavesti odgovorimo negativno (in osebno mislim, da kompjuterji, vsaj takšni, kakršne poznamo danes in kot si jih po vseh znanih scenarijih lahko zamislimo v bližnji prihodnosti, *ne* morejo imeti intencionalne zavesti, kaj šele sámozavedanja) – to še ne pomeni, da ne morejo biti avtonomni agensi« (Uršič, 2015: 479). Še več, takšne avtonomne, a ne(samo)zavedajoče se entitete niso zgolj možne, temveč tudi dejanske, saj jih – po vsem sodeč – najdemo v realnem svetu: živali so namreč »bitja, ki *so* avtonomni agensi, čeprav najbrž nimajo refleksivne zavesti (če izvajamo primata, morda). Pomislimo, recimo, na dinosavre: kako bi jim, če bi jih z biotehnoško genetiko znova oživili, vcepili kak 'etični kodeks', recimo, da ne bi ubijali človeških otrok, ki bi se gotovo radi igrali s temi pošastmi iz preteklosti v kakem realnem 'Jurskem parku'?« (prav tam).

¹³ Filozofi z izrazom »zombi« ne označujejo srh vzbujajočih »živih mrtvecev«, ki smo jih gledali in jih gledamo v številnih filmskih grozljivkah, marveč nekoga, »čigar vedenje je povsem naravno, pozorno, živahno in izrazno, a v resnici nima zavesti, temveč je nekakšen stroj« (Dennett, 2012: 97): »Glavna stvar pri filozofskem pojmu zombija je to, da med zombijem in normalno osebo ne moreš razlikovati samo na podlagi opazovanja zunanjega vedenja« (prav tam). Ali povedano natančneje: gre za *fenomenološke* zombije, za »bitja, ki so fizično istovetna ljudem, a so brez izkustev/doživetij: čeprav jih torej 'navzven' (na podlagi fiziološko-anatomskega ustroja in/ali vedenja) ni mogoče razlikovati od 'normalnih' ljudi, so 'navznoter' oropani vsakršnih fenomenoloških stanj« (Vörös, 2014: 266).

meljni vir Bostromovega zanimanja za inteligenco prav v »njemem potencialu, da rešuje praktične probleme in oblikuje prihodnost«, ¹⁴ kar je nekaj, kar bi nas moralo še posebej interesirati – in zaskrbljevati – z vidika varnosti. ¹⁵ Roboti, ki bi posedovali *avtonomno* zmožnost, da počnejo stvari, ki bi bile, denimo, uporabne za ljudi, bi, kakor so poleg Bostroma opozorili Colin Allen, Gary Varner in Jason Zinser (2000) ter mnogi drugi, imeli tudi zmožnost, da bi delali to, kar bi ljudem in drugim čutečim bitjem škodovalo. ¹⁶ To pa, kakor je dejala Rosalind Picard in si bomo natančneje ogledali v nadaljevanju, navsezadnje pomeni, da »večja kot bo svoboda stroja, bolj bo potreboval moralne standarde« (Picard, 1997: 134). ¹⁷

Možnih tehnoloških poti do superintelligence je več ¹⁸ (za nadrobnejšo razlago glej Bostrom, 2014: 22–51):

¹⁴ »Tudi če obstajajo druge vrste intelligence, so te domnevno manjšega ekonomskega in strateškega pomena kot 'kognitivne supermoči', ki jih Bostrom opisuje v 6. poglavju [knjige *Superintelligence*].« (Bensinger, 2015)

¹⁵ Ključen argument, ki ga podaja Bostrom, se da formulirati na način, ki bi se docela izognil rabi izraza »inteligenca« in bi ga, denimo, nadomestili »zmožnost programiranja« in »neka specifična sposobnost sklepanja« (Chalmers, 2010: 23). Nekaterim se zdi takšen argument celo prepričljivejši od Bostromovega (glej Bensinger, 2015).

¹⁶ Če nam uspe ustvariti umetno (splošno) inteligenco, potem bo to, kot že rečeno, nemara največja sprememba v človeški zgodovini in precej zlahka tudi njen konec. Ne pozabimo namreč, da bi šlo za inteligentne stroje, ki bi lahko reproducirali veliko večino dejavnosti naših možganov, pri čemer bi to delo celo opravljali mnogo hitreje in zanesljiveje od nas. Poleg tega bi lahko, kakor se bržkone razume samo po sebi, »za doseganje svojih ciljev /.../ uporabljali tudi spretnost prepričevanja, manipulacije in načrtovanja« (Bostrom, 2009). Če k temu dodamo še precejšnjo verjetnost »eksplozije intelligence«, ki bi sledila našemu stvarjenju umetnointeligentne entitete, potem bi se znalo zgoditi, da bi inteligenca strojev bliskovito tako zelo preseгла človeško, kot je človeška preseгла inteligenco hroščev ali črvov.

¹⁷ V kolikor bi imeli na neki točki kljub vsemu *dobre* razloge za verjetje, da smo uspeli ustvariti stroj, ki premore fenomenalno zavest, bi to porajalo nemalo zanimivih vprašanj. Ali bi bilo *moralno* sprejemljivo ugasniti ali uničiti stroj, ki bi bil zmožen zavedajočih se – ali kakršnihkoli – duševnih stanj (Dennett se npr. izizivalno sprašuje: »[Z]akaj naj bi bili razblinjeni upi 'zombija' manj pomembni od razblinjenih upov zavestne osebe?« (Dennett, 2012: 530))? Bi moral imeti takšen stroj pravice, kot so denimo pravica do življenja, svobode in varnosti? Bi ga bili »moralno obvezani priznati /.../ kot osebo in se posledično v odnosu do njega vzdržati vsega tistega, česar ne smemo storiti ljudem« (Harris, 2002: 29)? Bi bil v primeru, da bi bilo njegovo duševno življenje zadosti kompleksno, moralno odgovoren za svoje vedenje in bi si za svoje »moralne izbire« lahko resnično zaslužil grajo ali hvalo, nagrado ali kazen? Pripomniti velja, da se tudi ob vprašanju *dobrih* razlogov za prepričanje, da stroj poseduje zavest, zastavlja kopica vprašanj. Kakšni bi morali biti ti razlogi, da bi bili zadovoljivi? Bi za kaj takega *zadostovalo* dovolj sofisticirano – človeku podobno? – vedenje strojev? Če da, kdaj bi bilo določeno vedenje *dovolj* sofisticirano? In nenazadnje: mar ni možno (celo verjetno?), da bi robotovo (sámo)zavedanje vzniknilo, ne da bi to opazil kdorkoli izmed nas, ki robota (neizogibno) proučujemo zgolj »od zunaj«, brez vsakršnega dostopa do njegove notrine? Kdaj bi bilo, skratka, *upoštevajoč* vse te – filozofom razmeroma dolgo poznane – zagate, *racionalno* verjeti, da so stroji zavestni, in bi bilo do njih ustrezno privzeti to, kar Dennett imenuje »intencionalna drža« (ang. *intentional stance*), tako da bi »oddajalca zvoka /.../ obravnavaj[i] kot racionalni agens s prepričanj, željami in drugimi intencionalnimi ali 'naperjenimi' duševnimi stanji, katerega dejanja lahko razložimo (ali napovemo) na podlagi vsebine teh stanj« (Dennett, 2012: 101; moj poudarek)?

¹⁸ To dejstvo po Bostromovem mnenju že samo po sebi znatno poveča verjetnost prispetja na cilj po *najmanj* eni od njih.

- (1) eksponentno hiter (hardverski in softverski) razvoj računalnikov ter konstrukcija »močne« umetne (splošne) inteligence;
- (2) »emulacija« možganov, tj. stvaritev (super)intelligentnega softvera s pomočjo bolj ali manj filigranske preslikave/skeniranja »komputacijske« strukture bioloških možganov;
- (3) znatno izboljšanje delovanja bioloških možganov s pomočjo »selektivne vzreje« (ang. *selective breeding*) ali z biotehnološkim spreminjanjem genov (evgenika);
- (4) spajanje možganov/človeka in računalnika/stroja (nekakšna »kiborgizacija«);
- (5) obilen razvoj računalniških mrež in organizacij ter vznik »kolektivne superinteligence« ali celo »prebujenje« računalniških omrežij.

Bostrom meni, da sta najhitreje dosegljivi in hkrati najzanesljivejši poti močna umetna inteligenca in »emulacija« možganov, pripominja pa, da so šibke oblike superinteligence dosegljive tudi s pomočjo biotehnološkega nadgrajevanja kognicije, ki bi utegnili znatno pospešiti znanstveni in tehnološki napredek ter tako tudi pohitriti »vzlet« superintelligentne entitete ali celo množice njih. Manj obetavni se mu zdijo možgansko-računalniški vmesniki, ki bi po njegovem prepričanju v najboljšem primeru igrali podobno vlogo, kot jo igrajo »biološke poti« do nadčloveške inteligence.

Na tem mestu bi se rad, preden preidem k naznačitvi raznoraznih negotovosti, ki obdajajo vprašanje superinteligence, le še na kratko pomudil ob Bostromovem razpravljanju o »kinetiki« eksplozije inteligence. Pri tem gre za vprašanje hitrosti prehoda od stroja, ki bi dosegel človekovo raven kognitivnih sposobnosti, do superinteligence. Bostrom govori o treh možnih »tranzicijskih scenarijih«: (1) o počasnem scenariju, kjer govorimo o desetletjih ali stoletjih in ki pomeni idealno priložnost za relevantne družbenopolitične spremembe, za preizkušanje različnih pristopov, analiziranje situacije, skratka za temeljito pripravo na vznik superinteligence; (2) o zmernem scenariju, pri katerem gre za mesece ali leta (ter je časa za potrebno pripravo in spremembe nekoliko manj); (3) o hitrem scenariju, pri katerem govorimo o minutah, urah ali dnevih (kjer časa za poglobljene premisleke ter takšne in drugačne (proti)ukrepe tako rekoč ni). Vprašanje je torej, kako hitro po stvarjenju pametnega stroja, ki ga po intelektualnih sposobnostih ne bo več mogoče razlikovati od ljudi, bo prišlo do »vzleta« superinteligence kot »razuma, ki močno presega človeško splošno kognitivno zmožnost«. Bo šlo pri tem za minute, ure ali dneve? Za tedne? Ali pa celo za stoletja, tisočletja? Sloviti futurist Vernor Vinge je nekoč zapisal, da je najboljši odgovor na vprašanje, ali bodo računalniki kdajkoli tako pametni kot ljudje, najverjetneje sledeč: »Da, toda le začasno!« (Vinge, v Muehlhauser in Salamon, 2012: 15) Tudi Bostromu se zdi precej verjetno, da bo omenjeni prehod »eksploziven« in da se bo že stvarjenje strojne inteligence, ki bo po splošnih kognitivnih sposobnostih primerljiva z »vaškim beščkom« (Bostrom, 2014: 70), skoraj zagotovo izkazalo za težji in dolgotrajnejši iz-

ziv kot pa (softversko, hardversko in vsebinsko) izboljšanje tega sistema do te mere, da bo postal nadčloveško inteligenen.

Avtor nadalje upravičeno poudarja, da smo, ko premlevamo o vprašanjih, dilemah in problemih, povezanih s superinteligenco, »obkroženi z gosto tančico negotovosti« (prav tam: 255). Tisto, kar je pri vsem tem še posebej zaskrbljujoče, pa je, da je neznansko težko, če ne celo povsem nemogoče s kakršnokoli gotovostjo napovedati vedenje superinteligentne entitete. Lahko zgolj domnevamo, da bodo njena pota precej verjetno nenavadna, sploh gledano z naše razmeroma (ali če smo nekoliko bolj pesimistični, izredno) omejene perspektive. Razumevanje ozadja te ali one superinteligentne dejavnosti in ravnanja bo verjetno hudo presehalo meje naših umskih sposobnosti, in to *navkljub* temu, da je razumno predpostavljati, da obstajajo določeni instrumentalni cilji (npr. samoohranitev, kognitivna izboljšava, tehnološka izpopolnitev in pridobivanje virov), za katere si bo prizadeval tako rekoč sleherni inteligentni agens, vključno s superinteligentnim (glej Bostrom, 2012; Bostrom, 2014: 105–114).

A ne smemo pozabiti, da imamo v našem primeru bistveno prednost: *mi* bomo namreč tisti, ki bomo, če sploh bomo, stvaritelji »naših lastnih naslednikov«. To pomeni, da obstaja možnost, da določimo končne cilje superintelligence na ta način, da bo varovala človeške vrednote in delovala v skladu z veliko večino naših interesov. Toda kako doseči kaj takega? Pri tem nam nikakor ne more biti v pretirano uteho prepričanje o inherentni povezavi med inteligenco in moralnostjo, po katerem je superinteligentna entiteta po nekakšni nujnosti že tudi »supermoralna«. To po vsem sodeč ni nič drugega kot antropomorfní predsodek. Še več, kot opozarjata Bostrom in Muehlhauser, gre za »antropomorfizem«, ki v resnici ne drži niti za nas, saj »zlahka najdemo ljudi, ki so precej inteligentni, a nemoralni, ali takšne, ki so neinteligentni, toda popolnoma dostojni [*decent*]« (Bostrom in Muehlhauser, 2014: 44). Poleg tega se moramo zavedati, da »popolna vednost o moralni teoriji ne zagotavlja delovanja v skladu z njo« (Allen, Varner in Zinser, 2000: 260): »Lahko si zamislimo agensa, ki je docela zmožen razsojanja o tem, ali so določena dejanja moralna, a je hkrati povsem nemotiviran, da bi tudi deloval moralno. Takšen agens bo pogosto deloval na način, za katerega vé, da je [moralno] napačen. Pravzaprav je to opis nekaterih sociopatov.« (prav tam) Toda kakšno *zagotovilo* imamo, da superinteligentne entitete v tem oziru ne bodo podobne sociopatom? Zdi se, da prav nobenega.¹⁹

¹⁹ Bostrom poudarja, da bi nam bila lahko UI v svojih motivacijah precej manj podobna, kot bi nam bil, denimo, »zeleni luskasti Nezemljan« (Bostrom, 2014: 106), ki bi bil – predpostavimo – biološko bitje, ki se je, tako kakor mi, razvilo skozi evolucijski proces, tako da bi bilo upravičeno pričakovati, da bo imel »vrste motivacij, značilne za *razvita* [*evolved*] bitja« (prav tam; moj poudarek). Zato ni nič paradoksnega v možnosti, da bi bil edini končni cilj, ki bi ga utegnili imeti superinteligenta, »preštevanje zrnc peska na otoku Boracay« (prav tam: 107). Toda problem je v tem, da bi bil lahko celo tovrsten, navidez povsem nedolžen končni cilj *usoden* za človeštvo, saj bi imel preštevajoči agens v mnogih situacijah »konvergentni instrumentalni razlog«, da pridobi neomejeno količino fizičnih virov in da, če je možno, odstrani

Kakor opozarja Rosalind Picard, bo sistem, ki deluje v tako kompleksnem in nepredvidljivem okolju, kakršno je naše, potreboval več kot zgolj takšne ali drugačne zakone (npr. tako ali drugače modificirane tri slavne Asimove zakone robotike) – »bistveno bo potreboval vrednote in načela, moralni kompas, nemara celo religijo« (Picard, 1997: 134). K temu velja pripomniti, da obstaja dober razlog za to, da *ne* ustvarimo nadčloveških strojev, ki bi z nami delili vse naše trenutne vrednote. Zakaj? Zamislimo si, da bi se že antični Grki soočili z »znanstveno-tehnološkim Prehodom«, o katerem govorimo v tem članku, in bi uspeli v stroj nekako vprogramirati svoje tedanje vrednote kot njegov končni cilj. »Z naše perspektive bi to,« kakor upravičeno poudarjata Bostrom in Muehlhauser (2014: 45), »rezultiralo v tragediji, saj menimo, da smo bili od antičnih Grkov dalje priča moralnemu napredku (npr. prepoved suženjstva).« Pri tem ne smemo spregledati, da smo najbrž tudi sami še vedno zelo daleč od kakršnekoli etične popolnosti (kaj je vendar bolj očitno od tega?), zaradi česar bi bilo superinteligenci pametno dopustiti nadaljnjo možnost moralnega napredovanja.

Kot vzpostavljevalci »začetnih pogojev« superinteligence se torej, na kratko rečeno, znajdemo pred kar najbolj kompleksnim *etično-filozofskim* vprašanjem – kakšen vrednostni sistem vprogramirati superinteligenci?²⁰ Ali z Uršičevimi besedami: »[K]atere cilje oz. smotre ali 'maksime' bi vpisali v 'začetne pogoje' superinteligence, če bi jih lahko še pred njenim prebujenjem (so)oblikovali?« (Uršič, 2015: 455). In nenazadnje: *kdo* naj o tem odloča? Jaz, »mi«, »vi«? Bodo to zagovorniki etike vrlin, deontologi, konsekvencialisti, kar vsi omenjeni ali kdo drug? Kako razrešiti brezštevilna nestrinjanja, s katerimi imamo, vsaj kar zadeva najzahtevnejše (filozofske) dileme, opraviti ne samo v filozofskih, temveč tudi vseh drugih logih? Menim, da Picard razpira ključen problem, s katerim se moramo soočiti, ko se sprašuje o tem, kdo bi imel *in* bi moral imeti »moralno avtoriteto nad računalniki, roboti, softverskimi agenci« (Picard, 1997: 134). Čeravno drži, da je takšna avtoriteta trenutno v rokah teh, ki razvijajo in programirajo računalnike (in/ali njihovih plačnikov), pa je to navsezadnje vsekakor »vprašanje za družbo kot celoto« (prav tam) – vprašanje, ki zadeva družbo *kot celoto* in ne le *nekaj* posameznikov znotraj nje.

potencialne nevarnosti« (prav tam: 116), vključno s človekom, ki sestoji »iz atomov, ki jih [superinteligence] lahko uporabi za kaj drugega« (Yudkowsky, 2008: 333).

²⁰ Ryan Tonkens opozarja na še eno skrajno zanimivo, a zvečine spregledano razsežnost problema, ki sem se mu skušal približati v tem članku. Poudarja pomembnost vprašanja, ali bi bilo dani etični okvir – kakršenkoli bi ta že bil – sploh *zaželeno* in *moralno* implementirati v robote. Ker se zdi kantovski, deontološki, na dolžnosti(h) utemeljeni pristop mnogim najobetavnejši, se je Tonkens, da bi s tem spodbudil razpravo o tem, kar se mu zdi bistveno, osredinil nanj in pokazal, da bi bilo stvarjenje »kantovskih« umetnih agensov, ki bi izvajali etično relevantna dejanja, v resnici *antikantovsko* in – z vidika same etike, »vprogramirane« v te agense – *nemoralno* (za lucidno izpeljavo tega v marsičem presenetljivega sklepa glej Tonkens, 2009).

Zaključek

Če kaj, potem Bostromova knjiga vsekakor nazorno pokaže, da preprostih rešitev problema nadzorovanja superinteligence preprosto ni. Hkrati nas opominja na to, da je največkrat »mnogo lažje prepričati samega sebe, da si našel rešitev, kot pa jo dejansko najti« (Bostrom, 2014: 124). V tem oziru gre nedvomno za »pametno zamišljeno svarilo in napotek, da moramo začeti vsaj *premišljevati* o vprašanjih, povezanih z nadzorom vse sodobne vrhunske tehnologije, ne samo jedrske tehnologije ter biotehnologije, temveč tudi kompjuterske tehnologije ter informatike v najširšem pomenu« (Uršič, 2015: 480; moj poudarek).

Najbolj gotova stvar, ki jo po Bostromu lahko napovemo o prihodnosti, je, da se bodo dogajale spremembe. K temu pripominja, da kakor hitro »skušamo [te spremembe] natančneje opisati, takoj postanejo manj gotove« (Bostrom, 2009). Takšne misli dodobra odsevajo avtorjevo pogosto prav zavidljivo previdnost, s katero se loteva zapletenih tematik in problematik, ki jih obravnava. Tako na primer že uvodoma priznava, da je verjetno marsikatera njegova teza zmotna in poudarja, da »epistemska skromnost« ne zadostuje, temveč mora biti nadomeščena z nič manj kot »sistemskim priznanjem negotovosti in zmotljivosti« (Bostrom, 2014: viii). Lahko bi celo rekli, da gre pri Bostromu vseskozi za zavedanje tega, da je »edina futurološka napoved, ki je brez dvoma pravilna, /.../ da bodo v prihodnosti današnji futurologi videti trapasto« (Pinker, 2009: 96). Mar ni tovrstna – neprebrodjljiva – negotovost, s katero smo soočeni,²¹ že sama po sebi zadosten razlog za prebujenje iz brezskrbnega dremeža in za skrbno, poglobljeno premlevanje o superinteligenci, ki ostaja odprta možnost in bi nekoč, kot nas prepričljivo svari Bostrom, utegnila biti vzrok za znatno poslabšanje blaginje človeštva ali celo za eksistenčno katastrofo?²²

²¹ Kot pravi Dennett: »Nihče – vključno z mano – nima pregleda nad vsemi problemi in podrobnostmi, zato smo vsi v veliki meri zavezani k momljanju, ugibanju in pleteničenju.« (Dennett, 2012: 307)

²² Četudi se večina proučevalcev osredotoča na človeške živali in njihovo potencialno eksistenčno katastrofo (tj. izumrtje), ki bi bila posledica pojava superinteligence, v resnici seveda še zdaleč ne gre le zanje – v obzir je treba vzeti tudi trpljenje vseh ostalih čutečih bitij, ki bi ga utegnile povzročiti superinteligentne entitete (a ne nujno na tako, recimo temu, nepretanjene načine, kot smo ga do zdaj nemalokrat povzročali mi, pripadniki vrste *homo sapiens*).

The Challenge of "Superintelligence": A Comment on Bostrom's Book *Superintelligence: Paths, Dangers, Strategies* (2014)

The article outlines a series of complex challenges posed by the possibility of "superintelligence." In line with some of the key emphases made by the Swedish philosopher Nick Bostrom in his work *Superintelligence: Paths, Dangers, Strategies* (2014), I put forward some pressing philosophical (metaphysical, epistemological and ethical) questions that emerge when faced with the possibility of artificial agents surpassing human cognitive abilities in virtually all important areas of life. After a brief summary of Bostrom's "The Unfinished Fable of the Sparrows," I provide the preliminary definitions of the main concepts in the field of inquiry (i.e. intelligence, intelligence explosion, and superintelligence). I then explain why a distinction must be made between the question about the intelligence of an entity and the question about this entity also possessing conscious mental states, and go on to point out the possible technological paths towards superintelligence and the various uncertainties about related issues. The pressing nature of these questions and the fact that they still remain unresolved suggest that there are in fact no simple solutions, and that we are dealing with a serious problem that requires further thorough philosophical, scientific, as well as wider social engagement.

Keywords: superintelligence, artificial intelligence, consciousness, Nick Bostrom

Literatura

- Agar, N. (2016). »Don't Worry about Superintelligence«. *Journal of Evolution and Technology*, 26(1), str. 73–82. Dostopno na spletnem naslovu: <http://jetpress.org/v26.1/agar.htm>
- Allen, C., Varner, G. in Zinser, J. (2000). »Prolegomena to Any Future Artificial Moral Agent«. *Journal of Experimental & Theoretical Artificial Intelligence*, 12, str. 251–261.
- Bensinger, R. (2015). »Davis on AI Capability and Motivation«. Dostopno na spletnem naslovu: <https://intelligence.org/2015/02/06/davis-ai-capability-motivation/>
- Bostrom, N. (1998). »How Long Before Superintelligence?«. *International Journal of Future Studies*, 2, str. 1–13.
- Bostrom, N. (2003a). »Are You Living in a Computer Simulation?«. *Philosophical Quarterly*, 53(211), str. 243–255.
- Bostrom, N. (2003b). »When Machines Outsmart Humans«. *Futures*, 35(7), str. 759–764.
- Bostrom, N. (2005). »In Defense of Posthuman Dignity«. *Bioethics*, 19(3), str. 202–214.

Bostrom, N. (2009). »Znanstveni prerok«. *Mladina*, št. 4, str. 58–60. Dostopno na spletnem naslovu: <http://www.mladina.si/46052/znanstveni-prerok/>

Bostrom, N. (2012). »The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents«. *Minds and Machines*, 22(2), str. 71–85.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brown, R. (2012). »The Brain and its States«. V Edelman, S., Fekete, T. in Zach, N. (ur.), *Being in Time: Dynamical Models of Phenomenal Experience*, Amsterdam: John Benjamins, str. 211–230.

Brundage, M. (2015). »Taking Superintelligence Seriously: Superintelligence: Paths, Dangers, Strategies by Nick Bostrom (Oxford University Press, 2014)«. *Futures*, 72, str. 32–35.

Chalmers, D. J. (2010). »The Singularity: A Philosophical Analysis«. *Journal of Consciousness Studies*, 17(9–10), str. 7–65.

Davis, E. (2015). »Ethical Guidelines for a Superintelligence«. *Artificial Intelligence*, 220, str. 121–124.

Dennett, D. C. (2012). *Pojasnjena zavest*. Ljubljana: Krtina.

Harris, J. (2002). *Vrednost življenja: Uvod v medicinsko etiko*. Ljubljana: Krtina.

Hofstadter, D. R. (1990). »Refleksije«. V Hofstadter, D. R. in Dennett, D. C. (ur.), *Oko duha: Fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 380–389.

Jain, L. C., Quteishat, A. in Lim, C. P. (2007). »Intelligent Machines: An Introduction«. *Studies in Computational Intelligence*, 70, str. 1–9.

Legg, S. (2008). *Machine Super Intelligence (PhD dissertation)*. Lugano: University of Lugano. Dostopno na spletnem naslovu: http://www.vetta.org/documents/Machine_Super_Intelligence.pdf

Markič, O. (2011). *Kognitivna znanost: Filozofska vprašanja*. Maribor: Aristej.

Muehlhauser, L. in Helm, L. (2012). »Intelligence Explosion and Machine Ethics«. V Eden, A., Søraaker, J., Moor, J. H. in Steinhart, E. (ur.), *Singularity Hypotheses: A Scientific and Philosophical Assessment*, Berlin: Springer, str. 101–129.

Muehlhauser, L. in Salamon, A. (2012). »Intelligence Explosion: Evidence and Import«. V Eden, A., Søraaker, J., Moor, J. H. in Steinhart, E. (ur.), *Singularity Hypotheses: A Scientific and Philosophical Assessment*, Berlin: Springer, str. 15–40.

- Muehlhauser, L. in Bostrom, N. (2014). »Why We Need Friendly AI«. *Think*, 13(36), str. 41–47.
- Nagel, T. (1990). »Kako je biti netopir?«. V Hofstadter, D. R. in Dennett, D. C. (ur.), *Oko duha: Fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 399–410.
- Picard, R. (1997). *Affective Computing*. Cambridge: MIT Press.
- Pigliucci, M. (2014). »Mind Uploading: A Philosophical Counter-Analysis«. V Blackford, R. in Broderick, D. (ur.), *Intelligence Unbound: The Future of Uploaded and Machine Minds*, Chichester, UK: Wiley-Blackwell, str. 119–130.
- Pinker, S. (2009). *Kako deluje um*. Tržič: Učila International.
- Searle, J. R. (1990). »Duhovi, možgani in programi«. V Hofstadter, D. R. in Dennett, D. C. (ur.), *Oko duha: Fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 361–379.
- Tonkens, R. (2009). »A Challenge for Machine Ethics«. *Minds and Machines*, 19(3), str. 421–438.
- Uršič, M. (1987). *Matrice logosa: Filozofsko-logični eseji in študije*. Ljubljana: Državna založba Slovenije.
- Uršič, M. (2015). *Štirje časi: Filozofski pogovori in samogovori. Zima: Četrty čas, preludij: O sencah*. Ljubljana: Cankarjeva založba.
- Vörös, S. (2014). »Okrogel trikotnik ali miljo visok monocikel? O možnosti naturalizacije fenomenologije«. *Phainomena*, 23(90/91), str. 265–294.
- Yampolskiy, R. V. (2016). *Artificial Superintelligence: A Futuristic Approach*. Boca Raton: CRC Press.
- Yudkowsky, E. (2008). »Artificial Intelligence as a Positive and Negative Factor in Global Risk«. V Bostrom, N. in Čirković, M. M. (ur.), *Global Catastrophic Risks*, New York: Oxford University Press, str. 308–345.
- Yudkowsky, E. (2013). *Intelligence Explosion Microeconomics*. Berkeley, CA: Machine Intelligence Research Institute. Dostopno na spletnem naslovu: <https://intelligence.org/files/IEM.pdf>

Brin Blatnik

Filozofska fakulteta Univerze v Ljubljani

Problem racionalnega agensa

Koncept racionalnega agensa je pomemben temelj ekonomske teorije. V članku sledim njegovemu razvoju od zgodnjega razumevanja racionalnega agensa kot subjekta, ki sledi svojemu samointeresu, do poznega razumevanja racionalnega agensa kot subjekta, ki je konsistenten. Ob tem skušam pokazati, kako je ta teoretski razvoj vplival na izsledke ekonomske teorije in kakšne posledice ima na delovanje inštitucij, ki uporabljajo ekonomsko teorijo kot vodilo svojega delovanja.

Ključne besede: racionalni agens, samointeres, konsistentnost, teorija ekonomije

1. Zastavitev

V članku bom poskusil odgovoriti na dve vprašanji: kako je potekal razvoj teorije racionalnega agensa in kakšne so posledice sprememb teorije racionalnega agensa? Preden lahko odgovorim na ti vprašanji, pa je nujno, da razumemo, kaj je teorija racionalnega agensa in zakaj je pomembna.

1.1 Kaj je teorija racionalnega agensa?

Iz besede »agens«¹ je razvidno, da teorija govori o subjektu, iz pridevnika »racionalni« pa, da ima ta subjekt neke specifične lastnosti. Subjekt, o katerem govorimo, je subjekt, ki nastopa v ekonomskih modelih. Lastnosti tega subjekta, ki jih zaznamuje beseda »racionalni«, so osrednji predmet razprave v pričujočem članku in se skozi čas spreminjajo.

1.2 Zakaj je teorija racionalnega agensa pomembna?

Ekonomska teorija ima velik vpliv na delovanje družbenih ustanov. Njeni izsledki se uporabljajo pri odločitvah o državni potrošnji, pri določanju pravil bančne politike in ponavadi tudi pri upravljanju podjetij. Toda mnogo ljudi, ki te izsledke uporablja, nima časa za raziskavo teoretskih osnov, na katerih so osnovani. De-

¹ Beseda izhaja iz latinske *agere*, kar pomeni »nekdo, ki deluje«, in torej predstavlja nasprotje nečesa, kar ne deluje, je pasivno, je objekt.

lovne odločitve morajo biti namreč dovolj hitre, da se okoliščine, v katerih so narejene, ne spremenijo, in dovolj preproste, da jih lahko razumejo tudi ljudje, ki o ekonomiji ne vedo ničesar.

Takšen pristop pa ima tudi svoje slabosti. Odločitve, ki niso v celoti reflektirane, imajo pogosto nepričakovane posledice. Te posledice imajo lahko na naše življenje zelo negativen vpliv, kot se je denimo izkazalo v finančni krizi leta 2008. Tako se je vredno posvetiti raziskavi teoretskih osnov ekonomskih odločitev in tistim, ki za to nimajo časa, omogočiti boljši vpogled vanje.

Teorija racionalnega agensa je ena od takšnih pogosto zanemarjenih teoretskih osnov, saj je na njej osnovana večina preostale ekonomske teorije. Spremenjen pogled nanjo nam tako lahko spremeni tudi pogled na delovanje institucij, ki za svoje delovanje uporabljajo izsledke ekonomske teorije.

2. Razvoj teorije racionalnega agensa

2.1 Izvorno razumevanje racionalnosti kot maksimizacija samointeresa

Izvorno in najbolj poznano pojmovanje racionalnosti ekonomskega subjekta je *maksimizacija samointeresa*. Takšno pojmovanje srečamo že pri Adam Smithu, očetu mikroekonomije. V *Bogastvu narodov* denimo zapiše:

Obeda si ne obetamo zaradi mesarjeve, pivovarjeve ali pekove dobrohotnosti, marveč zato, ker se trudijo za svoj prid. Ne nagovarjamo njihove človečnosti, marveč sebičnost, in nikoli jim ne govorimo o svojih potrebah, marveč o njihovih koristih. (Smith, 2010: 22)

Morda najbolj neposredno pa idejo predstavi Francis Edgeworth v delu *Mathematical Physics*, kjer pravi: »Prvi princip ekonomije je, da vsakega agensa vodi le samointeres« (Edgeworth, 1881:16). Pri citatu sta za nas zlasti pomembna dva poudarka:

1. Edgeworth trdi, da je maksimizacija samointeresa *prvi* princip ekonomije. Trdi torej, da na tej predpostavki temelji ekonomija kot taka. Maksimizacija samointeresa ni le formalnost, temveč je temeljno razlagalno sredstvo ekonomske teorije. Brez te predpostavke za Edgewortha ekonomija ne obstaja, saj je samointeres edina sila, ki žene ekonomski subjekt.

2. Samointeres je *psihološko zaznamovan* pojem. Edgeworth ga uporablja v duhu klasičnega benthamovskega utilitarizma, saj je zanj maksimizacija samointeresov istovetna maksimizaciji ugodja oz. prijetnih občutkov nasploh (Edgeworth, 1881: 65). Tako je zgodnja teorija racionalnega agensa neločljivo povezana s subjektivnimi izkustvi.

2.2 Težave z razumevanjem racionalnosti kot samointeresa

Klasično razumevanje racionalnosti temelji na vprašljivi predpostavki. Če verjamemo, da je racionalnost maksimiziranje lastnega užitka, moramo najprej predpostaviti, da vse užitke lahko poenotimo v nekakšno skupno mero. Toda to je vse prej kot preprosta naloga: izkaže se, da ljudje v praksi enega užitka ali bolečine nismo sposobni primerjati z drugimi. Vstopiti v zapuščeno hišo, v kateri domnevno straši, nam povzroča nelagodje; prav tako pa nam povzroča nelagodje, ko se spečemo na vroči kuhalni plošči. Toda zdi se nemogoče eksaktno odgovoriti na vprašanje, koliko sekund bi morali držati roko na vroči plošči, da bi bilo nelagodje enako kot v primeru, ko vstopamo v hišo duhov. Dogodka se nam zdita različna, in tudi če poskušamo uganiti, kako naj bi se glasil odgovor, bodo rezultati, ki jih dobimo, zelo različni. Ravno to in nešteto drugih podobnih odločitev pa naj bi v svoji klasični različici opravljal racionalni agens.

Na prvi pogled se morda zdi, da gre pri vprašanju le za empirični problem, tj. da moramo opraviti čim več raziskav, v katerih ljudem ponudimo različne izbire in nato zabeležimo rezultate ter povlečemo ustrezne zaključke. Toda v praksi se je težava izkazala za nerešljivo. Sugden je o tem denimo zapisal:

Spomnimo se, da so prvi neoklasični ekonomisti, ki so delovali v poznem devetnajstem stoletju, razumeli racionalno izbiro v instrumentalnih pojmi: predpostavili so, da za vsakega posameznika obstaja enodimenzionalna izmera psihološkega ugodja, »util«, in da so racionalnost definirali kot maksimizacijo utilov. Te predpostavke so ekonomistom omogočale, da so uporabljali močna matematična orodja omejene maskimizacije in da so snovali elegantne splošne teorije iz preprostih premis. Ta način teoretiziranja je bil v esenci *a priori*, toda bil je osnovan na empiričnih predpostavkah o človeški naravi. Na žalost, predpostavljene enodimenzionalne mere ugodja niso našli (in stoletje za tem, še vedno ni najdena). To je pustilo sramotno vrzel v ekonomskih razlagah sveta. (Sugden, 2001: 127)

Ekonomisti so se problema skušali lotiti na več različnih načinov. Institucionalna ekonomska šola je na primer trdila, da ekonomija ne bi smela preiskovati posameznika, temveč institucije. Zanje je ideja o posamezniku kot o nekemu, ki maksimizira korist, zmotna. Posameznika namreč ne razume kot nekoga z vnaprej danimi lastnostmi, saj institucije, ki jih ustvari (tako namerno, na primer podjetje, kot nenamerno, na primer denar), neprenehoma vplivajo nanj in ga rekonstruirajo. Ali kot je to izrazil Veblen: »Želje in potrebe, cilji in nameni, poti in načini, moč in smer posameznikovega obnašanja, so funkcije institucionalne spremenljivke, ki je izredno kompleksna in povsem nestabilna« (Veblen, 1919: 242).

Primer takega sooblikovanja posameznika in institucij so oglaševalska podjetja. Posamezniki takšna podjetja ustvarijo in upravljajo na podlagi lastnih prepričanj in preferenc. Vendar pa oglaševalci vplivajo na prepričanja in preference posameznikov. Ta pojav imenujemo dvosmerna vzročnost in videti je, da je v nasprotju s

statičnim pogledom na posameznika, ki ga gojijo ortodoksne ekonomske šole (Hodgson, 2000).

2.3 Racionalnost kot konsistentost

Če so bili takšni alternativni pristopi nekaj časa razširjeni in priznani, pa so v sodobnem času potisnjeni na akademsko obrobje. Razlog za to ni le v teoretskih težavah, ki jih pestijo, temveč tudi v duhu časa, v katerem so nastajali. V zgodnjem 20. stoletju je namreč v Evropi cvetel logični empirizem, ki je ekonomiste spodbujal k temu, da so svojo teorijo poskušali obdržati empirično preverljivo ter ločeno od psiholoških faktorjev (Giocoli, 2003: 43; Creath, 2014).

Dober primer tega novega (pozitivističnega) pristopa je besedilo *Foundations of Economic Analysis*, ki ga je napisal Paul Samuelson. Samuelson že v samem uvodu kritizira takratne neempirične pristope in jim protipostavlja svojega:

Vzemimo malo slabe psihologije, kanček slabe filozofije in etike, dodajmo še dobro mero slabe logike in vsak ekonomist bo lahko dokazal, da je krivulja povpraševanja za dobrino padajoča /.../ s smiselnim teoremom mislim preprosto na hipotezo o empiričnih podatkih, ki bi jo lahko zamisljivo ovrgli v idealnih okoliščinah. (Samuelson, 1947: 4)

Samuelson v delu trdi, da preferenc sicer ne moremo opaziti ali izmeriti, lahko pa opazimo in izmerimo, kako posamezniki ravnajo v dani situaciji. Te meritve so zanj edini sprejemljivi temelj ekonomske teorije. V tekstu mu uspe pokazati, da lahko skoraj vse izsledke dotedanje teorije, osnovane na psiholoških faktorjih, razložimo z enim samim aksiomom, in sicer s *šibkim aksiomom razkritih preferenc*. V matematični notaciji ga izrazimo na naslednji način:

$$\sum_1^n p_i(x'_i - x_i) \leq 0 \Rightarrow \sum_1^n p'_i(x'_i - x_i) < 0$$

Znak »p« predstavlja ceno, znaka »x« in »x'« predstavljata množice dobrin, Σ pa nam pove, da pravilo drži za vse množice cen in dobrin. Iz aksioma sledi, da mora posameznik, ki je enkrat izbral množico dobrin x namesto množice dobrin x' , tako postopati vedno, tj. vedno bo izbral množico dobrin x namesto množice dobrin x' . Če kadarkoli izbere množico dobrin x' , ko bi si lahko privoščili množico x , je aksiom prekršen (Giocoli, 2005: 187). Povedano drugače, racionalni posameznik mora ravnati konsistentno s svojimi prejšnjimi izbirami. Razvidno je, da je ta nova teorija v celoti osnovana na elementih, ki jih lahko opazujemo in merimo, namreč cenah, dobrinah in prihodkih. Hkrati pa so razložene tudi vse empirične raziskave, ki so bile osnovane na stari teoriji. Zdi se, kot da so stare psihološke predpostavke odveč.

2.4 Težave s pojmovanjem racionalnosti kot konsistentnosti

Ko se spremenijo teoretski temelji teorije, se spremenijo tudi njene implikacije. V prejšnjem razumevanju racionalnosti je bilo človeško ravnanje razloženo kot posledica psiholoških vzgibov, običajno samointeresa. Če racionalnost pojmuje kot konsistentnost, pa ta obnašanja ne razlaga, temveč ga le opisuje. Lahko bi rekli tudi, da klasični ekonomisti s svojim konceptom racionalnosti kot samointeresa razlagajo, *zakaj* se ljudje obnašajo, kot se, neoklasični ekonomisti pa v svoji novi teoriji racionalnosti kot konsistentnosti razlagajo zgolj, *kako* se ljudje obnašajo (Guilhem, 2013: 47).

S to spremembo se je ekonomija oddaljila od svojih družboslovnih temeljev in se približala naravoslovnim. Naravoslovne znanosti opazujejo naravne pojave in iščejo stalne vzorce v njihovem delovanju. Te vzorce imenujejo znanstveni zakoni in jih uporabljajo za napovedovanje prihodnjih dogodkov. Enako poskušajo sodobni ekonomisti poiskati univerzalne zakone in konstante, ki naj bi usmerjale človeško delovanje.

Takšno mišljenje pa spremljajo tudi določene težave. Eno od njih je izpostavila t.i. *avstrijska šola ekonomije*, ki je bila dokaj vplivna v zgodnjem 20. stoletju, a je danes izgubila na ugledu, deloma zaradi njenih specifičnih pogledov na naravo ekonomskega subjekta. Pomislek, ki ga je na novo koncepcijo racionalnosti naperila avstrijska šola, se je glasil, da v primeru, ko so raziskovalni predmet ekonomske raziskave ljudje, univerzalnih konstant preprosto ne moremo najti, saj je obnašanje ljudi odvisno od njihovih prepričanj, vrednot in namenov, ti pa so svobodni in posledično nepredvidljivi.

Statistična raziskava, ki bi denimo iskala povezavo med povišanjem cene in količino opravljenih nakupov, bo odločitve vseh posameznikov poenostavila na eno samo številko. Ko pa se osredotočimo na delovanje *posameznih* ljudi, bomo opazili, da se na spremembo cene odzivajo zelo različno. Povišana cena bo morda koga prepričala, da je izdelek predrag in se tako za nakup ne bo odločil, kdo drug bo višjo ceno vzel kot znak kvalitete, tretji pa je morda še opazil ne bo. Tudi če obstajata dve ogromni skupini ljudi, kjer prva povsem preneha z nakupi, druga pa svoje nakupovanje poveča za trikrat, bo statistična raziskava pokazala le nek skupni trend, denimo naraščanje nakupov s padcem cen. To pa nam ne pove ničesar o tem, *zakaj* so ljudje sprejeli odločitve, ki so jih sprejeli, in ali bodo v prihodnje ravnali enako (Butler, 2010: 12–13).

Rezultate takšnih ekonomskih raziskav lahko tako uporabimo za napovedovanje prihodnosti le v primeru, če predpostavimo, da subjekti, ki jih preiskujemo, ne spreminjajo svojih vzorcev obnašanja in da bodo tudi v prihodnjih okoliščinah naredili enake izbire oz. da bodo ravnali konsistentno. Kot sem razložil v prejšnjem razdelku, je to sicer skladno s *predpostavkami* nove teorije racionalnega agensa, ni pa skladno z dejanskim delovanjem posameznikov. Tako se zdi, da sodobna

ekonomska teorija ustvarja rezultate, ki se skladajo z modelom, namesto da bi poskušala razložiti dejansko delovanje sveta.

3. Posledice sprememb

Če se ukvarjamo s sodobno ekonomsko teorijo, lahko opazimo, da obstaja v njej edinstvena teoretska napetost. Po eni strani modeli, ki jih razvijajo ekonomisti, temeljijo na matematičnih pravilih in osnovnih aksiomih. Z njimi se ekonomija poskuša oddaljiti od subjektivnih dejavnikov, kot so vrednote in nameni, ter pridobiti duh objektivnosti. Po drugi strani pa je predmet njenega raziskovanja ravno obnašanje posameznikov ali skupnosti, ki jih sestavljajo posamezniki. Ker je obnašanje posameznikov odvisno od subjektivnih faktorjev, kot so vrednote in nameni, ekonomijo večno zasleduje duh subjektivnosti.

Sprememba temeljne značilnosti racionalnega agensa iz samointeresa v konsistentnost je imela več pozitivnih posledic. Glavna med njimi je, da so lahko zaradi te spremembe ekonomski rezultati matematizirani in tako bolj nedvoumni in lažje preverljivi. Toda sprememba je imela tudi negativne posledice: bolj ko se ekonomski modeli matematizirajo in objektivizirajo, bolj so odmaknjeni od subjekta, ki ga raziskujejo, dokler ne dosežejo točke, ko se zdi, da preiskujejo nekaj povsem drugega.

3.1 Odgovor na težavo?

Če je težava v tem, da se novi ekonomski agens preveč razlikuje od obnašanja posameznikov, bo teorija uspešna le, v kolikor obravnava nekoga, ki se obnaša tako, kot se obnaša ekonomski agens, ki je v novi teoriji predpostavljen.

Kot je razvidno iz kritik, se dejanski ljudje od novega racionalnega agensa precej razlikujejo. Avstrijska in institucionalna šola sta dva primera takih kritik, obstajajo pa tudi mnoge druge. Ortodoksna teorija predpostavlja, da ima racionalni agens vnaprej določene preference, ki se ne spreminjajo, to pa za subjekta v svetu ne drži.

Kljub temu pa obstaja določen vzorec. Tisti, ki so se v šoli naučili ekonomskih modelov, pogosto verjamejo, da je sledenje takšnim modelom dobra odločitev. Posledično so njihove odločitve bližje tem, ki bi jih naredil novi racionalni agens. Toda povezava tu ni dovolj močna. Obstaja namreč drugi tip akterja, ki je racionalnemu agensu še bližje in ima na našo družbo še večji vpliv.

V mislih imam podjetja in druge institucije, ki sledijo ekonomskim načelom (na primer banke in določeni državni uradi). Za vodstvo teh so navadno zadolženi

ljudje s poslovno ali ekonomsko² izobrazbo. To so ljudje, ki so izšolani po ekonomskih načelih, osnovanih na teoriji racionalnega agensa. Hkrati pa so ti ljudje ravno v svoji službeni, poslovni funkciji najprevidnejši in se obnašajo tako, kot bi se moral obnašati racionalni agens, saj so bili naučeni, da je takšno obnašanje v poslovnem okolju, v katerem delujejo, primerno in potrebno. K temu velja dodati še, da obstaja tudi velika verjetnost, da bodo, če se previdno in konsistentno držijo načel, ki jih predpisuje ekonomska teorija, na svojem položaju napredovali in v vodstvu svojih organizacij prevzeli več vpliva. Njihova uspešnost se namreč meri ravno na podlagi zgoraj omenjenih ekonomskih modelov, ki so osnovani na teoriji racionalnega agensa.

Ta povezava je manj izražena v majhnih podjetjih. Tam vodstvo načela ekonomske teorije pozna slabše ali pa se za njih ne zanima. Toda večja in bolj vplivna, kot je institucija, bližja bo po svojem vedenju obnašanju novega racionalnega agensa, saj jo vodijo ljudje, ki njegove lastnosti poznajo bolje in jim sledijo natančneje.

Najbolj ekstremen primer te ideje so centralne banke, ki upravljajo z večjimi količinami denarja kot katerokoli podjetje, hkrati pa jih lahko upravljajo le ljudje, ki so bili slabo desetletje šolani o ekonomskih principih, osnovanih na teoriji novega racionalnega agensa. Ti ljudje so večinoma z raziskovalnim delom pomagali pri razvoju teh načel in bi v primeru, da bi od njih odstopali ali pa jih ne bi poznali, v trenutku izgubili službo. Toda podobno situacijo lahko opazimo tudi v večjih podjetjih, ki bi znotraj tržnega sistema ne postala velika, če njihovo vodstvo ne bi razumelo načel, na katerih je ta sistem osnovan in ravnalo v skladu z njimi.

Teorija, ki bo napovedovala obnašanje racionalnega agensa, bo tako pogosto pravilno napovedala obnašanje podjetij in institucij, saj se te sistemsko držijo pravil racionalnega agensa. Ekonomska teorija je s tem torej proizvedla svoj lasten objekt proučevanja in mu na podlagi njegovega delovanja naprej predpisuje napotke, po katerih naj se razvija.

Če privzamemo ta pogled na ekonomsko teorijo, se nam zastavita dve vprašanji. Prvo se glasi: »Ali sprejemamo *metodo* neoklasične ekonomske teorije?«, drugo pa: »Ali sprejemamo *rezultate* neoklasične ekonomske teorije?«.

² Z »ekonomsko izobrazbo« imam v mislih univerzitetne programe, ki poudarjajo učenje teorije ekonomije, s »poslovno izobrazbo« pa univerzitetne programe, ki dajejo poudarek na učenju praktične izvedbe.

4 Zaključek

Kot je moč razbrati iz našega kratkega prikaza razvoja racionalnega agensa, se je teoretska metoda ekonomije skozi zgodovino drastično spreminjala. Z besedami Giocolija:

[N]eoklasična ekonomija je logična, ne pa empirična znanost; njen vzornik je matematika, ne fizika; njena družbena struktura in sistem vrednot sta vzeta iz skupnosti matematikov. Vsi napredki teorije izbire in vrednosti v zgodnjem 20. stoletju, so v svojem bistvu formalni, a se pogosto skrivajo za fasado empirije. (Giocoli, 2003:45)

Sočasno z ostalo ekonomsko teorijo se je torej tudi racionalni agens spremenil v logični in ne v empirični konstrukt. Težava pri tem pa je, da se kljub temu še vedno predpostavlja ravno obratno, tj. da gre za empirični in ne za logični subjekt.

Če je za nas ekonomska teorija sprejemljiva le v kolikor raziskuje in predvideva obnašanje subjekta v svetu, bo odgovor na naše prvo vprašanje negativen, tj. metode neoklasične ekonomije ne bomo pripravljeni sprejeti. Še vedno pa ostaja možnost za drugačen pogled na ekonomsko teorijo.

Namesto da nanjo gledamo kot na vedo, ki preiskuje obnašanje subjekta, jo lahko pojmuje kot praktično orodje, ki napoveduje in predpisuje obnašanje ekonomskih sistemov. Tako razumljena ekonomska teorija bi lahko imela podobno vlogo, kot jo ima logika pri oblikovanju komunikacijskih sistemov in programiranju ali kot jo ima matematika pri šifriranju. Ne prva ne druga ne predpisuje ciljev, temveč služi le kot uporabno orodje za njihovo dosego.

Povedano drugače, delovanje denarja, bank, državnih financ itn. – vse to so izredno kompleksni procesi. Ker pa so to hkrati tudi izredno pomembni procesi, si zaslužijo, da jih obravnavamo dosledno in premišljeno. Neoklasična ekonomska teorija v njihovi analizi morda ni popolna, a je kljub napakam boljša od alternativ, ki smo jih preizkusili, in kaosa, ki bi nastal, če ti procesi ne bi imeli nikakršne teoretske podlage. Tako se po mojem mnenju prvo vprašanje, ki smo si ga zastavili: »Ali sprejemamo metodo neoklasične ekonomske teorije?«, pravzaprav zvede na drugo: »Ali sprejemamo rezultate neoklasične ekonomske teorije?«.

Tudi na to vprašanje ni mogoče podati dokončnega odgovora. Morda verjamemo, da sledenje ekonomskim načelom povzroča nesprejemljive ekonomske neenakosti med ljudmi in se nam zato zdi slabo. Toda kdo drug bo morda verjel, da so posledice dobre, saj v dobre namene izkoriščajo pohlep, ki je nespremenljiva osnova človeške narave, in na ta način zagotavlja blagostanje za vse. O posledicah sodobne ekonomske teorije pa lahko menimo tudi kaj tretjega. Zagovarjati eno ali drugo stran razprave je nekaj, v kar se na tem mestu ne bom spuščal; namesto tega bi rad poudaril nekaj drugega. Sprejemljivost teh posledic je nekaj, česar skozi prizmo sodobne ekonomske teorije ne moremo presojati. Ta namreč pokaže le,

kako maksimirati ugodje novega racionalnega agensa, ne subjekta v svetu. Morda se njegovo ugodje ujema z našimi vrednotami, toda na to nam ekonomska teorija ne daje odgovorov. Iz tega sledi, da potrebujemo od ekonomije neodvisno vedo, ki poseduje orodja, s katerimi lahko presoja o tem, kaj je dobro delovanje in kaj so zaželenе posledice. Takšna veda bi lahko postavljala cilje dobrega delovanja, ekonomska teorija pa bi lahko služila kot orodje za doseg te ciljev znotraj obstoječega sistema.

The Rational Agent Problem

The concept of a rational agent is a cornerstone of economic theory. In this paper, I follow its development from the early understanding of the rational agent as a subject that is realizing its self-interest to the later understanding of the rational agent as a consistent subject. I try to show how this evolution of theory influenced the findings of economic theory and what consequences it has on the workings of institutions that use economic theory as guidance for their decision making.

Keywords: rational agent, self-interest, consistency, theory of economic

Literatura

Butler, E. (2010). *Austrian Economics: A Primer*. London: Adam Smith Institute, str. 5–31.

Creath, R. (2014). »Logical Empiricism«. V Zalta, E.N. (ur.), *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition). Dostopno na: <http://plato.stanford.edu/archives/spr2014/entries/logical-empiricism/>

Edgeworth, F. (1881). *Mathematical Physics*. London: Routledge.

Giocoli, N. (2003). *Modeling Rational Agents: From Interwar Economics to Early Modern Game Theory*. Northampton, Mass.: Edward Elgar Publishing Limited.

Giocoli, N. (2005). »Modeling Rational Agents the Consistency View of Rationality and the Changing Image of Neoclassical Economics«. *Papers in Political Economy*, 49, str. 177–208.

Guilhem, L. (2013). »Reconciling Behavioural and Neoclassical Economics«. Raziskovalno delo 2013-09; [<hal-00819763>]. Dostopno na: <https://hal.archives-ouvertes.fr/hal-00819763>

Hodgson, G. (2000). »What Is the Essence of Institutional Economics?«. *Journal of Economic Issues*, 24(2), str. 317–328.

Samuelson, P. (1947). *Foundations of Economic Analysis*. London: Oxford University Press.

Smith, A. (2010). *Bogastvo narodov*. Ljubljana: Studia humanitatis.

Sugden, R. (2001). »The evolutionary turn in game theory«. *Journal of Economic Methodology*, 8(1), str. 113–130.

Veblen, T. (1919). *The Place of Science in Modern Civilisation and Other Essays*. New York: B. W. Huebsch, str. 231–251.

Borut Trpin

L. A. Paul: Transformative Experience

(Oxford: Oxford University Press, 2014; 208 strani)

Racionalno o prelomnicah

Filozofinja L. A. Paul, profesorica na ameriški UNC Chapel Hill in škotski univerzi St Andrews, je leta 1999 doktorirala na temo vzročnosti pod mentorstvom Davida Lewisa. Po vrsti člankov na temo vzročnosti in protidejstvenikov ter o metafizičnih vprašanjih, je pred kratkim v soavtorstvu z Nedom Hallom izdala monografijo *Causation: A User's Guide* [slov. Vzročnost: Priročnik za uporabnike] (2013), le leto kasneje pa še avtorsko monografijo *Transformative Experience* (2014) [slov. Transformativne izkušnje]; obe monografiji sta izšli pri Oxford University Press.

V slednji se avtorica odmakne od metafizike in filozofije znanosti na področje teorije odločanja in filozofije duha. Sooči nas s problemom transformativnih izkušenj, tj. tistih izkušenj, ob katerih se nepovratno in nepredvidljivo spremenimo. Monografija predstavlja pomemben in, predvsem, izviren doprinos k teoriji odločanja, saj problemu transformativnih izkušenj na tak način prej še ni bilo posvečene pozornosti.

Osnovno izhodišče, na katerem avtorica gradi svoje delo, je teorija racionalnega odločanja (TRO). Problem transformativnih izkušenj postane posebej zanimiv šele v tem okviru, saj za TRO predstavlja oviro. TRO je normativna teorija, ki racionalno odločanje opredeli kot odločanje, ki upošteva zakonitosti verjetnostnega računa v kombinaciji s predvidenimi koristmi (ang. *expected utility*). Če in samo če pri odločitvi upoštevamo zakonitosti verjetnostnega računa in predvidenih koristi, bo naša odločitev optimalna. Odločitve o transformativnih izkušnjah pa predstavljajo dvojno težavo, saj se po odločitvi nepredvideno spremenimo: ne moremo predvideti, kakšna je verjetnost določenega izida, niti kakšna je predvi-

dena korist. Ali to pomeni, da je TRO zaradi protiprimera transformativnih izkušenj pokopana?

To bi težko trdili. Avtorica v monografiji večkrat omeni, da je zagovornica TRO in da išče rešitev problema prav v tem okviru. Ne zanima je, ali je TRO utemeljena na pravih predpostavkah, temveč skuša naslovne transformativne izkušnje spraviti v kontekst racionalnega odločanja. Pri tem pa se zdi TRO nemočna. Odločevalka namreč nikakor ne more predvideti, kakšna je verjetnost oz. korist neke odločitve, saj se bo tudi sama spremenila, ko bo doživela izbrano izkušnjo.

Problem transformativnih izkušenj se morda sliši precej abstraktno, zato avtorica v ilustracijo najprej ponudi večje število miselnih eksperimentov, ki jih kasneje podkrepi s primeri iz resničnega življenja. S tem želi pokazati, da transformativne izkušnje ne predstavljajo zgolj zanimivega protiprimera TRO, temveč se nas tudi intimno dotikajo. Poglejmo enega izmed njenih miselnih eksperimentov.

Avtorica predlaga, da si zamislimo nevroznanstvenico Mary, ki je bolj znana iz Jacksonovega (1986) miselnega eksperimenta o kvalijah. Mary ima na razpolago vse podatke o tem, kako zaznavamo barve, vendar je že od rojstva zaprta v črno-beli sobi. Za razliko od Jacksonovega miselnega poskusa, ima Mary tokrat mnogo prijateljev in sorodnikov, ki živijo zunaj njene sobe, in ji poročajo o tem, kako čudovite so barve. A Mary je zadovoljna s svojim življenjem v črno-beli sobi. Drugačnega življenja si niti ne zna predstavljati. Avtorica zato provokativno vpraša: Ali naj Mary sprejme ponudbo, da stopi ven iz sobe, in – predvsem – na podlagi česa, naj se odloči?

Mary se odloča o naslovni transformativni izkušnji – izkušnji, ki jo spremeni v več vidikih. Dva ključna vidika, ki ju avtorica izpostavlja, sta spoznavni in osebni vidik. Pri odločanju o izhodu iz sobe Mary trči na spoznavno čer. Ne more si pomagati s preteklimi izkušnjami, saj barv še ni videla. Prav tako si ne more pomagati s kognitivnim modeliranjem – nikakor si ne more zamisliti, *kako je biti* Mary v barvnem svetu. Nima niti možnosti, da hipno pokuka skozi špranjo in se odloči, ali so ji barve vseh ali ne. Ko bo barve enkrat zagledala, se ne bo več mogla vrniti v stanje nepobarvanega lista. Ker take izkušnje še ni imela, tudi ne more presoditi, kolikšna je verjetnost, da ji bodo barve vseh, in kolikšna, da ne. Kako naj se torej racionalno odloči?

Čeprav miselni poskus z Mary podaja zanimivo vprašanje in vsaj na prvi pogled predstavlja protiprimer TRO, pa se problem transformativnih izkušenj še vedno zdi razmeroma nerelevanten. Avtorica omenja več podobnih dilem o tem, kako naj se odločimo, ali naj postanemo nesmrtni vampirji, ali naj postanemo zasvojeni (in domnevno zadovoljno omamljeni) in ali naj poskusimo durian, smrdljiv eksotični sadež, ki je nekaterim izredno okusen.

Teh primerov ne navaja zgolj za zabavo, temveč z njimi vzpostavi izhodišče za obravnavo transformativnih izkušenj v vsakdanjem življenju. Jedro njene knjige

je prav v aplikaciji teh ilustracij na velike življenjske prelomnice, s katerimi se vsi prej ko slej srečamo. Tako se med ostalimi prelomnicami podrobneje posveti vprašanju o izbiri poklica in načrtovanju družine. Kot šaljivo pripomni, to vprašanje z vampirji nima skupnih le neprespanih noči.

Bralcem, ki nas vseskozi drugoosebno nagovarja, postavi vprašanje, ali se lahko racionalno odločimo, da bomo imeli otroka. Podobno kot pri odločitvi o tem, da postanemo vampirji, gre za nepovratno odločitev, pri kateri si z zamišljanjem *kako je biti starš* ne moremo pomagati (pri tem se nanaša na Nagelovo vprašanje, *kako je biti netopir*; Nagel, 1974), saj nam je to izkustvo pred uresničitvijo subjektivno nedostopno, prav tako pa tudi nimamo izkušenj s starševstvom.

Avtorica opozarja, da se pri odločitvi tudi ne moremo opreti na pričevanja ostalih staršev. Mnogi starši namreč poročajo, da je bilo rojstvo oz. posvojitev otrok(a) najboljša odločitev v njihovem življenju. Življenja brez otrok(a) si ne znajo predstavljati, prinaša pa jim tudi neusahljiv vir zadovoljstva. Znanstvene raziskave, ki jih navaja avtorica, pa na drugi strani kažejo, da med zadovoljstvom parov z otroki in pari brez otrok ni razlik oz. da so starši celo nekoliko manj zadovoljni. Racionalna odločitev iz tretjeosebne perspektive bi torej bila, da živimo brez otrok ne glede na naše želje in (očitno iluzorna) pričakovanja. A vendar je staršem po transformativni izkušnji z otrokom običajno nepredstavljivo, da bi se vrnili v prejšnje stanje. Odpoved prvoosebnemu stališču naj bi bila torej racionalna, s čimer pa se avtorica ne strinja. Njena temeljna predpostavka, ki pa jo utemelji le površno (navaja npr. vrednoto avtentičnosti v odločanju), je, da moramo v odločanju vedno upoštevati (tudi) prvoosebno stališče. V kombinaciji te predpostavke in TRO zato pri odločanju poziva k formalni kognitivni fenomenologiji. Termin sicer uvede v analogiji s formalno epistemologijo, čeprav ga le površno omeni.

Avtorica skozi celotno knjigo vpeljuje primere, s katerimi želi pokazati, kako pogoste in pomembne so odločitve o transformativnih izkušnjah, in kako je racionalna odločitev v teh primerih nemogoča. V zaključku (izvorno: *Afterword*), ki obsega več kot četrtno knjige, presoja različne pristope k reševanju problema, npr. preko pojmov nezavedanja (ang. *unawareness*), neujemljivih dispozicij in hierarhičnih bayesovskih modelov. Ključen pa je avtoričin predlog, da k problemu pristopimo na drugačen način. Ne moremo vedeti, kako-je-biti po transformativni izkušnji in kakšna je verjetnost (ne)ugodnega razpleta, lahko pa problem postavimo v drugačen okvir. Namesto vprašanja, ali naj se odločim za transformativno izkušnjo, se vprašamo, ali si želimo izvedeti, kaj ta izkušnja prinaša. Ali si želimo razodetja ali pa nam trenutno stanje ustreza?

Prav v tem spoznanju se skriva glavno sporočilo obravnavane knjige: (racionalne) odločitve o transformativnih izkušnjah zahtevajo drugačno obravnavo kot »običajne« odločitve. Racionalno se lahko odločimo zgolj, ali si želimo neznane iz-

kušnje, kar naj bi bilo tudi racionalno vodilo v problemih odločanja o velikih življenjskih odločitvah.

Čeprav obravnavano delo uvaja zanimiv problem za teorijo racionalnega odločanja, prav zaradi svoje pionirskosti prinaša tudi kar nekaj težav. Ena izmed težav je predpostavka racionalnosti (kot optimalnosti v odločanju), ki ji je avtorica vsekozi zavezana. Intuitivno se zdi nerazumljivo, da bi se odrekli rojstvu otrok zaradi tretjeosebni znanstvenih izsledkov o zmanjšanem zadovoljstvu staršev v primerjavi s pari brez otrok. A hkrati bi strogo vztrajanje pri racionalnosti od odločevalke zahtevalo prav to. Avtorica pa se zaradi tega ne odpove principu racionalnosti, temveč sklene, da so znanstveni izsledki za posameznika nerelevantni, ker je vsak posameznik vpet v drugačen kontekst. Trdi, da znanost trenutno oz. morda nikoli ne bo tako individualizirana, da bi lahko empirično predpisala odločitev za vsakega posameznika. Pri tem se poraja vprašanje – če avtorica v taki meri poudarja dejanske posameznike, ali ne bi potem morala poudarjati tudi, da racionalnost v strogo normativnem smislu odpade? In ker se kot dejanski akterji ne odločamo v skladu z matematičnimi normami TRO, ali so transformativne izkušnje sploh relevanten problem za posameznike? Če niti znanstveni izsledki niti pričevanja o določeni transformativni izkušnji niso dovolj relevantni za odločitev, na podlagi česa se potem sploh lahko odločamo? Prav zaradi teh problemov bi na podlagi avtoričine obravnave sledilo, da se ne in se niti ne moremo racionalno odločiti. A v tem sklepu se nahaja za TRO najbolj ogrožajoča točka. TRO je pred odločitvami o transformativnih izkušnjah po avtoričinem mnenju bistveno nemočna, a zdi se, da je vsaka odločitev v nekem smislu odločitev o transformativni izkušnji.

V ilustracijo: predpostavimo, da se Akter pri slaščičarju odloča o preprosti izbiri. Ali naj izbere torto ali sladoled? Ob vseh dosedanjih obiskih mu je bila torta ljubša. Avtorica bi sklenila, da gre za odločitev o netrtransformativni izkušnji. Vendar pa bi z njenim lastnim pristopom lahko tudi v tem primeru govorili o transformativnosti. Kako lahko Akter ve, da se mu ni spremenil okus (notranja sprememba) ali pa da ni slaščičar ta dan namesto sladkorja pomotoma uporabil sol (zunanja sprememba)? Jasno se nam zdi, da se bo racionalno odločil, če bo izbral torto (kar zapoveduje tudi TRO), saj se zanaša npr. na dosedanje zanesljivost, a vendar bi avtorica – če bi svoji zamisli striktno sledila – morala sprejeti pomisleke. V čem se torej Akterjeva odločitev bistveno razlikuje od odločitve o rojstvu otroka?

Zdi se torej, da avtorica implicitno loči dva različna problema, kot transformativno odločitev pa implicitno sprejema le drugega:

- a) problem odločitev z nepredvidljivim izkupičkom, in
- b) problem »velikih« odločitev z dolgotrajnimi posledicami.

Če govorimo o transformativnih odločitvah samo takrat, ko ima odločitev dolgotrajne in povsem nepredvidljive posledice, kako lahko v trenutku odločanja sploh vemo, da se odločamo o transformativni odločitvi? Odločitve, ki so privedle do transformativne izkušnje, se kot take nujno izkažejo šele retrospektivno in v trenutku odločanja ne predstavljajo težave za TRO. Nevroznanstvenica Mary, ki je v obravnavani knjigi v kontaktu s prijatelji izven svoje črno-bele sobe, se racionalno odloči, da stopi ven, ker ji vsi prijatelji in drugi viri pričajo o tem, da je barven svet prijeten. Po drugi strani pa ji je tudi življenje v sobi po mnogih letih že povsem naravno in prijetno. Predpostavimo, da so ji barve, ko stopi ven, vseč in zasovraži življenje v črno-beli sobi. Čeprav doživi transformativno izkušnjo, pa dokler je bila v sobi, nikakor ni mogla vedeti, da se odloča o transformativni izkušnji, saj bi njeno stališče lahko ostalo nespremenjeno. Kot transformativna se izkušnja torej pokaže šele za nazaj. Če torej sprejmemo avtoričino zamisel, da je treba te odločitve preuokviriti, smo spet pri težavi, da v trenutku odločanja ne vemo zanesljivo, kdaj odločitev preuokviriti.

Poleg tega avtorica vseskozi govori o individualnih odločitvah, vendar pa so mnoge transformativne izkušnje skupinske. Tak je že centralni primer obravnavanega dela, odločitev o otroku. Ko se par odloča za otroka, bosta transformativni izkušnji podvržena oba bodoča starša. Na tem mestu se porajata dve vprašanji, ki ostaneta neodgovorjeni:

1) Prvoosebno edninsko: Kako lahko v svoje odločanje zajamemo nekoga drugega, ki bo podvržen transformativni izkušnji?

in

2) Prvoosebno dvojinsko oz. množinsko: Kakšen vpliv, če sploh, imajo transformativne izkušnje v procesih skupinskega odločanja?

Pod (2) bi poleg avtoričinega odločanja o otroku vsekakor padle tudi velike zgodovinske skupinske odločitve, npr. plebiscit.

Obravnavano delo torej prinaša pomemben uvid v to, da ima klasična TRO v jedru omejitve, ki jih je potrebno nasloviti. Avtorica sicer pri tem ni prva. Tako že sama omeni matematično nedoločljive koristi v t.i. pasadenski igri (Nover in Hájek, 2004), kjer TRO naleti na paradoksalne posledice (izkupiček iste igre je lahko hkrati pozitiven, negativen in celo neskončno velik). Vseeno pa, kot je zgoraj na kratko orisano, ostaja vprašanje, ali transformativne izkušnje dejansko predstavljajo problem enake magnitude kot matematično nedoločljive koristi. Zdi se, da se proti TRO postavljajo na neki povsem drugi (nematematični) višji ravni.

Problem transformativnih izkušenj se vsemu navkljub kaže kot pomemben. Le da se moramo (za razliko od obravnavanega dela) odločiti, na kateri način ga bomo naslovili. Na tem mestu vidimo dve možnosti:

- 1) Normativna, strogo racionalna smer: razvijanje novih formalnih pristopov na način, ki ga avtorica imenuje formalna kognitivna fenomenologija. Pri tem »bonus« točke v odločanju pripišemo avtentičnosti in posameznikovi fenomenologiji, omejimo pa se na idealne racionalne akterje.
- 2) Mehkejša smer: govorimo o običajnih ljudeh. O odločitvah, ki privedejo v transformativne izkušnje, govorimo, ko nam odločitev bistveno zamaje poprej predvideno oceno koristi. Ker se zdi, da svoje stališče do neke mere spremenimo po vsaki odločitvi, bi bil tako bolj na mestu stopnjevit pristop: od povsem transformativnih izkušenj do tistih, kjer se predvidena korist sploh ne spremeni.

Glede na to, da trenutno pod vodstvom avtorice na University of Notre Dame in University of North Carolina, Chapel Hill poteka obsežen interdisciplinarni raziskovalni projekt na temo transformativnih izkušenj s podporo fundacije Templeton, se nadajamo, da bodo naši pomisleki v prihodnje razrešeni.

Literatura

- Jackson, F. (1986). »What Mary Didn't Know«. *Journal of Philosophy*, 83, str. 291–295.
- Nagel, T. (1974). »What Is it Like to Be a Bat?«. *The Philosophical Review*, 83, str. 435–450.
- Nover, H. in Hájek, A. (2004). »Vexing Expectations«. *Mind*, 113, str. 237–249.