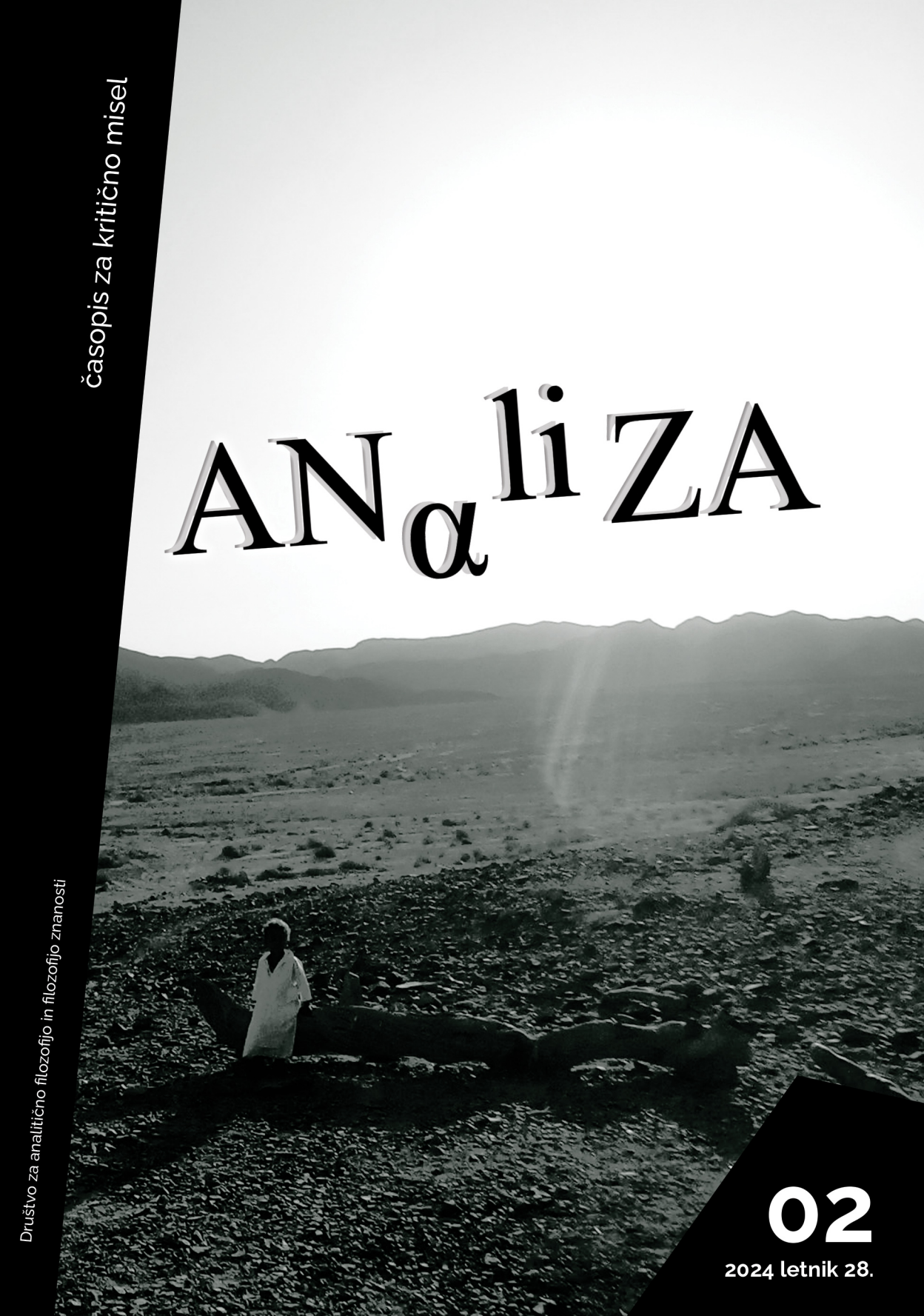
časopis za kritično misel

# ANαliZA



Društvo za analitično filozofijo in filozofijo znanosti

02
2024 letnik 28.

# ANαliZA

časopis za kritično misel

ANαliZA

časopis za kritično misel

**Glavni in odgovorni urednik** Tadej Todorović
(Univerza v Mariboru, Filozofska fakulteta, Slovenija)

**Lektoriranje** MODUS TOLLENS, Tadej Todorović s.p.

**Tehnični urednik** Jan Perša
(Univerza v Mariboru, Univerzitetna založba, Slovenija)

**Oblikovanje ovitka** Jan Perša
(Univerza v Mariboru, Univerzitetna založba, Slovenija)

**Grafika na ovitku** Mlada Benduinka, 20 km zahodno od Marsa Alam, Egipt, foto: Sata Šešerko, 2023

**Grafične priloge** Viri so lastni, razen če ni navedeno drugače.

**Vrsta revija** E-revija (on-line)

**Dostopno na** https://journals.um.si/index.php/analiza

**SEDEŽ UREDNIŠTVA**
**Analiza**
**Društvo za analitično filozofijo in filozofijo znanosti**
Aškerčeva cesta 2
1000 Ljubljana
e-pošta: info-daf@guest.arnes.si, https://daf.splet.arnes.si/analiza/

**ZALOŽNIK**
**Univerza v Mariboru, Univerzitetna založba**
Slomškov trg 15, 2000 Maribor, Slovenia
e-pošta: zalozba@um.si, https://press.um.si/, https://journals.um.si/

**ISSN 1408-2969 (tiskana izdaja)**
**ISSN 2712-4916 (On-line)**

**Revija je indexirana v:** The Philosopher's Index, Ulrich's, Philosophy Documentation Center (International Directory of Philosophy).

**Analiza: časopis za kritično misel** na leto izide v dveh zvezkih. Cena zvezka v prosti prodaji je 6,00 EUR, za naročnike 4,00 EUR oz. letna naročnina 8,00 EUR (z vključenim DDV). Naročila sprejemamo na naslov uredništva.

**Rokopise** nam lahko pošljete na elektronski naslov info-daf@guest.arnes.si, kjer jih sprejema glavni in odgovorni urednik revije.

Univerzitetna založba
Univerze v Mariboru

Izid te revije je sofinancirala
Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije

Javna agencija za znanstvenoraziskovalno
in inovacijsko dejavnost Republike Slovenije

ANαliZA
časopis za kritično misel

**Letnik XXVIII** **Številka 2** **December 2024**

# Politična filozofija

**ANALIZA**
Let. 28, št. 2, str. 153–166, december 2024

ANαliZA

# MED PRAVILI IN SVOBODO: RAWLS IN ILLICH O VLOGI INSTITUCIJ V DRUŽBI

**Sprejeto**
1. 11. 2024

**Pregledano**
16. 12. 2024

**Izdano**
31. 12. 2024

JANA VRDOLJAK
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
jana.vrdoljak@gmx.com

DOPISNI AVTOR
jana.vrdoljak@gmx.com

**Izvleček** Prispevek primerja Illichovo in Rawlsovo razumevanje vloge institucij pri oblikovanju pravične družbene ureditve in zagotavljanju posameznikove avtonomije. Rawls družbene institucije obravnava kot ključni temelj pravične in stabilne družbene strukture, medtem ko Illich opozarja na njihove omejitve, zlasti v zvezi z legitimnostjo ter politično, ideološko in gospodarsko nevtralnostjo. Po Illichovem mnenju institucije lahko ogrožajo posameznikovo svobodo, ko se iz sredstev za dosego ciljev spremenijo v cilje same po sebi. Kot alternativo predlaga družbo sožitja, ki zavrača tista orodja, torej tiste sisteme, tehnologije in institucije, ki bi se lahko izrodile v nadzor ali omejevanje svobode. Vendar pa tudi ta rešitev ni brez tveganj, saj lahko zapade v enake težave kot institucije, tako da prevzame določene oblike nadzora in omejevanja, značilne za tradicionalne skupnosti. Kljub tem omejitvam Illichova kritika opozarja na potrebo po reformi institucij, da bi te dejansko služile zaščiti posameznikove svobode in avtonomije.

**Ključne besede**
Illich,
Rawls,
družbene institucije,
svoboda,
pravičnost,
posameznikova
avtonomija

https://doi.org/10.18690/analiza.28.2.153-166.2024


Univerzitetna založba
Univerze v Mariboru

ANαliZA

**ANALIZA**
Vol. 28, no. 2, pp. 153–166, December 2024

# BETWEEN RULES AND FREEDOM: RAWLS AND ILLICH ON THE ROLE OF INSTITUTIONS IN SOCIETY

**Accepted**
1. 11. 2024

**Revised**
16. 12. 2024

**Published**
31. 12. 2024

JANA VRDOLJAK
University of Maribor, Faculty of Arts, Maribor, Slovenia
jana.vrdoljak@gmx.com

CORRESPONDING AUTHOR
jana.vrdoljak@gmx.com

**Abstract** The article compares Illich's and Rawls's understanding of the role of institutions in shaping a just social order and ensuring individual autonomy. Rawls regards social institutions as the essential foundation of a just and stable societal structure, while Illich highlights their limitations, particularly concerning legitimacy and political, ideological, and economic neutrality. According to Illich, institutions can threaten individual freedom when they shift from being the means for achieving goals to becoming ends in themselves. As an alternative, he proposes a convivial community that rejects tools – namely systems, technologies, and institutions – that could degenerate into mechanisms of control or restrictions on freedom. However, even this solution is not without risks, as it may fall into the same problems as institutions by adopting certain forms of control and restriction characteristic of traditional communities. Despite these limitations, Illich's critique underscores the need for institutional reform to ensure that they truly serve the protection of individual freedom and autonomy.

**Keywords**
Illich,
Rawls,
social institutions,
liberty,
justice,
individual autonomy

https://doi.org/10.18690/analiza.28.2.153-166.2024


University of Maribor Press

**Primerjava Rawlsove dobro urejene družbe in Illichove družbe sožitja**

Ena od začetnih težav je, da se Illich nikjer v svojih delih ne sklicuje na Rawlsovo družbeno teorijo. Druga težava pa je, da oba obravnavata družbeno teorijo, vendar izhajata iz različnih izhodišč: medtem ko Rawls ponuja idealno teorijo pravičnosti, se Illich osredotoča na dejansko družbeno stanje. Rawls (1999) predlaga univerzalno teorijo pravičnosti, medtem ko Illich zagovarja lokalne, geografske, kulturno specifične in okolju prijazne rešitve v obliki *sožitnosti* (Illich, 1973b). Rawls ponuja univerzalno teorijo družbe, ki pa je Illichu zaradi njene abstraktne narave in neskladja z dejanskimi družbenimi pogoji neizvedljiva. Prepričan je, da teh sprememb ni mogoče povsem zajeti v abstraktni teoriji, temveč jih lahko doživimo le skozi dejansko prakso (Illich, 1973a, str. 17). Rawls zagovarja zaprto družbo, ki jo opisuje kot »zaprt sistem, izoliran od drugih družb« (Rawls, 1999, str. 7), medtem ko Illich podpira odprto in gostoljubno skupnost (Cayley, 2005).

Skupno obema avtorjema je ukvarjanje z družbenimi institucijami, vendar pa preseneča, da Rawls ne obravnava ključnih področij, kot so zdravstvo, transport, šolstvo in zakonodaja. Ravno te institucije igrajo osrednjo vlogo v Illichovi kritiki, saj se sprevržejo v trenutku, ko postanejo same sebi namen, namesto da bi ostale sredstva za dosego širših ciljev (Illich, 1973b). V tem se razkriva ključna razlika med Illichovim in Rawlsovim pojmovanjem družbe: Rawls postavlja idealne pogoje delovanja institucij, ki naj bi zagotavljali trajno delovanje pod poštenimi pogoji, kar pa v tem pogledu deluje utopično (Rawls, 2011, str. 33). Illich trdi, da nobeni institucionalni instrumenti ne morejo zagotoviti trajnosti poštenega sodelovanja, saj institucionalizacija in vzdrževanje obstoječih družbenih struktur pogosto vodita v reproduciranje neenakosti in omejevanje posameznikove svobode. S primerom Cerkve, ki naj bi bila vzor in imela višje moralne standarde, Illich opozarja, da lahko institucije povzročijo škodljive posledice za celotno družbo, kar poudari z izrazom »corruptio optimi pessima[1]« (Cayley, 1992; Cayley, 2003; Cayley, 2005). Cerkev je poskušala zagotoviti krščansko sporočilo ljubezni tako, da je to ljubezen poskušala jamčiti z »institucionalizacijo, povzdigovanjem v zakonodajo, uzakonitvijo in njenim varovanjem preko kriminalizacije njenega nasprotja« (Cayley, 2003). Kot primer izkoriščanja navaja misijonarstvo kot ideološko orodje v službi globalnih bojev: »Priznati moramo, da so lahko misijonarji le figure v globalnem ideološkem boju in da je bogokletno uporabljati evangelij kot oporo za kateri koli družbeni ali politični

[1] Pokvarjenost najboljših je najslabša.

sistem« (Illich, 1973a, str. 55). Za razliko od Rawlsovega utopičnega prepričanja v dobro urejeno družbo, Illich na primeru institucije Cerkve izpostavi prisotnost zla, saj trdi, da čeprav so institucije pogosto ustanovljene z dobrim namenom, lahko postanejo orodje za zlorabo moči in zatiranje posameznikove svobode. Po Illichovem prepričanju bi se morala Cerkev odpovedati svoji »moči za delanje dobrega«, če bi želela znova uresničiti pristno sporočilo evangelija (Illich, 1973a, str. 84).[2]

Poleg Cerkve je imel Illich kritično stališče tudi do obveznega šolstva, saj je prepričan, da »niti učenje niti pravičnost nista spodbujena s šolanjem. Večina učenja poteka spontano /.../ in ni rezultat načrtovanega poučevanja« ter trdi, da smo se vsega, »kar vemo, /…/ naučili izven šole« (Illich, 1972). Illich ni proti šolstvu kot takemu, temveč proti obveznemu šolstvu, ki se lahko izrodi v orodje, s katerim se celotna družba skozi generacije manipulira v poslušne državljane. Ti sprejemajo ideološko sliko sveta kot nekaj samoumevnega, naravnega in neizpodbitnega, podobno kot zrak, ki ga dihajo, ali vodo, ki jo pijejo. S skrbno oblikovanim učnim načrtom, disciplino in rutino šolstvo vsiljuje določene vrednote, norme in poglede na svet. Učenci se v tem procesu učijo sprejemati te ideje kot naravne in samoumevne, brez kritičnega premisleka ali zavestnega preizpraševanja, s čimer šolstvo utrjuje obstoječe družbene hierarhije in ohranja prevladujočo ideologijo (The Golden Thread, 2021). Zaradi tega razume predpostavko, da je »šolanje /…/ zagotovilo socialne integracije«, kot mit liberalizma (Illich, 1973a, str. 92). Ker šole ocenjujejo, hkrati delujejo kot mesta, kjer so državljani »šolani« v svoje vloge (Illich, 1973a, str. 97), kar povezuje šolanje s posameznikovo samopodobo (Illich, 1973a, str. 103). Po njegovem mnenju »šola ni le nova svetovna religija, temveč tudi najhitreje rastoči trg dela na svetu« (Illich, 1972). Zato meni, da v družbi sožitja, o kateri bomo razpravljali v nadaljevanju, ni prostora za obvezno šolstvo:

> »V družbi sožitja bi bilo treba zaradi pravičnosti izključiti obvezno in neskončno šolanje. Starostno specifično, obvezno tekmovanje na neskončni lestvici za vseživljenjske privilegije ne more povečati enakosti, temveč mora favorizirati tiste, ki začnejo prej, so bolj zdravi ali bolje podprti zunaj učilnice.

[2] Illich je prenehal aktivno sodelovati na II. vatikanskem koncilu, ker je nasprotoval temu, da Cerkev v dokumentu *Gaudium et Spes*, ki obravnava njeno delovanje v sodobnem svetu, ne obsoja vlad, ki posedujejo jedrsko orožje – sredstvo za izvajanje genocida (Cayley, 2003). V eseju »*The Vanishing Clergyman*« rimokatoliško Cerkev, kot največjo nevladno birokracijo, po njeni učinkovitosti in uspešnosti primerja z General Motors Company in Chase Manhattan Bank (Illich, 1973a, str. 61).

> Neizogibno razslojuje družbo v številne plasti neuspeha, pri čemer vsako plast naseljujejo tisti, ki so izpadli iz sistema in so bili naučeni verjeti, da si tisti, ki so pridobili več izobrazbe, zaslužijo več privilegijev.« (Illich, 1973b)

Illich opozarja, da institucije same po sebi ne morejo zagotoviti pravičnega delovanja posameznikov, ki delujejo znotraj njihovih okvirov. Njegova kritika cilja na institucionalizacijo kot proces, ki pogosto reproducira sistemske neenakosti, instrumentalizira posameznike ter jim onemogoča avtonomijo odločanja in svobodo izbire, kar vodi v odtujenost in povečanje individualizma. Po Illichu institucije ne le da omejujejo posameznikovo svobodo, temveč pogosto postanejo orodja za vzdrževanje obstoječih socialnih in gospodarskih struktur, ki ohranjajo neenakost. Pri tem se osredotoča na nenehno gospodarsko rast, ki vzdržuje to stanje. V tem kontekstu Illich pravi:

> »Družba se lahko uniči, ko nadaljnja rast masovne proizvodnje naredi okolje sovražno, ko pogubi svobodno uporabo naravnih sposobnosti članov družbe, ko loči ljudi drug od drugega in jih zapre v umetno lupino, ko spodkopava strukturo skupnosti s spodbujanjem skrajne družbene polarizacije in drobljenja specializacij, ali ko rakava pospešitev nalaga družbene spremembe s hitrostjo, ki izključuje pravne, kulturne in politične presedane kot formalne smernice za trenutno vedênje. Korporativna prizadevanja, ki tako ogrožajo družbo, ne morejo biti tolerirana.« (Illich, 1973b)

Rawlsove moralne zmožnosti, poštenost in občutek za dobro so v osnovi nezadostne za reševanje Illichovih pomislekov o pravičnosti. Rawls sam opozarja na problem sistemske pristranskosti institucij, kar poskuša nasloviti s konceptom izhodiščnega položaja za tančico nevednosti. S tem posameznike abstrahira od njihovih konkretnih življenjskih okoliščin, da bi zagotovil enake ali poštene pogajalske pogoje (Rawls, 1999). Vendar pa to daje slutiti dojemanje človeške narave, ki je tako zelo zaznamovana s sebičnostjo, da naj bi bila sposobna le racionalne preračunljivosti in ne sočutja do drugih, kar spominja na Hobbesovo stališče o naravi človeka. Če za doseganje pravičnosti potrebujemo tako skrajno obliko abstrakcije, se pojavi vprašanje, zakaj bi privilegirani posamezniki sploh privolili v takšne pogoje. To razkriva širšo težavo Rawlsove teorije – vprašanje, ali ta resnično odraža kompleksnost človeške narave in ali lahko naslovi dinamične odnose moči in interesov, ki prevevajo institucionalna okolja.

Illichova kritika institucij se osredotoča na njihovo proceduralno naravo, ki temelji na ustaljenih vzorcih delovanja in natančno določenih pričakovanjih. Illich poudarja, da morajo institucije ostati odprte za nepredvidljive interakcije posameznikov, saj je prav v tej spontanosti ključni dokaz človeške svobode (Cayley, 2021). V tem kontekstu Rawlsov izhodiščni položaj, kjer pogodbeniki za tančico nevednosti dogovarjajo načela pravičnosti, pritegne pozornost. Čeprav Rawls posameznikom pripisuje moralne zmožnosti, kot sta čut za pravičnost in sposobnost pojmovanja dobrega, je vprašljivo, zakaj jih obenem reducira na abstraktne, atomistične subjekte. Illich, z osredotočenostjo na telesnost in konkretne življenjske okoliščine, to razume kot zanikanje človekove polnosti. Nadalje se pridružujemo kritiki Sena in drugih, ki dvomijo, da bi lahko edini možni rezultat takega dogovora bila oblika dveh načel pravičnosti, ki ju zagovarja Rawls. Zdi se, da je nepredvidljivost za Rawlsa nesprejemljiva, saj jo poskuša odpraviti z vzpostavitvijo stabilnega institucionalnega sistema, kjer so pravila in načela jasno določena. V tem pristopu daje večji poudarek institucijam kot posamezniku, kar vodi k zanemarjanju njegove aktivne in svobodne vloge znotraj sistema. Illich pa ravno v nepredvidljivosti vidi ključni dokaz človekove svobode in jo povzdiguje na raven misterija. Za njega je človekovo delovanje najpristnejše, kadar presega ustaljene okvirje in se izraža v spontanem odzivu na drugega – ne kot izpolnjevanje abstraktnih dolžnosti, temveč kot odgovor na konkretni klic. Hkrati se zaveda, da svoboda ni nujno povezana z dobrim ali z izvrševanjem dolžnosti. Ker se izogiba izrazom, kot je »dolžnost«, priznava tudi možnost, da posameznik ne odgovori na klic in da sočutje preprosto ni prisotno. Takšna možnost je za Rawlsa težko predstavljiva, zato morda, ker ne najde ustrezne rešitve, naivno predpostavlja, da se bodo vsi vedno ravnali v skladu z dogovorjenimi načeli. Rawlsova domneva, da »vsak ravna pravično in prispeva svoj delež k ohranjanju pravičnih institucij« (Rawls, 1999, str. 8), vzbuja dvom v poenostavljeno dojemanje človeškega vedênja, kot to trdi Sen (2009). Rawls zanemarja vprašanje, kako njegovo pojmovanje pravičnosti kot poštenosti vpliva na ljudi in njihovo medsebojno sodelovanje. Sen meni, da bi moral Rawls načela pravičnosti povezati z dejanskim ravnanjem ljudi (Sen, 2009, str. 69). Ker je »vedênje ljudi v celoti skladno z zahtevami za ustrezno delovanje teh institucij« (Sen, 2009, str. xi), Sen v preveliki osredotočenosti na institucije vidi pomanjkljivost Rawlsove teorije. Po njegovem mnenju bi se teorija morala osredotočiti na dejanska življenja ljudi, »kar bi imelo številne daljnosežne posledice za naravo in obseg pojma pravičnosti« (Sen, 2009, str. xi). Institucije in proceduralna pravičnost so pomembne, vendar ne morejo zamenjati ali nadomestiti vloge in pomembnosti človeških življenj in človeških izkušenj, na osnovi katerih so institucije oblikovane, saj »uresničena stvarnost sega

daleč onkraj organizacijske slike in vključuje življenja, ki jih ljudje uspejo – ali ne uspejo – živeti« (Sen, 2009, str. 18).

Rawls preveč zaupa v neomajno proceduralno pravičnost institucij, pri čemer pozablja ali zanemarja dejstvo, da so te, čeprav morda brezhibno zasnovane, v svojem delovanju v veliki meri odvisne od človeške aktivnosti, »od dejanskih vzorcev vedenja in vzajemnega političnega in družbenega vplivanja« (Sen, 2009, str. 354). Podobno misel razvija tudi Gray, vendar poudarja, da Rawlsov liberalizem s svojo osredotočenostjo na pravičnost institucij hkrati odpira nevarnosti šibke vladavine prava, ki brez moči države ne more delovati. Gray namreč trdi, da je »legalistični liberalizem, ki je prevladoval v zadnji generaciji, s spregledovanjem političnih pogojev, ki omogočajo vladavino prava, uspel predstaviti pravo kot samostojno institucijo. Uspelo mu je prezreti dejstvo, da institucija prava vedno temelji na moči države« (Gray, 2000, str. 131–132).

Dodajmo, da Rawls priznava, da ideal pravičnosti ni nikoli povsem dosegljiv, in sicer iz dveh razlogov. Prvi razlog je, »da obstaja neskončno mnogo ozirov, na katere se je mogoče sklicevati v izhodiščnem položaju, pri čemer v vsakem alternativnem pojmovanju pravičnosti nekateri oziri govorijo v njegov prid, drugi pa mu nasprotujejo« (Rawls, 2011, str. 174). Da bi to dilemo omilil, predlaga zaprt seznam mogočih razlogov. Drugi razlog, ki ga navaja, pa je, »da se končni izid tehtanja razlogov opira na presojo /…/ ki jo oblikuje in usmerja sklepanje« (Rawls, 2011, str. 174). Tudi tukaj Rawls predlaga podobno rešitev kot pri prvem razlogu – vzpostavitev zaprtega seznama razlogov (Rawls, 2011, str. 175). Rawls prav tako predpostavlja družbo kot zaprto, kar pa je v nasprotju ne le s Popperjevim razumevanjem odprte družbe, temveč tudi z Illichovim vidikom, ki poudarja odprtost in spontanost človeških odnosov.

Na eni strani imamo Senov očitek, da Rawls zapade v »institucionalni fundamentalizem« (Gaus, 2016, str. 21), na drugi strani pa Hayekovo stališče, ki dodatno spodkopava Rawlsovo zasnovo družbenih institucij in koncept proceduralne pravičnosti. Hayek namreč trdi, da je struktura sodobne kompleksne družbe nastala kot spontani red in je ni mogoče načrtovati:

> »Zato, ker ni bila odvisna od organizacije, temveč je zrasla kot spontani red, je struktura sodobne družbe dosegla stopnjo kompleksnosti, ki jo ima, in ki daleč presega tisto, ki bi jo bilo mogoče doseči s namernim organiziranjem. V resnici

> pravila, ki so omogočila rast tega kompleksnega reda, sprva niso bila zasnovana z namenom doseganja takšnega rezultata; toda tisti ljudje, ki so se slučajno držali ustreznih pravil, so razvili kompleksno civilizacijo, ki se je nato pogosto širila tudi na druge.« (Hayek, 1998, str. 50–51)

Trditev, da je družbo treba načrtovati, ker je postala preveč kompleksna, je torej napačna. Hayek meni, da zaradi tega niti članov družbe ni mogoče neposredno usmerjati, ampak je edino smiselno izboljšati pravila, ki spodbujajo nastanek spontanega reda. Popper (2013, str. 22) izraža stališče, podobno Hayekovemu, o spontanem in nenačrtovanem nastajanju družbenega reda in institucij: »Le manjšina družbenih institucij je zavestno oblikovanih, medtem ko je velika večina nastala z 'razvojem', kot nepredviden rezultat človeških dejanj.« Tako Hayek kot Popper poudarjata, da družbene institucije niso rezultat hitrega, načrtovanega delovanja, temveč dolgotrajnega in pogosto spontanega procesa prilagajanja. Podobno tudi Illich prepoznava postopnost in kompleksnost pri razvoju družbenih struktur, kar se kaže v njegovem razumevanju razvoja Cerkve v antiki, ki jo dojema kot zametek moderne države (Cayley, 2003).

Illich se v svojih delih sicer neposredno ne ukvarja z Rawlsovo teorijo pravične in dobro urejene družbe, vendar je iz njegove kritike institucij[3] razvidno, da zavrača ne samo idealne, ampak tudi univerzalne teorije. Predpostavljamo, da bi zavrnil tudi Rawlsovo teorijo pravičnosti, ki se osredotoča na institucionalno pravičnost, saj zanemarja konkretne medčloveške odnose in osebno odgovornost, ki jo Illich razume kot temelj pravičnosti. Illich osvetljuje moralo, ki je izrazito telesna, konkretna, osebna, lokalna, pristna in neposredno prisotna – moralo, ki je odgovor na klic in presega normativne kategorije ter institucionalne okvire, na katerih Rawls gradi svojo teorijo. V tej perspektivi se Illichova kritika lahko dopolni z Grayevo kritiko liberalizma, še posebej Rawlsovega sklicevanja na univerzalnost načel pravičnosti. Gray v svoji kritiki izpostavi, da je Rawlsovo razumevanje pravičnosti izrazito kantovsko, saj univerzalnost pojmuje kot neodvisno od tradicije ali zgodovine. Po Rawlsu naj bi bila načela pravičnosti zasnovana zgolj na racionalnih zmožnostih avtonomnega posameznika, kar jih postavlja onkraj kontingentnosti družbenih in političnih kontekstov. Vendar Gray opozarja, da je taka univerzalnost iluzorna, saj noben posameznik ne more povsem izstopiti iz družbeno-zgodovinskih okvirov, v katerih se nahaja. Zanj je že samo Rawlsovo sklicevanje na Kanta

[3] Gl. Illich 1972, 1973a, 1973b in 1976.

kulturološko pogojeno: »Vsako stališče, ki ga sprejmemo, izhaja iz določene oblike življenja in zgodovinskih praks, ki jo nadaljujejo; to je izraz človeške identitete, ki je zgodovinsko specifična, ne pa univerzalno in splošno človeška.« (Gray, 1995, str. 119)

Gray zato kritizira Rawlsov pristop kot povsem oddaljen od realnosti političnega in družbenega življenja. Rawlsova teorija, osredotočena na univerzalna načela, ne odraža konkretnih izzivov, s katerimi se soočajo posamezniki in skupnosti, ampak ostaja abstraktna in idealizirana. Gray to opiše kot paradoks kantovskega liberalizma, ki želi biti praktičen, a je hkrati brez politične relevantnosti:

> »Posebnost, pravzaprav absurdnost, tega novega kantovskega liberalizma – ki se je osvobodil tradicionalnih skrbi filozofije, da bi dosegel politični cilj praktičnega dogovora – je v tem, da se hkrati razvija na veliki razdalji od političnega življenja v realnem svetu. Teoretiki novega kantovskega liberalizma ne zastopajo nobenega političnega interesa ali volilne skupine, niti v liberalnih demokracijah, na katere so njihovi razmisleki usmerjeni /.../ Zato misli novih liberalcev ne odmevajo v nobeni od liberalnih demokracij.« (Gray,1995, str. 4)

Gray in Illich se torej srečata v kritiki univerzalizma, čeprav izhajata iz različnih kontekstov. Medtem ko Gray izpostavlja kulturno in zgodovinsko umeščenost posameznika, Illich poudarja lokalnost, konkretnost in moralno avtonomijo v odnosih. Oba pa opozarjata, da teorije, ki zanemarjajo človeško izkušnjo in kontingentnost, ne morejo biti temelj pravičnosti.

V osredju Illichove kritike institucij je prepričanje, da te omejujejo človekovo svobodo in s tem omalovažujejo človekovo dostojanstvo. Illich trdi, da nobeno omejevanje posameznikove svobode, pa tudi nobeni predpisi ali načela, ne morejo preprečiti zlorabe moči ali izkoriščanja situacij. Ker institucije, kot so šolstvo, zdravstvo, transport, ekonomija in druge, posamezniku odvzemajo avtonomijo ter ga hkrati delajo od njih odvisnega, ne le da vzdržujejo svojo lastno eksistenco, temveč ustvarjajo hierarhijo odnosov, ki pogosto koristi obstoječi družbeni eliti. Takšni mehanizmi preprečujejo kreativnost, ustvarjajo umetne potrebe in omejujejo svobodno delovanje posameznika. Zaradi tega smatra, da je potrebno delovanje institucij omejiti ter omogočiti spontano delovanje in organizacijo manjših skupnosti, ki omogočajo večjo svobodo, avtonomijo in samospoštovanje, v

nasprotju z institucijami, ki posameznika podvržejo normam, predpisom in hierarhijam. Tovrstno obliko bivanja oziroma sožitja Illich poimenuje *sožitnost*. Izraz sožitno, ki ga pripisuje predvsem orodjem, ne pa ljudem, razume kot »tehnični izraz za označevanje moderne družbe, v kateri so orodja odgovorno omejena« (Illich, 1973b). V tej skupnosti se ceni vrlina skromnosti, »ki ne izključuje vseh užitkov, temveč le tiste, ki odvračajo od ali uničujejo osebno povezanost« (Illich, 1973b). Te skupnosti, čeprav tradicionalne, niso zaprte, temveč omogočajo prosto sodelovanje in ustvarjanje med enakopravnimi posamezniki: »Takšno družbo, v kateri sodobne tehnologije služijo politično povezanim posameznikom in ne menedžerjem, bom imenoval 'sožitnostna'.« (Illich, 1973b) Ključnega pomena pri tem je, da te skupnosti prepoznavajo razdirajoče učinke orodij na posameznikovo svobodo ter se zavzemajo za zaščito te svobode, pri tem pa visoko cenijo osebne odnose in kakovost življenja, izraženo v obliki »radostne igrivosti«. »Družba sožitja bi bila rezultat družbenih ureditev, ki vsakemu članu zagotavljajo najširši in najbolj prost dostop do orodij skupnosti, omejujejo pa to svobodo le v korist enake svobode drugega člana« (Illich, 1973b).

Tako Illich kot Rawls pojmujeta pravično družbo kot takšno, v kateri so posameznikove svoboščine omejene zgolj z enakimi svoboščinami drugih. V tem kontekstu Rawls opredeli prvo načelo pravičnosti: »Vsaka oseba ima isto nedotakljivo pravico do popolnoma ustreznega sistema enakih temeljnih svoboščin, ki je združljiv z enakim sistemom svoboščin za vse.« (Rawls, 2011, str. 67) Illich pa trdi: »Pravična družba bi bila taka, v kateri je svoboda ene osebe omejena le z zahtevami, ki jih ustvarja enaka svoboda za drugo osebo.« (Illich, 1973b) Vendar se že naslednjem koraku odkriva temeljna razlika v njunem pristopu. Rawls nikoli ne postavi pod vprašaj legitimnosti institucij in predpostavlja njihovo politično, ekonomsko in ideološko nevtralno delovanje. Zaradi tega lahko v revidirani formulaciji načela razlike zapiše, da so družbene in ekonomske razlike dopustne le, če imajo vsi državljani enake pogoje in možnosti do služb in položajev in doda, da morajo neenakosti koristiti tistim članom družbe, ki so v najslabšem položaju (Rawls, 2011, str. 67). Za razliko od Rawlsa pa je Illich prepričan, da je predpogoj za pravično družbeno ureditev dogovor o izključitvi vseh »orodij«, ki preprečujejo posameznikovo avtonomno delovanje (Illich, 1973b). Med »orodja« Illich ne uvršča le predmetov, »ki se uporabljajo pri preprostejšem delu« (orodje, 2018), temveč pod njim pojmuje vse, kar človeku omogoča oblikovanje okolja in interakcijo z drugimi, s specifičnim namenom spreminjanja sveta. Tako »orodja« niso omejena le na fizične predmete in stroje, ampak vključujejo tudi sisteme, tehnologije in institucije. Če se

ta orodja lahko uporabljajo za svobodne in ustvarjalne namene posameznika, če za njihovo uporabo ni potreben certifikat, in če njihova uporaba ni omejena z namenom in pričakovanji njihovega oblikovalca, potem to vrsto orodij, kot je na primer telefon, uvršča med orodja za sožitnost. Industrijska orodja pa so tista, ki niso splošno dostopna in imajo specifično opredeljeno uporabnost, ki omejujejo kreativnost uporabnika (Illich, 1973b). Takšna klasifikacija pa je vprašljiva, saj telefon v svoji digitalni in pametni obliki omogoča širok spekter uporabe, a kljub temu potrebuje infrastrukturo in v veliki meri služi namenu njegovega izumitelja. Poleg tega je dostop do njega in njegova uporabnost pogosto omejena s specifičnimi družbenimi, ekonomskimi in tehnološkimi pogoji.

Illich zavrača večino orodij oziroma sistemov in sodobnih tehnologij, ki ogrožajo posameznikovo avtonomijo, zavirajo kritično razmišljanje, povečujejo družbene neenakosti in odtujenost ali grozijo z ekološko katastrofo:

> »Nekatera orodja so uničujoča, ne glede na to, kdo jih ima v lasti – bodisi mafija, delničarji, tuje podjetje, država ali celo delavska zadruga. Omrežja večpasovnih avtocest, dolge razdalje širokopasovnih oddajnikov, površinski rudniki ali sistemi obveznega šolstva so takšna orodja.« (Illich, 1973b)

Težko je sprejeti tako radikalno stališče, saj zanemarja pozitivne vidike, ki jih sodobna tehnologija in družbeni razvoj lahko prinesejo. Orodja, če so ustrezno regulirana in reformirana, imajo potencial za izboljšanje kakovosti življenja, zmanjšanje družbenih neenakosti in trajnostno uporabo naravnih virov. Illichovo stališče s svojo absolutnostjo zapira prostor za razpravo o možnih alternativnih pristopih, ki bi združevali tehnološke dosežke z etičnimi in ekološkimi načeli.

Illich se zaveda, da je njegova vizija družbe sožitja težko dosegljiva, vendar jo ne predstavi kot preprosto izbiro ali predlog, temelječ na pogodbenih dogovorih ali drugih podobnih teorijah. Namesto tega družbo sožitja ponuja kot odprto alternativo, ki ni zavezujoča in ne temelji na prepričanjih, ki bi jo izenačevala z obljubami o pravičnosti ali družbeni enakosti. Illich meni, da preprosto prikazovanje te družbene oblike kot bolj privlačne ali pravične ne zadostuje, da bi jo postavili kot resnično alternativno pot. Šele s spoznanjem, da je takšna družba edina možnost za preživetje človeštva, se odpira resnična priložnost za njen nastanek. Kot opozarja: »Potrebujemo način, kako prepoznati, da je za preživetje vseh ljudi nujno obrniti sedanje politične cilje.« (Illich, 1973b)

Rawls uvršča temeljne človekove pravice in svoboščine med primarne dobrine (Rawls, 1999, str. 54). Predpostavljamo, da bi Illich bil prav tako kritičen do temeljnih človekovih pravic, ki naj bi v demokratični pogodbeni družbi nadomestile moralo. Te so lahko razumljene kot instrumenti institucionalizacije posameznika in omejevanja njegove svobode. Kritika človekovih pravic je usmerjena predvsem v to, da te pravice posameznika uokvirjajo v njegovem delovanju in ne omogočajo prostora za svobodno kreativnost na področju moralnega delovanja. Illich je prepričan, da če je človekovo moralno delovanje do potankosti določeno, ni prostora za presenečenje ali dar sočutja, saj so obveznosti in dolžnosti jasno opredeljene. V takšni družbi bi bilo sočutje zgolj izraz dolžnosti in ne globokega sočutja ter empatije do drugega, kar Illich vidi kot oslabljenje ne le posameznika, ampak celotne družbe. Čeprav posameznik ohranja svobodo odločanja, se človekove pravice pogosto uporabljajo kot poenostavljena rešitev za reševanje dilem v medosebnih odnosih, kjer se sklicevanje nanje pogosto smatra za pravilno, četudi to ne pomeni, da so odločitve vedno v skladu z moralnimi vrednotami, ki so del celovitih etičnih naukov. Na ta način Illich dojema moralo podobno kot Gauthier, ki jo razume kot moralo po dogovoru, kateri bi Illich očital, da ji manjka razsežnost obžalovanja in odpuščanja (Cayley, 2021).

Illichovo pojmovanje krščanske etike, ki se v mnogih pogledih opira na njegovo eksegezo Prispodobe o usmiljenem Samarijanu,[4] se zavzema za osebno etiko, ki je globoko telesna in konkretna. Telesnost te etike jo sicer lokalno omejuje, vendar jo hkrati naredi pristno in neposredno. Gre za etiko, ki ni osredotočena na izpolnjevanje univerzalnih moralnih dolžnosti, temveč je predvsem odziv na klic, ki se izraža v sočutju, ki presega okvire družbenih norm (Illich, 1972; Cayley, 2003). Illich ne ponuja idealne ali univerzalne vizije družbe, temveč se osredotoča na sožitnost – življenjski način, ki je odprt za spontano, konkretno in nepričakovano medsebojno delovanje. To je etika, ki ne verjame v to, da bi lahko ustaljene institucije

[4] Tedaj je vstal neki učitelj postave, in da bi ga preizkušal, mu je rekel: »Učitelj, kaj naj storim, da dosežem večno življenje?« On pa mu je dejal: »Kaj je pisano v postavi? Kako bereš?« Ta je odgovoril: »*Ljubi Gospoda, svojega Boga, iz vsega srca, z vso dušo, z vso močjo* in z vsem mišljenjem, in *svojega bližnjega kakor samega sebe*.« »Prav si odgovoril,« mu je rekel, »to delaj in boš živel.« Ta pa je hotel sebe opravičiti in je rekel Jezusu: »In kdo je moj bližnji?« Jezus je odgovoril: »Neki človek je šel iz Jeruzalema v Jeriho in je padel med razbojnike. Ti so ga slekli, pretepli, pustili napol mrtvega in odšli. Primerilo pa se je, da se je vračal po tisti poti domov neki duhovnik; videl ga je in šel po drugi strani mimo. Podobno je tudi levit, ki je prišel na tisti kraj in ga videl, šel po drugi strani mimo. Do njega pa je prišel tudi neki Samarijan, ki je bil na potovanju. Ko ga je zagledal, se mu je zasmilil. Stopil je k njemu, zlil olja in vina na njegove rane in jih obvezal. Posadil ga je na svoje živinče, ga peljal v gostišče in poskrbel zanj. Naslednji dan je vzel dva denarija, ju dal gostilničarju in rekel: ›Poskrbi zanj, in kar boš več porabil, ti bom nazaj grede povrnil.‹ Kaj se ti zdi, kateri od teh treh je bil bližnji tistemu, ki je padel med razbojnike?« Oni je dejal: »Tisti, ki mu je izkazal usmiljenje.« In Jezus mu je rekel: »Pojdi in ti delaj prav tako!« (Lk 10, 25–37).

ali predvidljive procedure zagotovile pravičen izid, saj se zla ne more vedno preprečiti z mehanizmi kontrole in nadzora. Nasprotno pa Rawlsova teorija temelji na idealizirani in univerzalni podobi dobro urejene družbe, kjer so institucionalni okviri, procedure in pravila natančno določeni, zaradi česar je rezultat – pravičnost – predvidljiv in pričakovan. Illichova etika, v nasprotju z Rawlsovo, odraža prepoznavanje človeške ranljivosti in odprtosti do nepredvidljivosti, kjer ni mogoče zagotoviti popolne varnosti ali predvidljivosti, temveč se svoboda in odgovornost izražata v konkretnih, nezavednih in včasih kaotičnih medsebojnih dejanjih.

Illichova vizija družbe sožitja se sooča z vprašanjem, kako se izogniti grožnji anarhije. Vendar Illich anarhije ne razume kot družbeno ureditev brez pravil, temveč prevzema stališče Paula Goodmana, ki trdi, da anarhija ni zavračanje prava, ampak zavračanje moči. Kot pravi Cayley, imajo anarhisti »edinstveno spoštovanje dostojanstva prava, saj le oni verjamejo, da je pravo naravni del človeških družb in da zato ni odvisno od prisile« (The Golden Thread, 2021). Illich tako ne zavrača samega prava, temveč se osredotoča na odstranitev centralizirane moči, ki omejuje posameznikovo avtonomijo in svobodo. Njegova ideja družbe sožitja temelji na prepričanju, da bi družba morala omogočiti prosto sodelovanje in ustvarjalnost brez prekomerne hierarhije in kontrole.

Kljub temu pa Illich v svojem pristopu naleti na težavo, saj njegovo reševanje temelji na obnavljanju tradicionalnih oblik družbenih struktur, ki jih vidi kot alternativne moderne institucionalizaciji. Toda pri tem zanemarja pomemben vidik: v tradicionalnih družbah skupnosti pogosto prevzamejo vlogo, ki jo v industrijskih družbah opravljajo institucije. Te skupnosti določajo posameznikovo vlogo in postavljajo pričakovanja, ki lahko prav tako omejijo posameznikovo avtonomijo. Tako se izkaže, kljub Illichovemu upanju, da bi tradicionalne oblike družbene organizacije pripomogle k večji svobodi posameznika, da te lahko v resnici ohranjajo podobne oblike nadzora in omejevanja kot sodobne institucije. Illichova kritika institucionalizacije se torej zdi krožna, saj namesto iskanja pravega izhoda iz omejevanja avtonomije sledi logiki omejevanja institucij, vendar se s tem ponovno sooča z omejitvami, ki jih prinašajo tradicionalne oblike družbenega nadzora.

K tej kritiki lahko dodamo še Rawlsovo teorijo, ki ponuja bolj pragmatičen pristop k zagotavljanju pravičnosti v družbi. Rawls namreč ne zanemarja potrebne vladavine in uveljavitve zakona, saj meni, da je v dobro urejeni družbi potrebna tudi prisilna suverenost za zagotavljanje pravičnosti, čeprav sankcije morda nikoli ne bodo

potrebne: »Z izvajanjem javnega sistema kazni vlada odstrani osnovo za prepričanje, da drugi ne spoštujejo pravil. Iz tega razloga je verjetno vedno potrebna prisilna suverenost, četudi v dobro urejeni družbi sankcije niso stroge in jih morda nikoli ne bo treba uvesti.« (Rawls, 1999, str. 38) Rawls razume hierarhične in pravne strukture kot nujne za dosego stabilnosti in pravičnosti v družbi, saj omogočajo zaščito temeljnih pravic in zagotavljajo pravico do enakosti za vse.

## Zaključek

Na podlagi analize Rawlsa in Illicha je mogoče sklepati, da nobena od njunih vizij ni zadostna za odgovore na vprašanja sodobne družbe. Ključna naloga ostaja v vzpostavljanju kritičnega odnosa do institucij, predvsem v smislu njihove legitimnosti in nepristranskosti. Treba je raziskati njihove politične, gospodarske in ideološke vloge, pri tem pa se osredotočiti na vprašanje, kako te dejansko vplivajo na neenakost in svobodo. Namesto zanašanja na univerzalne teorije se zdi smiselno stremeti k reformam, ki bodo naslovile konkretne težave institucionalne ureditve, zmanjšale koncentracijo moči, povečale odgovornost ter omogočile pogoje za večjo avtonomijo posameznika v okviru družbenega delovanja.

**Literatura**


Cayley, D. (1992). *Ivan Illich in Conversation*. Anansi.

Cayley, D. (2003). *Ivan Illich in Memoriam*. https://www.davidcayley.com/transcripts

Cayley, D. (2005). *The Rivers North of the Future: The Testament of Ivan Illich*. Anansi.

Cayley, D. (2021). *Ivan Illich: An Intellectual Journey*. The Pennsylvania State University Press.

Gaus, G. (2016). *The Tyranny of the Ideal: Justice in a Diverse Society*. Princeton University Press.

Gray, J. (1995). *Enlightenment's Wake: Politics and Culture at the Close of the Modern Age*. Routledge.

Gray, J. (2000). *Two Faces of Liberalism*. Polity Press.

Hayek, F. (1998). *Law, Legislation and Liberty*. Routledge.

Illich, I. (1972). *Deschooling Society*. Harper & Row.

Illich, I. (1973a). *Celebration of Awareness: A Call for Institutional Revolution*. Penguin Books.

Illich, I. (1973b). *Tools for Conviviality*. Marion Boyars.

Orodje. (2018). V *Slovar slovenskega knjižnega jezika 2018*. ZRC SAZU. https://doi.org/10.3986/9789610501640

Popper, K. (2013). *The Open Society and Its Enemies (New One-Volume Edition)*. Princeton University Press.

Rawls, J. (1999). *The Theory of Justice (Revised Edition)*. The Belknap Press of Harvard University Press.

Rawls, J. (2011). *Pravičnost kot poštenost*. Krtina.

Sen, A. (2009). *The Idea of Justice*. The Belknap Press of Harvard University Press.

Sveto pismo. (b. d.). *Slovenski standardni prevod*. https://www.biblija.net

The Golden Thread (2021, 1. junij). *Ivan Illich, Part Moon, Part Travelling Salesman: Conversation with David Cayley (1989)* [video]. YouTube. https://www.youtube.com/watch?v=VXghb5xOD8s&list=PLEZ9EYTtCpN0B46hzH2inM9VrYzaD7UvE&index=1

# Ontologija fikcije

**ANALIZA**
Let. 28, št. 2, str. 169–186, december 2024

# KAJ RESNIČNEGA LAHKO POVEMO O FIKCIJSKIH LIKIH? PROBLEM REFERIRANJA FIKTIVNIH IMEN V FIKCIJSKEM DISKURZU

**Sprejeto**
9. 9. 2024

**Pregledano**
7. 12. 2024

**Izdano**
31. 12. 2024

MAJA NEMEC
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
maja.nemec@student.um.si

DOPISNI AVTOR
maja.nemec@student.um.si

**Izvleček** Članek obravnava, kako teorija pretvarjanja in artefaktična teorija fikcije delujeta v internem in eksternem govoru o fikciji in zakaj oba, fikcijski antirealizem in realizem, naletita na težave. Predstavljena je rešitev Amie Thomasson (2003b), ki ohranja obstoj fikcijskih likov kot abstraktnih artefaktov, a v intrafikcijski diskurz vseeno sprejme pretvarjanje, da razreši nejasnosti, ki nastanejo, ko uporabljamo fiktivna imena v različnih kontekstih. Izpostavljena je problematičnost *de re* pretvarjanja, ki zahteva pretvarjanje, da ima abstraktni objekt konkretne lastnosti. Če *de re* pretvarjanje zavrnemo in v govor o fikciji sprejmemo *de dicto* pretvarjanje, izgubimo uniformnost teorije in trdimo, da fiktivna imena v delu diskurza referirajo, v delu pa ne. Cilj članka je nakazati, da vpeljava pretvarjanja v realistično teorijo fikcije, kljub temu da zahteva dvojno rabo fiktivnih imen, bolje razloži naše razumevanje govora o fikciji kot antirealistične teorije.

**Ključne besede**
fikcija,
fikcijski diskurz,
fiktivna imena,
abstraktni artefakti,
pretvarjanje

https://doi.org/10.18690/analiza.28.2.169-186.2024


Univerzitetna založba Univerze v Mariboru

AN$_\alpha$LIZA

**ANALIZA**
Vol. 28, no. 2, pp. 169–186, December 2024

# WHAT CAN WE REALLY SAY ABOUT FICTIONAL CHARACTERS? THE PROBLEM OF FICTIONAL NAMES REFERENCING IN FICTIONAL DISCOURSE

MAJA NEMEC
University of Maribor, Faculty of Arts, Maribor, Slovenia
maja.nemec@student.um.si

CORRESPONDING AUTHOR
maja.nemec@student.um.si

**Accepted**
9. 9. 2024

**Revised**
7. 12. 2024

**Published**
31. 12. 2024

**Abstract** The article discusses how the theory of pretense and the artefactual theory of fiction work in internal and external talk about fiction, and why both fictional anti-realism and realism run into difficulties. It presents Amie Thomasson's (2003b) solution, which preserves the existence of fictional characters as abstract artefacts, but still adopts pretense into the intrafictional discourse to resolve the ambiguities that arise when using fictional names in different contexts. The problematic nature of *de re* pretense, which requires pretending that an abstract object has concrete properties, is highlighted. If we reject *de re* pretense and accept *de dicto* pretense in the fictional discourse, we lose the uniformity of the theory and claim that fictional names refer in one part of the discourse but not in the other. The aim of this paper is to suggest that, although it requires a dual use of fictive names, the introduction of pretense into a realist theory of fiction does a better job of explaining our understanding of talk about fiction than anti-realist theories.

**Keywords**
fiction,
fictional discourse,
fictional names,
abstract artefacts,
pretense

https://doi.org/10.18690/analiza.28.2.169-186.2024


University of Maribor Press

## 0 Uvod

Debata, ki prevladuje v filozofiji fikcije, izhaja iz vprašanja, na kaj se nanašamo, ko govorimo o fikcijskih likih. Oblikovali sta se dve nasprotujoči si poziciji: fikcijski antirealizem, ki zavrača obstoj fikcijskih likov, in fikcijski realizem, ki na tak ali drugačen način priznava obstoj tovrstnih bitnosti. V sklopu fikcijskega realizma prevladuje teorija abstraktnega artefakta (Thomasson, 1999), ki fikcijske like določi kot objekte, ustvarjene od avtorja (avtorjev), tj. odvisne abstraktne artefakte. Ker teorija priznava obstoj fikcijskih likov, priznava tudi resničnost stavkov, ki vsebujejo fiktivna imena. Temu nasprotno, antirealististično stališče predstavlja teorija pretvarjanja (Walton, 1990), ki zagovarja, da gre pri izrekanju propozicij, ki vsebujejo fiktivna imena, za neke vrste igro pretvarjanja oz. hlinjenja (*make-believe*).

Pod izrazom fikcijski lik razumemo vse literarne junake, kraje, bitja in predmete, ki se prvič pojavijo v nekem literarnem delu in jim (vsaj) izven področja literature oz. umetnosti ne moremo pripisati obstoja (Salis, 2014, str. 1). Vsak fikcijski lik ima fiktivno ime. Ta imena uporabljamo v fikcijskem diskurzu, ko se pogovarjamo o vsebini filmov in knjig, o najljubših pravljičnih junakih, analizi del, primerjavi različnih likov ipd. Amie Thomasson (2003b) je fikcijski diskurz uporabila kot izhodišče za reševanje neskladnosti, ki se pojavijo v vsakdanjem govoru o fikciji, če sprejmemo, da so fikcijski liki abstraktni artefakti. Da te pojasni, loči med internim in eksternim govorom o fikciji. Najustreznejša teorija fikcije bo tista, ki jo je mogoče razširiti na vse vrste diskurza z najmanj spremembami. To je za Thomasson teorija, ki hkrati ohrani obstoj fikcijskih likov in v del diskurza vpelje pretvarjanje. Zdi se, da upravičeno opozarja na neustreznost antirealističnih teorij fikcije za razlago metafikcijskega diskurza. Skušala bom braniti idejo o sprejetju pretvarjanja v interni fikcijski diskurz, čeprav nas to privede do nekaterih precej nenavadnih zaključkov glede na vrsto pretvarjanja, ki ga sprejmemo. Če trdimo, da fiktivna imena tudi v internem dikcijskem diskurzu referirajo, sprejmemo, da je vse naše pretvarjanje, ki ga predpisuje zgodba, o neki abstrakciji. Razložiti bo treba, ali si znamo predstavljati, kako bi abstrakcija (fikcijski lik) uprimerjala lastnosti kot konkretne osebe. Če nasprotno zagovarjamo, da fiktivna imena v fikcijskem diskurzu ne referirajo, pridemo do sklepa, da isto fiktivno ime v različnih kontekstih deluje različno. Enkrat ima referenta, tj. abstraktni objekt, ki nastane z rabo tega imena, spet drugič, v govoru o vsebini zgodbe, pa to isto ime deluje kot prazno ime. V tem primeru mora

fikcijski realist pojasniti, zakaj pride do dvojne rabe in ali je sploh smiselno vpeljevati nove entitete v ontologijo, če te ne omogočijo uniformne teorije.

## 1 Teorije fikcije in vrste fikcijskega diskurza

### 1.1 O čem sploh govorimo, ko govorimo o fikciji?

Kaj, če sploh kaj, so referenti fiktivnih imen, je izhodiščno vprašanje filozofov fikcije. Glede na odgovor sta se v grobem oblikovali dve poziciji. Fikcijski antirealizem, ki trdi, da fiktivna imena ne referirajo, in fikcijski realizem, ki trdi nasprotno, torej da fikcijska imena referirajo. Med antirealističnimi teorijami je morda najodmevnejša teorija pretvarjanja (Walton, 1990, 2010), ki zagovarja, da gre pri izrekanju propozicij, ki vsebujejo fiktivna imena, za neke vrste igro pretvarjanja oz. hlinjenja (*make-believe*). Pravila igre pretvarjanja določa vsebina določenega literarnega dela. Teorija ne priznava obstoja fikcijskih likov in trdi, da imajo stavki, ki vključujejo fiktivna imena, *fikcijske resničnostne pogoje* in jim je možno določiti le *fikcijsko resničnostno vrednost*, njihova dejanska resničnostna vrednost pa je neresnična.

Nasprotno pa fikcijski realisti (Kripke, van Inwangen, Thomasson idr.) trdijo, da fikcijski liki obstajajo. Amie Thomasson (1999, 2003a, 2003b) fikcijske like definira kot abstraktne artefakte, tj. abstraktne objekte, ustvarjene od avtorja (oz. avtorjev). Ker teorija priznava obstoj fikcijskih likov, priznava tudi resničnost stavkov, ki vsebujejo fiktivna imena. Pri tem izpostavi tri lastnosti tovrstnih bitnosti, in sicer da so fikcijski liki (Thomasson, 1999, str. 6–14): ustvarjeni, odvisni in minljivi. Ustvari jih avtor, odvisni so od avtorja in literarnega dela, v katerem obstajajo, ter lahko prenehajo obstajati, torej niso platonistične abstrakcije. Pri tem se opira na Searlovo (1996) pojmovanje institucionalnih in kulturnih bitnosti, kot sta, denimo, denar ali poroka. Fikcijski liki so po tej teoriji kulturne bitnosti, odvisne od človekove intencionalnosti.

Ne glede na to, če fikcijski liki obstajajo ali ne, se zavedamo, da stavke, ki vsebujejo fiktivna imena, izrekamo v določenih kontekstih. Pišemo zgodbe, s prijatelji se pogovarjamo o usodah naših najljubših literarnih junakov, v analizi romana določamo glavne in stranske like, otrokom pojasnjujemo, da pošasti iz grozljivk ne obstajajo. Vse to delamo v kontekstu fikcijskega diskurza. Vanj so uvrščeni vsi stavki, ki v takšnem ali drugačnem smislu vključujejo imena, ki se pojavijo le v fikciji.

## 1.2 Vrste fikcijskega diskurza

Raba fiktivnih imen tudi v vsakdanjem govoru ni vedno enaka. Če nas zanima, kaj se zgodi z Lepo Vido, govorimo o vsebini fikcije. Ko se sprašujemo, kdaj se je v slovenski književnosti prvič pojavil lik Lepe Vide, pa nas ne zanima več vsebina dela, ampak se na fiktivno ime nanašamo eksterno. Thomasson (2003b, str. 206–207) ločuje med štirimi vrstami fikcijskega diskurza:

1) diskurz znotraj fikcijskih del (stavki, ki jih zapiše avtor v zgodbi);
2) interni diskurz bralcev o vsebini fikcijskih del;
3) eksterni diskurz bralcev in kritikov o likih kot fikcijskih likih, okoliščinah njihovega nastanka, zgodovinska povezanost z drugimi literarnimi figurami itd. in
4) stavki z negativnimi eksistenciali.

V diskurz 1) uvrščamo stavke, ki so jih avtorji zapisali v literarnih delih. Na primer, v *Študiji v škrlatnem* je Arthur Conan Doyle zapisal: »Sherlock Holmes je vstal in si prižgal pipo.« Stavki diskurza 2) vsebujejo implicitni operator »glede na zgodbo«. Gre za diskurz o sami vsebini fikcijskih del. Primer tega je pogovor o vsebini literarnega dela med poukom književnosti. Ko profesor učence vpraša: »Kdo je najboljši prijatelj Sherlocka Holmesa?« in učenci odgovorijo: »Holmesov prijatelj je Watson,« se pogovarjajo o tem, kaj je res v določeni zgodbi (vsebini literarnega dela). V diskurz 3) se uvrščajo stavki, ki niso vezani na samo vsebino literarnega dela, pač pa na fikcijske like kot take. Sem sodijo stavki tipa »Glavni lik v *Študiji v škrlatnem* je Sherlock Holmes« ali »Tragikomedija *Čakajoč Godota* ima pet likov«. V diskurz 4) pa so uvrščeni stavki, ki poleg fiktivnega imena vključujejo negativni eksistencial, npr. »Sherlock Holmes ne obstaja«.

Diskurz 1) in 2) sta zelo podobna. V obeh primerih govorimo o intrafikcijskem diskurzu, saj se stavki obeh vrst naslanjajo na vsebino fikcijskega dela. Razlika je le v tem, da je v 1. vrsti tvorec diskurza avtor fikcije, ki delo z izrekanjem tudi ustvarja. V 2. vrsti diskurza ne gre več za ustvarjanje dela, ampak za govor o že nastali vsebini. Stavki 3. in 4. diskurza so metafikcijski stavki, ker se ne sklicujejo na vsebino fikcije.[1]

[1] V prispevku se osredotočam predvsem na razlago internega in eksternega govora o fikciji. Vprašanja negativnih eksistencialov ne obravnavam.

## 1.3 Nejasnosti v govoru o fikciji

Če podrobneje razmislimo o stavkih, ki jih izrekamo v kontekstu fikcijskega diskurza, hitro opazimo, da pogosto pritrjujemo med seboj izključujočim si trditvam. Prijatelju razlagamo, da je Sherlock Holmes briljanten detektiv, ki rad kadi, torej konkretna oz. dejanska oseba. Hkrati v eseju o analizi del Arthurja Conana Doyla pišemo, da je Sherlock Holmes glavni lik njegovih zgodb, torej ga obravnavamo kot abstraktni objekt, vsaj po mnenju fikcijskih realistov. V obeh primerih se nam zdi, da govorimo resnico, čeprav sta trditvi protislovni, kajti ni mogoče, da bi neka stvar bila dejanska oseba in abstrakcija hkrati. Točno na to opozarja Thomasson (2003b), ki izpostavlja 3 vrste nejasnosti v govoru o fikciji. Če se želimo izogniti sklepu, da ljudje zavedno pritrjujemo protislovjem, moramo tovrstne nejasnosti razčistiti.

Kako torej določiti, kaj, če sploh kaj, kar povemo o fikciji, je resnično? Zdi se, da teorija pretvarjanja ponudi najenostavnejšo rešitev – vsaka trditev, ki vključuje fiktivno ime, ima neresnično resničnostno vrednost, lahko pa ji določimo fikcijsko resničnostno vrednost glede na pravila igre pretvarjanja, ki jih določa fikcija, v kateri se ime pojavi (Walton 1999, 2010). Stavek »Sherlock Holmes je briljanten detektiv, ki rad kadi«, je po teoriji pretvarjanja dobesedno neresničen (ker vključuje fiktivno ime, to pa nima referenta), je pa fikcijsko resničen, ker je v skladu s tem, kar piše v zgodbah o Sherlocku Holmesu. Te zgodbe narekujejo pravila našega pretvarjanja in kar koli je v skladu z njimi, je tudi fikcijsko resnično. Ker je glede na zgodbe, ki jih je o Holmesu napisal Arthur Conan Doyle, res, da je Holmes detektiv in da rad kadi, ima resnično fikcijsko resničnostno vrednost. Nasprotno so stavki, ki trdijo nekaj, kar je v nasprotju z vsebino zgodbe, fikcijsko neresnični. Zato je izjava »Sherlock Holmes je klovn« fikcijsko neresnična. Obe trditvi pa sta dobesedno neresnični, ker govorita o nečem, kar po mnenju fikcijskih antirealistov ne obstaja. Razlaga je vabljiva, ker ne zahteva vpeljave novih bitnosti v našo ontologijo in hkrati ne pušča prostora za dvom pri določanju resničnostne vrednosti stavkom s fiktivnimi imeni. V nadaljevanju bom poskušala pokazati, zakaj fikcijski realisti pravilno opozarjajo, da teorija pretvarjanja vseeno naleti na težave z metafikcijskim diskurzom.

Če, nasprotno od antirealistov, sprejmemo, da fikcijski liki obstajajo in fiktivna imena imajo referente, moramo pojasniti, kako določimo, kateri od takih stavkov so resnični in kateri ne. Thomasson (2003b, str. 205) s primeri pokaže, kje v govoru o

fikciji naletimo na težave. Kar želi fikcijski realist doseči, lahko prikažemo v tabeli z intuitivnimi pripisi resničnostnih vrednosti:

**Tabela 1: Avtorstvo fikcijskih likov**

| Trditev | Resničnostna vrednost | Resničnostna vrednost glede na zgodbo |
|---|---|---|
| 1 »Frankensteinovo pošast je ustvarila Mary Shelley«. | R | N |
| 2 »Frankensteinovo pošast je ustvaril dr. Frankenstein«. | N | R |

**Tabela 2: Identiteta fikcijskih likov**

| Trditev | Resničnostna vrednost | Resničnostna vrednost glede na zgodbo |
|---|---|---|
| 1 »Sherlock Holmes je detektiv«. | N | R |
| 2 »Sherlock Holmes je fikcijski lik«. | R | N |
| 3 »Sherlock Holmes je klovn«. | N | N |

Artefaktična teorija fikcije nima težav s stavki metafikcijskega diskurza (»Frankensteinovo pošast je ustvarila Mary Shelley« in »Sherlock Holmes je fikcijski lik«). Problem se pojavi pri intrafikcijskih stavkih (»Frankensteinovo pošast je ustvaril dr. Frankenstein« in »Sherlock Holmes je detektiv«). Vsi ti stavki so v skladu z artefaktično teorijo neresnični, ker fikcijske like ustvarjajo avtorji (ne drugi liki) in ker niso konkretne bitnosti. Vseeno imamo občutek, da so tudi ti stavki resnični, saj je v romanu Frankensteinovo pošast res ustvaril dr. Frankenstein in Sherlock Holmes je glede na vsebino zgodbe res detektiv. Hkrati se zdi, da stavek »Sherlock Holmes je klovn« ni neresničen, ker je Sherlock Holmes fikcijski lik in ne človek, temveč zato, ker je v zgodbah res, da je detektiv in ne klovn.

Problem za teorijo abstraktnega artefakta torej pomeni ločevanje med internim in eksternim fikcijskim diskurzom (prim. Friend, 2007, str. 151–152). Trditev »Lepa Vida je fikcijski lik« (eksterni govor) je resnična, »Lepa Vida utone« (interni govor) pa je resnična samo glede na vsebino zgodbe, torej je dobesedno neresnična, ker abstraktni predmet ne more utoniti. Denimo, da po branju pesnitve izjavim: »Žal mi je za Lepo Vido.« To se zdi resnično, ampak te trditve ne moremo razumeti kot dobesedno resnične, ker je glede na eksterni diskurz Lepa Vida abstraktni predmet, ki ne more trpeti. Tovrstne trditve imajo lahko le fiktivno resničnostno vrednost

(resnične so le glede na vsebino zgodbe). Dobesedno resničnostno vrednost lahko torej pripisujemo le propozicijam eksternega fikcijskega diskurza.

Po teoriji pretvarjanja intrafikcijskim trditvam lahko določimo fikcijsko resničnostno vrednost glede na njihovo skladnost s pravili igre pretvarjanja, tj. z vsebino zgodbe, v kateri se pojavijo (Walton, 1990, 2010). Vse takšne trditve so sicer dobesedno neresnične, vendar so resnične v sklopu igre pretvarjanja, ki jo predpisuje neka določena fikcija. Delo omogoča, da so te trditve le fikcijsko resnične, za to pa ni potrebe po vpeljavi posebnih entitet v našo ontologijo. Fikcijski diskurz prevedemo v pogovor o rekvizitih igre pretvarjanja. To, kar si moramo predstavljati, je odvisno od rekvizita (zgodbe) in od pravil igre (»načela generiranja«) (Giovanelli, 2009, str. 589). Zato lahko trdimo, da je glede na zgodbo res, da je Frankensteina ustvaril dr. Frankenstein ter da je Sherlock Homes detektiv in ne klovn. Zakaj ne bi zavrgli artefaktične teorije in pretvarjanja razširili na vse vrste fikcijskega diskurza?

### 1.4 Problemi teorije pretvarjanja

Cilj vsakega filozofa fikcije, naj bo ta realist ali antirealist, je zasnovati teorijo, ki jo lahko apliciramo na celoten govor o fikciji. Ker teorija pretvarjanja dobro razloži intrafikcijski diskurz, je smisleno, da jo poskušamo prenesti tudi na metafikcijske trditve (»Frankensteinovo pošast je ustvarila Mary Shelley« in »Sherlock Holmes je fikcijski lik«).

Thomasson (2003b, str. 208) s primerom izpostavi, zakaj je teorija pretvarjanja neustrezna za eksterni diskurz o fikcijskih likih. V primeru prikaže pogovor med dvema policistoma, ki rešujeta nek primer. Prvi policist izjavi: »Ta primer je zelo zapleten. Morala bi poklicati Sherlocka Holmesa, da nama pomaga.« Drugi policist odvrne: »Oseba kot je Sherlock Holmes ne obstaja, Holmes je samo fikcijski lik.« Medtem ko lahko sprejmemo, da se v prvi izjavi policist pretvarja, da je Holmes realna oseba, ki jo je mogoče poklicati na pomoč, se za drugo izjavo ne zdi, da se je policist kakor koli pretvarjal. Ravno nasprotno, namen drugega policista je bil »izstopiti« iz igre pretvarjanja in izjaviti dobesedno (in ne fikcijsko) resnico o Holmesu.

Tu je jasno razviden problem teorije pretvarjanja. Zdi se, da pri eksternem govoru o fikciji nimamo nobene namere za pretvarjanje. Policist, ki razlaga, da je Holmes fikcijski lik, opozarja na razliko med fikcijsko in dejansko resničnostjo (Thomasson, 2003b, str. 208). Noben literarni teoretik, ki raziskuje okoliščine nastanka nekega fikcijskega lika, verjetno ne bi trdil, da piše o nečem, za kar se zgolj pretvarja, da obstaja.

Ne samo, da za pretvarjanje v eksternem diskurzu nimamo intence in teorija zahteva dodajanje *ad hoc* iger pretvarjanja, če pretvarjanje sprejmemo v vseh vrstah fikcijskega diskurza, se zdi, da dobimo tudi nova protislovja. Da razložimo stavke, kot je »Sherlock Holmes je fikcijski lik«, bi se morali najprej pretvarjati, da fikcijski liki obstajajo (so neke vrste bitnosti) in da Sherlock Holmes je taka bitnost. Ko pa se pogovarjamo o vsebini zgodbe Arthurja Conana Doyla, se pretvarjamo, da je Holmes detektiv. Torej bi se hkrati pretvarjali, da je Sherlock Holmes fikcijski lik (neka posebna vrsta bitnosti) in detektiv, ki kadi pipo (oseba). Lahko se pretvarjamo, da je neka stvar nekaj drugega (npr. da je kepa blata čokoladna pita) ali da obstajajo stvari, ki jih v resnici ni (pošasti, čarovnice, govoreče mačke itd.), ne moremo pa se pretvarjati, da obstaja nekaj, kar je hkrati konkretnost in abstraktnost (oseba in fikcijski lik). Walton tega ugovora verjetno ne bi sprejel, ker bi dejal, da imamo preprosto dve različni igri pretvarjanja o istem fiktivnem imenu. Od konteksta je odvisno, katero uporabimo. S tem se sicer izogne zgoraj omenjenemu protislovju, vseeno pa ne reši problema z intenco pri eksternem govoru o fikciji.

Če sprejmemo zgornje kritike, potem niti artefaktična teorija niti teorija pretvarjanja ne moreta v celoti razrešiti nejasnosti v govoru o fikciji. Zdaj imamo za odgovor na vprašanje o fikcijskem diskurzu dve možnosti. Lahko sklenemo, da je narava fikcijskega diskurza taka, da se nekaterim nejasnostim ne moremo izogniti. Če v vsakdanji govor ne želimo sprejeti protislovij, pa moramo pristati na to, da bo eno izmed teorij treba spremeniti. Thomasson (2003b) predstavi rešitev problema z vključitvijo pretvarjanja v svojo verzijo artefaktične teorije fikcije.

## 2 Vpeljava pretvarjanja v teorijo abstraktnih artefaktov

Nekateri fikcijski realisti (npr. Kripke, 1980; van Inwagen, 2003; Thomasson, 2003b) so sprejeli, da le del fikcijskega diskurza lahko razumemo dobesedno, v preostalem govoru o fikciji pa se zgolj pretvarjamo, da izrekamo resnične trditve. V prispevku

obravnavam teorijo, ki jo je oblikovala Amie Thomasson (2003b). Slednja je artefaktično teorijo fikcije razširila tako, da je sprejela pretvarjanje v internem fikcijskem diskurzu (tj. v vsebini fikcijskega dela in v govoru o vsebini fikcijskega dela), v eksternem diskurzu pa je ohranila možnost dobesedne resničnosti trditev. S tem ohrani ontološki status fikcijskih likov, ki jih definira kot od avtorja ustvarjene, minljive in odvisne abstraktne objekte (Thomasson 1999, 2003a), a hkrati sprejme, da o njih lahko izrekamo dobesedno resnico le v eksternem govoru o fikciji. Ko se v govoru nanašamo na vsebino zgodbe, o fikcijskih likih več ne govorimo kot o abstrakcijah, pač pa se pretvarjamo, da so Sherlock Holmes, Frankensteinova pošast, Emma Woodhouse, Lepa Vida ipd. konkretne bitnosti z lastnostmi, ki jih določa vsebina literarnega dela.

Sprejetje pretvarjanja v artefaktično teorijo omogoča, da razrešimo nejasnosti, omenjene v prejšnjem poglavju. Ker pretvarjanja ne razširimo na eksterni govor, sprejmemo, da je o fikcijskih likih možno povedati nekaj dobesedno resničnega. Literarni teoretiki se v raziskavah o likih na te lahko nanašajo dobesedno, ko razmišljajo o okoliščinah njihovega nastanka, o vprašanju avtorstva ipd. Pretvarjanje sprejmemo le v govor o fikciji, ki se nanaša na vsebino del. Thomasson (2003b, str. 207) trdi, da pri vseh takih stavkih uporabljamo implicitni operator *glede na zgodbo*. Stavkov »Frankensteinovo pošast je ustvaril dr. Frankenstein«, »Sherlock Holmes je detektiv« in »Lepa Vida utone« ne smemo razumeti dobesedno, ampak v kontekstu pravil pretvarjanja, ki jih narekuje zgodba, v kateri obstajajo. Dobesedno so torej neresnični, znotraj pretvarjanja pa vseeno ohranjajo fikcijsko resničnostno vrednost. To razloži tudi, zakaj se nam zdi, da je stavek »Sherlock Holmes je detektiv« resničen, »Sherlock Holmes je klovn« pa neresničen. Prvi je namreč v skladu s pravili igre pretvarjanja, ki jih narekujejo zgodbe o Holmesu, drugi pa ne, zato ni neresničnen samo dobesedno, ampak tudi fikcijsko (glede na zgodbo).

Menim, da se taka teorija sklada z našo intuicijo o fikcijskih likih in s koncepti, ki so jih o njih oblikovale literarne prakse. Zdi se nam, da so nekaj, vendar obstajajo drugače kot mi. O njih lahko razmišljamo dobesedno (da so abstrakcije, ustvarjene od avtorja v določenem času in določenem literarnem delu) ali pa v kontekstu zgodbe, v kateri so zapisani. Tam se pretvarjamo, da gre za konkretne osebe, stvari in bitja, ki uprimerjajo lastnosti. Vsebina zgodbe je tista, ki nam omogoči, da določamo tudi med tem, katere lastnosti o liku bi si naj predstavljali in katerih ne. Z vpeljavo fikcijske resničnostne vrednosti je razložena intuicija, zakaj se nekatere

intrafikcijske trditve zdijo resnične, druge pa ne, čeprav so vse dobesedno neresnične. Resnična se zdi izjava »Sherlock Holmes je detektiv«, za izjavo »Sherlock Holmes je klovn« pa bomo takoj dejali, da je neresnična, čeprav sta obe trditvi dobesedno neresnični. Vendar za drugi stavek ne menimo, da je neresničen zato, ker pripisuje konkretno lastnost abstrakciji, ampak zato, ker »biti klovn« ni ena izmed lastnosti, ki jih Sherlocku Holmesu pripisujejo zgodbe o njem. Ravno to razliko razjasni fikcijska (ali tudi hlinjena) resničnostna vrednost, ki velja samo v sklopu pretvarjanja. V nadaljevanju bom obravnavala nekaj vprašanj oz. problemov v primeru sprejetja takšne teorije.

### 2.1 Problem referiranja fiktivnih imen v internem govoru o fikciji

Če kot Thomasson (2003b) sprejmemo, da interni diskurz o fikciji vsebuje pretvarjanje, moramo pojasniti, kako to pretvarjanje sploh poteka. Filozofi fikcije sprejemajo dva odgovora: ali trdimo, da gre v intrafikcijskem diskurzu za pretvarjanje o nečem (*de re* pretvarjanje) ali gre za pretvarjanje, da obstajajo osebe, stvari objekti itd., ki jih v resnici ni (*de dicto* pretvarjanje). Vprašanje je pomembno zaradi referiranja fiktivnih imen. Če izberemo prvo možnost (*de re* pretvarjanje), trdimo, da fiktivna imena tudi v internem fikcijskem diskurzu referirajo in s tem ohranimo pomenljivost stavkov, kot je »Sherlock Holmes je vstal in si prižgal pipo«. Če izberemo *de dicto* pretvarjanje, nasprotno trdimo, da fiktivna imena (vsaj) v intrafikcijskem diskurzu ne referirajo in s stavki, ki jih vsebujejo, ničesar ne zatrjujemo.

Za fikcijske antirealiste je odgovor preprost. Ker ne priznavajo ontološkega statusa fikcijskih likov, bo njihov odgovor *de dicto* pretvarjanje. Tako se, ko govorimo o lastnostih Sherlocka Holmesa, pretvarjamo, da obstaja nek človek z imenom Sherlock Holmes, ki je detektiv in kadi pipo.

Za fikcijske realiste je naloga težja. Če že vpeljemo nove entitete v našo ontologijo in sprejmemo, da fiktivna imena referirajo v eksternem govoru o fikciji, se zdi naravno najprej pomisliti, da verjetno morajo referirati tudi v internem govoru. To bi pomenilo, da vedno, ko uporabljamo neko fiktivno ime (iz fikcije), referiramo na abstraktni objekt. V stavku »Sherlock Holmes je vstal in si prižgal pipo« je referent imena Sherlock Holmes abstraktni artefakt, ki ga je s pisanjem dela ustvaril Arthur Conan Doyle. Menim, da sta prednosti *de re* pretvarjanja naslednji: teorija ostane enotna v sprejemanju ontologije abstraktnih objektov in ohranimo pomenljivost

intrafikcijskih stavkov. Čeprav je ta sklep sprva zelo mamljiv (*de re* pretvarjanje v zgodnjih delih zagovarja tudi Thomasson (1999)), naleti na precej nelagoden problem.

Problem nastane, ko se vprašamo, glede česa se pretvarjamo, ko izjavimo, da fikcijski liki so (prim. Friend, 2007; Sainsbury, 2009; Sawyer, 2002; Thomasson, 2003b). Tu predstavljam obliko kritike, kot jo izpostavlja Friend (2007, str. 152). Predstavljajmo si Lizzie, fikcijski lik, ki ga je po teoriji abstraktnega artefakta ustvarila Jane Austen v delu *Prevzetnost in pristranost*. Ker je Lizzie abstraktni artefakt, ne more uprimerjati lastnosti »biti ženska« ali »biti trmasta«. Uprimerja le lastnost »biti abstraktni artefakt«. Stavek »Lizzie je trmasta« zato ne more biti dobesedno resničen. Če v intrafikcijski diskurz sprejmemo *de re* pretvarjanje, moramo sprejeti, da so fikcijski liki nekaj drugega, kot to, kar se pretvarjamo, da so. Fikcijski liki so po artefaktični teoriji abstraktni objekti. To pomeni, da nimajo nobene lastnosti, ki jim jih pripisujejo zgodbe, pač pa so ustvarjene abstraktnosti, za katere se med branjem fikcije le pretvarjamo, da imajo lastnosti, ki lahko pripadajo samo konkretnim objektom (Friend 2007, str. 152). Če to sprejmemo, je Sherlock Holmes abstraktni artefakt, za katerega se pretvarjamo, da uprimerja lastnost »kaditi pipo«. *De re* pretvarjanje v tem primeru od nas zahteva, da se o abstraktnih objektih pretvarjamo, da lahko uprimerjajo lastnosti konkretnih stvari (da abstrakcije lahko kadijo, so jezne, se rodijo itd.). Da poudari čudnost te zahteve, Sarah Sawyer (2002, str. 192) predstavi analogijo s števili:

> *»Poskusite se pretvarjati, da je resničen naslednji stavek: 'Številka dve je človek, ki rad igra kroket.' Ni jasno, da imamo kakršno koli idejo o tem, kako to storiti.«*

Čeprav tudi sama priznava kritiko *de re* pretvarjanja, Thomasson (2003b, str. 212) skuša problem nekoliko omiliti s primerjavo z gledališkimi igralci, ki se *de re* pretvarjajo, da so mačke, angeli ali celo števila. Pri pretvarjanju, da je neka oseba mačka, nimamo težav, ker se pretvarjamo, da je ena konkretnost nekaj drugega. Pri ostalih dveh primerih, kjer se igralec pretvarja, da je bodisi angel bodisi število tri, pa lahko skušamo najti rešitev za *de re* pretvarjanje. Namreč, če je res mogoče, da se realna oseba pretvarja, da je abstrakcija, bi moralo biti mogoče tudi, da se za abstrakcije pretvarjamo, da so konkretne bitnosti. Vendar se igralec, ki se pretvarja, da je angel, ne pretvarja, da je nek abstraktni objekt, ampak kvečjemu, da je konkretnost, ki obstaja v fiktivnem svetu igre. Igralec se ne pretvarja, da je

abstrakcija, ki ne obstaja v prostoru in času, ravno nasprotno, pretvarja se, da je nekaj konkretnega, kar ima drugačne lastnosti od človeka (lahko leti, ima čudežne moči itd.). Enako velja za otroka, ki se v igri o številih pretvarja, da je število 3. Verjetno bo prej držalo, da se ta otrok pretvarja, da je številka, kot jo zapišemo v zvezku (npr. ima kostum takšne oblike). Lahko nam na odru fizično prikaže, kaj pomeni, če številu tri prištejemo dve. Zelo težko pa bi na odru prikazali nekaj, kar ne obstaja v prostoru in času.

## 2.2 *De dicto* pretvarjanje in vprašanje dvojnosti

Da se izognejo problemu predstavljanja abstraktnega objekta s konkretnimi lastnostmi, so nekateri fikcijski realisti (npr. Searle, 1979; Kripke 1980, 2013) ubrali drugo pot razlage internega fikcijskega diskurza in sprejeli *de dicto* pretvarjanje. Stavki, ki se nanašajo na vsebino zgodb, od nas ne zahtevajo, da se pretvarjamo, da so fikcijski liki (abstrakcije) realne osebe, ampak samo to, da je nekoč obstajala neka oseba z določenim fiktivnim imenom (Thomasson 2003b, str. 212):

> *» /…/ v stavkih Holmesovih zgodb se za Holmesa (fikcijski lik, abstraktni artefakt) ne pretvarjamo, da je detektiv, temveč ima pretvarjanje obliko: Nekoč je živel človek, ki je se je imenoval 'Holmes', bil je detektiv, bil je zelo pameten itd. Na podlagi takšnih (de dicto hlinjenih) pripisov se lahko poznejši bralci, kritiki in zgodovinarji de re sklicujejo na fikcijski lik Sherlocka Holmesa in o njem govorijo, na primer, da ga je ustvaril Arthur Conan Doyle, da je najslavnejši lik viktorijanske književnosti itd.«*

Prvi problem *de dicto* pretvarjanja je pomenljivost trditev s fiktivnimi imeni. Če imena nimajo referenta, kako izražajo pomenljive trditve? Možen odgovor na to je, da intrafikcijski stavki sploh ne izrekajo nobenih trditev, ampak se le pretvarjajo, da jih (Kripke, 2013; Searle, 1979). Searle (1979, str. 68) razliko med intrafikcijskim diskurzom in dobesedno rabo jezika vidi ravno v tem, da interni diskurz o fikciji ni vezan na resnico. Avtor, ki piše fikcijo, ne opisuje dejanskih dogodkov, zato ni vezan na resničnost svojih izjav. S pripovedovanjem zgodbe ne trdi ničesar, le pretvarja se, da nam pripoveduje nekaj o nekomu. Ilokucijsko dejanje, ki ga izreka (piše) avtor, je le hlinjeno. Hlinjeno ilokucijsko dejanje ustvari tako, da uporabi fiktivno ime in tvori stavke, kot da referira na realno osebo. S tem sicer ustvari fikcijski lik (abstrakcijo), vendar v stavkih, ki jih zapisuje, ne referira na abstraktni objekt, ta nastane kot posledica tvorjenja hlinjenih trditev o osebi, za katero se pretvarja, da obstaja. Vse

to je združljivo s teorijo Thomasson, da so fikcijski liki abstraktni artefakti, ki jih z izrekanjem ustvari avtor.

To pojasni 1. vrsto intrafikcijskega diskurza (avtorjeva raba fiktivnih imen). Kaj pa lahko rečemo o izjavah bralcev fikcije o vsebini del? Tudi ti se le pretvarjajo, da izrekajo trditve. Razlika je v tem, da ima njihovo pretvarjanje določena pravila, ki jih predpisuje delo, o vsebini katerega se pogovarjajo. Ko izrečemo »Sherlock Holmes je detektiv«, je to fikcijsko res zato, ker se je tako pretvarjal avtor, ko je pisal zgodbo o Holmesu.

Zdi se, da *de dicto* pretvarjanje bolje pojasni interni govor o fikciji (k tej ugotovitvi se nagiba tudi Thomasson, 2003b, str. 214). Od nas ne zahteva, da razložimo, kako si predstavljati abstraktni predmet s konkretnimi lastnostmi. Morda lahko rečemo celo, da ta razlaga pretvarjanja bolje opiše postopek nastajanja fikcijskega dela, ker je težko trditi, da ima pisatelj ob pripovedovanju zgodbe v mislih kakršen koli abstraktni objekt, o katerem se pretvarja, da pripoveduje zgodbo. Nujna posledica sprejetja *de dicto* pretvarjanja samo v interni fikcijski diskurz pa je dvojnost. Nimamo več enotne teorije, ki pojasni vse vrste fikcijskega diskurza, ampak ločeni razlagi za metafikcijske in intrafikcijske trditve (v metafikcijskem diksurzu fiktivna imena referirajo, v intrafikcijskem pa ne).

Če je bil naš prvotni cilj oblikovati teorijo, ki bo uniformna za vse vrste fikcijskega diskurza, nam je spodletelo. V ontologijo smo vpeljali nove entitete, o njih lahko povemo le malo resnice, hkrati pa teh entitet sploh ne moremo uporabiti za razlago intrafikcijskega diskurza.

Thomasson (2003b, str. 214) zagovarja, da fiktivna imena vsaj v eksternem govoru morajo referirati, ker brez referiranja fiktivnih imen vsaj v metafikcijskem diskurzu ne rešimo nejasnosti, ki nastanejo, ko govorimo o fikciji (podobno trdi tudi Kripke, 2013). Mislim, da je njen ugovor, zakaj v eksternem diskurzu ne moremo priznati pretvarjanja, smiseln. Ko izjavimo: »Prvo delo, v katerem se pojavi lik Sherlocka Holmesa, je *Študija v škrlatnem*«, izrekamo resnico o nastanku lika. Pri takšnih in podobnih metafikcijskih trditvah smo vezani na dobesedno resnico, ne na igro pretvarjanja, ki jo zahteva fikcija, niti si sami ne izmišljamo hlinjenih trditev.

Poleg tega teorija dobro razloži našo intuicijo o tem, da se zdi resnično izreči »Sherlock Holmes je fikcijski lik« in »Sherlock Holmes je detektiv«. V prvi izjavi res izrekamo dobesedno resnico o Holmesu, v drugi pa se preprosto strinjamo z avtorjem zgodbe o Holmesu in se kot avtor pretvarjamo, da zatrjujemo resnico o neki realni osebi (za katero vemo, da v resnici ne obstaja).

Nathan Salmon (1998, str. 298), je *de dicto* pretvarjanju ugovarjal, češ da se fikcijskih likov, ko smo jih že vzpostavili kot posebne vrste entitete, ne zdi smiselno uporabljati v samo enem kontekstu. Sawyer (2002) tej kritiki, mislim, da upravičeno, ugovarja, ker da napačno enači intrafikcijsko in metafikcijsko rabo fiktivnih imen. Točno to trdi tudi Searle (1979), ki zagovarja, da v intrafikcijskem diskurzu izrekamo hlinjena ilokucijska govorna dejanja, v metafikcijskem diskurzu pa gre za dobesedno rabo jezika.

Če sprejmemo tezo, da se metafikcijski in intrafikcijski diskurz razlikujeta ravno po tem, da v enem nekaj dobesedno zatrjujemo, v drugem pa se le pretvarjamo, da nekaj zatrjujemo, je smiselno sprejeti drugačno rabo fiktivnih imen v obeh vrstah diskurza. Vseeno je nenavadno, da lahko fiktivno ime enkrat referira, drugič pa funkcionira kot prazno ime. Trdimo, da z izrekanjem intencionalnih hlinjenih trditev (z izrekanjem stavkov s praznimi imeni, za katera se pretvarjamo, da referirajo) avtorji ustvarjajo abstraktne objekte, na katere lahko *de re* referiramo v eksternem diskurzu.

Čeprav je to morda nekoliko čudno, se zdi, da takšnemu konceptu fikcijskih likov pritrjujejo literarne prakse. V enem od argumentov zakaj vpeljati fikcijske like kot posebne entitete, Thomasson izhaja natanko iz tega, da literarne prakse, ki ves čas uporabljajo izraze, kot je fikcijski lik, določajo pogoje za obstoj takih bitnosti. Ti pogoji pa so po Thomasson tako minimalni, da je nesmiselno ne sprejeti fikcijskih likov v našo ontologijo. V skladu z literarnimi praksami in našimi splošnimi prepričanji avtorje obravnavamo kot tiste, ki ustvarijo fikcijske like. Pogoj, da avtor ustvari fikcijski lik, je samo to, da se pretvarja, da referira na realno osebo in o njej hlinjeno trdi stvari kot del ustaljene tradicije pretvarjanja pri pripovedovanju zgodb (Thomasson, 2003a, str. 148–150).

Jezikovne prakse nam res dovoljujejo, da fiktivna imena uporabljamo tako, da je njihov referent abstraktni fikcijski lik in če fikcijski antirealist ne najde ustreznega pojasnila, kako brez vpeljave novih bitnosti razložiti razliko med intrafikcijskim in

metafikcijskim diskurzom, se zdi, da artefaktična teorija kljub zahtevi po dvojni rabi imen ponuja boljšo rešitev.

## 3 Zaključek

V vsakodnevnem govoru nam ni čudno trditi, da je Sherlock Holmes odličen detektiv in hkrati trditi, da je lik Holmesa prototip za veliko modernejših literarnih junakov. Radi bi se izognili sklepu, da ljudje vsakodnevno pritrjujemo protislovjem, za katera ne samo, da ne vemo, da so protislovja, ampak imamo celo občutek, da ne izražajo ničesar protislovnega. Reševanja tega problema so se najbolj zagreto lotili fikcijski realisti. Amie Thomasson z delitvijo fikcijskega diskurza skuša najti način, ki bi čim bolj ohranil naše intuitivno prepričanje o fikcijskih likih, vseeno pa bi omogočil, da je nekaj našega govora o fikciji dobesedno resničnega. Zdi se, da upravičeno trdi, da je sprejeti nove entitete, o katerih lahko povemo nekaj resničnega, bolj intuitivno (in manj zahtevno), kot sprejeti, da se pretvarjamo, da take abstraktnosti obstajajo, čeprav v metafikcijskih izjavah nimamo nobene intence po pretvarjanju.

Problem nastane, ko želimo vzpostaviti teorijo, ki bo enotna za celoten fikcijski diskurz. Da to storimo, moramo predpostavljati, da je referent fiktivnih imen nek abstraktni objekt tudi, ko se pretvarjamo, da je konkretna oseba. *De re* pretvarjanje je vabljivo, ker ohrani enotno teorijo in ne zahteva dvojne rabe imen, vendar je težko razložiti, kako se za abstrakcije pretvarjamo, da so konkretne. A morda smo prestrogi, glede na to, da govorimo o fikciji, ki sama temelji na domišljiji in izmišljanju.

Skušala sem nakazati, zakaj je *de dicto* pretvarjanje manj problematično, čeprav se na prvi pogled zdi, da nas vodi v nasprotno od tega, kar bi radi dosegli. Z *de dicto* pretvarjanjem v internem diskurzu nimamo več enotne razlage govora o fikciji, fiktivna imena ne referirajo v internem govoru o fikciji, niti z njimi ne izrekamo pravih propozicij, ampak le hlinjene. Se za tako malo resnice še vedno splača sprejeti obstoj fikcijskih likov? Moj namen je bil pokazati, zakaj se splača na vprašanje odgovoriti pritrdilno. Prvi razlog poda že Thomasson, ki opozarja na neustreznost pretvarjanja v metafikcijskem diskurzu. Drugi razlog je mogoče iskati v tem, da je zahteva po enotnosti teorije odvečna, ker se interni in eksterni govor o fikciji popolnoma razlikujeta. Tretjo prednost za fikcijskega realista pa vidim v literarnih

praksah, ki največ uporabljajo pojme, kot je fikcijski lik. Literarni teoretiki bi se verjetno strinjali, da o fikcijskih likih kot takih izrekajo dobesedne trditve. V teoretičnih člankih o književnih delih zasledimo stavke kot »Šeligo je ustvaril lik hrepeneče in nepotešene ženske za vse čase, zato razbija klasično enotnost časa, saj srečujemo Vido kot predvojno meščanko, kot boginjo in demonsko bitje« (Tergeslav, 1999, str. 292). Vidimo, da tudi literarni teoretiki avtorje dojemajo kot ustvarjalce in vsaj zdi se, da se ne vključujejo v nobeno pretvarjanje, ko trdijo, da so liki nekaj, kar je ustvarjeno, pač pa izrekajo dobesedno resnico. Če sprejmemo argument Thomasson, da so pogoji za obstoj fikcijskih likov minimalni, fikcijskemu realizmu dobro kaže.

**Reference**


Friend, S. (2007). Fictional characters. *Philosophy Compass*, *2*(2), 141–156. https://doi.org/10.1111/j.1747-9991.2007.00059.x

Giovanelli, A. (2009). Walton, Kendall L(ewis). V S. Davies, K. M. Higgins, R. Hopkins, R. Stecker in D. E. Cooper (ur.), *A companion to aesthetics* (67) (str. 588–591). Blackwell Publishing.

Kripke, S. A. (1980). *Naming and Necessity*. Basil Blackwell.

Kripke, S. A. (2013). *Reference and Existence: The John Locke Lectures*. University Press.

Sainsbury, R. M. (2009). Fikcijske bitnosti so abstraktni artefakti. *Analiza: časopis za kritično misel*, *13*(4), 93–114.

Salis, F. (2014). Fictional entities. *Online Companion to Problems of Analytic Philosophy, 2013 Edition*. http://hdl.handle.net/10451/10860

Salmon, N. (1998). Nonexistence. *Noûs*, *32*(3), 277–319.

Sawyer, S. (2002). Abstract artifacts in pretence. *Philosophical papers*, *31*(2), 183–198.

Searle, J. R. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Searle, J. R. (1996). *The construction of social reality*. Penguin Books.

Terseglav, M. (1999). Zgodba nesrečne Lepe Vide v sodobni slovenski dramatiki. *Traditiones*, *28*(2), 289–297.

Thomasson, A. L. (1999). *Fiction and metaphysics*. Cambridge University Press.

Thomasson, A. L. (2003a). Fictional characters and literary practices. *The British Journal of Aesthetics*, *43*(2), 138–157.

Thomasson, A. L. (2003b). Speaking of Fictional Characters. *Dialectica*, *57*(2), 205–223.

van Inwagen, P. (2003). Existence, Ontological Commitment and Fictional Entities. V M. J. Loux in D. W. Zimmerman (ur.), *The Oxford Handbook of Metaphysics* (str. 131–157). Oxford University Press.

Walton, K. L. (1990). *Mimesis as make-believe: On the foundations of the representational arts*. Harvard University Press.

Walton, K. L. (2010). Strah pred fikcijo. *Analiza: časopis za kritično misel*, *14*(4), 75–92.

# Modalni katapulti

**ANALIZA**
Let. 28, št. 2, str. 189–208, december 2024

ANαliZA

# MODAL CATAPULTS AND THE LIMITS OF MODAL LOGIC

**Accepted**
28. 2. 2024

**Revised**
10. 5. 2024

**Published**
31. 12. 2024

DANILO ŠUSTER
University of Maribor, Faculty of Arts, Maribor, Slovenia
danilo.suster@um.si

CORRESPONDING AUTHOR
danilo.suster@um.si

**Abstract** I explore modal "catapults," a variety of closure principles for modal operators. Consider a proposition *p* that logically implies, entails, strictly implies, modally implies, materially implies, …, a proposition *q*. According to the appropriate catapult for a modal operator *M*, if *Mp*, then also *Mq*. Modal catapults play a crucial role in the logical analysis of traditional philosophical arguments, such as fatalism and incompatibilism. Additionally, standard deontic paradoxes and moral dilemmas involve a deontic modal catapult in some form. In the realm of deontic logic, I advocate for a solution grounded in actualism and counterfactuals (Jackson, Goble). In considering whether it ought to be that *A* we should look particularly at what would be the case, were *A* the case. This approach explains the failures of closure while still acknowledging its central role in modal reasoning. Modal catapults are indispensable to the logic of modalities, but they also delineate the boundaries of this approach.

**Keywords**
normal modal logic, closure principles, the consequence argument, deontic logic, actualism, moral dilemmas

https://doi.org/10.18690/analiza.28.2.189-208.2024


University of Maribor Press

AN$_\alpha$LIZA

**ANALIZA**
Vol. 28, no. 2, pp. 189–208, December 2024

# MODALNI KATAPULTI IN MEJE MODALNE LOGIKE

DANILO ŠUSTER
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
danilo.suster@um.si

DOPISNI AVTOR
danilo.suster@um.si

**Sprejeto**
28. 2. 2024

**Pregledano**
10. 5. 2024

**Izdano**
31. 12. 2024

**Izvleček** Modalni »katapulti« so osnovne logične sheme sklepanj za modalne operatorje (»zaprtost« za logično posledico). Denimo, da propozicija *p* logično implicira, strogo implicira, modalno implicira ali materialno implicira …, propozicijo *q*. V skladu z ustreznim »katapultom« za modalni operator *M* velja prenos modalnosti: če *Mp*, potem tudi *Mq*. Modalni katapulti imajo pomembno vlogo pri logični analizi tradicionalnih filozofskih argumentov, kot sta fatalizem in inkompatibilizem. Tudi pri deontičnih paradoksih in v moralnih dilemah imajo katapulti osrednjo vlogo. Zagovarjam pristop, ki temelji na aktualizmu in protidejstvenem razmišljanju (Jackson, Goble): v razmišljanju o tem, ali bi *moralo* biti res, da *p*, upoštevamo, kaj *bi* bilo res, če *p*. Ta pristop dobro pojasni, kdaj prenos deontične nujnosti ne deluje in hkrati razloži njegovo uporabnost. Modalni katapulti so nepogrešljivi v logiki modalnosti, vendar tudi zarišejo meje tega pristopa.

**Ključne besede**
normalna modalna logika, načela »zaprtosti«, argument iz posledic, deontična logika, aktualizem, moralne dileme

https://doi.org/10.18690/analiza.28.2.189-208.2024


Univerzitetna založba Univerze v Mariboru

## 1 Modal logic and philosophy

Tradition has it that the phrase "Let no one ignorant of geometry enter" was engraved at the door of Plato's Academy. Humberstone (2005, 534) proposes that philosophy departments of a broadly analytical stripe would do well to post a similar inscription: "Let no one who is ignorant of modal logic enter here." Philosophical trends change in time, and so do their inscriptions. However, it is still true that the various branches of modern philosophical logic cannot be understood without some basic knowledge of modal logic. According to MacFarlane (2020, xv), "philosophical logic" today encompasses two main areas: (a) the philosophical investigation of the basic notions of logic and (b) the deployment of logic to help with philosophical problems. In the second sense, it consists mainly in the formal investigation of alternatives and extensions to classical logic. Modal logic is particularly significant, as many traditional philosophical problems—such as fatalism and free will, realism and knowledge, and moral dilemmas—entail modal notions and reasoning.

The book under discussion (Šuster, 2023) is divided into two parts. The initial "motivational" section comprises six essays addressing traditional philosophical problems, each centred around a specific modal argument or rule. The book's second part provides a formal "toolbox" for the first part: a standard introduction to normal (propositional) modal logic and possible world semantics. The final chapter presents a detailed account of non-monotonic logic and semantics of counterfactual conditionals – they are, after all, the pillars of hypothetical thinking and the royal road to the empire of modality in general.

Overall, the book fits the program of "hard-core" analytic philosophy perhaps best exemplified by Williamson in his methodological "sermon" (2007, 288):

> "Much even of analytic philosophy moves too fast in its haste to reach the sexy bits. Details are not given the care they deserve: crucial claims are vaguely stated, significantly different formulations are treated as though they were equivalent, examples are under-described, arguments are gestured at rather than properly made, their form is left unexplained, and so on. /…/ The fear of boring oneself or one's readers is a great enemy of truth. Pedantry is a fault on the right side. /…/ Precision is often regarded as a hyper-cautious characteristic. It is importantly the opposite. Vague statements are the hardest

> to convict of error. Obscurity is the oracle's self-defence. To be precise is to make it as easy as possible for others to prove one wrong."

Some might say that a hard-core analytic philosophy with its insistence on logic, clarity, and detail is slightly out of date, "recently old philosophy is like recently old fashion: old enough to be dowdy but not old enough to be romantic" (Saunders, 2022). Williamson and I would both disagree. The standards are not just recently old ;they were set by the first grandmaster of logic and philosophy, Aristotle. Consider his famous discussion on necessity, time, logic and freedom in *De Interpretatione* (cf. the third chapter of Šuster, 2023). Moreover, they are likely to stay with us; I was always impressed by the precision, carefulness and attention to modal details in van Inwagen's formulation of *the consequence argument* for classical incompatibilism (if determinism is true, then no one is or ever was able to do otherwise). Much of my thinking about modal rules of inference, their interconnections, and implications in general has been shaped by the analysis of the reasoning that underpins this specific argument. However, in the book itself, I did not explain the actual *title*. In this article,I aim to address this omission and elucidate the connections between the various topics discussed in the book.

## 2 Modal catapults

Fischer and Ravizza coined the name "modal slingshot" (1996, 213):

> "The basic idea of the modal principle is that if some state of affairs S1 obtains and one does not have any choice about (or control over) S1's obtaining, and if S1 implies S2 and one does not have any choice about (or control over) the fact that if S1 obtains, then S2 obtains, then it follows that S2 obtains and one does not have any choice about (or control over) S2's obtaining. The modal principle works as a kind of modal slingshot: it projects the modal property of "powerlessness" from one state of affairs (S1) to another (S2)."

They were discussing van Inwagen's consequence argument (CA) and the role of the modal principle *Beta* (no choice about *p*, no choice about if *p*, then *q*, therefore no choice about *q*). Let us say that a particular principle corresponds to a "slingshot,"as a smaller handheld device for launching "powerlessness" in the case of CA."Catapults," on the other hand, are not supposed to indicate "siege weapons"(perhaps for attacking philosophical arguments), but should be associated with

larger, more complex devices used to launch a projectile over a distance. "Modal catapults" is a metaphorical designation of general logical weapons for projecting modal properties of propositions. My use of this term is closely related to the more familiar logical concept of *closure*. In standard metalogic, the logical closure of a set of propositions is the set of all propositions that logically follow from those propositions. Closure under *entailment* ensures that a given set of propositions includes all propositions that are logically entailed by it. When a set of propositions is closed under a *rule*, this means that applying the rule to any propositions within the set will produce propositions that are also within the set. One can specify a system of modal logic as a logically closed set of propositions (closure of characteristic modal axioms under the specific inference rules). The closure of a modal operator under a rule ensures that the modal properties of propositions are transferred.

Suppose that a proposition *p* logically implies, entails, strictly implies, modally or **M**-implies, materially implies, ..., a proposition *q*. Then, according to rule *R*, if *Mp*, then also *Mq*. Some catapults *seem* to be valid (if *p* is possibly true and *p* entails *q*, then *q* is also possibly true), and some are uncontroversially invalid (if *p* is necessarily true and *p* materially implies *q*, then *q* is necessarily true). But many are in-between, and this is where most of the philosophical action is (Ought we do something whenever our doing it logically follows from our doing something else that we ought to do?). The paradigm case of a modal catapult is the defining rule of normal modal logic. According to Chellas (1980, 114–115) normal systems of modal logic can be characterized in terms of the schema: »$\Diamond\phi =_{def} \sim\Box\sim\phi$« and the rule of inference:

**RK.** From $\vdash (\phi_1 \,\&\, \phi_2 \,\&\, \ldots, \phi_n) \supset \phi$ infer $\vdash (\Box\phi_1 \,\&\, \Box\phi_2 \,\&\, \ldots, \Box\phi_n) \supset \Box\phi$

Rules of inference, in this case, preserve the theoremhood: the conclusion of a rule is a theorem if each of its hypotheses is. **RK** expresses a general rule of modal consequence: a proposition is necessary if it is a consequence of a collection of propositions each of which is necessary. When $n = 1$ we get the core representative of modal catapults, the principle of closure under logical consequence:

**RM.** From $\vdash \phi \supset \psi$ infer $\vdash \Box\phi \supset \Box\psi$

Closure is also a fundamental principle in logics of counterfactual conditionals of the form "if it had been the case that ϕ, then it would have been the case, that ψ" or '$\phi > \psi$'. A conditional variant of **RK** is:

**RCK.** From $\vdash (\psi_1 \,\&\, \ldots \,\&\, \psi_n) \supset \chi$ infer $\vdash (\phi > \psi_1 \,\&\, \ldots\, \phi > \psi_n) \supset (\phi > \chi)$

Modal catapults, in my sense, encompass a family of modal principles (variously called "the principles of distribution," "closure principles," "the principles of inheritance," etc.). Some typical modal catapults are:

- If a subject S knows that *p*, and *p* entails *q*, then S also knows that *q*.
- If it is inevitably the case that *p* and it is also inevitably the case that, if *p*, then *q*, then it is inevitably the case that *q*.
- If something is obligatory and it necessarily entails something else, then that something else should also be considered obligatory.

Let me formally introduce some typical instances. Let 'Np' stand for: *p* is true and no one has or ever had any choice about *p*. There is an enormous discussion about the following rule ("slingshot"), which plays a central role in CA:

**Beta** $\quad N\phi, N(\phi \supset \psi) \vdash N\psi$

A variation, also much discussed in the literature on CA, is:

**Beta 2** $\quad \Box(\phi \supset \psi) \vdash N\phi \supset N\psi$

Modal catapults have a venerable logical tradition. Diodorus Cronus offered the so-called "Master Argument" in the form of an inconsistent triad (Duncombe, 2024):

(MA1) Every past truth is necessary;
(MA2) The impossible does not follow from the possible;
(MA3) There is a possible truth which neither is true nor will be.

Diodorus denied (MA3) and affirmed, "All possible truths are either true or will be true." This was supposed to yield a form of *fatalism*, since only what is now true or will be true is possible, *unrealized* possibilities are excluded. A key premise is the

principle (MA2) that "Nothing impossible follows from the possible." In terms of modal logic, the principle expresses the closure of the possible over entailment (von Wright 1979, 302):

$$(\Diamond\phi \,\&\, \Box\,(\phi \supset \psi)) \supset \Diamond\psi$$

The variations of this principle are catapults in the form:

$$\Box\,(\phi \supset \psi) \supset (\Diamond\phi \supset \Diamond\psi)$$
$$\Box\,(\phi \supset \psi) \supset (\Box\phi \supset \Box\psi)$$

Kapitan (2002, 130) refers to the variations being discussed in the debate over CA as 'Diodoran Principles'. They are often used by incompatibilists to demonstrate the incompatibility of free will and determinism. Where *P* is any truth, *H* a proposition expressing the complete state of the world at a time in the distant past, *L* a conjunction of the laws of nature and '□' expresses broad logical necessity, it is a consequence of determinism that:

$$\Box[(H \,\&\, L) \supset P]$$

Consider now *the Simple argument* for *Incompatibilism*, an instance of **Beta 2**:

$$N(H \,\&\, L), \Box[(H \,\&\, L) \supset P] \vdash NP$$

Since a conjunction of laws and history is "out of anybody's control," it follows, given the truth of determinism, that, no one has a choice about any true proposition at all ('NP'). Should the compatibilists, therefore, deny the validity of Diodoran principles? Let me first notice that Chrysippus, the ancient compatibilist, really denied the closure principle (MA2): "Nothing impossible follows from the possible", and so do some modern compatibilists. Here is Perry's counterexample (2004, 247). Let *R* be the proposition that Joe raises his hand at *t*, where *t* is some future time. Let *Q* be a conjunction that Joe raises his hand at *t* and that Joe's mother ate a cookie in 1950. Since *Q* includes *R* as one of its conjuncts, *Q* entails *R* (so '□ (Q ⊃ R)'). Suppose also that Joe's mother did *not* eat a cookie in 1950. Joe can render *R* false by not raising his hand at *t* (so ~NR), but Joe cannot render *Q* false (therefore NQ), since *Q* was rendered false by his mother back in 1950.

The incompatibilists are ready to offer some "tweaks" to defend their argument. van Inwagen (1983, 68) defines "S can render *p* false" (a denial of 'Np') as "It is within S's power to arrange or modify the concrete objects that constitute his environment in some way such that it is not possible in the broadly logical sense that he arrange or modify those objects in that way and the past have been exactly as it in fact was and *p* be true." He is aware of the fact that, according to this definition, one *can* render false untrue propositions about the past: "I can render the proposition that Socrates died of old age false, since it is not possible that the past should have been exactly as it in fact was and Socrates have died of old age." According to van Inwagen, Joe can render it false that Joe's mother ate a cookie in 1950 after all, contrary to Perry's judgment!

Here, I am not interested in all the twists and turns of this particular philosophical discussion (see chapter 4 of Šuster, 2023), but it is clear that the modal slingshot (an instance of a modal catapult) is at the centre of the debate. The notion(s) of the ability to act otherwise in the free will debate (a denial of 'N' in the slingshot) are often explicated in terms of *counterfactual* conditionals. Thus Kapitan (2011, 135): S is able at *t* to see to it that ~P iff there is a course of action X such that at *t* (i) S is able at *t* to do X, and (ii) were S to do X, then ~P. According to the broadest possible understanding, all that is required from an agent S to have such an ability is that from S's doing X it may be inferred that *P* is false. Even contradictions and past falsities are such that one is able, at *t,* to see to it that they do not obtain in this sense: they are (now) false, and whatever S does, they (still) remain false.

But the compatibilists usually prefer an *active* reading of "If S were to do A, then *P*". The antecedent's being true in some sense "requires" *P* to be true ("makes it the case", "brings it about" that the consequent true). The *Simple argument* is invalid in this interpretation: nobody can make it the case that the thesis of determinism, '□[(H & L) ⊃ P]' is false and nobody can bring about a different combination of the past and the laws of nature, 'N(H & L).' The premises are true. But surely one can make it the case that one's hand, which was actually unraised, is raised (so '~NP')? This is quite clear from the early compatibilistic refutations of CA. Slote (1982, 19) interpreted 'N' as "selective" necessity – as being determined in a particular sort of way, which selects "some factor that brings about the unavoidable thing without making use of (an explanatory chain that includes) the desires, etc., the agent has around that time". The avoidability of *P* (or ability to render false the proposition

that *P*) then involves something like "the-agent-including-explanatory-chain" that *brings* about that not-P, or *active* ability. Let *P* be some "particular about-to-be-performed action" of an agent S (say raising her hand). Slote (1982, 20) argues *against* the inference:

$$N_s (H \& L), N_s[(H \& L) \supset P] \vdash N_s P$$

Premises are true (unavoidable for S, independent of her desires, etc.), but the conclusion is false – the relevant action is *brought about* and explained through S's desires, abilities, etc. Strictly speaking, this principle is closure of 'N' under implication, but Slote was right when he spoke about "closure under logical implication" or **Beta 2,** which really *fails* for active (in Slote's terminology, *selective*) ability.

Pruss tries to rehabilitate CA with the help of counterfactuals and claims that the following catapult principle is a plausible axiom (2013, 433):

**WEAKEN.** $p > q, \Box(q \rightarrow r) \vdash p > r$

Pruss reads 'Np' as the claim that there is nothing that anyone can (ever) do that would falsify *p*. He then derives **Beta 2** ('$\Box(\phi \supset \psi) \vdash N\phi \supset N\psi$') from **WEAKEN** and views this result as a decisive defence of CA based on the Simple argument for incompatibilism.

According to Chellas (1980, 269), counterfactual conditionals can be conceived of as expressions of *relative* necessity: the proposition expressed by ψ is in some way necessary with respect to the condition expressed by ϕ, or '[ϕ]ψ,' where the antecedent forms a unary modal operator. So, we get:

**RCK'.** From $\vdash (\psi_1 \& \ldots \& \psi_n) \supset \chi$ infer $\vdash ([\phi]\psi_1 \& \ldots [\phi]\psi_n) \supset [\phi]\chi$

Conditionality assumes the aspect of a propositionally indexed modality. '[ϕ]ψ' holds at a possible world *w* just in case ψ holds at all possible worlds possible with respect to the given one, relative to the proposition expressed by ϕ. Expressed in terms of relative necessities, **WEAKEN** becomes a not so innocent catapult-like principle:

$$[p]q, \Box(q \rightarrow r) \vdash [p]r$$

The compatibilist will likely remain unimpressed by this principle in the same way as they are by **Beta 2**, both will be declared invalid. They will draw attention to their *active* reading of conditionals, "If S were to do A, then *P*." Pollock (1984, 111) already made a distinction between *simple* subjunctive conditionals "If P had been true, then Q would have been true" ('P > Q') and the *necessitation* conditionals ('P >> Q'). A simple subjunctive conditional can be true because there is a connection between *P* and *Q,* such that *P*'s being true in some sense "requires" *Q* to be true but it can also be true because *Q* is already true and *P*'s being true would not interfere with this("even if P, still Q"). Pollock notices that a *Catapult* principle fails for the necessitation: "If P >> Q is true and Q entails R, then P >> R is true." For instance: If I had pushed the button, it would have rung – pushing the button necessitates that the doorbell rings. That the doorbell rings entails that the doorbell exists. But pushing the button does *not* make it true that the doorbell exists. The compatibilists will point out that WEAKEN *fails* for Pollock's necessitation conditional, which is just the notion employed in their "active" analysis of ability (the ability to make it the case, to produce, to bring about, etc.). The catapult does not work and Pruss' defense of CA fails.

I accept compatibilistic interpretations, but I think that logic *alone* will not settle the issue between the broad and the active understanding of the ability to act otherwise in CA. However, it will point out different properties of abilities and conditionals in question and make the disagreement much more precise, in line with the hard-core program of analytic philosophy (cf. Šuster, 2021).

## 3 Deontic catapults

Standard deontic logic (SDL) is a well-explored system of normal modal logic. Let 'O' stand for "it is obligatory that …" One of the central principles of SDL is the "Inheritance Principle":

**ROM.** From $\vdash \phi \supset \psi$ infer $\vdash O\phi \supset O\psi$

And a variant:

**ROM.'** From $O\varphi$ and $\Box(\varphi \supset \psi)$ infer $O\psi$

Brink (1996, 111) uses yet another variation, called *the obligation execution principle*:

$$(O\varphi \,\&\, \Box(\psi \supset \sim\varphi)) \supset O\sim\psi$$

One is obliged not to do anything that would interfere with the execution of our (original) obligations. This principle plays a crucial role in a debate on moral dilemmas (Šuster, 2023, chapter 6) – many have objected to its validity, but it is not easy to give it up. According to Goble (2009), it is the task of deontic logic to explain what follows from a statement that one ought to do something and to explain what other normative propositions follow. For example (Goble, 2009, 469):

"If the law of a nation states that every person aged 18 must register for national service, then Irwin, who has just turned 18, is surely entitled to infer that he must register for national service. The law does not explicitly say that he must; Irwin is not mentioned by name in any law of that nation. Nevertheless, that Irwin ought to register is surely implied by what the law does explicitly say."

Yet nearly all of the so-called paradoxes of deontic logic, in one way or another, involve **ROM** and its variants. Consider Ross's Paradox: "It is obligatory that the letter is mailed (M)," therefore, "It is obligatory that the letter is mailed (M) or the letter is burned (R)." A natural regimentation would be:

$$OM, \Box(M \supset (M \vee R)) \vdash O(M \vee R)$$

The conclusion is highly counterintuitive. Or take the Good Samaritan paradox:

(1) It is obligatory that Jones help (H) Smith who has been robbed (R).
(2) Necessarily, if Jones helps Smith who has been robbed, then Smith has been robbed.
(3) It is obligatory that Smith has been robbed.

In the form of a catapult, we get:

$$O(H \,\&\, R), \Box((H \,\&\, R) \supset R)) \vdash OR$$

The conclusion is unacceptable (McNamara & Putte, 2022). But does it really *follow*? In the first version of his *Logics*, Nolt (1997, 362) claims that the "Inheritance Principle" is not valid in modal deontic logic. He gives the example of the Bad Bart, who is bent on murdering me and can be stopped only by being killed. Then I may reason as follows:

> I should live (L).
>
> It is necessarily the case that if I live (L) Bad Bart dies (~B).
>
> Bad Bart should die (~B).
>
> Or: $OL, \Box(L \supset {\sim}B) \vdash O{\sim}B$

A simple possible worlds semantics model with two worlds is supposed to be enough to demonstrate the invalidity: in the actual world, I do not live, and Bart does not die, whereas in the one and *only deontic* alternative to the actual world, we are both alive. But there is a *caveat*: the deontic alternative is not relatively *possible* with respect to the actual world.[1] There are *two* accessibility relations in this model: one reflexive, defining the notion of (alethic) possibility, and another (S) serial, specifying the notion of (deontic) permissibility (for any two worlds $i$ and $j$, $iSj$ if and only if $j$ is morally permissible relative to $i$). Both premises are true in the model: I am alive in all the deontic alternatives in the model, and the conditional premise is true in both worlds. The conclusion is false, since Bart is alive in the deontic alternative to the actual world. According to Nolt, this is the case of a *moral tragedy*: what *ought* to be the case *cannot* be the case. It ought to be the case that no life is lost, but given the circumstances, this cannot be the case (the world where we are both alive is not possible with respect to the actual world).

The idea that permissible worlds need not be a subset of possible worlds is sometimes offered as a solution for some of the so-called deontic paradoxes (Morscher, 2002). In the standard "Kripke-style" possible world semantics, $p$ is obligatory in the actual world @ iff it holds in all the ideal worlds from the standpoint of @ (worlds $w$ such that everything obligatory at @ holds in $w$). In this

[1] Borut Čerkovnik (University of Ljubljana) pointed out this explanation for the invalidity of **ROM** in a discussion note on Šuster, 2023.

semantics, only *two* types of worlds are distinguished in a model: actual and ideal ones. All ideal worlds are automatically possible. This is reflected in typical theses of SDL, which say that the (logically) impossible cannot be obligatory, and if the impossibility is interpreted broadly enough, we even get: "it is possible that all (relevant) normative demands are met" as a thesis of SDL (cf. McNamara & Putte, 2022). Nolt can only get his countermodel to the "Inheritance Principle" if there are *three* types of worlds in the model: the actual world, ideal (permissible) worlds, and worlds that may or may not be possible relative to the actual world. Kant's famous *dictum* that "ought" implies "can" is then violated but, according to Nolt, modal deontic logic allows for such moral tragedies.

Suppose that we ought to preserve the Earth's biosphere, yet we are nevertheless fated to destroy it. We should but we cannot, according to Nolt. Yet what does "fated to …" mean here? Consider a person tied to a tree on the shore of a lake, unable to move. Does she have an *obligation* to help a child drowning in the lake? I do not think so. Is the *ought* of ecology really different from the *ought* of saving the drowning child? And is the "*fated*" of ecology completely different from the *cannot* of saving the child? Difficult to say without ethical and metaphysical investigations. Even Nolt admits that, for consequentialists, all permissible worlds must also be possible, thus excluding the very possibility of moral tragedies. Given the intuitiveness of Kant's principle, I prefer to reserve the notion of a *moral tragedy* for standard moral *dilemmas*, situations where an agent's obligations conflict. The Bad Bart case can then be formulated as:

| | |
|---|---|
| 1. OL | a |
| 2. OB | a |
| 3. □ (L ⊃ ~B) | a |
| 4. O~B | 1, 3 ROM |
| 5. O~B ⊃ ~O~~B | Dd' |
| 6. ~O~~B | 5, 6 MP |

The derivation of a contradiction (lines 2 and 6) uses **ROM** and a fairly uncontroversial deontic principle, **Dd'**. In a moral dilemma, (i) the agent is required to do each of two actions; (ii) the agent can do each of the actions; yet (iii) the agent cannot do both actions. No matter what she does, she will do something wrong. The possibility of moral dilemmas is automatically excluded by the axioms of SDL (**Dd'**), but I think they are inevitable in our moral life and normative practice in general.

## 4 Paradoxes, dilemmas and actualism

Deontic paradoxes "… were the booster rocket that provided the escape velocity deontic logic needed from subsumption under normal modal logics, thus solidifying deontic logic's status as a distinct branch of logic" (McNamara & Putte, 2022). Many solutions have been proposed, I prefer the modifications of modal catapults based on the insights of the logic of counterfactual conditionals (analysed in chapter 9, Šuster, 2023).

A system of modal logic is monotonic if it is closed under **RM** (Chellas, 1980, 234). Counterfactuals do not obey the principle of monotonicity – witness the failure of the principle of *Antecedent Strengthening*. If I had scratched the match, it would have lighted; but it is not true that if I had scratched and drenched the match, then it would (still) have lighted. According to the Stalnaker-Lewis possible worlds semantics, we test whether the conditional "If it were the case that A, then it would be the case that B" (or 'A > B') is true in a possible world *w* by considering the closest possible worlds to *w* where *A* is true. The conditional is then true in *w* just in case *B* is true at all in the closest possible worlds to *w* where *A* is true. Closeness is standardly explicated in terms of similarity: *B* is true throughout some class of A-worlds that beat all competitors in respect of how like the actual world *w* they are. The closest "scratched match" world is the one where the match lights. The closest scratched and drenched match world, which is required for the evaluation of the conditional with the strengthened conditional, is a *different* world, further away from actuality and the consequent is false in *that* world.

How does this help with obligations? Well, there is a counterfactual element in deontic modality – in considering whether it ought to be that *p* we should look to what would be the case in the closest *p*-world to the actual world. Recall the Good Samaritan. It ought to be that Jones helps Smith, who has been robbed. But "Smith is helped by Jones" entails that Smith has been robbed. Yet it seems false that it ought to be that Smith has been robbed. Suppose that Smith has actually been robbed. Then what makes it true that it ought to be that he is helped by Jones is that the closest "robbed & helped" world is *better* than the closest "robbed & not-helped"world. And this is consistent with the fact that what would have been the case had he *not* been robbed in the first place is better than what is actually the case. The closest "robbed" world is worse than the closest "not -robbed" world, so the

conclusion "It ought to be that Smith is robbed" is false, and the initial catapult is "broken."

Jackson (1985, also Goble, 1990) argues for *contrastivism* and *actualism* about deontic claims. Contrastivism is the thesis that ought-sentences have their truth-conditions relative to a class of alternatives. It ought to be that *p* is true just when *p* is better than the relevant alternative propositions (mutually exclusive but not necessarily jointly exhaustive) alternatives. But the implicitly suggested reference class of alternatives is *changing*. As with counterfactual antecedents: in considering whether it ought to be that *A* we should look particularly to what would be the case were *A* the case, and what would be the case were *A* the case might be importantly different from would be the case were *A* & *C* the case. Obligations are *relative*, they concern what ought to be out of a range of exclusive alternatives. It ought to be that A *out* of {A, $A^1$, $A^2$. . .} iff what would be the case were A true (the closest A-world) is better than what would be the case were $A^i$ true, for each *i* (Jackson 1985, 185). To simplify (Blumberg & Hawthorne, 2023, 86):

> OA is true in $w$ iff the closest A-world to $w$ is better than the closest ~A-world.

Actualist semantics that spells out the comparison of obligations in terms of similarity to the actual world explains the invalidity of ROM-like catapults. According to Jackson, the set of alternatives to which "ought" is relative can change at each stage in the conversation. The phenomenon is particularly clear in the so-called Sobel sequences of counterfactuals (Bennett, 2003, 160):

(1) If you had walked on the ice, it would have broken.
(2) If you had walked on the ice while leaning heavily on the extended arm of someone standing on the shore, the ice would have broken.
(3) If you had walked on the ice while leaning heavily on the extended arm of someone standing on the shore but slipped, the ice would have broken.

(1) and (3) are true and (2) is false, but we could easily continue the sequence of adding extra conditions to the counterfactual antecedent, leading to changes in the truth value. "A > B" is true since the closest A-world to the actual world *w* is a B-world. However, "(A & C) > B" is false because the closest "A & C" world differs

from the closest A-world. Jackson (1985) invokes this type of explanation in his account of the invalidity of ROM. Suppose Smith has actually been robbed and helped. Ought it be the case that he is robbed? The relevant alternatives are: {Smith is robbed, Smith is not robbed, Smith is robbed and helped, Smith is robbed and not helped}. The "not robbed" case is the best out of *this* set. Ought it be the case that he is robbed and helped? The relevant alternatives are now: {Smith is robbed and helped, Smith is robbed and not helped}. The "robbed and helped" case is the best in *this* set of alternatives. Once again, what would be the case were A true (the closest A-world) differs from what would be the case were both, A and C true.

Nevertheless, two tasks remain. How does this solution work in the case of moral dilemmas? Secondly, modal catapults are not so easy to dismiss. Utilising these principles, we may persuade moral agents that they are committed to the logical consequences of their moral principles. One should also be able to explain the reasonableness of this pattern! Let me start with the second task.

Many have noticed that the plausibility of strengthening the antecedent in the case of counterfactual conditionals can be restored after all. Consider the sequence above – when asserting the truth of (1), we *ignored* the possibility of leaning heavily on the extended arm of someone standing on the shore as irrelevant, but if this possibility is not ignored, both (1) and (2) will be false. We can generalise: relative to any given *fixed* set of alternative possibilities, (1), (2), and (3) have the same truth-value. A contextually variable *strict* conditionals analysis of counterfactuals was always an option (Lowe, 1990, 83):

$$A > B \text{ is true in a context } c \text{ in a world w iff } \Box_{cw} (A \supset B)$$

'$\Box_{Cw}$' is a necessity operator meaning: "In every possible world sufficiently similar to *w* as determined by the context *c*, it is true that ... ." Within a given context, relative to the same set of possibilities, any time 'A > B' is true, so is '(A & C) > B'. The moves in the Sobel sequence are marked by changes in the context *c*.

Stalnaker (1984, 125) was already aware that one could defend a strict conditional account of counterfactuals as an alternative to the variably strict account. The principal difference will then be in the demarcation between semantics and pragmatics, determining at what level of abstraction one's notion of validity is

defined. I address some of these issues (Šuster, 2023, chapter 2) in discussing Stalnaker's distinction between valid and reasonable inferences. I think we might also adopt the contextual approach as a solution of deontic paradoxes, which nevertheless respects the inferential potential of deontic closure principles. Consider a variation of Goble's example. According to laws in Slovenia, an identity card must be held by a citizen older than eighteen if he or she does not have another valid official identification document with a photograph issued by a public authority. Therefore, Ana, who is nineteen and does not have any other official identification document, is obliged to have an identity card. Given the general law, the particular is implied. Clearly, there are *no* contextual changes in this case. If *A* entails *B*, then if one ought to do *A,* then one ought to do *B*, provided we consider the *same* set of contextually relevant worlds. In the previous case of Smith, we first consider the case of his actually being robbed (he ought to be helped!), but in the conclusion, we do not envisage the *same* range of possibilities (he should not be robbed at all!).

However, a contextual move does not help in the case of moral dilemmas. It seems evident that there are situations where an agent's obligations conflict in the *same* context of relevant possibilities. There are situations in which some state of affairs both ought to be and ought not to be. For instance, I ought to help my friend even when this obligation is in conflict with the obligation to my community. Yet principles from deontic logic can be used to argue *against* the very existence of moral dilemmas. To simplify, take **Dd',** which immediately gives (via a plausible rule that logically equivalents are interchangeable): $\sim(O\varphi \,\&\, O\sim\varphi)$.

Does the analogy with the logic of counterfactuals help to solve this conundrum? Inconsistent "oughts" look like impossible antecedents. Recall the semantics of conditionals:

> A > B is true at *w* iff some (accessible) A and B-world is closer to w than any A and ~B-world, if there are any (accessible) A-worlds.

According to Lewis (1986, 18):

> If A is *impossible*, A > B is vacuously true regardless of the consequent B.

In deontic logic combined with actualism, we are now also supposing that among possible situations in which a particular proposition, that *A*, is true (false), some are closer to the actual case than others. In genuine dilemmas we seem to have both: the closest A-world to *w* is better than the closest ~A-world and the closest ~A-world to *w* is better than the closest A-world. Rather than accept all such obligations as vacuously true or introduce impossible worlds or lean on paraconsistent logic, I prefer to understand situations like these as lying *beyond* the scope of standard deontic logic. SDL is applicable only to domains in which it is presupposed that there are no such conflicts. Thus, I adopt an elegant solution proposed by Goble (2005, 2009) and restrict the scope of a catapult to *normal*, non-conflicted obligations (Šuster 2023, 117–121):

**ROMu.** From $\vdash \phi \supset \psi$ infer $\vdash P\phi \supset (O\phi \supset O\psi)$

**ROMu'.** From $P\phi$, $O\phi$ and $\Box(\phi \supset \psi)$ infer $O\psi$

If $\phi$ entails $\psi$, then if one ought to do $\phi$, then one ought to do $\psi$, provided that $\phi$ is *permitted* by the normative system. In other words, if $\phi$ is an unconflicted obligation and it entails $\psi$, then $\psi$ too is obligatory. The principles of deontic logic are modified to allow for the possibility of genuine normative conflicts, but we still keep the logical core of modal logic in "normal" situations (of course, we still have to be aware of contextual shifts and their impact on the validity of catapults). It seems to me that the inapplicability of normal modal logic and moral theory does not imply the end of rationality in some broader sense. Nagel (1979, 135) points us to Aristotle and practical wisdom, "which reveals itself over time in individual decisions rather than in the enunciation of general principles." However, I am aware that moral dilemmas might often be described as *deep disagreements* internalised within the agent. For example, I may feel torn between the loyalty I owe to a friend and my obligations towards the community. In such cases, I often suggest resorting to the more flexible tools of *informal* logic. The formal solutions I propose for addressing the failures of deontic closure – actualism, contextualism, and restrictionism – appear to be an inelegant patchwork. However, these issues are notoriously difficult, and nothing decisive has been proposed in the extensive discussion of the subject.

## 5 Conclusion

It may seem that I introduce and explore modal catapults solely to highlight their limitations as instruments for philosophical analysis. Why insist on normal modal logic and the general closure principles for modalities (necessity, ability, obligation, etc.)? Because modal catapults are, in one form or another, indispensable to the logic of modalities; discarding these principles is equivalent to relinquishing the entire framework of modal logic. Pollock, for instance, was well aware that the "active" necessitation conditional is of more interest to philosophers than the simple subjunctive. Nevertheless, he makes a crucial observation when he notes that the necessitation conditional satisfies virtually *no* logical laws (Pollock, 1984, 111 ).

Generally, at least from the perspective of standard logical approaches, there is little to discuss regarding a modal operator that is not closed under entailment.

Does this observation warrant a pessimistic view towards the endeavours of "hard-core" analytic philosophy? On the contrary, much like in other scientific fields, we develop new tools and explore new formal approaches (beyond the scope of the book in discussion). Moreover, any form of logic, whether normal or otherwise, cannot replace philosophical reflection. Formal logic is a catapult that propels our initial beliefs with modal principles, shaping the trajectory of their journey. It neither dictates the starting point nor the philosophical interpretations of the landing points.

**References**

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford University Press.

Blumberg, K. and Hawthorne, J. (2023). Inheritance: Professor Procrastinate and the Logic of Obligation. *Philosophy and Phenomenological Research, 106*(1): 84–106.

Brink, D. O. (1996). »Moral Conflict and Its Structure«. In Mason H. E. (ed.), *Moral Dilemmas and Moral Theory* (pp. 102–126). Oxford University Press.

Chellas, B. F. (1980). *Modal Logic: An Introduction*. Cambridge University Press.

Duncombe, M. (2024). Diodorus Cronus. In E. N. Zalta and U. Nodelman (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2024 Edition)*. The Metaphysics Lab, Stanford University. https://plato.stanford.edu/archives/sum2024/entries/diodorus-cronus/

Goble, L. (1990). A Logic of Good, Should, and Would: Part I. *Journal of Philosophical Logic, 19*(2), 169–99.

Goble, L. (2005). A logic for deontic dilemmas. *Journal of Applied Logic, 3*, 461–483.

Goble, L. (2009). Normative conflicts and the logic of 'ought'. *Noûs, 43*(3), 450–489.

Fischer, J. and Ravizza, M. (1996). Free will and the modal principle. *Philosophical Studies, 83*, 213–230.

Humberstone, L. (2005). Modality. In Jackson, F., Smith, M. (ed.), *The Oxford Handbook of Contemporary Philosophy* (pp. 534–614). Oxford University Press.

Jackson, F. (1985). On the Semantics and Logic of Obligation. *Mind, 94*, 177–195.

Kapitan, T. (2002). A Master Argument for Incompatibilism? In Kane, R. (ed.), *The Oxford Handbook of Free Will* 1st (pp. 127–157). Oxford University Press.

Kapitan, T. (2011). A Compatibilist Reply to the Consequence Argument. In Kane, R. (ed.), *The Oxford Handbook of Free Will* 2nd (pp. 131–150). Oxford University Press.

Lewis, D. (1986). *Philosophical Papers Vol. II*. Oxford University Press.

Lowe, E. J. (1990). Conditionals, Context, and Transitivity. *Analysis,* 50(2), 80–87.

MacFarlane, J. (2020). *Philosophical Logic: A Contemporary Introduction*. Routledge.

McNamara, P. & Putte, F. (2022). Deontic Logic. In E. N. Zalta and U. Nodelman (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2022 Edition)*. The Metaphysics Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/logic-deontic/

Morscher, E. (2002). The Definition of Moral Dilemmas: A Logical Confusion and a Clarification. *Ethical Theory and Moral Practice, 5*(4), 485–91.

Nagel, T. (1979). *Mortal Questions*. Cambridge University Press.

Nolt, J. (1997). *Logics*. Wadsworth Pub. Co.

Perry, J. (2004). Compatibilist options. In Campbell, J. K., O'Rourke, M. and Shier, D. (ed.), *Freedom and Determinism* (pp. 231–254). MIT Press.

Pollock, J. L. (1984). *The Foundations of Philosophical Semantics*. Princeton University Press.

Pruss, A. R. (2013). Incompatibilism proved. *Canadian Journal of Philosophy, 43*(4), 430–437.

Saunders, A. (8. 4. 2022). *Honourable intentions*. The Philosopher's Zone, ABC Radio National. https://www.abc.net.au/listen/programs/philosopherszone/honourable-intentions/3935166

Slote, M. (1982). Selective Necessity and the Free-Will Problem. *The Journal of Philosophy, 79*, 5–24.

Stalnaker, R. (1984). *Inquiry*. The MIT Press.

Šuster, D. (2021). Arguing about free will: Lewis and the consequence argument. *Croatian journal of philosophy*, *21*(63), 375–403.

Šuster, D. (2023). *Modalni Katapulti: Uvod v Filozofsko Logiko*. Univerza v Mariboru, Univerzitetna založba.

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

von Wright, G. H. (1979). The 'Master Argument' of Diodorus. In Saarinen, E. Hilpinen, R., Niiniluoto, I. and Hintikka, M. P. (ed.), *Essays in Honour of Jaakko Hintikka* (pp. 297–302). Springer.

Williamson, T. (2007). *The Philosophy of Philosophy*. Blackwell.

**ANALIZA**
Let. 28, št. 2, str. 209–228, december 2024

ANαliZA

# WHAT ONE CAN KNOW: FITCH'S ARGUMENT AND ITS CONSEQUENCES

**Accepted**
3. 3. 2024

**Revised**
6. 7. 2024

**Published**
31. 12. 2024

NENAD SMOKROVIĆ
University of Rijeka, Rijeka, Hrvaška
nenad.smokrovic@efri.hr

CORRESPONDING AUTHOR
nenad.smokrovic@efri.hr

**Abstract** The paper, motivated by the chapter in Šuster's book, considers the aspect of the so-called Fitch's argument (FA) that seriously challenges the verificationist theory. Contrary to Šuster's view, it is throughout the paper that I'm pursuing the idea that most of the attempts that intend to vindicate verificationism from the grip of Fitch's argument, including Edgington's theory, fail in their intention. Concerning the attempts to mitigate the effect of Fitch's argument to verificationism in the framework of classical logic (Eddington as the most important representative), I'm siding with their critics (Williamson, Percival) and claim that they fail in their intention. Regarding the attempts to block the effect of Fitch's argument in the framework of non-classical (intuitionistic, relevant, dialetheist, and so on) logic, they do it by introducing principles that invalidate some of the basic classical rules and principles, usually introducing trivial worlds. In that case, the verificationist principle (as well as all inferences included in Fitch's argument) is vacuously valid, which seems to be unsatisfactory. In any case, there is no decisive evidence that either classical or any of the non-classical approaches can avail the verificationist anything to escape out of the grip of FA.

**Keywords**
verificationism,
anti-realism,
Fitch's paradox,
classical logic,
non-classical logic

https://doi.org/10.18690/analiza.28.2.209-228.2024


University of Maribor Press

AN$_{\alpha}$LIZA

**ANALIZA**
Vol. 28, no. 2, pp. 209–228, December 2024

# KAJ LAHKO VEMO: FITCHEV ARGUMENT IN NJEGOVE POSLEDICE

NENAD SMOKROVIĆ
Univerza v Reki, Reka, Hrvaška
nenad.smokrovic@efri.hr

DOPISNI AVTOR
nenad.smokrovic@efri.hr

**Sprejeto**
3. 3. 2024

**Pregledano**
6. 7. 2024

**Izdano**
31. 12. 2024

**Izvleček** Članek, ki je motiviran na osnovi poglavja v Šusterjevi knjigi, obravnava vidik tako imenovanega Fitchevega argumenta (FA), ki predstavlja resen izziv za verifikacijsko teorijo. V nasprotju s Šusterjevim stališčem v celotnem prispevku zasledujem stališče, da večina poskusov, ki nameravajo verifikacionizem ubraniti izpod primeža Fitchevega argumenta, vključno z Edgingtonovo teorijo, ne uspe. Kar zadeva poskuse omilitve posledic Fitchovega argumenta za verifikacionizem v okviru klasične logike (Eddington kot najpomembnejši predstavnik), se postavljam na stran njihovih kritikov (Williamson, Percival) in trdim, da jim njihova namera ni uspela. Kar zadeva poskuse blokiranja vpliva Fitchevega argumenta v okviru neklasičnih (intuicionističnih, relevantnih, dialetheističnih in tako naprej) logik, to počnejo z uvajanjem načel, ki razveljavljajo nekatera osnovna klasična pravila in principe, pri čemer običajno uvajajo trivialne svetove. V tem primeru je verifikacionistično načelo (kot tudi vsi sklepi, vključeni v Fitchev argument) prazno resnično, kar ni zadovoljivo. V vsakem primeru pa ni prepričljivih dokazov, da bi bodisi klasični bodisi kateri koli od neklasičnih pristopov verifikacionistu lahko pomagal, da bi se rešil iz primeža FA.

**Ključne besede**
verifikacizem,
antirealizem,
Fitchev paradoks,
klasična logika,
neklasična logika

https://doi.org/10.18690/analiza.28.2.209-228.2024


Univerzitetna založba
Univerze v Mariboru

## 1 Introduction

In his excellent book *Modal Catapults* (Šuster, 2023), which, regarding the modal logic, I consider to be the most valuable publication in the area, Danilo Šuster devotes considerable attention to *Fitch's argument* (Fitch, 1963, also called *Church-Fitch's paradox*). The importance of Fitch's argument (FA) lies in the fact that it seriously challenges the verificationist theory, a distinguished form of anti-realism. Fitch's argument is, therefore, at the very centre of the fierce discussion between verificationists and their opponents, the realists. Though Šuster's book addresses many other interesting topics, this paper focuses only on this issue. Šuster provides an elegant and clear presentation of the argument and, when providing the interpretation of attempts to make the argument less harmful for the verificationist position, he focuses on Edgington's proposal (Edgington, 1985). Šuster himself (Šuster, 2023, 92) seems to be inclined to attempts that intend to rescue verificationism from the grip of Fitch's argument, particularly Edgington's solution. He holds (Šuster, 2023, 92) that "it [Edgington's solution] is the closest and most in the spirit of the semantics of modal logic". I agree with this qualification, and I also accept that among various attempts tending to save verificationism, Edgington's solution is "most in the spirit of semantics of modal logic" in the framework of classical logic. However, many other attempts to vindicate verificationism are trying to find their way into the framework of non-classical, non-standard logic. Nonetheless, regarding all these attempts, it is throughout the paper that I'm pursuing the idea that most of them, including Edgington's theory, either fail[1] in their intention to vindicate verificationism, or, intending to block the effect of Fitch's argument in the framework of non-classical (intuitionistic, relevant, but most often para-consistent and para-complete) logics, do so by introducing principles that invalidate some of the basic classical rules and principles. In any case, there is no decisive evidence that any of the non-classical approaches can help verificationists escape the grip of FA.

[1]It is interesting that some of the leading logicians suggest different non-classical logics as the only way for blocking FA. For instance, Williamson claims, "How else might a verificationist escape from Fitch's argument? One way would be to substitute intuitionistic for classical logic. It may even be the only way" (Williamson, 1987, 261). Priest claims something quite different, "We have seen that Fitch's argument may be blocked by an appeal to dialetheism. Moreover, it is the *only way* [my italic] that we have found in which the argument may be blocked (In Salerno, 2009, 100–101). However, I'm suggesting that there is no plausible way to vindicate verificationism.

However, before entering the topic more deeply, a few notes concerning the general importance of the argument are required. What is nowadays known as Fitch's argument[2] (FA) is also referred to as the knowability paradox by many authors. In this form, it is an unavoidable topic in contemporary formal epistemology, particularly in the context of what, in principle, one can know. The key question in this regard is: *are all true propositions knowable*? Answering this question, realists and anti-realists in epistemology determine their opposing positions. Realists, holding that humans are non-omniscient, answer negatively. On the other hand, anti-realists answer positively: if something is true then it can be known. Let us call this the *knowability thesis* (KT). Formally:

(KT) $\forall\varphi\ (\varphi \supset \Diamond K\varphi)$.

A much stronger, but unreasonable, even *silly*[3] form of verificationism is that all truths are in fact known:

(SV) $\forall\varphi\ (\varphi \supset K\varphi)$.

Both variants of verificationism should endorse omniscience concerning the knower. The alternative is, quite generally, that there are no omniscient agents and that for any agent S there is the truth she does not know (not all truths are known) and that possibly no one will ever know. Let us call this claim *non-omniscient*. Formally:

(Non-omniscient) $\exists\varphi\ (\varphi \wedge \sim K\varphi)$.

The (KT), $\forall\varphi\ (\varphi \supset \Diamond K\varphi)$, is at the heart of verificationism. Furthermore, it is often considered to be the *quintessential implication* of semantic anti-realism[4] (Kvanvig, 2006, 56).

---

[2] What I call Fitch's argument is usually referred to in the literature as Fitch's or Church-Fitch's paradox. Some authors, however, deny that the argument is paradoxical (see Williamson, 2000), claiming that it is only surprising. To avoid this discussion, I prefer a more neutral term, Fitch's argument.

[3] Williamson (2000, 272) illustrates the "silliness" by inviting us to imagine a situation in which his office contained either an even number of books at some time *t* in the past or not. Nobody knows, as a matter of contingent truth. Thus, either it is an unknown truth that it was an even number of books at *t*, or it is an unknown truth that it was an odd number. Either way, there is an unknown truth and strong verificationism is false.

[4] The knowability thesis follows immediately from the verificationist's claim that a proposition $\varphi$ is true if and only if it is possible to prove (or verify) $\varphi$. If it is possible to prove $\varphi$, it is possible to know that $\varphi$.

As mentioned, verificationists should accept the positive answer to the question, "are all truths knowable?". In short, verificationism constrains truth epistemically. It equates truth (or meaningfulness) with a cognizer's cognitive ability, an ability to know, to believe, to verify, to confirm. Realism is, on the other hand, ontologically committed. Some object (a real number, for example) exists independently of our cognitive ability to grasp them. This independence of reality from our cognitive grasping limits our knowledge in the sense that there are unknown, even unknowable propositions. We are non-omniscient both in the mathematical realm (Gödel's undecidability) as well as in the realm of contingent, non-mathematical propositions. Accordingly, realism is bound to hold that some truths are not known and some of them possibly not knowable.

There are various reasons for accepting anti-realism. Besides well-founded motivations in the history of philosophy, issuing either from the intention to avoid scepticism (from Berkeley to American pragmatists) or from the "meaning as use" doctrine (Wittgenstein, more recently Dummett), the optimistic idea that all truths are knowable, at least in principle, is one of the tenets deeply entrenched in the modern scientific worldview, which is rarely questioned. It is widely accepted by physicalists (to be more precise, the knowability thesis fits particularly well with materialists or methodological physicalists, as Kvanvig named them (2006, 43–47). The modern idea of verificationism goes back to American pragmatists and logical empiricists. C. S. Pierce, for instance, claimed that truth is what the scientific community would agree on in the long run. For logical empiricists, truth (or meaning) is tied to what we are capable of verifying. In the long run, the proponents of the modern worldview firmly believe that all truths will be known and that the important problems will be ultimately solved.

The optimistic standpoint is, of course, worth holding, but its epistemic background is doubtful, at least according to FA. The moderate verificationism (expressed as KT) gets in trouble precisely because FA presents a relatively simple proof, given in a few lines, arguing that verificationism is inconsistent. This is certainly bad news for the anti-realists' optimism. Namely, the claims that all truths are knowable is provably equivalent to the omniscient-like claim that *all truths are known*. Let us call it the knowability principle (KP). We can put it formally:

(KP) $(\varphi \supset \Diamond K\varphi) \vdash (\varphi \supset K\varphi)$.

The claim on the left of the formula (KT), or, as Kvanvig notes, "the quintessential implication of the semantic anti-realism", is equivalent to the problematic claim on the right side of the formula that all truths are, in fact, known (Williamson (2000) calls this claim strong verificationism). This claim seems unreasonable, even silly.

This devastating effect of FA, of course, puts convinced anti-realists (verificationists) on high alert, but also unsettles those who feel inclined to the appeal of realism, however, at the same time hold that the optimistic worldview is worthy enough to be vindicated[5]. Therefore, it is not surprising that van Bentham (2004, 105) says, "Much of the literature on Fitch's Paradox seems concerned with averting a disaster, and saving as large a chunk of verificationism". Faced with this peril, a verificationist has a choice. She might try to save as much of verificationism as possible by formulating the counter-argument either in the framework of classical[6] (CL) or in the frame of the non-classical logical (NCL) systems. A suitable tool seems to be intuitionistic logic. Namely, as Williamson claims (1987), KP can hardly be refuted by means of classical logic, but with the extended operators in intuitionistic logic, it might be possible. However, in this paper, I claim that despite numerous endeavours, observing them generally, various projects of saving verificationism expressed in the slogan "what is true can be knowable", do not succeed. It is arguably so concerning vindications as expressed in the frame of classical logic, whereas when expressed in non-classical (para-complete and/or para-consistent) logics, they can block the effect of FA, but at the price of postulating trivial worlds. This is certainly unacceptable for ones who endorse classical logic. However, it might also be too costly for a verificationist who accepts a non-classical logic.

In what follows, FA is going to be briefly presented. This paragraph is followed by an overview of different strategies that aim to block, in a way, the threat of FA for anti-realism. There are numerous attempts to refute or at least minimize the effect of FA on verificationism. I'm proposing to divide them into two main groups of strategies, the *restrictionist* and the *revisionist* ones. The restrictionist strategy, exemplified by its most important representative, Dorothy Eddington, is presented at length and in more detail, while the revisionist ones are briefly indicated.

[5] This uncomfortable situation is nicely put forward by Šuster: "Although realism is close to me, I nevertheless think that moderate anti-realism, which represents a complex set of concepts and belongs to the historical treasury of philosophical ideas, goes with a developing conceptual complex, and it is hard to believe that it will be buried by a few lines of normal modal logic" (Šuster, 2023, 92, translated from Slovenian by the author).

[6] By classical logic I mean the propositional and predicate calculus, but also normal modal and epistemic logic.

## 2 Fitch's Argument

Let us start with the informal characterization of the argument. Let us take that every true proposition is knowable, and suppose, for reductio, that there is a proposition, $\varphi$, which is true but not known, $\varphi \wedge \sim K\varphi$. Then it must be possible to know $\Diamond K (\varphi \wedge \sim K\varphi)$. It is now easy to prove that it is possible to both know $\varphi$ and not know it, $\Diamond(K\varphi \wedge \sim K\varphi)$, which is a contradiction (see [13]).

Before presenting the formal proof, let us recall the two variants of verificationism mentioned above, both supposing the *knowing agent to be omniscient*. One of them is the knowability thesis, a reasonable assumption, that all true propositions can be known by someone sometime:

1. $\forall\varphi\ (\varphi \supset^{7} \Diamond K\varphi)$, *knowability thesis (KT).*

The other, less reasonable, if reasonable at all, is that all truths are (actually) known. This strong form is obviously false and unacceptable, similar to Berkeley's solipsism (*esse est percipi*):

2. $\forall\varphi\ (\varphi \supset K\varphi)$, *(SV) or "silly".*

In opposition to these assumptions is the realistic, non-omniscient idea:

3. $\exists\varphi(\varphi \wedge \sim K\varphi)$.

The various forms of proof proceed by employing additional assumptions. Omitting the quantifiers, we can group them as follows and formalize them accordingly:

*Epistemic rules*

(*Fact):* $K\varphi \supset \varphi$, saying that *knowledge is necessarily factive.*
(Dist): $K(\varphi \wedge \psi) \supset (K\varphi \wedge K\varphi)$, *distribution over conjunction.*

[7] Wherever the argument is presented in the setting of classical logic, I am using the "horseshoe", indicating material implication. Always when the argument is treated non-classically, the arrow sign (→) is used.

*Aletheic modal rules*

*(LNC):* $\sim\Diamond(\varphi \wedge \sim\varphi)$[8], a *law of noncontradiction.*
*(Close):* $(\Diamond\varphi \wedge (\varphi \supset \psi)) \supset \Diamond\psi$, modal formulation of *closure principle.*

*The rule of inference is*

$\vdash\varphi \supset \Box\varphi$, *Rule of Necessitation.*

The proof,[9] briefly presented, proceeds as follows. It starts with two contrasting propositions, *KT* ($\varphi \supset \Diamond K\varphi$) and (unknown) $\varphi \wedge \sim K\varphi$. Substituting the (unknown) in (KT) as the value of $\varphi$, we get, by modus ponens:

1) $\Diamond K (\varphi \wedge \sim K\varphi)$.

Assuming $K(\varphi \wedge \sim K\varphi)$ and applying (*dist*) over conjunction we get:

2) $K\varphi \wedge K\sim K\varphi$.

By application of the *Fact* to the second conjunct of 2), it yields:

3) $K\varphi \wedge \sim K\varphi$. By *reductio:*
4) $\sim K(\varphi \wedge \sim K\varphi)$.

Application of the Rule of Necessitation, $\vdash\varphi$ / $\Box\varphi$, yields:

5) $\Box\sim K (\varphi \wedge \sim K\varphi)$.

---

[8] Given that the dual of $\Box\sim\varphi$ is $\sim\Diamond\varphi$, we have: $\vdash \sim\varphi \vdash \sim\Diamond\varphi$.

[9] In presenting FA, I am following Kvanvig's simple and elegant formalization. See slightly different formalization in Wansig's (2002):

1) $\varphi\wedge\sim K\varphi$ assumption
2) $\Diamond K(\varphi \wedge \sim K\varphi)$ 1, WV
3) $\Diamond(K\varphi \wedge K\sim K\varphi)$ 2, Dist
4) $\Diamond(K\varphi \wedge \sim K\varphi)$ 3, A3
5) $\sim\Diamond(K\varphi \wedge \sim K\varphi)$ A3
6) $\sim(\varphi \wedge \sim K\varphi)$ 1, 4, 5

Expressing $\Box\sim \varphi$ by its dual, $\sim\Diamond \varphi$, we eventually have:

6) $\sim\Diamond K (\varphi \wedge \sim K\varphi)$

Line 6 contradicts line 1. The verificationists must deny that there are truths we do not know (that we are non-omniscient), which leaves us with:

7) $\sim\exists\varphi (\varphi \wedge \sim K\varphi)$. The conclusion is that all truths are actually known:

8) $\forall\varphi (\varphi \supset K\varphi)$.

Showing that KT collapses to SV, FA brings verificationism into trouble. But does this result conclusively refute any possible vindication of verificationism? Is it fatal for verificationism? Many think it is not. As Kvanvig says, there is a long way from Fitch's argument to the refutation of any form of verificationism, and, expectedly, many philosophers have sought to free verificationism from the commitment to KP $(\varphi \rightarrow \Diamond K\varphi) \vdash (\varphi \rightarrow K\varphi)$ in order to avoid a refutation of anti-realism by FA.

## 3 Can verificationism be vindicated?

To provide an overview of the recent discussion concerning possible avoidances of the effect of AF, as well as their critics, on verificationism, let me utilize C. Jenkins (inPriest, 2009, 304) who offers a succinct formulation of the given problem in three questions:

i) Does Church–Fitch's argument really refute global anti-realism?

ii) If it does not, is this because the argument is fallacious, or because anti-realists are not in fact committed to KT [10]?

iii) If anti-realists are not committed to KT, how should their doctrine of epistemic accessibility[11] be expressed?

---

[10] KT ($\varphi \supset \Diamond K\varphi$) Jenkins calls WVER, weak verificationism, following Williamson.
[11] By the doctrine of *epistemic accessibility,* Jenkins (in Salerno, 2009, 302) means the ani-realist idea that "because of its mind-dependent nature, all of reality is epistemically accessible to us".

To answer the question negatively, *i)* would obviously be too hasty, so we are going to assume a positive answer, namely that FA really refutes verificationism. In this case, the presentation of the problem regarding ii) might proceed either to claim a) that the proof procedure in FA is erroneous or to come back to b) claiming that verificationism is not, in fact, committed to KT ($\varphi \supset \Diamond K\varphi$). Concerning *iia,* we agree with Kvanvig, who claims that "the logic of the paradox is not in any simple way problematic" (2006, 14). However, to inspect more closely where the proof actually might go wrong, one can locate it either in the very steps of the proof or one can cast doubt on the additional assumptions *(epistemic, alethic-modal rules and the rule of necessitation)*. The steps of the proof are doubtlessly correct. Concerning the *assumptions*, it turns out that it is (dist) *the distribution over conjunction* that some authors indicate as a weak point. Williamson (2000), for instance, has claimed that knowledge need not be distributed over conjunction, but as a full-blooded realist he convincingly claims that this can hardly help the verificationist in avoiding FA because "the anti-verificationist argument can be reconstructed in at least two ways[12]. The verificationist cannot escape by denying distribution" (2000, 84–85).

This leaves us with the second disjunct of *ii)*, that verificationist (anti-realist) is not, in fact, committed to KT. This being the case, the answer to question *iii)* requires a kind of taxonomy of various strategies that hope to resist the threat of FA. A rough taxonomy that more or less corresponds to what most of the authors (see Brogaard & Salerno, 2019, 2002; Jankins, 2009; Kvanvig, 2006) propose is to divide various attempts into two groups. The widely accepted division is on *restriction* and *revision* strategies. According to my understanding of the taxonomy, *restriction* strategies mostly accept classical logic (CL) as a framework for dealing with FA, while the very strategy consists of restricting the scope of the universal quantifier in: *for all x, x can be known.* A bit more specifically, (Jankins, 2009, 305), "not all true propositions are supposed by anti-realist to be knowable, but only some". In this paper, considerable attention will be paid to Edgington's proposal, which suggests that only *actual* truths are knowable. I'm going to show, relying on the criticism of the thesis (Williamson, 1987; Percival, 1990) the proposal does not succeed.

[12] Relying on two variants of verificationism, the weak ($p \rightarrow \Diamond Kp$) and the strong one ($p \rightarrow Kp$), he (Williamson, 1993, 84) claims that one can reconstruct the anti-verificationist argument without relying on distribution, either by a) arguing that verificationists are committed to something stronger than ($p \rightarrow \Diamond Kp$) or by b) deducing something weaker than ($p \rightarrow Kp$) from ($p \rightarrow \Diamond Kp$).

Revisionist strategies, on the other hand, concern the question of whether the "proper" logic of knowability is the classical logic and, if not, whether the substitution of classical logic with some non-classical logic (NCL) can, by invalidating Fitch's reasoning, help the verificationist to vindicate her standpoint? Given that FA stands or falls with the logical principles we referred to as the *additional assumptions* (epistemic, alethic-modal rules and the rule of necessitation) that are rendered as valid in the framework of CL, while some of them (or all) are invalidated in different forms of NCL, the revisionist typically considers CL as inappropriate for avoiding FA. The NLC logics proposals are of the paracomplete or the paraconsistent kind. Concerning the intuitionistic logic, in which several important verificationist attempts (Dummett, Tennent) are grounded, it should be noted that it is counted as paracomplete, holding that p and ~p can both be false. It invalidates the law of excluded middle (LEM: $\varphi \vee \sim\varphi$). The question is whether substituting CL with NCL can help verificationism in vindicating its standpoint: *all truths are knowable.*

## 4 D. Edgington: Trans-worlds knowability

In the restrictionist camp, Dorothy Edgington (1985) ingeniously proposed a solution to the devastating effect of FA, offering a modification of the verification principle, ($\varphi \supset \Diamond K\varphi$). Accepting that in the present setting, KP is inconsistent with ($\varphi \wedge \sim K\varphi$), she introduced a variant reading of the claim "every truth is knowable", arguing that, under a *suitable interpretation*, the assumption that all actual truths are knowable, ($\forall\varphi$) ($\varphi \supset \Diamond K\varphi$), and the assumption that some actual truth is not known, ($\exists\varphi$) ($\varphi \wedge \sim K\varphi$), are consistent. The crux in Edgington's endeavour is to make the distinction between the situation in which *one knows* and the situation *one knows about*. In light of this twist, as Williamson observes, the trouble with KP "is that it conflates the situation *s* in which *p* with the situation *s*' in which it is known that, *in s, p*" (Williamson, 1987, 256). To offer a brief exposition of Edgington's understanding of *suitable interpretation*, a few preliminary steps are needed. It should first be noted that, instead of possible worlds, she speaks of the concept of "possible situations," which are close enough but less specific than the "possible worlds" concept. Next, to discriminate the situation in which *one knows* from the situation in which *one knows about*, she introduces a new, tense operator S. Applied to the contradiction ($\varphi \wedge \sim K\varphi$), the new form, SK ($\varphi \wedge \sim K\varphi$), is introduced, where S means "it will be or is or was the case that." The main idea can now be presented.

Expressed in the notation of *possible situations* (let's call this *quantifying context,* QC), the above claim can be formulated as:

(QC) $\forall s$ ((in $s$, $p$) $\rightarrow \exists s'$ (in $s'$, K(in $s$, $p$)),

meaning that, for every situation $s$, if $p$ is true in $s$, it is known in $s'$ that $p$ is true in $s$, (Edgington, 1985, 367). Now, there is no reason "why it should not be known in $s'$ that in $s$, it is unknown truth that $p$" (as Williamson formulates this in 1987, 256).

However, the situation is not as simple as that. There is a problem arising in this proposal and it concerns the relationship between knowledge in two kinds of situations, in situation $s'$, "from" which one observes the truth, and situation $s$ where the observed truth is located. Namely, the knowledge in situation $s'$ and in situation $s$ must be the same knowledge, which apparently is not the case.[13] Attempting to solve the problem, Edgington seems to see a way out by introducing a temporal analogy to QC, where instead of evaluating sentences at situations ($s$ and $s'$), tense sentences should be evaluated at time points t and t'. Let us call this the *temporal context*:

(TC) For every $p$, if $p$ is true at $t$, then ($\exists$t') (someone knows at $t$ 'that $p$ is true at $t$).

This formulation, hopefully, makes it easier to establish the relationship between knowledge contexts in the temporal analogy than is the case in the quantified, QC variant. This being the case, the claim goes like this: if the relationship between, for instance, the thought "It was raining" expressed at seven o'clock and the thought "It is raining" expressed at six o'clock can be established (in TC), then it can be established for the relationship between the situations $s'$ and $s$, in QC formulation.

To accomplish this brief review of Edgington's proposal, the above quantification over situations should be reconciled with the original *modal version* of KP, ($\forall\varphi$ ($\varphi \supset \Diamond K\varphi$), the principle she wants to vindicate. To do that, Edgington equates the

[13] The only characteristic of the concept of knowledge we need at this point is its standard realistic account, namely that knowledge is factual. That means that an agent $s$ knows that $p$ only if $p$. Accordingly, two persons have the same factual knowledge iff they know the same factual contents, and their content-bearers (thought, expressed sentence) express the same factual content iff the content-bearers have the same truth-conditions. This is the case iff they would be made true by the same fact where they are both true. It is obvious that in our situations, $s'$ and $s$ knowledge can not be the same because the facts known are different (compare Percival, 1991)

situation *s* (in QC, or time point *t*, in TC), the situation in which one knows that *p,* with the *actual* situation designated with the operator A. In terms of possible worlds, it is now consistent with the claim that someone in some other world (call it *non-actual* world *w*) knows that in the *actual* world, $w_A$ one knows a proposition. In formal notation, we get: $\Diamond KA\ (\varphi \wedge \sim K\varphi)$, meaning that there is some world *w* in which it is known that it is true in the actual world $w_A$ that $\varphi$ is true but not known (compare Kvanvig, 2006, 57). Finally, it is suggested that the contradiction can be resolved by:

(A) $A\varphi \supset \Diamond KA\varphi$. This seems to be consistent. But is it so?

However, there are several convincing criticisms concerning Edgington's argument, we will pursue two of them. Williamson's (1987) and Percival's (1991) criticisms seem to be particularly compelling. Williamson's criticism, in fact, identifies three weak points in Edgington's proposal. The first one has to do with the difference between necessary and contingent propositions. Both are supposed to be known (by the agent) in most versions of verificationism, but it is dubious whether it is so in Edgington's proposal. The second issue deals with the problem of identification of knowledge across worlds (or situations). For the formulation in (A) to work, the knower should have the same content of knowledge at the actual and at the non-actual situation (world). However, this seems to be problematic as well. The final point aims to challenge the supposed analogy between the temporal knowledge in TC (knowledge identity across time) and the knowledge in modal contexts (knowledge identity across worlds) in which the proposal is formulated. The two last points are common targets to both, Williamson's and Percival's criticism.

a) The crucial question for the first critical point refers to the kind of propositions that are supposed to be knowable according to the same (A), $A\varphi \supset \Diamond KA\varphi$. According to the original, *knowability principle* ($\varphi \supset \Diamond K\varphi$), the obvious answer is that such propositions are to be *contingent* as well as *necessary* ones. However, Edgington's defence of verificationism by transforming ($\varphi \supset \Diamond K\varphi$) to ($A\varphi \supset \Diamond KA\varphi$) represents, as Williamson (1987, 257) claims, a "surprisingly weak form of verificationism". Namely, $A\varphi$ is true at *s* (or at *w* in the modal variant) iff $\varphi$ is true in the actual situation (world). To be true in the actual, $\varphi$ has to be true in all accessible situations (worlds). Therefore, $A\varphi$ and $\Box A\varphi$ are equivalent. Thus, Williamson (1987, 258) claims, "In *s*, *p*', (for most values for *s* and *p*) is therefore

*necessarily true*, if true at all". Therefore, it follows that this variant of verificationism is committed only to the *knowability of necessary truths*. This being the case, verificationism would not be in a position to require any *contingent* truth to be knowable[14].

b) It has been mentioned that the crucial idea of Edgington's proposal is that all truths can be known only if a truth is known "from" a situation other than that in which it is located. The schema (A) claims that it is a *non-actual* situation (world) from where a proposition *p* in an *actual* situation is known. The difference between situations is crucial because without determining the relationship between actual and non-actual knowledge, the scheme (A) would be in danger of collapsing "into the obviously silly schema $A\varphi \rightarrow AKA\varphi$" (Williamson, 1987, 260), in which case Edgington's solution would be back to the initial problem. Having established the difference between situations, a clear account for the *relationship* between thoughts (taking it to be a content-bearer of knowledge) in the non-actual and the actual situation should be characterized. Even more, the defence of the schema (A), $A\varphi \rightarrow \Diamond KA\varphi$, depends on giving such an account. But that is exactly where the problem lies.

Let the truth *p* in our example be "the earthquake of the low intensity happens in a situation *s* and no one knows that". In terms of QC, it can be known "from" the non-actual situation *s'*. The formulation QC requires that the truth that *p* expresses in *s* is the same as the truth which "In *s*, *p*" expresses in *s'*". The problem is to

[14]I am grateful to an anonymous referee who drew my attention to the Rabinowicz's and Segerberg's (1994) paper proposing a response to the criticisms of Edgington's version of verificationism. Due to the space limitations, I will only briefly present their views. Particularly, my focus is on whether their proposed response threatens Williamson's first objection to Edgington's proposal. As is claimed, Williamson (1987) argues that Edgington s variant of verificationism is committed to the knowability of necessary truths only. Namely, according to the standard truth-conditions for actuality, $A\varphi$ is true at w iff it is true at all actual worlds. Therefore, actuality is equivalent to necessity, $A\varphi \leftrightarrow \Box A\varphi$. Rabinowicz and Segerberg recognize the seriousness of the problem and admit that it cannot be solved with standard semantics' resources. The source of the problem, they claim, lies in the inability of standard semantics (standard truth-conditions), used by Edgington "to mix the actuality operator and the epistemic operator" (1994, 104). In standard semantics, the perspective from which one knows that p (actual world) is true in the non-actual world has been considered as fixed. Rabinowicz and Segerberg have proposed a solution that consists in the introduction of utterly new semantics, such that the standard semantics with the fixed actual world is replaced with the two-dimensional, variable-perspective semantics (1994, 104). In the two-dimensional semantics, "a formula is being evaluated not just at one point, *v*, but at an ordered pair of points, (*w, v*), with w being the point of perspective and *v* the point of reference." [1994, 104]. However, the implementation of new semantics, variable-perspective ones, changes in a considerable manner the meaning of the necessity operator, as well as the actuality and knowability operator. As a consequence, it is only under *this interpretation* that the solution to FA (the denial of the knowability principle) seems to be plausible. Accordingly, Rabinowicz's and Segerberg's analysis cannot be considered as a treat to Williamson's criticism because it is entirely situated in standard semantics.

determine what counts as the same knowledge in respectively different situations (or worlds). Percival (1990) calls this problem the "knowledge-identity across contexts". Edgington, being, of course, aware of this, admits that if, for instance, an agent A had non-actually had a thought, expressed in words "it is actually the case that *p*", A would not have been expressing the thought of the requisite kind, since his use of "actually" would refer rigidly to his own situation, not to B's. Without going much deeper into it, Edgington hopes that it can be resolved if a) the analogy between the modal (expressed either in terms of possible situations or in worlds) and temporal formulation (TC) can be established and b) if the account of knowledge between non-actual and actual situation can be given in the temporal context, it is also justified in the modal context. The criticism in both Williamson (1987) and Percival (1991) admit that the analogy between the modal and the temporal can be established. But even if that is the case, the suitable relationship between knowledge in timepoints *t* and *t* 'can be claimed only if there is a causal relation between knowledge in *t* and *t*'. However, the same relationship does not hold in the modal context, and therefore, the argument that (A) can refute KP fails.

The brief reconstruction of Edgington's requirement a) can go like this: the requisite relationship between non-actual and actual can be re-established in the frame of the temporal context (TC), in terms of "now" (the temporal analogy of the non-actual situation; for the sake of example, at seven o'clock) and "then" (actual situation; at six o'clock). Although agent A, at seven o'clock, would not be able to express the thought "it is now raining (six o'clock)", since A's use of the word "now" will refer to seven o'clock and not to six o'clock, the relationship between the thought at seven o'clock about the thought at six o'clock can be saved if two thoughts were casually connected. Williamson (1987) admits that there can be such a connection and specifies the possible connection in this way: "The only apparent way in which one might think at seven o'clock the thought that one could have expressed at six o'clock in the words 'It is now raining' is by remembering six o'clock" (1987, 257). But, he continues, "Though it might be that the causal link of remembering can work in temporal context, there is "no causal link between the non-actual and the actual" (1987, 258). Williamson's point is that "there is an insuperable fact that this analogy does not work in the case of "actual" and "non-actual" (1987, 258). A similar point is made by Kvanvig (2006, 57–58):

> "[I]t is questionable whether it is possible to entertain any proposition about some singular, individual world not identical to the world one is in. Any way of describing the world will apply equally to a number of different worlds, and any sort of different reference to other worlds would seem to be impossible in virtue of lack of any causal link between possible worlds."

In addition to this, Percival (1991) gives a sophisticated and long exposition of why it does not work. We will quote at this point only his indication of why Edgington's endeavour does not succeed:

> "I think she is misled by the formal similarities between standard semantics for tense- and modal-logics. But in reasoning analogically from the temporal to the modal case she does emphasize that the knowability principle and its temporal analogue, the thesis that all present truths were, are, or will be known, cannot be assessed without establishing what counts as the same knowledge in, respectively, different worlds and times." (Percival, 1991, 82–3)[15]

It is reasonable to conclude this section of the paper by emphasizing that Edgington's, I would say a heroic attempt to vindicate verificationism against FA in the framework of classical logic, did not succeed.

**Revisionist strategies**

The family of *revisionist* strategies concerns primarily the problem of the proper logic for the knowability paradox. According to my classification, what is meant by *revisionist strategies* are those attempts aiming to undermine FA, which replace classical logic (CL) with some of the non-classical logics (NCL). The line of demarcation between revisionist and restrictionist strategies concerns, therefore, a kind of logic supposed to be suitable for weakening the FA effect on verificationism. To avoid any misunderstanding, it should be noted that a particular revisionist strategy may

[15] As a response to criticisms (particularly concerning Williamson s problem b)), Edgington published a paper (2010) in which she strengthened the arguments made in the original paper (1985) and offered new ones. She further develops her argument for counterfactual knowledge (in 1985, 563). Williamson s answer to Edgington s response followed in 2021. Concerning the argument from counterfactual knowledge, which, according to Edgington, blocks FA, Williamson argues that the argument is subject to trivialization. To avoid trivialization, Williamson claims, Edgington needs to offer a more general account of how the knower is allowed to specify a counterfactual situation for the purposes of her argument, and it is unclear how to do so.

be restrictionist in character in the sense that it proposes the restriction of the scope of the knowledge operator, but making it in terms of NCL. The restrictionist strategy, on the contrary, restricts the scope of the knowledge operator, making it in the terms of CL.

The NCL that we are interested in are either paracomplete or paraconsistent,[16] or both. Among those non-standard logics, the *intuitionistic modal logic* (IML) is particularly suitable for most kinds of anti-realism,[17] exceptionally so in Dummett's[18] (1977) interpretation. As Percival (1990) notes, "Dummett has repeatedly argued, on grounds independent of Fitch's proof, that anti-realism warrants intuitionistic rather than classical logic and an assertability-conditional rather than a truth-conditional theory of meaning." (1990, 182–183). One of the prominent differences between classical modal logic and IML is the reinterpretation of the operator K (which in CML means "it is known that") as "it is verified that". Adding the possibility operator $\Diamond$, the idea is expressible (in Williamson, 1992) as:

> $\varphi$ *iff it is possible that it is verified that* $\varphi$.

This general intuitionistic formulation, applied to the context of FA, substitutes the knowability thesis, $\varphi \supset \Diamond K\varphi$, becoming $\vdash \varphi \leftrightarrow \Diamond K\varphi$ (expressing exactly: $\varphi$ iff it is possible that it is verified that $\varphi$). In this interpretation, however, the possibility operator $\Diamond$ changes and takes on a peculiar meaning. Namely, the schema $\vdash \varphi \leftrightarrow \Diamond K\varphi$ yields a "disastrous schema $\vdash \Diamond\varphi \rightarrow \varphi$" (see Williamson, 1992, 67), where the notion of possibility collapses into actuality.[19] We are not going to go deeper into the characteristics of IML, but will instead turn to one prominent representative of the intuitionistic interpretation of KT. We are going to briefly present the main points in Neil Tennant's proposed solution for blocking the knowability principle $((\varphi \supset \Diamond K\varphi) \vdash (\varphi \supset K\varphi))$.

---

[16] Loosely speaking, a paracomplete logic is a logic in which a proposition and its negation can both be false, while paraconsistent is the one in which a proposition and its negation can both be true.

[17] It, for example, nicely accommodates Putnam's *internal realism*.

[18] In his works, Dummett (2001) does not explicitly address Fitch's paradox in an intuitionistic setting, but he offers a solution to KP without relying on intuitionism.

[19] Schema $\vdash \Diamond\varphi \rightarrow \varphi$, which is highly controversial, is the exact opposite of the non-controversial schema $\vdash \varphi \rightarrow \Diamond\varphi$, meaning: if $\varphi$ obtains, it is possible.

Here is a brief indication of his (Tennant, 1997) variant of an intuitionistic answer to FA. It consists of the modification of the knowability thesis (KT), in which the epistemic operator K is restricted. The main idea of the restriction strategy is that not all but only a restricted domain of propositions is knowable. The proposition to be knowable should be *Cartesian,* and a proposition φ is such if and only if the contradiction does not follow from Kφ. This claim is summarised in Brogaard & Salerno (2002, 146) as:

*Every true statement A is knowable, where 'K(φ) 'is not self- contradictory.*

Williamson (2009) formalises this statement as follows:

*(◊KC) φ; ergo ◊Kφ, where φ is Cartesian.*

Informally, (◊KC) says that truth entails knowability except when Fitch's problem occurs. This proposal seems to be all too simple. As it is said byBrogaard & Salerno (2002), "A defence of this clause is all that is needed to block the problematic substitution of 'B & ~K(B) 'for 'A 'in the knowability principle. After all, K(B & ~K(B)) is self-contradictory". Tennant's restriction was subject to numerous criticisms, for instance, by Brogaard & Salerno (2002), Kvanvig (2006), and Williamson (2009). At first glance, many would agree that the idea of a Cartesian restriction of the epistemic operator is "desperately"[20] ad hoc. Whether or not this is the case, critics raise other, more severe objections. Since it is beyond the purposes of this paper to give an overall assessment of Tenant's proposal, I'm just going to indicate a general direction of the criticisms.

It is apparent that, for (◊KC) to hold, it should be decidable for every $\varphi$ whether it is Cartesian or not. In brief, the claim, "there is no undecidable statements", must hold. The undecidable statement (neither it nor its negation is known) is of the form: $\sim K\varphi \ \& \sim K\sim\varphi$. However, the problem of decidability, as well as the intuitionistic meaning of the possibility operator ◊, are the focus of critics (see Brogaard & Salerno, 2002; Williamson, 2009).

[20] See Williamson, 1993.

Pointing out the general idea of the criticisms, let me remind you of the "disastrous" schema $\vdash \Diamond\varphi \rightarrow \varphi$, which is a standard anti-realist meaning for $\Diamond$, making possibility factive. In epistemic reading, it is: $\Diamond K\varphi \rightarrow \varphi$. By contraposition, we get: $\sim\varphi \rightarrow \sim\Diamond\varphi$. It is easy from this to get: $K\sim\varphi \rightarrow \sim\Diamond\varphi$ (compere proof in Brogaard & Salerno, 2002). Hence, this form of anti-realism entails that there are no undecidable statements. However, ($\Diamond$KC) requires that it should be decidable for every $\varphi$ whether it is Cartesian or not. Since one cannot know in advance whether $\varphi$ is Cartesian, anti-realism is faced with a kind of paradox of decidability. Concluding their long analysis of Tennant's proposal, Brogaard & Salerno (2002) state that "the restriction strategies proposed thus far are insufficient to treat the real problem".

To conclude the discussion concerning strategies for avoiding the effect of FA on verificationism, the remaining part is a paraconsistent approach. Those logics, certainly important and intriguing as they are, still cannot significantly contribute to the vindication of verificationism. Due to the space limitation, we are going only to gesture toward its proposed solutions. Generally speaking, paraconsistent approaches suggest appealing either to the *truth-value gap* or to the *truth-value glut* to block the effect of FA on verificationism. In either way, the verificationist principle can be proved as valid. However, invalidating some of the relevant rules in Fitch's proof (in Priest's *dialetheist* proposal, for instance, it is the rule of *contraposition* (see Priest, 2009), and in proposed semantics introducing *trivial* worlds, $\alpha \vdash \Diamond K\alpha$ becomes *vacuously* valid and verificationism is (vacuously) vindicated. In principle, the same holds for other paraconsistent approaches (for example in Biell, 2009). It should be noted that appealing to the truth-value gap or the truth-value glut (or both, in some combinations) can block *any* proof or argument. The question is whether this manoeuvre of making the verificationist principle vacuously valid can really help the verificationist to win the battle against the FA challenge.

In any case, the proof supporting the knowability principle, $(\varphi \supset \Diamond K\varphi) \vdash (\varphi \supset K\varphi)$, valid in classical logic, is robust enough to resist any attempts, formulated in the frame of classical logic, to refute it. Approaches coming from the camp of non-classical logics can invalidate the knowability principle, but it is dubious whether they, in fact, can help anti-realism. Let me conclude with a much sharper verdict. As Williamson (1993, 204) notes, "The attempts on behalf of anti-realism to deal with the Fitch problem give every sign of a degenerating research programme."

**Acknowledgment**

This work has been partially supported by the University of Rijeka project, grant number: uniri-iskusni-human-23-147-3108.

**References**

Biell, J. (2009). Knowability and Possible Epistemic Oddity. In J. Salerno (ed.), *New Essays on the Knowability Paradox* (pp. 105–125). Oxford University Press.

Brogaard, B. & Salerno, J. (2002). Clues to the paradoxes of knowability: Reply to Dummett and Tennant. *Analysis*, *62*(2), 143–150.

Brogaard, B. & Salerno, J. (2019). Fitch's Knowability Paradox. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2019 Edition)*. The Metaphysics Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/fitch-paradox/

Dummett, M. (1977). *Elements of Intuitionism*. Oxford University Press.

Dummett, M. (2001). Victor's Error. *Analysis*, *61*(1), 1–2.

Edgington, D. (1985). *The Paradox of Knowability*. *Mind*, 94 (376), 557–568.

Edgington, D. (2010). Possible Knowledge of Unknown Truth. *Synthese*, *173*(1), 41–52.

Fitch, F. (1963). A logical analysis of some value concepts. Journal of *Symbolic Logic 28*(2),135–142.

Jankins, C. (2009). The Mystery of the Disappearing Diamond. In J. Salerno (ed.), *New Essays on the Knowability Paradox* (pp. 302–319). Oxford University Press.

Kvanvig, J. (2006). *The Knowability Paradox*. Oxford University Press.

Percival, P. (1990). Fitch and Intuitionistic Knowability. *Analysis*, *50*(3), 182–187.

Percival, P. (1991). Knowability, actuality, and the metaphysics of context-dependence. *Australasian Journal of Philosophy*, *69*(1), 82–97.

Priest, G. (2009). Beyond the Limits of Knowledge. In J. Salerno (ed.), *New Essays on the Knowability Paradox* (pp. 93–104). Oxford University Press.

Rabinowicz, W. & Segerberg, K. (1994). Actual Truth, Possible Knowledge. *Topoi*, 13, 101–115.

Salerno, J. (Ed). (2009). *New Essays on the Knowability Paradox*. Oxford University Press.

Šuster, D. (2023). *Modalni katapulti, Uvod v filozofsko logiko*. Univerza v Mariboru, Univerzitetna založba.

Tennant, N. (1997). *The Taming of the True*. Clarendon Press.

Van Bentham, J. (2004). What One May Come to Know? *Analysis*, *64*(2), 95–105.

Wansig, H. (2002). Diamonds are Philosopher's Best Friends. *Journal of Philosophical Logic, 31*, 591–612.

Williamson, T. (1987). On the Paradox of Knowability. *Mind*, *96*(382), 256–261.

Williamson, T. (1992). On Intuitionistic Modal Epistemic Logic. *Journal of Philosophical Logic*, *21*(1), 63–89.

Williamson, T. (1993). Verificationism and Non-Distributive Knowledge. *Australasian Journal of Philosophy*, *71*(1), 78–86.

Williamson, T. (2000). Knowledge and its *Limits, Oxford University Press.*

Williamson, T. (2009). Tennant's Troubles. In J. Salerno (ed.), *New Essays on the Knowability Paradox* (pp. 183–204). Oxford University Press.

Williamson, T. (2021). Edgington on Possible Knowledge of Unknown Truth. In L. Walters & J. Hawthorne (ed.), *Conditionals, Probability, and Paradox: Themes from the Philosophy of Dorothy Edgington* (pp. 195–211) Oxford: Oxford University Press.

**ANALIZA**
Let. 28, št. 2, str. 229–238, december 2024

# ON MODALITIES WITH POSSIBLE WORLDS

**Accepted**
1. 5. 2024

**Revised**
22. 7. 2024

**Published**
31. 12. 2024

ANDREJ ULE
University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia
andrej.ule@guest.arnes.si

CORRESPONDING AUTHOR
andrej.ule@guest.arnes.si

**Abstract** I am discussing the concept of possible worlds as it is used in modal semantics, which Danilo Šuster also discusses in his book *Modal Catapults*. I wonder how to understand possible worlds in the talk of the non-existent, for example, in explicitly imaginary discourses. Only the context of the discussion, where modal propositions about virtual (irreal) entities or virtual (irreal) states of affairs appear but not a priori judgments about what is in principle logically possible or impossible, can establish a meaningful speech about relevant possible worlds. This raises several problems, e.g., interpreting statements that refer to both real and virtual entities, especially interpreting some counterfactuals about real and virtual entities. It seems that in some instances, we cannot "get rid" of counterfactuality, e.g., by semantic conversion into ordinary general sentences about possible worlds of a certain kind.

**Keywords**
possible world, virtuality (irreality), imaginary discourse, the context of discussion, counterfactual

https://doi.org/10.18690/analiza.28.2.229-238.2024


University of Maribor Press

AN$_\alpha$LIZA

**ANALIZA**
Vol. 28, no. 2, pp. 229–238, December 2024

# O MODALNOSTIH Z MOŽNIMI SVETOVI

ANDREJ ULE
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
andrej.ule@guest.arnes.si

DOPISNI AVTOR
andrej.ule@guest.arnes.si

**Sprejeto**
1. 5. 2024

**Pregledano**
22. 7. 2024

**Izdano**
31. 12. 2024

**Izvleček** Razpravljam o konceptu možnih svetov, kot se uporablja v modalni semantiki, o katerem govori tudi Danilo Šuster v svoji knjigi *Modalni katapulti*. Sprašujem se, kako razumeti možne svetove v govoru o neobstoječem, na primer v eksplicitno imaginarnih diskurzih. Menim, da lahko smiseln diskurz o relevantnih možnih svetovih vzpostavi le kontekst razprave, v katerem se pojavljajo modalne propozicije o virtualnih (nerealnih) entitetah ali virtualnih (nerealnih) stanjih stvari, ne pa apriorne sodbe o tem, kaj je v principu logično možno ali nemogoče. Pri tem se pojavlja več težav, npr. interpretacija relativnih izjav, ki se nanašajo tako na realne kot na virtualne entitete, še posebej interpretacija nekaterih protidejstvenikov o realnih in virtualnih entitetah. Zdi se, da se v določenih primerih ne moremo "znebiti" protidejstvenosti, npr. s semantično pretvorbo v navadne generične stavke o možnih svetovih določene vrste.

**Ključne besede**
možni svet, virtualnost (irealnost), imaginarni diskurz, kontekst razprave, protidejstvenost

https://doi.org/10.18690/analiza.28.2.229-238.2024


Univerzitetna založba Univerze v Mariboru

In his book *Modal Catapults*, Danilo Šuster discusses various forms of modal arguments that are important for philosophy, e.g., the use of various types of conditionals in ethics, epistemology, and ontology (Šuster, 2023). For the most part, philosophy relies on today's standardized forms of modal logics and on various types of semantics of possible worlds for this task, which stem from Hintikka's and Kripke's "invention" of various semantics of this type in the 1950s (Kripke, 1963; Hintikka, 1957), although the very idea of interpreting modal statements with the help of possible worlds is much older, as it can be found already in Leibniz in the 17th century. With the help of such semantics, it is relatively easy to interpret various modal statements and entire modal logic systems. Despite the tremendous popularity and prevalence of this semantic methodology, this methodology also has some limitations. This was noticed already by W. V. O. Quine in his criticisms of modal statements and the semantics of possible worlds (Quine, 1960). Some, such as D. Lewis, try to strengthen the vague ontological status of possible worlds by giving them the status of alternative realities (Lewis, 1976) or reject them and allow them only the actual world (Mackie, 1973). Some think that beyond various types of modalities, we need a special status of actuality (Cowling, 2011). Others understand possible worlds only as verbalized imaginary possibilities (Rosen, 1990) or try to expand the concept of possible worlds with "impossible worlds" (Nolan, 2021), etc. In these ways, various authors try to solve paradoxes concerning modal sentences, especially counterfactuals. Some paradoxes of this type are also excellently presented by Danilo in his book (for more on different variants of the possible worlds' theory, see, e.g., Divers, 2002).

In my essay on the non-existence (Ule, 2019), I talked about how complex the logical formulation of the non-existence speech is. Assumed "possible worlds" are generally worlds that contain some non-existent, although at least possible virtual entities or possible virtual facts. We can assume that such entities or facts belong to certain worlds that are at least virtually, e.g., verbally, or imaginatively possible. We can ask whether such virtual worlds also belong to the possible worlds of modal logic. We cannot say anything about this in advance, since only the specific context of the discussion or argumentation indicates (but by no means uniquely implies) what belongs to the multitude of semantic possible worlds.

In discussions about the modal qualities of things, events, or sentences, we must assume some area of explicit or at least tacit agreement, i.e., we must assume some common context of the discussion, which sufficiently defines the "boundaries"

between the actual and only possible, as well as the boundaries between possible, necessary, and impossible. In this case, sudden shifts in the context may quickly occur, which lead us to various semantic and logical confusions.

Discussions about possible worlds in modern philosophy, especially in modal logic, are already quite extensive, branched, and even contradictory. It is not possible to find any at least relatively plausible conception of possible worlds that would not raise some weighty objections (see, e.g., Divers, 2002). Therefore, I will limit myself to an informal and initial definition of possible worlds as groups of objects, states of affairs, and events that can be assumed to be possible in certain real or fictitious circumstances. I don't want to commit to any more precise definitions of the mentioned "groups", such as, e.g., the widespread use of "maximality", e.g., maximally consistent sets of sentences, propositions, states of affairs or properties of things, etc., that is sets that would include all in principle possible non-contradictory combinations of sentences, propositions, states of affairs or properties of things, etc., because in our linguistic and mental practice, we only need limited sets, e.g., in a given discussion, the relevant set of descriptions, objects, states of affairs, and events.

Even our "current world" cannot be defined as a maximal set of the above type because every day we encounter a series of current situations where, at least practically, and sometimes also in principle, it is not possible to determine whether a sentence or a proposition is true or false, whether a state of affairs is a fact or not, whether a thing has or does not have certain properties, etc. For example, quantum physics features a whole series of situations where, even in principle, it cannot be determined whether some quantum objects at a given time interval have or do not have some properties (just think of the famous Schrödinger's cat in a quantum box, where it is in principle impossible to determine whether the cat is alive or dead before we "look at it") (Schrodinger, 1935). Therefore, I do not agree with Dale Jacquette's thesis, who tried to define the current or the actually existing world as the only possible world where we can speak, that it corresponds to a maximally consistent combination of sentences or propositions, as well as maximally consistent combinations of states of things or properties of things, while all other possible worlds should be non-maximal or submaximal, because in those we can always find objects where it is not possible to determine whether a property belongs to an object or not (Jacquette, 2005: 244–246).

One could say that such worlds know of "defining voids". So, at most, we could talk about the fact that the current world is at least "minimally" submaximal in relation to all other possible worlds, but this is also a questionable definition, because for every submaximal possible world, we could "find" a corresponding maximal possible world, where, for example, in cases of "defining voids", such voids were artificially filled by arbitrarily assigning some properties or their complements to previously "undetermined" objects. Even in the case of Schrödinger's cat, we can assume the existence (or better subsistence) of two possible worlds, in one the cat would be dead even before we opened the box, and in the other it would be alive. Well, this solution of the problem was suggested in the famous Everett's theory of many worlds, where every logically possible quantum-mechanical course of things corresponds to some alternative physical world, where exactly such a course takes place and in which there is some observer who notices the "corresponding" state of things, e.g., finds a dead or alive cat in the box (Vaidman, 2018).

I believe that the current world can only be determined as actual by pointing to it, i.e. by saying or implying that we exist in the same reality where we speak or think about, and where this existence includes some implicit totality of actually existing beings, like "all actual things", "everything that actually happens", etc. Such a determination is never absolutely certain, it depends, among other things it depends, on the context of the discussion or thinking, because this context also carries within it a distinction between what happens or exists *regardless of the context* and what exists *in regard of the context*. This also means that the distinction between the actual possible world and "merely" possible (e.g., virtual) worlds is relative and conditional, but not absolute. However, it is nevertheless important that such a distinction is possible according to each context (discussions, etc.), so we are always faced with it, whether we are aware of it or not. In this sense, I am talking about the inevitability of agreeing to some current world as actual and about the difference between the actual and non-actual possible worlds.

I would like to point out here that I adopt a somewhat non-standard conception of possible worlds, i.e. I adopt a non-fixed set of objects for possible worlds because they may change according to the different variants of the given context of the discussion.[1]

[1] Therefore, I don't accept the Kripke's postulate about necessary identities or about de re identity of objects in the set of possible worlds (Kripke, 1971). I accept only the limited postulate about necessary identities, namely identities are necessary only in those segments of the given context where it is possible to talk (think) about a certain object

In a discussion of some ancient mythological story, e.g., about the Pegasus story, I assume that Pegasus exists only in some Greek mythological stories, by no means in all, and of course not in the actual (real) world. However, in those stories where Pegasus appears, he represents the same being, regardless of whether he is imagined perhaps as an immortal divine or as a mortal semi-divine being (who, for example, died together with the hero Bellerophon when they wanted to climb Mount Olympus). The sentence "Pegasus may have had golden wings" is similarly a meaningful part of the discussion because Pegasus with golden wings "appears" in various contemporary artistic representations. These present a kind of extended mythological context of Pegasus story. In this sense, we can talk about a possible world by regarding the context of Greek mythology, where Pegasus has golden wings, although "usually" he is represented as a purely white horse with white wings.

Let us consider some sentences about meaningful combinations of fictitious and real essences and states of affairs. Take the sentence, "Hercules is stronger than Muhammad Ali." We cannot assign a clear meaning and truth value to this sentence because we simply do not know in what context we are talking about Hercules and Muhammad Ali, as there does not seem to be any common context of discussion that contains the stories of either character. Also, Hercules is a character from Greek mythology, who is considered a purely fictional character, while Muhammad Ali was a real person, one of the best boxers of all time. However, we think the sentence makes sense and is even true. I think that is the case because most of us unwittingly change the original, purely mythological context of talking about Hercules into an expanded mythological-real context, where we either imagine Hercules as a real person fighting Muhammad Ali or we attach Muhammad Ali to the myth of Hercules and imagine an extended mythological context where the two again clash with each other.

In any case, it probably seems to most of us that in such a duel, Hercules would win without a doubt because he is said to be endowed with a whole range of divine powers and qualities (e.g., invulnerability, divine strength, incredible speed of reaction to problems, etc.). Therefore, given the reasonably permissible beliefs about the mythological Hercules, we can mentally construct a multitude of possible worlds

---

or creature. In these segments we can talk about "this and that object (creature)". Kripke's proof of the necessity of identities applies only to true nominal identities, i.e. for identities that explicitly allow nominal predicates (e.g., the predicate "necessary (x = a)") but not for real identities that consider only predicates of real properties (e. g. the predicate "x is great") (see my paper "Ali je identiteta res nujna" (Is Identity Truly Necessary)(Ule, 2003).

in which a meeting between the mythological Hercules and the actual Muhammad Ali could take place, and in all these possible worlds, Hercules would defeat Muhammad Ali because Muhammad Ali does not possess divine qualities, which are said to be "owned" by Hercules. In this sense, the sentence "Hercules is stronger than Muhammad Ali" is even necessarily true, not just contingently true.

In short, in judgments of sentences about virtual entities, we must consider both completely real possibilities and various virtual possibilities. This also applies to sentences where fictional and factual entities are mixed.

Take for example the relational sentence, "Hamlet is more unhappy than Margaret II." (Margaret II is the current queen of Denmark). Here, we must proceed similarly to the case of Hercules and Muhammad Ali, in short, we must find some extended context of the story of Hamlet, in which the (real-life) Queen Margaret II would also belong. In doing so, we would make minimal changes to preserve the assumed identity of Hamlet and the identity of the real queen and see whether the sentence "Hamlet is more unhappy than Margaret II" would be non-trivially true given this context. I think anyone can easily imagine such a context, in which Hamlet would still be an extremely unhappy and divided man, and the Queen of Denmark relatively happy and content, although she might be very concerned about Hamlet's fate. You should never, e.g., make changes such that Margaret II. would become Hamlet's queen-mother, because then the assumption that Hamlet is more unhappy than Margaret II might be untrue, at least judging by what is supposed to happen to his mother in the play.

The situation is more complicated in the cases of counterfactuals with unrealistic ingredients. Here, it may happen that we cannot get rid of counterfactuality in any way, in short, we cannot switch to some non-contrafactual conditionals about possible worlds of a certain type.

Let's take our above example, "Hercules is stronger than Muhammad Ali". In no "normal" possible world of Greek mythology does the said duel occur. However, the given sentence seems to be true. Maybe it is true counterfactually. In this sense, the sentence "Hercules is stronger than Muhammad Ali" suggests a counterfactual:

> (1) "If Muhammad Ali met Hercules, then Muhammad Ali would lose the fight to Hercules."

We could argue (in the sense of Lewis, 2001) that in all nearby possible worlds, which, given the context of Greek mythology, are as similar as possible to the world where the actual Muhammad Ali lived and where the fight between Muhammad Ali and Hercules would take place, Muhammad Ali loses to Hercules.

If we try to semantically interpret this counterfactual, we must consider some "extended" set of possible worlds that *would* contain both the events of Greek mythology and the real world. We have to especially consider worlds that would be, in some sense, the closest to the worlds of Greek mythology where Hercules appears, but Muhammad Ali could also appear in them. We could consider these worlds, e.g., as possible worlds of "relevant" discussions on Greek Mythology and modern boxing champions.

However, even in this scenario, a new counterfactual is hidden again, namely:

> (2) If it would be permissible to discuss the worlds where the match between Hercules and Muhammad Ali would take place, then the discussions would result in agreeing that Hercules defeats Muhammad Ali.

We need then to presuppose something like the set of possible discussions HA that apply to the match of Hercules and Muhammad Ali, and which would be the nearest to the set of "relevant" discussions on Greek Mythology and the modern boxing champions. In accord with Lewis, we say the counterfactual (2) is true iff all possible discussions of HA agree with Hercules's defeat of Muhammad Ali.

However, this could not be any "factual" conditional on possible discussions because it has to consider hypothetical discussions on possible worlds that *would (could)* contain Greek Mythology and the real box champions and not only the real possible discussions on Greek Mythology and the modern box champions. So, we need the next counterfactual,

> (3) If it would be possible to discuss the discussions where Greek Mythology and the real world somehow "intersect", and which would discuss the match of Hercules and Muhammad Ali, then the discussions would result in agreeing that Hercules defeated Muhammad Ali.

This could not be the end of the story. We get the circle of more and more complicated counterfactuals regarding possible worlds of possible discussions, possible discussions of possible worlds, and so on to infinity. Merely the introduction of some "extended" set of virtual possible worlds as such does not help us to break out of this circle.

I think that similar dilemmas can be found in other areas of the logic of conditionals and in deontic logic, which Danilo beautifully presented in his book. At the end of the book, Danilo writes that modal logic is not interested in which possible worlds are close and which are further from reality, but only in what follows when it is determined, what is accessible (i.e., is close) and what is not (i.e., is far) (Šuster, 2023, p. 207). This is true, but often, it is not a simple fact of definiteness but a reference to irreducible counterfactuals that speak of what would follow in a given context of speech from something that we would assume to be possible.

**References**

Cowling, S. (2011). The Limits of Modality. *The Philosophical Quarterly*, *61*(244), 473–495.
Divers, J. (2002). *Possible Worlds*. Routledge.
Hintikka, J. (1957). Modality as referential multiplicity. *Ajatus*, *20*, 49–64.
Jacquette, D. (2005). Nonstandard Semantics for Modal Logic and the Concept of a Logically Possible World. *Philosophia Scientiæ*, *9*(2), 239–258.
Kripke, S. (1963). Semantical Considerations on Modal Logic. *Acta Philosophica Fennica 16,* 83–94.
Kripke, S. (1971). Identity and Necessity. In M. K. Muntz (Ed.), *Identity and Individuation* (pp. 135–164). New York University Press.
Lewis, D. (1986). *On the Plurality of Worlds*. Blackwell.
Lewis, D. (2001). *Counterfactuals*. Blackwell.
Nolan, D. (2021). Impossibility and Impossible Worlds. In Bueno, O. and Shalkowski, S. (Eds.), *The Routledge Handbook of Modality* (pp. 40–48). Routledge.
Quine, W. van O. (1960). *Word and Object*. MIT Press.
Rosen, G.(1990). Modal Fictionalism. *Mind*, *99*(395), 327–35.
Šuster, D. (2023) *Modalni katapulti*. Univerzitetna založba Univerze v Mariboru.
Ule, A. (2002). Ali je identiteta res nujna? *Analiza*, *6*(4), 114–111.
Ule, A. (2019). Govor o neobstoju brez dodatnih ontoloških predpostav. In O. Markič and M. Malec (Eds.), *Filozofska pot Andreja Uleta* (pp. 7 –16). Znanstvena založba Filozofske fakultete.
Schrödinger, E. (1935). Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften*, *23*(48), 807–812.
Vaidman, L. (2018). Many-Worlds Interpretation of Quantum Mechanics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*. The Metaphysics Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/qm-manyworlds/

# Epistemologija

**ANALIZA**
Let. 28, št. 2, str. 241–262, december 2024

ANαliZA

# HIGHER-ORDER EVIDENCE IN SCIENCE: SOME PROBLEMATIC CONSEQUENCES OF STEADFASTNESS AND LEVEL-SPLITTING

**Accepted**
1. 10. 2024

**Revised**
30. 11. 2024

**Published**
31. 12. 2024

MARTIN JUSTIN
University of Maribor, Faculty of Arts, Maribor, Slovenia
martin.justin1@um.si

CORRESPONDING AUTHOR
martin.justin1@um.si

**Abstract** Despite our best efforts, we often fail to act in a perfectly rational manner. Recently, some epistemologists have suggested that we should admit our failings and develop a modest epistemology that would take our fallibility seriously. This includes accounting for the role of evidence of our irrationality, usually called higher-order evidence. It seems intuitive that modest reasoners should take such evidence into account. However, it turns out that incorporating higher-order evidence into a principled theory of what rationality requires is not an easy task. In this paper, I first review the debate about higher-order evidence, describing in detail the puzzle of higher-order evidence and the main positions about it in the literature. Then, I provide two novel examples of higher-order evidence, taken from science. I argue that these examples put pressure on the views that reject the role of higher-order evidence. These views commit themselves to the conclusion that some common scientific practices, such as evaluating evidence in systematic reviews or even running null hypothesis significance tests, are irrational.

**Keywords**
higher-order evidence,
modest epistemology,
evidence,
rationality,
science

https://doi.org/10.18690/analiza.28.2.241-262.2024

AN$_\alpha$LIZA

**ANALIZA**
Vol. 28, no. 2, pp. 241–262, December 2024

# DOKAZI VIŠJEGA REDA V ZNANOSTI: NEKAJ PROBLEMATIČNIH POSLEDIC ODLOČNOSTI IN RAZDRUŽEVANJA RAVNI

MARTIN JUSTIN
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
martin.justin1@um.si

DOPISNI AVTOR
martin.justin1@um.si

**Sprejeto**
1. 10. 2024

**Pregledano**
30. 11. 2024

**Izdano**
31. 12. 2024

**Izvleček** Kljub prizadevanju po nasprotnem pogosto ravnamo iracionalno. V zadnjem času so nekateri epistemologi predlagali, da priznamo svoje napake in poskušali razviti *skromno* epistemologijo, ki bo v zakup vzela dejstva o naši iracionalnosti. Pomemben del takšne epistemologije predstavlja tudi opis vloge dokazil o naši iracionalnosti, t. i. dokazil višjega reda. Intuitivno se zdi, da bi morali pri tvorjenju prepričanj takšna dokazila upoštevati. Vendar pa se izkaže, da vključevanje dokazil višjega reda v teorijo racionalnosti ni lahka naloga. V tem prispevku najprej pregledam razpravo o dokazilih višjega reda, podrobno opišem problem dokazil višjega reda in predstavim glavne odgovore nanj. Nato predstavim dva nova primera dokazil višjega reda, vzeta iz znanosti. Trdim, da ta primera pod vprašaj postavljata stališča, ki zavračajo vlogo dokazov višjega reda. Ta stališča implicirajo, da so nekatere vsakdanje znanstvene prakse, kot sta vrednotenje dokazil v sistematičnih preglednih člankih ali celo izvajanje testov statistične značilnosti, iracionalne.

**Ključne besede**
dokazila višjega reda, skromna epistemologija, dokazila, racionalnost, znanost

https://doi.org/10.18690/analiza.28.2.241-262.2024


Univerzitetna založba Univerze v Mariboru

## 1 Introduction

One of the more uncontroversial claims in epistemology is that we should strive to have rational beliefs. Regardless of what we think this entails – confirming to our evidence, forming beliefs reliably or responsibly, conditionalizing on our priors and evidence – being perfectly rational is an ideal. We often fail to achieve it. We fail to apply appropriate rules of rationality, we fail to recognize correct rules of rationality, we make mistakes in our reasoning, we get misleading evidence, etc.

Recently, some epistemologists have suggested that we should admit our failings and develop a modest epistemology that would take our fallibility seriously (Dorst, 2020; Christensen, 2020; DiPaolo, 2019). Central to this project is to provide an account of how to deal with evidence of our rational failings. Such evidence abounds. We know that we reason more poorly when sleep-deprived, hungry, or under the influence of drugs; we are demonstratively bad at reasoning about probabilities and exhibit other cognitive biases.

In epistemological jargon, such evidence is usually referred to as higher-order evidence. There is no single agreed-upon definition of higher-order evidence. But to align our ideas: in contrast to evidence that directly concerns some question we try to answer ("I see that the cup on the table is empty."), higher-order evidence concerns our epistemic performance or evidential situation ("I visited my optometrist yesterday, and it turns out my sight is very unreliable.").

It seems intuitive that modest reasoners should take such evidence into account. However, it turns out that incorporating higher-order evidence into a principled theory of what rationality requires is not an easy task. It can lead us down some troubling paths, such as rejecting established epistemic principles, e.g., conditionalization, or admitting that norms of rationality simply give inherently contradictory advice. Modest epistemology has thus encountered resistance in the form of rejecting any bearing or higher-order evidence on the rationality of our beliefs.

In this paper, I will point to an undesirable consequence of rejecting the import of higher-order evidence that has not yet been discussed in the literature. I will argue that higher-order evidence sceptics commit themselves to the view that some

common scientific practices, such as evaluating evidence in systematic reviews or even running null hypothesis significance tests, are irrational. This will not provide a decisive argument against higher-order evidence scepticism. Rather, it will shift the burden of proof on such scepticism to reject the role of higher-order evidence.

The paper is organized as follows. Section 2 presents the problem of higher-order evidence, while Section 3 overviews its main positions. In Section 4, I present two cases of how higher-order evidence is used in today's scientific practice and spell out the consequences of rejecting these cases. The second part of this section also discusses and rejects two arguments against the admissibility of these cases. Section 5 concludes the paper.

## 2 Higher-Order Evidence: The Problem

To get the discussion going, let us consider the following case:

**Experiment.** Max is a PhD student in pharmacology. She is working in a team that develops novel antibiotics. As part of her research, she was tasked with testing a new promising compound, *c,* on some e-coli bacteria. To do this, she prepared two rows of Petri dishes: in one, she treated the bacteria with *c*, and in the other, she treated them with one of the existing antibiotics as control. She was doing this late in the evening, after a 10-hour shift, and under considerable time pressure. The next day, she went to check the Petri dishes and observed that there was no difference in the presence of e-coli between the rows. From this, she concludes that *c* is not better than the existing compound. At lunch, however, she read a newspaper article about a new study that found that in 50 % of the cases, researchers in circumstances Max was experiencing yesterday, messed up their experiments. Let us assume that Max was lucky this time and correctly applied the two compounds but has no way of checking that.

In the **Experiment**, Max received two kinds of evidence[1] that in some way concern her beliefs about the efficiency of the new compound. On the one hand, she has first-order evidence, consisting of the experimental results. On the other hand, she

[1] Note that, strictly speaking, one piece of evidence can have both first- and higher-order bearing. Consequently, evidence cannot be neatly separated into different kinds so we should be careful with such "kinds of evidence" talk. When discussing first- or higher-order evidence, I will thus generally mean "evidence that bears on *p* in a first-/higher-order way".

also has higher-order evidence, consisting of the newspaper article reporting on the study results. The problem arises when we consider the direction in which these two pieces of evidence are pointing. Max's first-order evidence suggests *q*: "The new compound is no more effective than the known ones." Meanwhile, her higher-order evidence throws doubt on *q* by suggesting that this belief was formed in an unreliable way. In other words, Max seems to be in a kind of epistemic bound. She should believe both: (1) that *q* and (2) that it is irrational to believe that *q*. Believing something like "*q,* but it's irrational for me to believe *q*" seems epistemically suspect, to say the least.[2]

There are two distinct but related ways of making the problem of higher-order evidence more precise. One presents it as a conflict between what could be called substantive and structural epistemic principles (this diction is from Whiting (2021); see Lasonen-Aarnio (2014), Worsnip (2018), Horowitz (2022), Ye (2022) for presentation along these lines). The other presents it as a problem of higher-order uncertainty (Dorst, 2024; Henderson, 2022). The upshot of both presentations is very similar – higher-order evidence can wedge a gap between our first- and higher-order doxastic states, which seems problematic – but they can differ in how they carve up the individual positions in the debate. In the remaining of the section, I will present both ways of understanding the problem.

Let's start with the view that the problem of higher-order evidence has to do with a conflict between different kinds of epistemic principles. Consider the following plausible substantive epistemic principle:

> *Evidentialism*[3]: S is justified in believing *p* at t if and only if S's evidence at t on balance supports *p*. (Feldman, 2009)

As we saw in the **Experiment**, Max's total evidence, which consists of both her first- and higher-order evidence, supports two propositions. One is *q*: "The new compound is no more effective than the known ones." The other one is something

[2] In jargon: it's epistemically akratic (Horowitz 2014).

[3] Evidentialist principle like *Evidentialism* is here used only as an example. We could in principle substituted it with any substantive epistemic principle of the form: "S is justified in believing *p* at t iff *p* satisfies some condition C for S at t" (Ye 2022)

like $q$*: "I am totally unreliable at determining $q$ on the basis of my evidence." Now consider a plausible structural epistemic principle:

> *Bridge*: It is irrational for a person to believe that p and to believe that it is irrational for them to believe that p. (Whiting, 2021; Ye, 2022)

It should be quite clear that having both $q$ and $q$* violates *Bridge*. But both these beliefs are justified by *Evidentialism*. Thus, it seems that either *Evidentialism* or *Bridge* must give.

Alternatively, the lesson of cases like the **Experiment** can be hashed out in terms of higher-order uncertainty (Dorst, 2024; see also Henderson, 2022, who uses the term conviction). First-order uncertainty refers to a familiar kind of uncertainty about whether some state of affairs obtains. For example, if my roommate promised to vacuum the apartment while I am away during the weekend, I might be uncertain whether the apartment is indeed vacuumed. I know my roommate is usually good at his word, but at the same time, he is a bit sloppy when it comes to cleaning the apartment. So, before I open the apartment door on Sunday afternoon, I might be only 0.7 certain that the apartment is vacuumed. Conversely, higher-order uncertainty refers to the uncertainty about whether my first-order belief or credence is rational. For instance, I might reflect more about my roommate and his cleaning habits and realize that I usually judge him too harshly – I systematically underestimate his zeal for cleaning. Consequently, I might become uncertain whether my credence of 0.7 in the proposition that the apartment is clean is rational, given that I am a biased judge in this case.

More specifically, Dorst (2024) defines higher-order uncertainty as "a unique and precise probability function that encodes the rational degree of belief," or $P(P(q) = t)$, where $P(p) = t$ is my first-order credence in a given proposition. Or, in the form of a principle:

> *Higer-Order Uncertainty*. It is rational to have higher-order uncertainty if and only if there is a proposition $q$ and a threshold $t$ such that you should be unsure whether you should be $t$-confident of $q$: $0 < \big(P(P(q)) = t\big) < 1$.

The (intuitive) idea that connects this notion with the debate on higher-order evidence states that higher-order evidence directly bears on this kind of uncertainty. Receiving negative higher-order evidence, e.g., reading the news story in the **Experiment**, increases higher-order uncertainty about the rationality of the first-order belief. Similarly, receiving positive higher-order evidence – if Max would learn, for example, that she is extremely reliable in performing her experiments – can decrease it. The question of higher-order evidence can thus be recast as questions about higher-order uncertainty: Can higher-order uncertainty be rational? And if yes, is there a connection between higher-order uncertainty and first-order beliefs?

As Dorst (2024) shows, this is a fruitful way of understanding this problem. Different positions on the existence and the role of higher-order uncertainty also align well with different answers to the puzzle of higher-order evidence, seen as a conflict between substantive and structural principles of rationality. However, since the notion of higher-order uncertainty is embedded into a specific formal framework, translating between the two ways of carving up the positions sometimes requires a bit more work. For simplicity, I will leave a more detailed discussion of the notion of higher-order uncertainty on the side in the rest of this paper and refer to it only in passing.

## 3 Higher-Order Evidence: The Positions

To sum up the discussion in the previous section, higher-order evidence presents us with a puzzle of the following pattern (Sliwa and Horowitz 2015):

(1) One's rational beliefs should reflect the bearing of one's (first-order) evidence.
(2) One's rational beliefs should reflect the bearing of one's (higher-order) evidence.
(3) One's rational first- and higher-order doxastic states should not be in tension.

Different views on higher-order evidence differ in how they respond to this puzzle. Steadfast views reject (2) (Tal, 2021; Littlejohn, 2018; Titelbaum, 2015; Kelly, 2005). They argue that higher-order evidence has no rational import, so we should simply ignore it. In higher-order uncertainty talk, they reject the rationality of such

uncertainty: if $P(q) = t$, then $P(P(q) = t) = 1$. Level-splitting views reject (3) (Lasonen-Aarnio, 2014; Worsnip, 2018). They argue that evidence has both first- and higher-order import on our beliefs but that there is no way to reconcile these different impacts. For level-splitters, beliefs like "*p* but it is irrational for me to believe *p*" are rationally permissible, thus (3) must give. In higher-order uncertainty talk, these views accept that any possible level of such uncertainty is permissible. Calibrationist views reject (1) (Ye, 2022; Elga, 2010; Christensen, 2010). They argue that higher-order evidence defeats or brackets the first-order bearing of our evidence. Thus, our first-order beliefs should follow or "calibrate with" our higher-order attitudes. If we are uncertain that our belief that *p* is rational, this belief should be revised to reflect this uncertainty. In the higher-order uncertainty talk, these views accept the possibility of higher-order uncertainty but try to show that there is a systematic connection between it and the first-order attitudes. Finally, Dilemma views accept the puzzle wholeheartedly (Christensen, 2016a; Schoenfield, 2015a, 2015b). They argue that rationality presents us with genuinely incoherent requirements.

While some might be more intuitively appealing than others, all these options are surprising in some way. The rest of this section will briefly present the main motivations and appeal of each of the views and point to some of their problems.

### 3.1 Steadfast views

Steadfastness is mainly motivated by the idea that these views follow naturally if we take two principles of rationality seriously (see Field 2019 for a concise summary; she, however, does not endorse steadfastness). One of these principles is a form of *Bridge*, already discussed above. Titelbaum (2015) presents the following version:

> *Akratic Principle*: No situation rationally permits any overall state containing both an attitude A and the belief that A is rationally forbidden in one's current situation.

The other principle is a kind of meta-principle. It asserts that principles of rationality apply universally to all agents in all situations.

> *Universality*: Requirements of rationality apply universally to all agents regardless of their situation.

These two principles both seem very intuitive. However, taken together, they have an interesting consequence. They are in tension with the idea that making mistakes about what rationality requires is rationally permissible. Consider this situation. Agent A is in a situation S. A also believes in principle R: "When in S, you are permitted to believe *p*". A (correctly) recognizes that she is in S and that R applies to her (given *Universality*). Thus, she concludes that *p*. Given the *Akratic Principle,* A should not be uncertain whether R is a genuine requirement of rationality. Such uncertainty would imply that she is not permitted to believe *p* in S, which would violate the *Akratic Principle*. In other words, A should be certain that she is correct about what rationality requires of her. That holds for every belief that S justifiably holds. Thus, to justifiably hold any belief, she should think that she never makes any mistakes about what rationality requires of her.

This tension between *the Akratic Principle, Universality,* and the possibility of making rational mistakes motivates Steadfastness. Nevertheless, defenders of Steadfastness still need to provide a story about why we should reject the possibility of making rational mistakes instead of the *Akratic Principle* or *Universality* or rejecting the dilemma altogether.[4] As it turns out, providing this story requires very strong commitments about the access or competence of agents. Titelbaum (2015), for example, argues that agents have a priori insight into what rationality requires of them in a given situation. In his words:

> "Every agent possesses a priori, propositional justification for true beliefs about the requirements of rationality in her current situation. An agent can reflect on her situation and come to recognize facts about what that situation rationally requires. Not only can this reflection justify her in believing those facts; the resulting justification is also empirically indefeasible." (Titelbaum, 2015)

[4] In an interesting turn, Skipper (2021) for example shows that rejecting the possibility of higher-order uncertainty is compatible with Calibrationism, i.e., the view that our beliefs should reflect the higher-order bearing of our evidence.

Littlejohn (2018), on the other hand, argues that the prohibition of mistakes about what rationality requires is implied by us being epistemically competent.

Additionally, Steadfastness also gives unintuitive answers in cases such as the **Experiment.** According to this view, Max should ignore any higher-order bearing of her evidence and remain certain that her belief about the experiment is rational. I will explore this worry that Steadfastness leads to problematic conclusions about such cases in more detail in Section 3.

### 3.2 Level-Splitting views

Level-Splitting accepts both first- and higher-order import of our evidence but rejects the idea that our doxastic states on these different levels must cohere in the way required by principles such as the *Bridge* or *Akratic Principle*. This view is usually argued for by carefully considering and rejecting other possible answers to the puzzle of higher-order evidence that try to reconcile the three claims.

Let us look in a bit more detail at Lasonen-Aarnio's (2014) presentation of this view. After presenting the above puzzle, she outlines three possible ways in which a theory of rationality could deal with such conflicting recommendations. The first one includes introducing what she calls an "über rule": an overarching epistemic rule that would determine the correct rational response for every possible epistemic circumstance. If such a rule exists, then it could not happen that one would apply it perfectly and at the same time get evidence that one's belief is flawed – the rule would already pre-empt the possibility of receiving such evidence and provide an appropriate response for it. While the existence of such a rule would indeed solve the dilemma, Lasonen-Aarnio is sceptical of the possibility and desirability of such rules. First, the existence of an "über-rule" would imply that all other epistemic rules, if taken to hold universally, are, in fact, wrong. Second, the über-rule would push us towards an undesirable kind of epistemic particularism, where epistemic recommendation would be limited to carefully applying the rule to each individual epistemic situation. Consequently, nothing general could be said about epistemic guidance. In other words, the über-rule cannot play the kind of guidance we expect from epistemic rules.[5]

[5] For a push back against Lasonen-Aarnio's analysis of über-rules, see Kappel (2019). He argues that a version of *Evidentialism* could be considered as a kind of (feasible) über-rule.

Second, Lasonen-Aarnio (2014) considers the view that epistemic rules are hierarchically ordered; thus, in every situation, one of the conflicting rules will override others. Under this picture, an epistemic system would consist of two elements: (1) a set of correct epistemic rules and (2) an ordering relation on these rules, a kind of meta-rule that tells us which rule to follow. This picture might seem promising – in cases like the **Experiment**, such meta-rules could tell us whether, given the specific circumstance, we should follow our first-order or our higher-order evidence. However, since meta-rules are just epistemic rules, we can acquire evidence that we made a mistake in applying them. If this is the case, then we would need another meta-rule to tell us how to resolve this conflict between lower-level meta-rules. As Lasonen-Aarnio convincingly argues, the hierarchical picture of epistemic rules thus either ends up in an infinite regress or posits a kind of über-rule that cannot be defeated.

The third option that Lasonen-Aarnio (2014) presents argues that we could simply admit that situations like the one in the **Experiment** present genuine epistemic dilemmas, where an agent is damned to do something they ought not to. I will discuss Dilemma views in more detail in the next subsection. Here, I will just note that both Lasonen-Aarnio (2013) and Worsnip (2018) argue that such views fail to sufficiently explain why we should take different "oughts" to concern the same, rather than different normative domains. In other words, they admit that first- and higher-order evidence can both have normative force, as dilemma theorists would have it. However, they disagree that these normative forces act in the same domain, namely epistemic rationality.

This finally brings us to the Level-splitting views of higher-order evidence. As already mentioned, these views are similar to the Dilemma views in that they want to preserve both the first- and higher-order bearing of our evidence. But where Dilemma views also accept *Bridge,* the principle that doxastic attitudes on different levels must cohere, Level-Splitting rejects it. As Lasonen-Aarnio puts it, she thinks "that subjects who fail to revise their beliefs in putative cases of defeat are criticizable from an epistemic point of view: they are being unreasonable by failing to take into account evidence about their own cognitive imperfections" (Lasonen-Aarnio 2014). However, she rejects the idea that such epistemic failings are also failings of rationality: "There are epistemic oughts that a subject can violate without thereby being epistemically irrational." Similarly, Wornsnip (2018) argues that the tension

between doxastic states can be rational because *Evidentialism* and *Bridge* operate on different normative domains. *Evidentialism* is a narrow requirement that concerns individual states' (ir)rationality. On the other hand, *Bridge* is a broader requirement that guides reasoning more broadly and does not issue recommendations regarding particular situations.

The main problem with the Level-Splitting Views is that they reject intuitive principles like the *Bridge* or the *Akratic Principle*. Consequently, these views allow sets of beliefs that were above recognized as epistemically suspect, for example: "p but I ought not to believe that p." While not strictly incoherent, many consider such sets problematic enough to reject Level-Splitting on these grounds (Henderson, 2022; Ye, 2022).

### 3.3 Dilemma Views

Like Level-Splitting, Dilemma views concede that there is no elegant way to accommodate both first- and higher-order evidence into our picture of rationality (Schoenfield, 2015b, 2015a; Christensen, 2016a, 2016b). In contrast to Level-Splitting, they try to preserve *Bridge*. To unpack this difference, consider how the two views evaluate the combined belief "p, but I ought not to believe that p." As explained above, Level-Splitting understands this combined belief as rationally permissible, so agents are not rationally required to revise it. On the other hand, Dilemma views concede that it is irrational; however, agents cannot revise it in a way that would satisfy all the requirements of rationality. In other words, examples such as the **Experiment** put agents in an epistemic bind where every course of action is irrational.

Despite this gloomy outlook, defenders of the Dilemma views argue that some options are less bad than others. In other words, even though an agent who receives misleading higher-order evidence cannot act rationally, there is still one *best* epistemic response available to them. Both Christensen and Schoenfield tie this to the notion of accuracy: if none of the available beliefs is rational, then agents can at least aim for accuracy. Or, as Christensen (2016b) puts it:

> "if [an agent] has very strong anti-reliability evidence about herself, she will see that she is faced with two possibilities: either (a) believing something likely to be inaccurate, or (b) believing something irrational. What should such an agent do? She will not aim for (a), since having high confidence that P is too close to having high confidence that a belief that P is accurate. So, of course, she will aim for (b): she'll aim for accuracy over rationality."

Although the Dilemma views manage somehow to preserve all three desiderata involved in the puzzle of higher-order evidence, they present an unusual picture of rationality. As Ye (2022) points out, the Dilemma theorists might underappreciate the amount of higher-order evidence we acquire in everyday lives. If we try to generalize the view, it thus turns out that we are very frequently engaged in such rational binds and dilemmas.

### 3.4 Calibrationist views

Calibrationist views argue that, in cases such as the **Experiment,** higher-order evidence overrides or defeats first-order evidence (Ye, 2022). Specifically, in such cases, our beliefs should be revised to reflect the bearing of our higher-order evidence.

These views are primarily motivated by an appeal to intuitions in cases such as **Experiment.** To recall, in **Experiment**, Max, the scientist, received higher-order evidence about her unreliability in conducting experiments. To many involved in this debate, it seems intuitive that this evidence should somehow impact Max's belief about the experiment's results. Building on this intuition, defenders of Calibrationism argue that such higher-order evidence defeats the rational import of our first-order evidence. Specifically, this defeat works by "bracketing" our original first-order evidence. The thought here is that higher-order evidence presents us with evidence that either our epistemic performance or evidential situation is in some way problematic. Since we do not have perfect epistemic access and cannot easily know exactly what has gone wrong, the best policy is to bracket the suspect reasoning and evidence.

On an alternative but connected picture, higher-order defeat acts by evidence disposition: when we get higher-order evidence, we no longer possess the original first-order evidence (González de Prado 2020). As Ye summarizes, it is usually the case that for a proposition to serve as our evidence, we must satisfy some condition with regard to this proposition. For example, De Prado (2020) defends the following condition:

> *Competence:* If an agent is not in a position to competently treat that p as evidence that q, she does not possess that p as evidence that q.

If an agent receives higher-order evidence of unreliable reasoning from *p*, then this agent can be seen as violating *Competence*. Consequently, under the right understanding of evidence, higher-order evidence can be seen as dispossessing agents of evidence.

Regardless of the exact understanding of higher-order defeat, defenders of Calibrationism argue that it forces us to adopt the belief that it would be rational for us to have independently of the bracketed (or dispossessed) first-order evidence. Exactly which belief this is is a contentious matter. Schoenfield (2015a), for example, presents this simple principle:

> *Calibrationism*:[6] If, independently of the first-order reasoning in question, your expected degree of reliability concerning whether *p* at time t is *r*, *r* is the credence that it is rational for you to adopt at t.

Ye (2022) presents a more elaborate picture, which she calls Evidence-Discounting Calibrationism. In contrast to *Calibrationism*, Ye's account does not require us to calibrate our credences. Rather, it states that we should calibrate the degree to which we rely on our first-order evidence in forming these credences. In a principle form:

> *Evidence-Discounting Calibrationism:* If, independently of the first order reasoning in question, your expected degree of reliability concerning whether *p* at time t is *r*, the degree to which you rely on your first-order evidence in forming a credence in *p* that is rational for you to adopt at t should cohere with *r*.

[6] This simple model is also sometimes called the "Thermometer Model" (White, 2009).

I will not go into details about Ye's account here but let us at least clarify what she means by "the degree to which one relies on one's first-order evidence." The idea here is that in forming the new credence, one should aggregate two credences: one that would be rational if one would take first-order evidence into account and the other that would be rational if one would ignore this evidence. The expected reliability then determines how the two credences should be aggregated. If the expected reliability of an agent equals 1, an agent is perfectly reliable. Consequently, all the weight should be given to evidence-based credence. If, on the other hand, the expected reliability of an agent is 0.5, the agent is completely unreliable. Consequently, all the weight should be given to evidence-ignoring credence. Formally: $C_1(H) = xC_0(H|E) + (1-x)C_0(H)$.

Despite their intuitive appeal, the Calibrationist views of higher-order evidence are not uncontroversial. For example, they seem to force us to ignore our evidence (Kelly, 2010; Eder and Brössel, 2019). This is epistemically suspect, especially in cases when we receive misleading higher-order evidence, such as in the **Experiment**. Horowitz (2019) additionally argues that we can know a priori that higher-order evidence will often be misleading, which puts added pressure on Calibrationist views. Additionally, both Christensen (2010) and White (2009) have argued that principles like *Calibrationism* conflict with *Conditionalization*:

> *Conditionalization*: When getting new evidence, one's new credence in a proposition should match one's old credence conditional on the evidence.

Since *Conditionalization* is fundamental to Bayesian epistemology (Lin, 2023) – the dominant framework of dealing with graded beliefs – this is often seen as problematic for Calibrationist views.

## 4 Higher-Order Evidence and Science

The previous section overviewed the existing positions on the role of higher-order evidence in our epistemic lives. As shown, all positions come with both upsides and problems. In what follows, I will try to put some additional pressure on the views that deny the import of higher-order evidence (claim (2) from the above summary of the puzzle). Specifically, I will present two cases taken from scientific practice in which higher-order evidence is thought to play an important epistemic role. I will

argue that Steadfastness and Level-Splitting commit themselves to the view that scientists, in these cases, act irrationally.

Consider the following case:

**Review**: Based on anecdotal data and some observational studies, which all show this, Max believes that taking medicine M has a side effect p. This is the correct assessment of where the evidence is pointing. Max then reads a systematic review of evidence about this side effect of M. The review notices that, indeed, all available evidence agrees that the incidence of p is, on average, higher in people who also take M. However, all evidence is of extremely low quality. Specifically, the review notes that because of the low quality of evidence, "the true effect is likely to be substantially different from the estimate of effect," that is, from p.

Anecdotal data and the studies present Max's first-order evidence about p, the side effect of treatment M. On the other hand, she has some evidence about this first-order evidence – the systematic review. This review presents no new evidence about the proposition "p is a side effect of M". Rather, it simply evaluates the evidence that's already available to Max. If we understand higher-order evidence as "evidence which bears on a believer's rational capacities, epistemic performance, or evidential situation" (Horowitz, 2022), the review seems to be a clear example of such evidence.

Situations like the **Review** are not uncommon in science. In 2011, the UN's International Agency for Research on Cancer classified radiofrequency electromagnetic fields – the type of radiation emitted by mobile phones and other electronic devices – as possibly carcinogenic to humans. This decision was made based on some early observational studies. However, in 2024, a large systematic review of all high-quality observational studies, which looked at the available evidence and did not present any new data, concluded that "exposure to RF from mobile phone use likely does not increase the risk of brain cancer" (Karipidis et al., 2024).

Let us look at another case:

**Significance**. Max was tasked with analysing data from experiments that looked at the efficiency of a new compound, *c**, in reducing the growth of e-coli bacteria. She ran the analysis and found that compared to control, *c*, on average,* reduced the growth of e-coli by 20 %. Based on this result, she concluded that *c** reduces the growth of e-coli". Assume that this is correct: *c** indeed reduces growth. Then she remembered an important thing from her training: she should run a null-hypothesis significance test to check whether the difference in the experimental and control group is statistically significant. In essence, these tests tell us how likely we would get the observed result (in this case, the difference in growth of bacteria in the presence of a compound c*) if the null hypothesis is true – meaning that there is actually no effect (in this case, that there is actually no difference between groups in growth of bacteria and the observed difference is due to chance[7]). Max ran the test and calculated the *p*-value of 0.2. She remembered that publications in her field usually require a value of 0.05 or lower to accept the results as plausible.

As in the **Review** example, Max has two kinds of evidence at play. The experimental data presents her first-order evidence about the effects of *c**. On the other hand, the results of the null hypothesis test represent her higher-order evidence. The additional test presented her with no new evidence directly bearing on the effects of *c**. Rather, it can be understood as evidence about the reliability of her first-order evidence. Like systematic reviews, null hypothesis testing is central to contemporary empirical sciences, with a p-value of 0.05 usually taken as the threshold for "significance" (Nuzzo, 2014; "Points of Significance", 2023).

**Review** and **Significance** thus both present cases in which higher-order evidence is thought to play an important role in belief formation. However, according to Level-Splitting and Steadfast views of higher-order evidence, taking the results of the systematic review in the **Review** or the significance test in the **Significance** seriously would be a mistake. Both the systematic review and the significance test are misleading since Max's evidence in both cases on balance still supports her original belief. Steadfasters should thus argue that Max should simply ignore her

[7] I am grateful to the anonymous reviewer for helping me clarify the notion of statistical significance here.

higher-order evidence in both cases. Level-Splitters, conversely, could concede that she could become more uncertain that her beliefs in the two cases are rational; nevertheless, they would insist that she can continue to hold these beliefs.

I find these conclusions quite problematic. They imply that scientists who take systematic reviews and p-values seriously act in an irrational manner. That is a strong commitment. It is not enough in itself to reject Steadfastness or Level-Splitting – we could also understand it as an exciting contribution of epistemology to science. However, it puts additional pressure on these views to strengthen their position. There are at least two objections that defenders of Steadfastness or Level-Splitting could make to defend themselves against this charge. Before concluding, I will review these objections and argue that they fail.

### 4.1 Objection 1: Undercutting rather than Higher-Order Defeat

Defenders of Steadfastness and Level-Splitting could argue that the **Review** and the **Significance** cases are dissimilar to other cases of higher-order evidence and that, consequently, their views do not apply to them. The argument goes as follows. In the **Experiment**, Max received higher-order evidence of her own cognitive failings – she learned about *her own* unreliability. In the **Review** and the **Significance**, on the other hand, she received higher-order evidence directly about *her evidence*. The systematic review and the significance test affected her evidence in a much more direct way than reading a study about her unreliability: while the study suggested that she might have misunderstood her evidence, the review and the significance test showed that what she thought counted as her evidence in the given situation was *actually not evidence at all*.

To put this in more technical terms, defenders of Level-Splitting and Steadfastness could argue that the **Review** and the **Significance** are not examples of defeat by higher-order evidence but examples of a defeat of a much more straightforward kind, that is, of undercutting defeat. In short, d is an undercutting defeat for S's belief that p if and only if d is a reason for S to believe that her reasons for believing p are inadequate (Graham and Lyons, 2021).[8] A classic example of undercutting defeat is learning that a red light is shining on a table that looks red: if sensory

[8] This notion is originally due to Pollock (1986).

information about the red table seems to justify the belief that the table is red, the new information about the red light "undercuts" this justification. Both the results of the systematic review and of the null hypothesis significance test can be seen as basically analogous to such red lighting: if Max's results first seemed to justify her beliefs, the results of the review and the test acted as reasons to believe that this justification is inadequate. Since Level-Splitting and Steadfastness do not reject the role of undercutting defeat, they do not claim that scientists should not take systematic reviews and null hypothesis tests seriously.

This objection points to a real ambiguity in defining higher-order defeat and separating it from other kinds of defeat. The difference between various kinds of defeaters is a complex issue that cannot be resolved here. However, I still think the situations in the **Significance** and the **Review** cases more closely resemble the situation in the **Experiment** than cases of normal rebutting defeat. As we saw above, cases of higher-order evidence are problematic because they, on balance, support both beliefs that "p" and "it would be irrational for me to believe that p". That is also the case in **Significance** and **Review**: in both cases, Max's evidence still supports her first-order beliefs. At the same time, she acquired some evidence that her evidence is unreliable and possibly misleading. Thus, her total evidence in both cases supports her first order belief and some degree of higher-order uncertainty.

But this is not the case in the example with the seemingly red table and the coloured lighting. In that example, one's total evidence stops supporting the belief that "The table is red." Rather, the same sensory information now supports a different belief, perhaps "The table is coloured by the lighting," rather than a combination of beliefs "the table is red" and "it would be irrational for me to believe that it's red". In any case, even if we accept that **Significance** and **Review** are indeed cases in which higher-order evidence acts as a simple undercutting defeater, it still rests on Level-Splitters and Steadfasters to explain why their position does not extend to this kind of defeat.

### 4.2 Objection 2: Wrong Conclusion

The second objection states that we are drawing the wrong conclusions about **Significance** and **Review**. Rather than understanding these cases as uncontroversial examples of ordinary scientific practice, we should see them as highly controversial.

The fact that Level-Splitting and Steadfastness suggest the same should thus be seen as an interesting upshot of these views.

The fact is that the role of systematic reviews and null hypothesis significance testing in science is controversial. Significance testing is often misunderstood (Wasserstein and Lazar, 2016), and the value of systematic reviews is often overstated (Uttley et al., 2023). However, most critics would shy away from recommending completely abandoning these practices. Quite the opposite, most often they suggest a more stringent and even stricter adherence to these practices (Ritchie 2020). The suggestion that scientists should simply ignore higher-order evidence such as systematic reviews or p-values is, as things stand, an interesting but still highly controversial claim.

## 5 Conclusion

This paper presented the problem of higher-order evidence, reviewed the main positions in the debate, and provided a new argument against the views that reject the epistemic importance of such evidence. Section 2 presented the problem of higher-order evidence, and Section 3 overviewed the main positions on it. In Section 4, I presented two novel cases of how higher-order evidence, taken from the scientific practice. I showed that at least two positions in the higher-order evidence debate give highly controversial answers in these cases. Although this is not reason enough to reject these views, it puts additional pressure on them to strengthen their position. The second part of Section 4 discussed and rejected two arguments against the admissibility of these cases.

**References**

Christensen, D. (2010). Higher-Order Evidence. *Philosophy and Phenomenological Research, 81* (1), 185–215. https://doi.org/10.1111/j.1933-1592.2010.00366.x

Christensen, D. (2016a). Conciliation, Uniqueness and Rational Toxicity. *Noûs 50*(3), 584–603. https://doi.org/10.1111/nous.12077

Christensen, D. (2016b). Disagreement, Drugs, Etc.: From Accuracy to Akrasia. *Episteme, 13*(4), 397–422. https://doi.org/10.1017/EPI.2016.20

Christensen, D. (2020). Akratic (Epistemic) Modesty. *Philosophical Studies*, *178*(7), 2191–2214. https://doi.org/10.1007/S11098-020-01536-6

DiPaolo, J. (2019). Second Best Epistemology: Fallibility and Normativity. *Philosophical Studies, 176*(8), 2043–66. https://doi.org/10.1007/S11098-018-1110-Y/METRICS

Dorst, K. (2020). Evidence: A Guide for the Uncertain. *Philosophy and Phenomenological Research, 100*(3), 586–632. https://doi.org/10.1111/PHPR.12561

Dorst, K. (2024). Higher-Order Evidence. In M. Lasonen-Aarnio and C. Littlejohn (Ed.), *The Routledge Handbook of the Philosophy of Evidence* (pp. 176–194). Routledge.

Eder, A. A. and P. Brössel. (2019). Evidence of Evidence as Higher-Order Evidence. In *Higher-Order Evidence* (pp. 62–83). Oxford University Press. https://doi.org/10.1093/oso/9780198829775.003.0003

Elga, A. (2010). "How to Disagree About How to Disagree." In R. Feldman and T. A. Warfield (Ed.), *Disagreement* (pp. 175–186). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199226078.001.0001

Feldman, R. (2009). Evidentialism, Higher-Order Evidence, and Disagreement. *Episteme, 6*(3), 294–312. https://doi.org/10.3366/E1742360009000720

Field, C. (2019). It's OK to Make Mistakes: Against the Fixed Point Thesis. *Episteme, 16*(2), 175–85. https://doi.org/10.1017/EPI.2017.33

González de Prado, J. (2020). Dispossessing Defeat. *Philosophy and Phenomenological Research, 101*(2), 323–40. https://doi.org/10.1111/phpr.12593

Graham, P. J. J. C. Lyons. (2021). The Structure of Defeat. In *Reasons, Justification, and Defeat* (39–68). Oxford University Press. https://doi.org/10.1093/oso/9780198847205.003.0003

Henderson, L. (2022). Higher-order Evidence and Losing One's Conviction. *Noûs, 56*(3), 513–29. https://doi.org/10.1111/nous.12367

Horowitz, S. (2014). Epistemic Akrasia. *Noûs, 48*(4), 718–44. https://doi.org/10.1111/nous.12026

Horowitz, S. (2019). Predictably Misleading Evidence. In M. Skupper and A. Steglich-Petersen, (Ed.) *Higher-Order Evidence: New Essays* (pp. 105–23). Oxford University Press. https://doi.org/10.1093/OSO/9780198829775.003.0005

Horowitz, S. (2022). Higher-Order Evidence. In E. N. Zalta and U. Nodelman (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2022 Edition)*. The Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/higher-order-evidence/

Kappel, K. (2019). Escaping the Akratic Trilemma. In M. Skupper and A. Steglich-Petersen (Ed.), *Higher-Order Evidence: New Essays* (pp. 124–43). Oxford University Press. https://doi.org/10.1093/oso/9780198829775.003.0006

Karipidis, K, D. Baaken, T. Loney, M. Blettner, C. Brzozek, M. Elwood, C. Narh et al. (2024). The Effect of Exposure to Radiofrequency Fields on Cancer Risk in the General and Working Population: A Systematic Review of Human Observational Studies – Part I: Most Researched Outcomes. *Environment International 191* (September), 108983. https://doi.org/10.1016/J.ENVINT.2024.108983

Kelly, T. (2005). The Epistemic Significance of Disagreement. In T. S. Gendler and J. Hawthorne (Ed.), *Oxford Studies in Epistemology, Volume 1* (pp. 167–196). Oxford University Press.

Kelly, T. 2010. Peer Disagreement and Higher-Order Evidence. In R. Feldman and T. A. Warfield (Ed.), *Disagreement* (111–74). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199226078.003.0007

Lasonen-Aarnio, M. (2013). Disagreement and Evidential Attenuation. *Noûs, 47*(4), 767–94. https://doi.org/10.1111/NOUS.12050

Lasonen-Aarnio, M. (2014). Higher-Order Evidence and the Limits of Defeat. *Philosophy and Phenomenological Research, 88*(2), 314–45. https://doi.org/10.1111/PHPR.12090

Lin, H. (2023). Bayesian Epistemology. In E. N. Zalta and U. Nodelman (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2023 Edition)*. The Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2023/entries/epistemology-bayesian/

Littlejohn, C. (2018). Stop Making Sense? On a Puzzle about Rationality. *Philosophy and Phenomenological Research, 96*(2), 257–72. https://doi.org/10.1111/phpr.12271

Nuzzo, R. (2014). Scientific Method: Statistical Errors. *Nature, 506*(7487), 150–52. https://doi.org/10.1038/506150A.

Points of Significance. (2023). *Nature Human Behaviour, 7*, 293–94. https://doi.org/10.1038/s41562-023-01586-w

Pollock, J. L. (1986). *Contemporary Theories of Knowledge*. Hutchinson .

Ritchie, S. (2020). *Science Fictions. Exposing Fraud, Bias, Negligence and Hype in Science*. The Bodley Head

Schoenfield, M. (2015a). A Dilemma for Calibrationism. *Philosophy and Phenomenological Research, 91*(2), 425–55. https://doi.org/10.1111/phpr.12125

Schoenfield, M. 2015b. Bridging Rationality and Accuracy. *Journal of Philosophy, 112*(12), 633–57. https://doi.org/10.5840/JPHIL20151121242

Skipper, M. (2021). Does Rationality Demand Higher-Order Certainty? *Synthese, 198*(12), 11561–85. https://doi.org/10.1007/s11229-020-02814-w

Sliwa, P and S. Horowitz. 2015. Respecting All the Evidence. *Philosophical Studies, 172*(11), 2835–58. https://doi.org/10.1007/s11098-015-0446-9

Tal, E. (2021). Is Higher-Order Evidence Evidence? *Philosophical Studies, 178*(10), 3157–75. https://doi.org/10.1007/s11098-020-01574-0

Titelbaum, M. G. (2015). Rationality's Fixed Point (or: In Defense of Right Reason). In T. S. Gendler and J. Hawthorne (Ed.), *Oxford Studies in Epistemology, Volume 5* (pp. 253–94). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198722762.003.0009

Uttley, L., D. S. Quintana, P. Montgomery, C. Carroll, M. J. Page, L. Falzon, A. Sutton and D. Moher. (2023). The Problems with Systematic Reviews: A Living Systematic Review. *Journal of Clinical Epidemiology, 156*, 30–41. https://doi.org/10.1016/J.JCLINEPI.2023.01.011

Wasserstein, R. L., and N. A. Lazar. (2016). The ASA Statement on P-Values: Context, Process, and Purpose. *The American Statistician,* 70(2), 129–33. https://doi.org/10.1080/00031305.2016.1154108

White, R. (2009). On Treating Oneself and Others as Thermometers. *Episteme, 6*(3), 233–50. https://doi.org/10.3366/E1742360009000689

Whiting, D. (2021). Higher-Order Evidence. *Analysis, 80*(4), 789–807. https://doi.org/10.1093/ANALYS/ANAA056

Worsnip, A. (2018). The Conflict of Evidence and Coherence. *Philosophy and Phenomenological Research, 96*(1), 3–44. https://doi.org/10.1111/phpr.12246

Ye, R. (2022). *Higher-Order Evidence and Calibrationism.* Cambridge University Press. https://doi.org/10.1017/9781009127332

**ANALIZA**
Let. 28, št. 2, str. 263–284, december 2024

ANαliZA

# CILJI OMEJENE EPISTEMSKE RACIONALNOSTI

**Sprejeto**
28. 9. 2024

**Pregledano**
1. 12. 2024

**Izdano**
31. 12. 2024

NASTJA TOMAT
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
nastja.tomat@ff.uni-lj.si

DOPISNI AVTOR
nastja.tomat@ff.uni-lj.si

**Izvleček** Pojmovanja racionalnosti znotraj epistemologije se pogosto naslanjajo na idealizirane modele epistemskih agentov in okolja, v katerem ti delujejo. Omejena epistemska racionalnost je koncept, ki se takšnim idealizacijam poskuša izogniti in se pri postavljanju norm raziskovanja naslanja na empirične podatke o človeški kogniciji in epistemskem okolju. Izhaja iz Simonove omejene racionalnosti, Gigerenzerjeve ekološke racionalnosti in McKennove ne-idealne epistemologije ter deluje kot hibriden koncept, ki vsebuje tako normativne kot deskriptivne elemente in služi kot izhodišče za ponujanje epistemskih vodil, ki bodo pripomogle k izboljševanju naše epistemske situacije in doseganju epistemskih ciljev. Znotraj epistemologije zasledimo različne epistemske cilje, na primer oblikovanje resničnih prepričanj, upravičenje, znanje in razumevanje. V prispevku povzemam izbrane vidike razprave o epistemskih ciljih in vrednotah s poudarkom na veritističnem monizmu in prevprašujem, katere cilje bi morala zasledovati omejena epistemska racionalnost. Zaključujem, da lahko omejena epistemska racionalnost zasleduje različne vrste kognitivnega uspeha, ne zgolj resnice, vendar se omejuje na raziskovanje o tematikah, ki so za posameznika na nek način relevantne.

**Ključne besede**
epistemska racionalnost, omejena racionalnost, ne-idealna epistemologija, epistemski cilji, epistemske vrednosti

https://doi.org/10.18690/analiza.28.2.263-284.2024


Univerzitetna založba Univerze v Mariboru

AN$_\alpha$LIZA

**ANALIZA**
Vol. 28, no. 2, pp. 263–284, December 2024

# THE GOALS OF BOUNDED EPISTEMIC RATIONALITY

NASTJA TOMAT
University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia
nastja.tomat@ff.uni-lj.si

CORRESPONDING AUTHOR
nastja.tomat@ff.uni-lj.si

**Accepted**
28. 9. 2024

**Revised**
1. 12. 2024

**Published**
31. 12. 2024

**Abstract** Notions of rationality in epistemology are often based on idealised models of epistemic agents and the environment in which they are embedded. The concept of bounded epistemic rationality aims to avoid such idealizations and proposes norms of inquiry based on empirical data about the limitations of human cognition and the features of the epistemic environment. Drawing on Simon's work on bounded rationality, Gigerenzer's notion of ecological rationality, and McKenna's programme of non-ideal epistemology, it functions as a hybrid concept that contains both normative and descriptive elements and offers epistemic advice that helps to improve our epistemic position and achieve our epistemic goals. Traditionally, epistemic goals include forming true beliefs and avoiding false beliefs, justification, knowledge, and understanding. In this paper, I summarise selected aspects of the debate about epistemic values and goals, focusing on veritistic monism, and reflect on the goals of bounded epistemic rationality. I conclude that bounded epistemic strives towards various forms of cognitive success, not only towards the truth, but limits rational inquiry to questions relevant to the epistemic agent.

**Keywords**
epistemic rationality,
bounded rationality,
non-ideal epistemology,
epistemic goals,
epistemic values

https://doi.org/10.18690/analiza.28.2.263-284.2024


University of Maribor Press

## 1 Uvod

Vprašanje, kaj pomeni biti epistemsko racionalen, je ena od osrednjih tematik epistemologije. Epistemska racionalnost se nanaša na epistemske naravnanosti, stanja in procese, kot so oblikovanje prepričanj, sklepanje, presojanje, načrtovanje in preudarjanje. V literaturi pogosto zasledimo razlikovanje med praktično in teoretično racionalnostjo. Medtem ko se praktična racionalnost primarno nanaša na aktivnosti, ki so na takšen ali drugačen način povezane z vedenjem – odločanje, načrtovanje, usmerjanje in izvedbo dejanj – se teoretična ali epistemska racionalnost primarno osredotoča na proces spoznavanja in na racionalnost doksastičnih stanj, predvsem prepričanj. Epistemsko racionalnost od praktične loči prav njena usmerjenost h kognitivnim ali epistemskim ciljem (Knauff in Spohn, 2021).

Tradicionalna analitična epistemologija se pri naslavljanju epistemoloških vprašanj prvega reda, kot je vprašanje o epistemski racionalnosti, pogosto naslanja na idealizirane modele epistemskih agentov, njihovih spoznavnih sposobnosti, interakcij med njimi in okolja, v katerem spoznavajo. Čeprav je uporaba idealizacij, poenostavitev in približkov neizogiben del tako znanstvenega kot filozofskega raziskovanja, lahko vodi v različne težave. Pretirano naslanjanje na idealizirane modele lahko pripomore k oblikovanju epistemskih norm in standardov, ki bodo za običajne epistemske agente nedosegljivi, poleg tega pa ne bodo pripomogli k izboljševanju našega epistemskega položaja (McKenna, 2023). Teorije epistemske racionalnosti so torej lahko pretirano idealizirane in psihološko nerealistične ter kot takšne ne morejo služiti kot epistemska vodila za izboljševanje našega raziskovanja in doseganje epistemskih ciljev.

Medtem ko je na področju praktične racionalnosti splošno sprejeto, da naše fizične omejitve vplivajo na to, katera dejanja bomo od agentov pričakovali, zahtevali in označili za racionalne, na področju preučevanja epistemske racionalnosti temu pogosto ni tako (Thorstad, 2024b). Zdi se, da so številna pojmovanja epistemske racionalnosti in norm racionalnosti še vedno precej oddaljena od empiričnih spoznanj o človeških kognitivnih procesih. Izhajajoč iz kritike idealiziranih teorij racionalnosti je Herbert A. Simon že v petdesetih letih prejšnjega stoletja predstavil koncept omejene racionalnosti (Simon, 1955, 1956, 1976, 1990, 1992). Zagovarjal je, da je treba racionalnost pojmovati na način, ki bo skladen z omejitvami človekovega kognitivnega sistema ter strukturo okolja, v katerem delujemo. Čeprav je Simonova omejena racionalnost pomembno vplivala na razumevanje racionalnosti znotraj

številnih področij, na primer vedenjskih znanosti, psihologije in ekonomije, se znotraj epistemologije nanjo naslanjajo le redki avtorji (Cherniak, 1986; Pils, 2022; Thorstad, 2022, 2024b, 2024a). Filozofski vidiki omejene racionalnosti torej še niso bili sistematično preučevani. Takšno preučevanje bi zajemalo pojasnitev in natančno opredelitev predpostavk koncepta omejene racionalnosti, na primer tega, na kakšen način so vanj vključeni normativni in deskriptivni elementi oziroma ali je to primarno normativen ali deskriptiven koncept, ter tega, na kakšen način se razlikuje od neomejene racionalnosti; ugotavljanje, na katerih področjih filozofije je omejena racionalnost uporabna in na kakšen način; ter naslavljanje vprašanj o teoretski in metodološki ustreznosti koncepta znotraj različnih disciplin, od filozofije do psihologije in ekonomije, ter njegov potencial za interdisciplinarno uporabo (Sturm, 2020).

V prispevku bom predstavila pojem omejene epistemske racionalnosti kot hibriden, ne-idealen koncept, ki vsebuje normativne elemente, hkrati pa upošteva empirična spoznanja o omejitvah človeške kognicije in o lastnostih epistemskega okolja, v katerem ljudje spoznavamo, ter se kot takšen izogne nekaterim težavam idealiziranih pojmovanj epistemske racionalnosti. Poglavitni namen prispevka je nasloviti vprašanje o ciljih omejene epistemske racionalnosti. Znotraj epistemologije poteka obširna razprava o tem, čemu bi morali pripisovati epistemsko vrednost in kaj bi morali smatrati kot epistemske cilje. Kot epistemski cilji se najpogosteje pojavljajo resnica, upravičenje, znanje in razumevanje; v teorijah omejene racionalnosti pa igra pomembno vlogo tudi ideja, da se zna organizem ustrezno prilagoditi na okolje. Pojavlja se torej vprašanje, kakšne cilje naj bi zasledovala omejena epistemska racionalnost, ki združuje tako elemente klasičnih pojmovanj epistemske racionalnosti znotraj epistemologije kot elemente omejene in ekološke racionalnosti.

Najprej bom prestavila program ne-idealne epistemologije, ki ga je razvil Robin Mckenna (2023) in ki predstavlja okvir za naslavljanje vprašanj o omejeni epistemski racionalnosti. Nato bom opisala Simonov koncept omejene racionalnosti in iz njega izhajajoč pojem ekološke racionalnosti, ki so ga razvili Gerd Gigenerzer in sodelavci. V drugem delu bom orisala koncept omejene epistemske racionalnosti, ki se naslanja na ne-idealno epistemologijo ter omejeno in ekološko racionalnost. Nazadnje bom povzela izbrane vidike razprave o epistemskih ciljih s poudarkom na cilju resnice ter se posvetila vprašanju o ciljih omejene epistemske racionalnosti.

## 2 Ne-idealna epistemologija

Robin McKenna (2023) piše, da se tradicionalna analitična epistemologija pri naslavljanju epistemoloških vprašanj, kot so vprašanja o upravičenju, znanju in raziskovanju (ang. *inquiry*), pogosto naslanja na idealizacije. Idealizacij je več vrst: lahko se nanašajo na epistemske agente, predvsem na zanemarjanje omejitev njihovih kognitivnih kapacitet; na interakcijo med epistemskimi agenti; na socialne institucije, na primer precenjevanje kapacitet ustanov, ki ustvarjajo in razširjajo znanje; in na epistemsko okolje, v katerem živimo, na primer podcenjevanje stopnje napačnih informacij, s katerimi smo soočeni. Ne-idealna epistemologija je pristop k naslavljanju epistemoloških vprašanj, ki se takšnim idealizacijam poskuša izogibati. McKenna poudarja, da je naslanjanje na poenostavitve in idealizirane modele pomemben in neizogiben del znanstvenega raziskovanja, vendar meni, da pretirano in nekritično naslanjanje na idealizacije lahko vodi do tega, da si nekaterih relevantnih raziskovalnih vprašanj sploh ne zastavljamo in zato ovira naše razumevanje pojavov. Kot primer navaja pristop k preučevanju znanstvenih institucij, kjer njihovo dejansko delovanje primerjamo z idealnim modelom in ocenjujemo, kje in v kolikšni meri se od tega ideala odklanjajo – vendar pa nam takšen pristop onemogoča, da bi razumeli in razložili, kako institucije delujejo v resničnem svetu. McKenna poudarja, da normativnih vprašanj ne moremo nadomestiti z deskriptivnimi, vendar meni, da je treba pri razlaganju pojavov izhajati iz dejanskega stanja stvari, ne pa iz idealiziranih modelov, ki se jim poskušamo približati. Meni, da idealizacije same po sebi niso nujno problematične, vendar je treba ugotoviti, katere idealizacije so pri katerih raziskovalnih problemih dopustne in katere ne. Idealen in ne-idealen pristop bosta tako vodila v različne epistemske norme in norme raziskovanja. Za primer lahko vzamemo nestrinjanje med vrstniki (ang. *peer disagreement*). Lahko se naslanjamo na idealizirane modele epistemskih agentov in interakcije med njima, kjer predpostavimo, da je primarna motivacija obeh agentov izboljšanje razumevanja tematike, o kateri poteka interakcija, pridobivanje resničnega prepričanja ali nek drug epistemski cilj. Oba imata približno enako znanje o tematiki, razvijata lastne argumente in poslušata argumente drugega. V resničnem življenju pa epistemski vrstniki pogosto niso enako dobro informirani, v razprave pa ne vstopajo le z namenom pridobivanja resničnih prepričanj ali drugih epistemskih ciljev. Norme, ki nam narekujejo, kako postopati v primeru nestrinjanja z vrstnikom – koliko časa in na kakšen način sodelovati v razpravi ter kako in če sploh spremeniti stopnjo zaupanja v svoje prepričanje – bodo drugačne, če izhajamo iz idealiziranih ali ne-idealiziranih modelov. V prvem primeru bi bilo za agente smiselno, da se udeležujejo

razprave in do določene mere posodobijo svoje prepričanje, v drugem pa temu ni nujno tako, sploh če vsaj eden od agentov v razpravi nima primarno epistemskih ciljev, temveč želi, na primer, le intelektualno nadvladati drugega. McKenna meni, da primarna težava norm, ki jih predlaga idealna epistemologija, ni dejstvo, da so za običajne, resnične epistemske agente pogosto nedosegljive, temveč da nam ne pomagajo pri izboljševanju naše epistemske situacije in torej služijo kot slaba epistemska vodila. Če spoznavamo po normah tradicionalne, idealizirane analitične epistemologije, v številnih situacijah tvegamo, da bomo svoj epistemski položaj še poslabšali – da ne bomo dosegli epistemskih ciljev (McKenna, 2023).

McKennova ne-idealna epistemologija upošteva empirična spoznanja psihologije, zgodovine, sociologije in drugih ved o tem, kako potekajo naši kognitivni procesi, kakšne so značilnosti socialnih institucij, na kakšen način socialna situiranost (npr. socialna vloga in identiteta ter razmerja moči) in značilnosti epistemskega okolja vplivajo na naše raziskovanje ter na naše epistemske obveznosti ter odgovornosti. Medtem ko je McKennov projekt eksplicitno etičen in političen, osredotočen na značilnosti socialnih institucij ter socialne situiranosti agentov ter na pojave, kot so epistemsko izključevanje in zatiranje, so za koncept omejene epistemske racionalnosti relevantni predvsem trije vidiki programa. Prvi je, da omejena epistemska racionalnost temelji na ne-idealiziranih modelih – za razliko od McKenne, ki se osredotoča na socialne in politične vidike epistemskih vprašanj, omejena epistemska racionalnost jemlje v zakup predvsem idealizacije kognitivnih sposobnosti epistemskih agentov in epistemskega okolja, v katerem delujejo. Drugi je, da omejena epistemska racionalnost ponuja norme raziskovanja (ang. *norms of inquiry*), ki upoštevajo empirična spoznanja o dometu in omejitvah človeške kognicije in kot takšne pripomorejo k doseganju izbranih epistemskih ciljev ter lahko služijo kot vodila za dobro spoznavanje. Tretji je, da omejena epistemska racionalnost, v skladu s programom ne-idealne epistemologije, zajema tako normativna vprašanja o tem, kako bi ljudje morali spoznavati, raziskovati, oblikovati in posodabljati prepričanja, kot tudi deskriptivna, empirična spoznanja o naši kogniciji in epistemskem okolju.

## 3 Omejena in ekološka racionalnost

Koncept omejene racionalnosti se je pojavil kot kritika ne-omejenih, idealiziranih pogledov na racionalnost, ki so zlasti v ekonomiji in ki so kot standard racionalnosti postavljali sledenje aksiomom teorije odločanja. Takšne teorije racionalnosti so

predpostavljale, da imajo agenti popoln in urejen set preferenc, poznavanje vseh možnih alternativ in znanje o tem, katera alternativa bo s kakšno verjetnostjo vodila do katerega izida (Neumann in Morgenstern, 1944). Agenti, ki nastopajo v takšnih teorijah, so opremljeni s kognitivnim aparatom, ki jim omogoča izjemno kompleksno procesiranje informacij, hkrati pa imajo na voljo vse relevantne informacije iz okolja, na podlagi katerih lahko pridejo do optimalne odločitve. Simon je menil, da je potrebno takšen pogled na racionalnost nadomestiti s pojmom racionalnosti, ki je kompatibilen z dostopom do informacij in kognitivnimi kapacitetami, ki jih ima človek v lastnem okolju v resničnem svetu. Uporabil je metaforo škarij, pri katerih eno rezilo predstavlja kognitivne kapacitete organizma, drugo pa strukturo okolja; da bi razumeli racionalnost, je treba upoštevati obe rezili. Omejena racionalnost ne zahteva optimalnih rešitev, temveč le rešitve, ki so dovolj dobre, kar je Simon poimenoval *satisficing*, poleg tega pa se ne osredotoča le na končni izid, na primer odločitev, temveč na proces, ki je do izida privedel. Simonova omejena racionalnost je torej proceduralne narave (Simon, 1955, 1956, 1976, 1990, 1992).

Koncept omejene racionalnosti se je od izvirnih Simonovih del do danes razvijal in nadgrajeval ter predstavlja pomembno ogrodje za preučevanje odločanja in racionalnosti (Viale, 2020). Eden od konceptov, ki izvira iz omejene racionalnosti, je ekološka racionalnost, ki jo preučujejo Gigerenzer in sodelavci (Gigerenzer, 2000, 2008; Gigerenzer in Gaissmaier, 2011; Gigerenzer in Todd, 2001). Po tej teoriji je določena strategija za reševanje problemov racionalna v tolikšni meri, kot je prilagojena strukturi naloge. (I)racionalnosti strategij za sklepanje, odločanje, reševanje problemov, oblikovanje prepričanj itn. ne bi smeli presojati glede na ujemanje z a priori sprejetimi normami, kot so sledenje pravilom logike, verjetnosti ali teorije odločanja, temveč glede na to, kako dobro se obnesejo pri specifičnih nalogah. Ko izberemo strategijo, ki v določenem okolju vodi v natančnejše presoje in napovedi v primerjavi z ostalimi strategijami, smo ekološko racionalni. Gigerenzerjev raziskovalni program se osredotoča na preučevanje hitrih in varčnih hevristik, ki so opredeljene kot strategije, ki ignorirajo del informacij z namenom, da odločanje postane natančnejše, hitrejše in varčnejše kot pri uporabi kompleksnejših metod. Ena od ugotovitev njegovega raziskovalnega programa je, da pri nekaterih nalogah enostavne hevristike vodijo do večjega deleža pravilnih presoj kot kompleksnejše strategije, ki upoštevajo večje število informacij in so računsko zahtevnejše. Eden od glavnih ciljev njegovega raziskovanja je preučiti, pri katerih nalogah oziroma v katerih pogojih bodo katere strategije uspešnejše od drugih – pri

čemer kot kriterij uspešnosti uporablja pravilnost presoje. Gigerenzer meni, da hevristike niso a priori iracionalne ter da jih ne bi smeli pojmovati kot ovire, temveč kot orodja mišljenja. Takšen pristop se torej odmika od standardnega pogleda na racionalnost (Stein, 1997) in racionalnost razume skozi interakcijo med kognitivno strategijo ter okoljem.

## 4 Omejena epistemska racionalnost

Omejena epistemska racionalnost[1] je koncept, ki spada v metodološki okvir neidealne epistemologije ter se naslanja na omejeno in ekološko racionalnost. Usmerjena je k epistemskim ciljem in upošteva spoznanja kognitivne psihologije o tem, da imamo ljudje omejeno računsko in napovedno moč, pozornost, delovni spomin in druge kognitivne sposobnosti. Uporablja *ought-implies-can* princip normativnosti (Wedgwood, 2013), ki narekuje, da so kognitivne operacije, način raziskovanja in oblikovanja prepričanj, ki jih zahtevamo od epistemskih agentov, le tista, ki so jih agenti v principu sposobni izvesti. Psihološka spoznanja o dometu in omejitvah človeške kognicije torej služijo kot okvir, znotraj katerega postavlja epistemske norme in norme raziskovanja. Poleg omejitev kognicije upošteva tudi značilnosti epistemskega okolja, v katerem spoznavamo, na primer razmerje med pravilnimi in napačnimi informacijami (Levy, 2021). Omejena epistemska racionalnost se ne osredotoča le na končna doksastična stanja, temveč na proces raziskovanja, kar je skladno tako s Simonovim proceduralnim pogledom na racionalnost kot tudi s tako imenovanim zetetičnim obratom (ang. *zetetic turn*) v epistemologiji. Zetetična epistemologija se za razliko od epistemskih norm, ki določajo pogoje za racionalnost prepričanj in ostalih doksastičnih stanj, osredotoča na norme raziskovanja, ki določajo, kakšen bi moral biti proces raziskovanja – kdaj ga začeti, kako vrednotiti dokaze, kdaj raziskovanje zaključiti in podobno (Friedman, 2019, 2020, 2023; Haziza, 2023; Kelp, 2021; Thorstad, 2021). Kot pri drugih pojmovanjih racionalnosti je tudi pri omejeni epistemski racionalnosti eno od poglavitnih vprašanj, kaj so njeni cilji. V nadaljevanju prispevka bom povzela nekatera vprašanja o epistemskih vrednotah in ciljih, zlasti o cilju resnice, in prevpraševala, h katerim ciljem naj bi bila usmerjena omejena epistemska racionalnost.

[1] Za podrobnejši oris omejene epistemske racionalnosti s poudarkom na odnosu med normativnimi in deskriptivnimi pristopi k preučevanju racionalnosti glej Tomat (2024).

## 5 Epistemske vrednosti in epistemski cilji

Razprava o epistemskih ciljih je neločljivo povezana z razpravo o epistemski vrednosti (ang. *epistemic value)*, ki jo opredeljujemo kot vrednost, pripisano različnim oblikam kognitivnega uspeha, na primer resničnim ali upravičenim prepričanjem, znanju in razumevanju (Bondy, b.d.). V razpravi o epistemskih vrednostih se pojavljajo različna vprašanja; eno od njih je, katere izmed pojavov, ki jim pripisujemo epistemsko vrednost, naj privzamemo kot epistemske cilje ali cilje raziskovanja.

Pogosto se kot primarni ali temeljni epistemski cilj eksplicitno ali implicitno pojavlja cilj resnice (ang. *truth goal*; npr. Alston, 1985; BonJour, 1985; David, 2001, 2013; Foley, 1987; Nozick, 1993; Pritchard, 2019, 2021). Sestavljen je iz dveh delov: imeti resnična prepričanja in ne imeti neresničnih prepričanj. Cilj je usmerjen v sedanjost, kar pomeni, da ocenjujemo le status prepričanj v sedanjosti, ne pa na primer v prihodnosti. Obstajajo številna prepričanja, ki bi na dolgi rok pripomogla k velikemu številu resničnih prepričanj, tudi če so sama neresnična. Poleg tega sta oba dela cilja, tako pridobivanje resničnih kot izogibanje neresničnim prepričanjem, enako pomembna. Le enemu delu bi bilo preprosto zadostiti: če bi želeli oblikovati čim več resničnih prepričanj, bi morali verjeti skoraj vsemu (in s tem tvegati, da poleg resničnih pridobimo tudi veliko število neresničnih prepričanj), če pa bi se želeli izogniti neresničnim prepričanjem, ne bi smeli verjeti skoraj ničemur, kar lahko vodi v skepticizem (Pritchard, 2018). Če želimo doseči cilj resnice, je torej treba oblikovati strategijo, ki bo zagotavljala ustrezno ravnovesje med izčrpnostjo in točnostjo sistema prepričanj, ki ga bomo oblikovali (Foley, 2011). Pozicija, ki zagovarja, da je resnica temeljna epistemska vrednost in da iz nje izvira tudi vrednost ostalih epistemskih ciljev ter pojmov, kot so upravičenje, znanje ali razumevanje, se imenuje veritistični monizem. Povedano drugače, po veritističnem monizmu epistemski cilji nimajo epistemske vrednosti, ki bi bila neodvisna od vrednosti resnice. Pritchard (2014) meni, da je resnica tudi sestavni cilj ustrezno izvedenega intelektualnega raziskovanja. Hkrati poudarja, da je veritistični monizem skladen s tem, da imajo intelektualna raziskovanja tudi številne druge cilje, tako epistemske kot ne-epistemske. Pritchard (2014) tako zagovarja dve medsebojno povezani tezi: da je resnica temeljna epistemska vrednost in sestavni cilj ustrezno izvedenega intelektualnega raziskovanja.

Za vlogo resnice kot temeljne epistemske vrednosti in primarnega epistemskega cilja obstajajo različni argumenti. Eden je, da je težko zanikati instrumentalno vrednost resničnih prepričanj. Resnična prepričanja nam, tudi če jim ne pripišemo končne ali intrinzične vrednosti, pomagajo dosegati različne, tudi ne-epistemske cilje, in to dosegajo ne glede na to, ali so upravičena in del znanja. Poleg tega lahko vrednost drugih epistemskih pojavov, kot je upravičenje, razlagamo na podlagi tega, da pripomorejo k doseganju resničnih prepričanj, obratno pa ne (npr. BonJour, 1985). Nekateri filozofi tudi menijo, da je treba epistemske pojme, kot so racionalnost ali upravičenje, opredeliti oziroma zasidrati v ne-epistemskih pojmih, da ne bi prišli do krožnih opredelitev – in resnica kot semantični koncept lahko igra to vlogo (David, 2001).

Eno od pomembnih vprašanj, ki se pojavlja v razpravi o epistemskih vrednostih, je, zakaj bi *morali* zasledovati ravno to vrednost oziroma cilj. Lahko trdimo, da je imeti resnična prepričanja dobro v etičnem smislu; da naša obveznost stremenja k cilju resnice izvira iz naših moralnih dožnosti in da bi nam moralo biti mar za resnico (Zagzebski, 2020); ali da je posedovanje cilja resnice del polnega, intelektualno izpolnjujočega življenja (Lynch, 2004). Še ena možnost je, da priznavamo, da je resnica kot temeljna epistemska vrednost preprosto določena ali privzeta. V tem primeru je resnica (ali karkoli drugega) kot temeljna epistemska vrednost zgolj predpostavka, iz katere izhajamo in na podlagi katere gradimo epistemske presoje – na primer o tem, ali je prepričanje epistemsko dobro ali slabo. Če to drži, potem ni treba zagovarjati, da so resnična prepričanja zmeraj nekaj, k čemur bi morali stremeti, ali da je to cilj, ki je pomemben vsem ljudem; dovolj je že to, da prepoznamo, da je resnica pomembna dovolj velikemu številu ljudi v dovolj širokem naboru situacij, da se je uveljavila kot temeljni kriterij za domeno epistemskih presoj (Sosa, 2007).

Poteka tudi razprava o tem, ali so epistemski cilji pogojeni s posameznikovimi željami. Pozicija, ki zagovarja, da so, mora utemeljiti protiintuitivno posledico – da so posamezniki brez tega cilja izvzeti iz sodb o racionalnosti. Zdi se, da nam epistemska racionalnost nekaj predpisuje oziroma vsebuje zapovedi (ang. *oughts*) na način, ki ni arbitraren in ki so mu agenti podvrženi ne glede na njihove želje (David, 2001). Foley (1987) na primer meni, da bi tudi v primeru, da bi obstajal posameznik, ki ne bi posedoval cilja resnice, raziskovanje in prepričanja tega posameznika epistemsko ovrednotili na enak način kot raziskovanje in prepričanja posameznika, ki pa ima cilj resnice.

Naslednje vprašanje je, v kakšnem odnosu so epistemski cilji s prilagoditvenimi cilji organizma. Nekateri filozofi (npr. Graham, 2012; Plantinga, 1993) vrednost resnice utemeljujejo na podlagi pravih funkcij. Razlage na podlagi pravih funkcij (ang. *proper functions accounts*) zagovarjajo, da je epistemski status prepričanja odvisen od tega, ali smo prepričanje oblikovali prek pravilno delujočih kognitivnih procesov oziroma struktur. Proces ali struktura P, ki proizvaja učinek U, ima pravo funkcijo, če so predhodniki P prav tako proizvajali U, in je U del razlage za to, da so organizmi, ki premorejo P, preživeli. Za primer lahko vzamemo srce: naši predniki so imeli srca, katerih funkcija je bila črpanje krvi, kar je ljudem omogočalo preživetje in razmnoževanje, in to pomeni, da je prava funkcija srca črpanje krvi. Isti argument lahko uporabimo za kognitivne sisteme: sistemi, ki so vodili do resničnih prepričanj, so del razlage za to, zakaj so naši predniki preživeli. Oblikovanje resničnih prepričanj je torej prava funkcija kognitivnih sistemov, kar pomeni, da imajo resnična prepričanja vsaj neko vrednost. Kljub temu pa potekajo razprave o tem, ali so prave funkcije normativne – ali je to, da je resnično prepričanje prava funkcija naših kognitivnih sistemov, že zadostno za to, da zaključimo, da imajo resnična prepričanja epistemsko vrednost in da bi morali stremeti k njim. Še ena težava utemeljevanja epistemske vrednosti resničnih prepričanj na podlagi prave funkcije je, da očitno obstajajo sistemi, ki sistematično vodijo v napačna prepričanja, pa so se še vedno ohranili (Hazlett, 2013). Podobno David (2001) meni, da je prava funkcija našega kognitivnega sistema zagotavljanje uspešne reprodukcije oziroma razširjanja genetskega materiala, in da temu lahko služijo tako resnična kot neresnična prepričanja. Kvanvig (2013) pa zagovarja, da sta znanje in razumevanje vsaj tako dobra kandidata za pravo funkcijo kognitivnih sistemov kot resnična prepričanja.

Glede na to, da ljudje zasledujemo številne cilje, na primer epistemske, moralne in praktične, se poraja vprašanje, v kakšnem odnosu so epistemski in ne-epistemski cilji. Pogosto se znajdemo v situacijah, kjer si bodo različni cilji nasprotovali, vendar to, da ne-epistemski cilji v določenih situacijah prevladajo nad epistemskimi, še ne pomeni, da so epistemski cilji takrat neobstoječi ali nepomembni – pomeni zgolj to, da niso absolutni (David, 2013). Tudi če sprejemamo veritistični monizem, to ne pomeni, da ne priznavamo obstoja drugih epistemskih in ne-epistemskih ciljev. Možno je, da resnica deluje kot primarni cilj le v epistemskem smislu, ne pa kot naš primarni cilj v splošnem (Pritchard, 2014).

Pojavlja se tudi vprašanje o tem, kaj je cilj raziskovanja in ali je takrat, ko dosežemo resnično prepričanje, točka, kjer naj bi raziskovanje ustavili. Pritchard (2014) meni, da je primeren čas za zaključek raziskovanja takrat, ko dosežemo razumevanje. Hkrati pa trdi, da to, da je zaključek raziskovanja razumevanje ali znanje, še ne pomeni, da cilj ne more biti resnica. Znanje ali razumevanje v odnosu do neke propozicije sta epistemski poziciji, ki nam omogočata, da presodimo, ali smo že dosegli cilj raziskovanja – resnično prepričanje. Podaja analogijo s procesom pridelave kave, katerega cilj je ustvarjanje odlične skodelice kave. V okviru procesa pridelave se udejstvujemo v aktivnostih, kot so izbira zrn, določanje načina praženja, izbira naprave za praženje kave in podobno. Na neki točki te aktivnosti dosežejo cilj – odlično skodelico kave. Kljub temu pa moramo nekako določiti, ali je bil cilj dosežen, kar naredimo tako, da na primer poskusimo skodelico kave in s tem pridobimo znanje o tem, da je odlična. Znanje o tem, ali je skodelica kave odlična, je torej to, kar zaključi naše aktivnosti, njihov cilj pa še vedno ostaja odlična skodelica kave sama po sebi.

Ena od kritik cilja resnice je, da je ohlapen in nedoločen – pogosto ne opredeli, na koliko in na katere propozicije se nanaša (David, 2001, 2013). V najširšem smislu bi se nanašal na vse možne propozicije ali na vse propozicije, o katerih imamo ljudje sposobnost razmišljati. Številni filozofi pa zagovarjajo, da se nanaša zgolj na podskupino prepričanj, ki so za agenta na nek način relevantna (npr. Alston, 2005; David, 2013; DePaul, 2001). Pozicijo, da imajo ne-trivialne resnice višjo epistemsko vrednost kot trivialne, pa nekateri smatrajo kot protiargument za veritistični monizem: če so ne-trivialne resnice epistemsko vredne več od trivialnih, to pomeni, da resnica ni edini in temeljni predmet naših epistemskih presoj (DePaul, 2001). Pritchard (2014) to zanika in meni, da je veritistični monizem skladen s tezo, da se intelektualno raziskovanje v primeru izbire med trivialnimi in ne-trivialnimi resnicami osredotoča na ne-trivialne. Izhajajoč iz dela Nicka (Treanor, 2012) zagovarja, da to, da nimajo vse resnice enake teže, ne pomeni, da resnica ne more imeti vloge temeljne epistemske vrednosti.

O vrednosti cilja resnice lahko razmišljamo iz perspektive teoretične epistemologije ali iz perspektive organizmov, ki posedujejo kognitivne sisteme, o katerih epistemologija razpravlja. Tudi če znotraj epistemologije privzamemo, da je resnica primarni in temeljni epistemski cilj, to ne bo nujno držalo iz perspektive epistemskih agentov, ki zasledujejo številne cilje, med katerimi so nekateri zanje pomembnejši in do neke mere neodvisni od cilja resnice – na primer blagostanje ali reproduktivna

uspešnost. Tu se poraja vprašanje o tem, ali so lahko cilji omejene epistemske racionalnosti tudi ne-epistemski oziroma ali lahko epistemske cilje presojamo tudi iz ne-epistemskih vidikov. Kvanvig (2013) na primer zagovarja, da je znotraj epistemologije treba kognitivni uspeh, kot je oblikovanje resničnih prepričanj, vrednotiti le z epistemskega vidika in da so vprašanja o tem, ali ta resnična prepričanja pripomorejo k evolucijskemu uspehu, blagostanju organizmov ali katerimkoli pragmatičnim ciljem, za epistemske presoje irelevantna. Nekatere pozicije znotraj epistemologije pa temu nasprotujejo. Ena od močnejših tez o vplivu pragmatičnih dejavnikov na epistemske presoje je teza pragmatičnega poseganja (ang. *pragmatic encroachment*). Zagovorniki te teze menijo, da je pripisovanje znanja, upravičenja ali racionalnosti funkcija ne le epistemskih, temveč tudi pragmatičnih dejavnikov. To pomeni, da se lahko agenta, ki sta v identični epistemski situaciji in sta prepričanje oblikovala z enako zanesljivim procesom in na podlagi enakih dokazov, se pa razlikujeta glede na praktične okoliščine, razlikujeta v tem, ali je njuno prepričanje upravičeno oziroma ali posedujeta znanje (Fantl in Mcgrath, 2007). Obstajajo pa tudi druge, šibkejše pozicije, ki epistemske in pragmatične dejavnike povezujejo na drugačen način. Harman (2004) na primer meni, da so praktični razlogi za epistemsko racionalnost pomembni, ker usmerjajo naše razmišljanje – povedo nam, kaj je za nas dovolj relevantno, da bi o tem sklepali, razmišljali in oblikovali prepričanja. Mišljenje porablja naše omejene časovne in računske vire in zato je treba je sklepati kompromise: če večino virov porabljamo za mišljenje o eni stvari, jih ne moremo za mišljenje o kateri drugi. Praktični razlogi nam povedo, koliko truda, časa in energije je smiselno porabiti za določeno tematiko in kdaj razmišljanje o njej zaključiti. V vsakdanjem življenju imamo številna prepričanja, ki so morda napačna, pa zaradi njih ne utrpimo nobenih praktičnih posledic in tako tudi ne bi bilo racionalno, da na račun preverjanja in posodabljanja teh prepričanj opuščamo razmišljanje o drugih, pomembnejših tematikah. Praktični dejavniki ne morejo služiti kot epistemski razlogi za oblikovanje določenega prepričanja, lahko pa nam povedo, na kateri točki začne mišljenje o določenem problemu porabljati preveč virov in je zato z njim smiselno zaključiti (Harman, 2004).

## 6 Cilji omejene in ekološke racionalnosti

Omejena racionalnost se izvorno nanaša na proces izbire oziroma odločanja in vedenjski izid odločitve. Eden od ciljev Simonovega projekta je bil oblikovanje modela racionalne izbire, ki bo za razliko od idealizirane, normativne teorije odločanja upošteval tako omejitve kognitivnih kapacitet organizmov kot omejitve

okolja, v katerem delujejo. Ne glede na to, ali nas zanimajo normativni ali deskriptivni vidiki modelov odločanja, lahko človeško racionalnost pojmujemo le kot grob, poenostavljen približek idealiziranih modelov racionalnosti, ki nastopajo na primer v teoriji iger. Za razliko od takšnih modelov, ki kot kriterij racionalnosti pojmujejo optimalno izbiro, na primer izbiro alternative z največjo pričakovano koristnostjo, Simon kot kriterij omejene racionalnosti postavi izbiro, ki ni nujno najboljša, temveč le presega določen prag in je kot takšna dovolj dobra (Simon, 1955).

Tako pri konceptu omejene kot ekološke racionalnosti se pojavlja pojem adaptivnosti. Simon adaptivnost opredeljuje kot sposobnost organizmov, da svoje vedenje prilagodijo zahtevam specifične naloge, s katero so soočeni. Adaptivnost organizma zajema tako evolucijske prilagoditve sistemov kot zavestno in namerno prilagajanje, na primer učenje in reševanje problemov. Ljudje smo torej adaptivni sistemi in naše vedenje je fleksibilno. To, kaj smatramo za adaptivno vedenje, je odvisno od zahtev okolja – vedenje (na primer strategija reševanja problemov), ki bo adaptivno v eni situaciji, ne bo nujno adaptivno tudi v drugi. Adaptivna sposobnost organizma pa je omejena tako s kognitivnimi omejitvami organizma kot s strukturo naloge, in zaradi teh omejitev odziv organizmov na naloge ni optimalen, temveč zgolj omejeno racionalen. Eden od primerov adaptivnega, omejeno racionalnega vedenja je uporaba hevristik za reševanje problemov – hevristika v določenih okoljih ob zmerni porabi virov namreč vodi do zadovoljivih rešitev. Model omejene racionalnosti torej razlaga, kako se organizmi z omejenimi kognitivnimi kapacitetami uspešno prilagajajo na izjemno kompleksne naloge v resničnem svetu.

Simon izhaja iz predpostavke, da imajo različni organizmi različne potrebe, želje in cilje, in da to določa, kaj so relevantni vidiki njihovega okolja. Organizmi imajo lahko več ciljev naenkrat, vendar pa med zasledovanjem enega in drugega prihaja do izmenjave – časovni in kognitivni viri, porabljeni za zasledovanje enega cilja, ne morejo biti porabljeni za zasledovanje drugega. Organizmi imajo relativno enostavne mehanizme oziroma postopke, po katerih določajo, kateri cilj ima prioriteto ozrioma katere cilje bodo zasledovali v sedanjem trenutku. Cilje organizma Simon opredeli ohlapno: lahko se nanašajo tako na zadovoljevanje fizioloških potreb, ki omogočajo preživetje, kot na številne druge potrebe in želje. Glede na to, koliko ciljev in koliko časa imajo organizmi, se spreminja tudi prag, kjer bodo presodili, da je cilju

zadovoljeno – če je ciljev več, časa pa manj, lahko organizmi problem razporejanja virov naslovijo tako, da znižajo prag (Simon, 1956).

Adaptivnost je ključen pojem tudi pri Gigerenzerjevi ekološki racionalnosti. Gigerenzer racionalnosti ne opredeljuje prek izbranih normativnih sistemov, temveč kot odnos med umom in okoljem oziroma med strategijo in med nalogo. Racionalnost razume kot adaptivno mišljenje in preučuje, kako se ljudje soočamo s situacijami, kjer moramo presojati, napovedovati, sklepati in se odločati, v kompleksnem in negotovem okolju. Gigerenzer hevristike, njihove sestavne dele (pravila, ki usmerjajo hevristično mišljenje) in kognitivne kapacitete, ki omogočajo njihovo uporabo (spomin, pozornost in podobno), imenuje torba z orodjem (ang. *adaptive toolbox*). Ne gre torej za to, da bi bila ena sama strategija reševanja problemov a priori (i)racionalna, temveč se njeno (i)racionalost presoja glede na to, ali v specifični situaciji vodi do želenih izidov – v tem primeru do točnejših napovedi kot ostale možne strategije. V nekaterih situacijah lahko uporaba preprostih hevristik, ki upoštevajo mnogo manj informacij kot kompleksni statistični modeli, na primer multipla regresija, vodijo do točnejših presoj ali sklepov. Adaptivnost – sposobnost izbora ustrezne strategije za določeno nalogo – je torej v jedru ekološke racionalnosti (Gigerenzer, 2000, 2008; Gigerenzer in Gaissmaier, 2011; Gigerenzer in Todd, 2001).

## 7 Cilji omejene epistemske racionalnosti

Omejena epistemska racionalnost je usmerjena v doseganje epistemskih ciljev, hkrati pa z upoštevanjem kognitivnih omejitev agentov in omejitev epistemskega okolja, v katerem se nahajajo, ne zahteva optimalnih, temveč le dovolj dobre rešitve. Kot epistemske cilje dopušča različne vrste kognitivnega uspeha, od oblikovanja resničnih prepričanj in izogibanja neresničnim, do natančnih in točnih napovedi, razlage in razumevanja. Je adaptivna v smislu, da pripoznava, da je (i)racionalnost posameznega načina raziskovanja oziroma strategije reševanja problemov, presojanja, oblikovanja prepričanja itn. odvisna od tega, ali ta strategija v specifični situaciji vodi do izboljšanja naše epistemske situacije oziroma do (dovolj dobrega) doseganja epistemskih ciljev. Je ekološka v Gigerenzerjevem smislu: ne privzema rigidnih norm ali pravil, temelječih na sledenju normativnim sistemom, kot so pravila logike in verjetnosti, temveč dopušča, da so racionalne tiste strategije oziroma načini raziskovanja, ki bodo zanesljivo vodili do (dovolj) resničnih prepričanj, dovolj celostnega razumevanja ali dovolj točne napovedi.

Norme omejene epistemske racionalnosti so torej ne-idealne norme raziskovanja, ki se nanašajo na to, kako določati, kaj so relevantni problemi, o katerih bomo raziskovali; kako prepoznavati zaupanja vredne vire informacij; kako presoditi, kdaj imamo dovolj dokazov, da zaključimo raziskovanje in oblikujemo prepričanje itn. Predpostavimo, da želimo oblikovati norme raziskovanja, ki bodo vodile do resničnih prepričanj o tem, da ljudje pripomoremo h globalnemu segrevanju. V idealnem primeru bi se te norme nanašale na epistemske agente, ki raziskujejo z namenom doseganja resnice; ki imajo ogromne kognitivne in časovne vire ter so motivirani za raziskovanje; in ki imajo enostaven dostop do pravih informacij o globalnem segrevanju. V resničnem svetu pa se agenti pogosto poslužujejo motiviranega sklepanja (ang. *motivated reasoning*), so pri raziskovanju pristrani in bolj kot doseganje resničnega prepričanja želijo, na primer, utrditi že obstoječe mnenje; imajo omejene kognitivne vire in čas, ki ga lahko namenijo za raziskovanje o globalnem segrevanju ter pogosto ne znajo pravilno prepoznati verodostojnih informacij; in raziskujejo v okolju, kjer je o globalnem segrevanju prisotno ogromno napačnih informacij. Način raziskovanja, ki ga omejeni agenti v resničnem svetu dejansko lahko izvedejo in ki bo vodil do resničnih prepričanj o globalnem segrevanju, je drugačen od načina raziskovanja, ki bi vodil do resničnih prepričanj pri idealnih agentih v idealnem okolju.

Omejena epistemska racionalnost upošteva, da je določanje, katere strategije bodo v katerih pogojih uspešnejše od drugih, empirično vprašanje, hkrati pa še zmeraj naslavlja normativna vprašanja o epistemsko dobrem spoznavanju. Ključno normativno vprašanje je, kako opredeliti, kaj pomeni »dovolj dobro« doseganje epistemskih ciljev. Ena možnost je, da privzamemo, da smo omejeno epistemsko racionalni, če se epistemskemu cilju približujemo v dovolj veliki meri, da nam to omogoča vedênje, ki nas bo vodilo do ne-epistemskega cilja, ki ima za nas intrinzično vrednost. V tem primeru imajo epistemski cilji zgolj instrumentalno vrednost in so v funkciji doseganja drugih, ne-epistemskih ciljev. Druga možnost je, da se naslanjamo na delo Catherine Elgin in njen koncept dovolj resničnega (ang. *true enough*; Elgin, 2004, 2017). Ko govorimo o cilju resnice kot enem izmed ciljev, ki jih zasleduje omejena epistemska racionalnost, lahko zagovarjamo, da je prepričanje dovolj resnično, če pripomore k doseganju širokega nabora vrst kognitivnega uspeha. Dovolj resnična prepričanja bodo izpolnjevala svoj kognitivni namen – približala nas bodo, na primer, razumevanju pojava na ravni natančnosti, kot si ga želimo. Elgin (2004, 2017) ne zagovarja, da bi morali cilj resnice opustiti, temveč da se moramo prevpraševati, kakšno vlogo naj igra v kompleksnem sklopu epistemskih

vrednot in ciljev, ki jih želimo zasledovati. Težava veritističnega monizma je, da je preveč omejujoč – če bi kot primarni epistemski cilj dopuščali le doseganje resnice, bi s tem izpustili širok nabor kognitivnih aktivnosti in dosežkov. Izhajajoč iz tega lahko pri omejeni epistemski racionalnosti dopuščamo pluralizem epistemskih ciljev in se ne omejujemo le na veritizem.

Še ena možnost je, da spoznanja o omejenosti naših kognitivnih in časovnih virov uporabimo kot argument za to, da kot cilje omejene epistemske racionalnosti pojmujemo le resnična prepričanja, razumevanje, pravilno presojo itn. o tematikah, ki so za posameznika relevantne. Kljub argumentom v prid epistemski vrednosti trivialnih resnic je omejena epistemska racionalnost zagotovo usmerjena le v oblikovanje resničnih prepričanj, ki so za posameznika na nek način relevantna, pomembna in zanimiva. Ravno zaradi omejenosti kognitivnih in časovnih virov je nemogoče, da bi od resničnih agetov zahtevali, da poizvedujejo o tematikah, ki so zanje popolnoma nerelevantne, in da oblikujejo ogromno število resničnih prepričanj ali posedujejo razumevanje praktično vsake tematike, o kateri je možno razmišljati. V vsakdanjem življenju se moramo neprestano odločati o tem, kateri problemi so dovolj pomembni, da jim bomo posvetili kognitivne in časovne vire. Omejena epistemska racionalnost tako ne zahteva, da bi agenti raziskovali in oblikovali prepričanja na način, ki bi posegal v vse ostale pomembne aktivnosti v življenju, temveč predpisuje le raziskovanje o relevantnih problemih. Podobno pozicijo privzemata Bishop in Trout, ki podajata teorijo strateškega reliabilizma, kjer racionalno mišljenje opredelita kot zanesljivo, stroškovno učinkovito in usmerjeno na relevantne probleme (Bishop in Trout, 2004). Če cilj resnice zamejimo na relevantne probleme, mora teorija epistemske racionalnosti opredeliti, kaj pomenijo relevantni problemi, kar pa ne bo popolnoma epistemsko, temveč bo tudi pragmatično ali moralno vprašanje. Omejena epistemska racionalnost je torej hibridna teorija, ki vsebuje tako epistemske kot ne-epistemske elemente.

Za cilje omejene epistemske racionalnosti sta relevantna še dva pomisleka. Prvi se nanaša na adaptivnost: čeprav je omejena epistemska racionalnost adaptivna v smislu, da upošteva, da je treba strategije prilagoditi značilnosti naloge, s katero smo soočeni, pa ne zasleduje adaptivnosti v evolucijskem smislu izboljševanja verjetnosti preživetja in razmnoževanja. Ne glede na to, ali utemeljevanje veritističnega monizma z vidika pravih funkcij drži, je evolucijska adaptivnost z vidika epistemskih presoj irelevantna. Posamezniki v vsakdanjem življenju zasledujejo številne cilje, ki si bodo med seboj včasih nasprotovali. V primeru, da posamezniki na račun

epistemskih izberejo zasledovanje ne-epistemskih ciljev, to ne pomeni, da so v splošnem iracionalni – pomeni pa, da so epistemsko iracionalni. Nezanesljivo raziskovanje in oblikovanje napačnih prepričanj o tematikah, tudi če to pripomore k blagostanju, uspešnosti ali drugim ne-epistemskim ciljem, torej ne bo omejeno epistemsko racionalno.

Drugi pomislek je, da sodbe o omejeni epistemski racionalnosti niso pogojene s posameznikovimi cilji in željami, kar pomeni, da lahko tudi posameznika, ki si ne želi zasledovati epistemskih ciljev, označimo za omejeno epistemsko iracionalnega. Agenti, ki morebiti ne posedujejo epistemskih ciljev, torej niso izvzeti iz sodb o epistemski (i)racionalnosti. Edino področje, kjer dopuščamo pogojenost presoj o (i)racionalnosti z nečim, kar je vezano na specifičnega agenta oziroma skupino agentov, je relevantnost ciljev. Vendar se moramo tudi v tem primeru, kot opozarjata že Bishop in Trout (2005), izogniti popolnemu subjektivizmu pri pojmovanju relevantnosti, ki bi dopuščal, da je relevanten vsak problem, ki ga posameznik smatra kot takega – tudi če gre na primer za preštevanje, kolikokrat se v knjigi pojavi črka L.

## 8 Zaključek

Omejena epistemska racionalnost se od pojmovanj epistemske racionalnosti znotraj tradicionalne analitične epistemologije razlikuje po tem, da se naslanja na delo o omejeni in ekološki racionalnosti ter se zanaša na empirične podatke o omejitvah človeške kognicije in značilnostih epistemskega okolja, v katerem raziskujemo. Omejena racionalnost je torej hibriden, ne-idealen koncept, ki se poskuša izogibati idelizacijam epistemskih agentov in epistemskega okolja ter želi ponuditi norme raziskovanja, ki bodo dosegljive za realne epistemske agente z vsemi njihovimi omejitvami in ki bodo pripomogle k doseganju epistemskih ciljev v okolju, v katerem se nahajajo. Usmerjena je k epistemskim ciljem, vendar predpisuje le raziskovanje o tematikah, ki so za posameznika na nek način relevantne. Ni nujno zavezana veritističnemu monizmu; zasleduje lahko različne vrste kognitivnega uspeha, na primer razumevanje, točnost, napoved in razlago, vendar ne zahteva optimalnosti niti v samem procesu raziskovanja niti v doseganju epistemskih ciljev. Zaradi omejenosti naših kognitivnih in časovnih virov ter nepopolnih informacij iz okolja se bomo pogosto lahko le približali idealom doseganja resnice, celostnega razumevanja pojava, popolne napovedi in podobno. Omejena epistemska racionalnost tako stremi k oblikovanju prepričanj, ki bodo »dovolj resnična« v smislu

Elgin, ali do dovolj poglobljenega razumevanja, da bomo na njem znali osnovati vedenje – ne zahteva na primer popolnega razumevanja mehanizmov globalnega segrevanja na vseh ravneh analize, temveč le stopnjo razumevanja, ki nam bo omogočila, da oblikujemo in izvajamo ukrepe, ki bodo globalno segrevanje zmanjševali. Deluje torej kot teorija, ki vsebuje tako epistemske kot pragmatične elemente ter izhodiščna točka za oblikovanje ne-idealnih norm raziskovanja, ki bodo pomagale izboljševati našo epistemsko situacijo.

**Literatura**


Alston, W. P. (1985). Concepts of Epistemic Justification. *The Monist*, *68*(1), 57–89. https://doi.org/10.5840/monist198568116

Alston, W. P. (2005). *Beyond "Justification": Dimensions of Epistemic Evaluation*. Cornell University Press. https://www.jstor.org/stable/10.7591/j.ctv2n7gqf

Bishop, M. A. in Trout, J. D. (2004). *Epistemology and the Psychology of Human Judgment* (J. D. Trout, Ed.). OUP USA.

Bondy, P. (b. d.). *Epistemic Value*. Internet Encyclopedia of Philosophy. Pridobljeno 1. 12. 2024. https://iep.utm.edu/epistemic-value/

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Harvard University Press.

Cherniak, C. (1986). *Minimal Rationality*. MIT Press.

David, M. (2001). Truth as the Epistemic Goal. V M. Steup (ur.), *Knowledge, Truth, and Duty* (str. 151–169). Oxford University Press.

David, M. (2013). Truth as the Primary Epistemic Goal: A Working Hypothesis. V M. Steup, J. Turri in E. Sosa (ur.), *Contemporary Debates in Epistemology, 2nd Edition* (str. 363–377). Wiley-Blackwell.

DePaul, M. R. (2001). Value Monism in Epistemology. V M. Steup (ur.), *Knowledge, Truth, and Duty* (str. 170–184). Oxford University Press. https://doi.org/10.1093/0195128923.003.0011

Elgin, C. Z. (2004). True Enough. *Philosophical Issues*, *14*, 113–131.

Elgin, C. Z. (2017). *True Enough*. MIT Press.

Fantl, J. in Mcgrath, M. (2007). On Pragmatic Encroachment in Epistemology. *Philosophy and Phenomenological Research*, *75*(3), 558–589. https://doi.org/10.1111/j.1933-1592.2007.00093.x

Foley, R. (1987). *The theory of epistemic rationality*. Harvard University Press.

Foley, R. (2011). Epistemic Rationality. V S. Bernecker (ur.), *The Routledge Companion to Epistemology* (str. 37–46). Routledge.

Friedman, J. (2019). Inquiry and Belief. *Noûs*, *53*(2), 296–315. https://doi.org/10.1111/nous.12222

Friedman, J. (2020). The Epistemic and the Zetetic. *Philosophical Review*, *129*(4), 501–536. https://doi.org/10.1215/00318108-8540918

Friedman, J. (2023). The Aim of Inquiry? *Philosophy and Phenomenological Research*, phpr.12982. https://doi.org/10.1111/phpr.12982

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford University Press.

Gigerenzer, G. (2008). Why Heuristics Work. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *3*(1), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x

Gigerenzer, G. in Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*, 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G. in Todd, P. M. (2001). *Simple heuristics that make us smart*. Oxford University Press.

Graham, P. J. (2012). Epistemic Entitlement. *Noûs*, *46*(3), 449–482. https://doi.org/10.1111/j.1468-0068.2010.00815.x

Harman, G. (2004). Practical Aspects of Theoretical Reasoning. V A. R. Mele in P. Rawling (ur.), *The Oxford handbook of rationality* (str. 45–56). Oxford University Press.
Haziza, E. (2023). Norms of Inquiry. *Philosophy Compass*, *18*(12), e12952. https://doi.org/10.1111/phc3.12952
Hazlett, A. (2013). *A luxury of the understanding: On the value of true belief*. Oxford University Press.
James, W. (1896). *The Will to Believe and Other Essays in Popular Philosophy*. Longmans Green & Co.
Kelp, C. (2021). *Inquiry, Knowledge, and Understanding*. Oxford University Press. https://doi.org/10.1093/oso/9780192896094.001.0001
Knauff, M. in Spohn, W. (2021). *The Handbook of Rationality*. The MIT Press. https://doi.org/10.7551/mitpress/11252.001.0001
Kvanvig, J. L. (2013). Truth is Not the Primary Epistemic Goal. V M. Steup in J. Turri (ur.), *Contemporary Debates in Epistemology* (str. 285–295). Blackwell.
Levy, N. (2021). *Bad Beliefs: Why They Happen to Good People*. Oxford University Press. https://doi.org/10.1093/oso/9780192895325.001.0001
Lynch, M. P. (2004). *True to Life: Why Truth Matters*. MIT Press.
McKenna, R. (2023). *Non-Ideal Epistemology*. Oxford University Press.
Neumann, J. V. in Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.
Nozick, R. (1993). *The Nature of Rationality*. Princeton University Press.
Pils, R. (2022). A Satisficing Theory of Epistemic Justification. *Canadian Journal of Philosophy*, *52*(4), 450–467. https://doi.org/10.1017/can.2022.38
Plantinga, A. (1993). *Warrant and Proper Function*. Oxford University Press.
Pritchard, D. (2014). Truth as the Fundamental Epistemic Good. V J. Matheson in R. Vitz (ur.), *The Ethics of Belief* (str. 112–129). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199686520.003.0007
Pritchard, D. (2018). *What is this thing called Knowledge? (4th ed.)*. Routledge. https://doi.org/10.4324/9781351980326
Pritchard, D. (2019). Intellectual Virtues and the Epistemic Value of Truth. *Synthese*, *198*(6), 5515–5528. https://doi.org/10.1007/s11229-019-02418-z
Pritchard, D. (2021). Veritism and the Goal of Inquiry. *Philosophia*, *49*(4), 1347–1359. https://doi.org/10.1007/s11406-021-00325-7
Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852
Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769
Simon, H. A. (1976). From substantive to procedural rationality. V T. J. Kastelein, S. K. Kuipers, W. A. Nijenhuis in G. R. Wagenaar (ur.), *25 Years of Economic Theory: Retrospect and prospect* (str. 65–86). Springer US. https://doi.org/10.1007/978-1-4613-4367-7_6
Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*, 1–19. https://doi.org/10.1146/annurev.ps.41.020190.000245
Simon, H. A. (1992). What is an "explanation" of behavior? *Psychological Science*, *3*(3), 150–161. https://doi.org/10.1111/j.1467-9280.1992.tb00017.x
Sosa, E. (2007). *A Virtue Epistemology*. Oxford University Press.
Stein, E. (1997). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford University Press.
Sturm, T. (2020). Towards a critical naturalism about bounded rationality. V R. Viale (ur.), *Routledge Handbook of Bounded Rationality* (str. 73-90). Routledge.
Thorstad, D. (2021). Inquiry and the Epistemic. *Philosophical Studies*, *178*(9), 2913–2928. https://doi.org/10.1007/s11098-020-01592-y
Thorstad, D. (2022). Two Paradoxes of Bounded Rationality. *Philosophers' Imprint*, *22*(n/a). https://doi.org/10.3998/phimp.1198
Thorstad, D. (2024a). *Inquiry Under Bounds*. Oxford University Press.
Thorstad, D. (2024b). Why bounded rationality (in epistemology)? *Philosophy and Phenomenological Research*, *108*(2), 396–413. https://doi.org/10.1111/phpr.12978

Treanor, N. (2012). Trivial Truths and the Aim of Inquiry. *Philosophy and Phenomenological Research*, *89*(3), 552–559. https://doi.org/10.1111/j.1933-1592.2012.00612.x

Tomat, N. (2024). Bridging the Gap between the Normative and the Descriptive: Bounded Epistemic Rationality. *INDECS, 22*(1), 107-121. https://doi.org/10.7906/indecs.22.1.6

Viale, R. (2020). *Routledge Handbook of Bounded Rationality*. Routledge.

Wedgwood, R. (2013). Rational "Ought" Implies "Can." *Philosophical Issues*, *23*(1), 70–92. https://doi.org/10.1111/phis.12004

Zagzebski, L. T. (2020). *Epistemic Values: Collected Papers in Epistemology*. Oxford University Press.

**ANALIZA**
Let. 28, št. 2, str. 285–310, december 2024

# SPOZNANJE IN KONVERZACIJA: JEZIKOVNE HIBE NA OZADJU GRICEOVE TEORIJE IMPLIKATUR

**Sprejeto**
3. 10. 2024

**Pregledano**
20. 12. 2024

**Izdano**
31. 12. 2024

NIKO ŠETAR
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
niko.setar@gmail.com

DOPISNI AVTOR
niko.setar@gmail.com

**Izvleček** V zadnjih letih je neformalna epistemologija dodobra osredotočena na teorijo spoznavnih vrlin in hib, ki izhaja iz Aristotelove etike vrlin, in se deli na dva ključna pristopa: reliabilizem in responsibilizem, ki nudita sorodne a nasprotujoče si poglede na naravo tovrstnih vrlin in hib. Delo Quassima Cassama je poglobilo splošno dojemanje hib v strokovni javnosti kot vedenj, ki onemogočajo dostop do znanja in se klasificirajo v kategorije glede na zvestobo. Teorija spoznavnih vrlin in hib je s seboj pripeljala tudi teorijo argumentacijskih vrlin, h kateri je najpomembneje prispeval Andrew Aberdein, ki analizira prenos znanja in prepričanja skozi argumentacijo. Argumentacijske hibe taista teorija opredeli kot lastnosti, ki vodijo v zmotno uporabo argumentacije in sprejem pomanjkljivih argumentov. Vendar pa Cohen opozarja, da vsi načini prenosa znanja, na primer enostavne inferenčne izjave ali učilniška konverzacija, niso nujno argumentativni. Ker je v teh primerih prenos znanja ali prepričanja vseeno lahko prisoten na neargumentacijski način, je smiselno poiskati argumentacijski-sorodno vrsto hibe, ki se utegne pojavljati v tovrstnem prenosu prepričanj. Možnost opredelitve takšnih hib najdemo v Griceovi teoriji implikatur, kjer pomanjkljive izjave sprožijo pragmatične implikacije, kar odpira prostor za raziskovanje manifestacije spoznavnih hib v jeziku v neargumentativnih kontekstih prenosa znanja ali prepričanja.

**Ključne besede**
spoznavna hiba, argumentacijska hiba, konverzacija, implikatura, epistemologija, filozofija jezika

https://doi.org/10.18690/analiza.28.2.285-310.2024

AN$_{\alpha}$li ZA

**ANALIZA**
Vol. 28, no. 2, pp. 285–310, December 2024

# KNOWLEDGE AND CONVERSATION: LINGUISTIC VICES IN THE CONTEXT OF GRICE'S THEORY OF IMPLICATURES

NIKO ŠETAR
University of Maribor, Faculty of Arts, Maribor, Slovenia
niko.setar@gmail.com

CORRESPONDING AUTHOR
niko.setar@gmail.com

**Accepted**
3. 10. 2024

**Revised**
20. 12. 2024

**Published**
31. 12. 2024

**Abstract** Recently, informal epistemology has centred on a theory of epistemic virtues and vices, rooted in Aristotle's virtue ethics and divided into two key approaches: reliabilism and responsibilism, which offer related but contrasting views. Quassim Cassam's work has advanced the perception of vices as behaviours that prevent access to knowledge, classified by fidelity. The theory of epistemic virtues and vices has also led to the theory of argumentative virtues, notably developed by Andrew Aberdein, who examines the transmission of knowledge and belief through argumentation. Argumentative vices, by the same theory, are characteristics that lead to fallacious argumentation and acceptance of flawed arguments. However, Cohen argues that not all knowledge transfer – such as inferential statements or classroom conversation – is necessarily argumentative. Since belief transfer can occur non-argumentatively, it is reasonable to seek an argumentation-related vice in such cases. Grice's theory of implicatures suggests flawed statements generate pragmatic implications, creating space to examine epistemic vices in language within non-argumentative belief transfer contexts.

**Keywords**
epistemic vice, argumentative vice, conversation, implicature, epistemology, philosophy of language

https://doi.org/10.18690/analiza.28.2.285-310.2024


University of Maribor Press

Neformalna epistemologija se zadnja leta vse več posveča teoriji spoznavnih vrlin in hib, ki izhaja iz temeljnih del Ernesta Sose (1991, 2007), Johna Greca (1999), Jamesa Montmarqueta (1993) in Linde Zagzebski (1996), oprtih na Aristotelove vrline opisane v Nikomahovi Etiki. Pri tem se krešeta predvsem dva poglavitna pristopa: reliabilizem (Sosa, Greco), ki spoznavne vrline v grobem opredeli kot zanesljive in natančne kognitivne in zaznavne zmožnosti, ki jih mora za spoznavno primerno (vrlo) ravnanje subjekt spretno uporabljati, in pa responsibilizem, ki spoznavne vrline opredeljuje kot vrste ravnanja, za katere je subjekt odgovoren in v primeru vrlega ravnanja zaslužen in hvalevreden, v nasprotnem primeru pa grajevreden. Iz responsibilizma izhaja do današnjega dne najbolj prodorno delo na temo spoznavnih hib, *Vices of the Mind* Quassima Cassama (2019), ki v svoji teoriji imenovani obstruktivizem spoznavne hibe opredeli kot načine intelektualnega vedenja, ki subjektu ovirajo dostop do znanja, in za katere je subjekt bodisi grajevreden, bodisi celovito kriv. Hibe v tem pojmovanju razvrsti v tri kategorije glede na zvestobo (tj. pogostnost pojavljanja): značajske lastnosti, kadar gre za hibe, ki se pri subjektu pojavljajo konsistentno in ne glede na temo spoznavnega postopka; stališča, kadar gre za hibe, ki se pri subjektu pojavljajo konsistentno v določenih kontekstih in tematikah; ter načine mišljenja, kadar gre za občasne ali izolirane primere kognitivnega lapsusa.

Teorija spoznavnih vrlin pa je s seboj prinesla še teorijo argumentacijske vrline in posledično argumentacijske hibe (Aberdein, 2010, 2013, 2014; Cohen, 2005, 2007), ki preučuje, kako se znanja oziroma prepričanja širijo skozi argumentacijo in vzpostavlja tipologijo argumentacijskih hib, pri katerih gre za lastnosti argumentatorja, ki vodijo v zmotno uporabo ali načrtno zlorabo argumentacijskih zmot, ter lastnosti sogovorca, ki vodijo v neupravičen sprejem tovrstnih zmot kot veljavnih argumentov. Dodatne podrobnosti teorije spoznavnih ter argumentacijskih vrlin so na voljo v navedenih virih in v Šetar (2024).

Do preobrata na tematiko tega prispevka pride v Cohenovem (2007) pomisleku, da vsi načini prenašanja znanja ali prepričanja še niso argumentativne narave. Cohen izpostavlja primer izjave »Tako Marie kot Pierre Curie sta bila fizika, torej je bila Marie Curie fizik,« za katero ugotavlja, da gre v najboljšem primeru za enostavno inferenco, ne pa tudi za argument. Še bolj jasen primer je način prenosa znanja, ki ga Cappelen in Dever (2019) imenujeta učilniška konverzacija, izražen v pogovoru med osebama A in B, pri čemer A vpraša kaj je najvišja gora Škotske, B pa odvrne, da je to Ben Nevis. Čeprav je brez dvoma prišlo do prenosa znanja, pa pogovor ne

vsebuje ničesar, kar bi nakazovalo na format argumentacije. Če se spoznavne hibe v postopku argumentacijskega prenosa znanja izražajo kot argumentacijske hibe, je smiselno predpostaviti, da se v postopku neargumentacijskega prenosa znanja v prostem pogovoru kažejo kot neke druga, sorodna hiba, ki ji lahko nadenemo delovno ime konverzacijska ali jezikovna hiba. A kaj bi lahko tovrstna hiba bila?

Izhodišče za odgovor na to vprašanje predstavlja Griceov članek (1975) *Logika in konverzacija,* ki temelji na preučevanju implikatur, tj. pragmatičnih implikacij nepopolnih ali pomanjkljivih izjav.

Za razumevanje delovanja konverzacijskih implikatur, torej pragmatičnih implikacij znotraj poteka pogovora oziroma konverzacije, Grice vzpostavlja opisno analizo idealne konverzacije in opredeljuje pravila, ki tovrstno konverzacijo narekujejo. Osnova teh pravi je Sodelovalno načelo, ki veli, da »naj bo vaš konverzacijski prispevek takšen, da bo po namenu ali smeri pogovora, v katerega ste vpleteni, ustrezal stopnji pogovora, na kateri se nahajate« (Grice, 1975, 45).

Podrobnejša pravila, ki podkrepijo Sodelovalno načelo Grice imenuje konverzacijske maksime, ki jih uvršča v štiri kategorije: maksimo količine, kakovosti, odnosa in načina. Vsaka izmed njih vsebuje eno ali več posebnih maksim. V kategorijo količine tako sodita dve maksimi, ki sta formulirani kot sledi:

1. »Naj bo vaš konverzacijski prispevek tako informativen, kot je potrebno (za trenutne namene izmenjave).
2. Naj vaš konverzacijski prispevek ne bo bolj informativen, kot je potrebno.« (ibid.)

Nadalje, »v kategorijo KAKOVOSTI sodi supermaksima – »Naj bo vaš konverzacijski prispevek resničen« - ter dve bolj specifični maksimi:

1. Ne govorite tistega, za kar mislite, da ni resnično.
2. Ne govorite tistega, za kar nimate ustreznih dokazov.« (ibid., 46)

Kategorija odnosa vsebuje eno samo maksimo, ki zapoveduje relevantnost glede na osrednjo rdečo nit pogovora, ali z drugimi besedami prepoveduje sunkovito in nesmiselno (glede na potek konverzacije) odstopanje od tekoče tematike.

Nazadnje, kategorija načina pod osrednjo »supermaksimo« jasnosti združuje štiri podrobnejše maksime načina:

1. »Izogibajte se nejasnemu izražanju.
2. Izogibajte se dvoumnosti.
3. Bodite kratki (izogibajte se nepotrebni dolgoveznosti).
4. Bodite urejeni.« (ibid., 46)

Konverzacijske maksime imajo odličen potencial, da iz njih izpeljemo konceptualizacijo jezikovne vrline. Poglejmo najprej obe maksimi količine. Prva zahteva primerno informativnost ter implicira, da je prenizka informativnost nezaželena oziroma konverzacijsko neprimerna. Druga maksima neposredno prepoveduje pretirano informativnost. To ustreza splošnemu modelu vrline kot »zlate sredine« med dvema hibama, ki sta njen primanjkljaj in presežek, po katerem se ravnajo tudi teorije spoznavnih vrlin in hib ter argumentacijskih vrlin in hib. V primeru maksim kakovosti je zahteva po resničnosti vsebine izjave že na prvi pogled sorodna z zahtevo po resničnosti izraženih prepričanj v okviru spoznavne iskrenosti – govorjenje neresnic za katere vemo, da gre za neresnice, se pravi laganje, se povezuje s spoznavno neiskrenostjo ali celo spoznavno zlonamernostjo, medtem ko se podajanje izjav, za resničnost katerih nimamo zadostnih dokazov, utegne lepo povezati z jezikovno figuro nakladanja (Frankfurt, 2005) in spoznavno brezbrižnostjo. Toda maksimi načina in odnosa modelu vrline kot osredinjene vrednosti med dvema hibama ne ustrezata. Bodite relevantni implicira, da smo lahko premalo relevantni oziroma nismo relevantni glede na temo ali rdečo nit pogovora, medtem ko ni mogoče, da bi bili pretirano relevantni. Če pa podamo preveč informacij, kršimo drugo maksimo količine in pri tem ni mogoče, da bi se te informacije pretirano nanašale na predmet razprave. Prav tako na prvi pogled ni mogoče trditi, da bi se lahko pretirano izogibali dvoumnostim, ali pa bili pretirano urejeni (čeprav je, konec koncev, mogoče biti prekratek, a bržkone to oporeka drugi maksimi količine, ne pa tretji maksimi načina). Na srečo nam na tej točki ni potrebno razglabljati o konceptu jezikovne hibe in vrline s tako omejenim izhodiščem – Grice namreč nadalje ponuja razdelavo načinov, na katere lahko udeleženka v konverzaciji spodleti pri zadoščanju omenjenim maksimam:

1. »Možno je, da tiho in prikrito PREKRŠI maksimo; tako je lahko v določenih primerih zavajajoča?.

2. Lahko se ODLOČI ZA ODSTOP od upoševanja maksime in KP [Sodelovalno načelo, zgoraj][vpelji to v uvodu!]; izreče, nakaže, ali jasno razkrije, da ne želi sodelovati v skladu z zahtevami maksime. Reče lahko, denimo »več ne morem povedati, moja usta so zapečatena.«
3. Lahko pride do TRKA: udeleženec je lahko denimo nezmožen upoštevati prvo maksimo Količine (bodi toliko informativen, kolikor je potrebno), ne da prekrši drugo maksimo Kakovosti (imej zadostne dokaze za to, kar rečeš).
4. Lahko PREZRE maksimo, kar pomeni, da ji ODKRITO ne zadosti. Ob predpostavki, da je govorec zmožen zadostiti maksimi ne da krši kakšno drugo maksimo (zaradi trka), se ne odloči za odstop, in kljub malomarnosti pri svojem ravnanju ne poskuša zavajati, se za poslušalca pojavi manjša težava: kako lahko to, kar je govorec rekel, uskladimo s predpostavko, da upošteva KP? Ta situacija karakteristično vodi v konverzacijsko implikaturo, in kadar konverzacijska implikatura nastane na takšen način, pravim, da gre za IZKORIŠČANJE maksime.« (Grice, 1975, 49)

Grice v prvem primeru nujno predvideva, da je bil govorec namenjen prekršiti maksimo na način, pri katerem sogovorcem ne da vedeti, da je to storil, zaradi česar je pogosto zavajajoč. Prav zaradi namernosti bomo predpostavili, da gre v tem primeru za nesporen primer jezikovne hibe: govorec je zavedno in namerno ravnal na tak način, da poslušalcu ni predal zadostnih, primernih, ali relevantnih informacij in s tem onemogočil njegov dostop do prepričanj, ki jih te informacije vsebujejo, jim služijo v dokaz, ali podobno.

V primeru odločitve za odstop, da govorec poslušalcu vedeti, da to, kar bo dejal vnaprej, ne bo ustrezalo konverzacijskim maksimam ali celo sodelovalnemu načelu v celoti, in da naj torej ne bo obravnavano kot nadaljevanje iste konverzacije. Pri tem se pojavi vprašanje, ali lahko gre za jezikovno hibo, ali ne. V odgovor na to si poglejmo Griceovo vzpostavitev koncepta konverzacijske implikature: »Za človeka, ki s tem, da (pri tem, ko) reče (ali prikaže kot rečeno), da p, implicira, da q, lahko rečemo, da konverzacijsko implicira da q, V KOLIKOR (1) predpostavljamo, da upošteva konverzacijske maksime ali pa vsaj sodelovalno načelo; (2) predpostavljamo, da se zaveda, ali misli, da je q potreben za to, da je njegov izrek oz. prikazan izrek p (oz. dejanje v TEM smislu) konsistenen s to predpostavko; in (3) govorec meni (in pričakuje, da poslušalec meni, da govorec meni), da je poslušalec

zmožen razdelati ali intiutivno dojeti, da je predpostavka omenjena v (2) RES potrebna« (ibid., 49-50). Kaj natanko ta razdelava zahteva, je mogoče predstaviti na naslednji način:

1. Govorec upošteva NAJMANJ sodelovalno načelo.
2. Govorec meni, da je implikat potreben za interno konsistenco izjave.
3. Govorec meni, da je poslušalec zmožen dojeti potrebnost predpostavke IN kaj in zakaj je implicirano, na podlagi:
   - Konvencionalnega pomena besed
   - Sodelovalnega načela in maksim
   - Splošnega konteksta izjave
   - Skupnih znanj v ozadju (govorca in poslušalca)
   - Zavedanja, da je vse zgornje na voljo obema udeležencema konverzacije.

Da je neka izjava jezikovno vrla tako ni potrebno, da upošteva vse konverzacijske maksime, marveč je lahko vrla tudi tedaj, kadar je ena izmed maksim prekršena, v kolikor veljajo vse zgornje točke. V kolikor je ena izmed njih prav tako prekršena, gre za jezikovno hibo.

Vrnimo se na primer tihe in prikrite prekršitve maksime z namenom zavajanja. Da gre pri tovrstni prekršitvi za jezikovno hibo ni utemeljeno zgolj s tem, da je prekršena ena izmed maksim, marveč tudi z načinom, na katerega je bila prekršena. Lahko je denimo prekršena tako, da se govorec sploh ne ozira na sodelovalno načelo, s čemer prekrši prvega izmed zgornjih pogojev. Bolj pomembno, pa je, da govorec ne le, da ne meni, da je implikat (kršitev, če se grobo izrazimo) potreben za interno konsistenco izjave, temveč se lahko in najverjetneje se v celoti zaveda, da kršitev škodi interni konsistenci izjave. Nadalje se lahko pri tem požvižga na to, ali je poslušalec zmožen dojeti karkoli glede kršitve. Hkrati se dodobra zaveda, da poslušalec nima potrebnega skupnega znanja.

Po drugi strani, odkrit in izražen odstop od sodelovalnega načela v prvi vrsti ne prekrši nobene maksime – izjava namreč neposredno sporoča, da vsaj eni izmed maksim ni mogoče ugoditi. Na meta nivoju je v tem dejanju vsem maksimam zadoščeno: maksimi količine, v kolikor je povedano, da določeni maksimi ni mogoče ugoditi; maksimi kakovosti, v kolikor je vsebina izjave resnična (npr. govorec je

zakonsko zavezan k molčečnosti in podobne okoliščine); maksimi odnosa, v kolikor je to, o čemer govorec ne more govoriti, relevantno za »rdečo nit« konverzacije; in maksimi načina, dokler je izjava podana na tak način, da ni dvoumnosti ali nejasnosti glede tega, da in zakaj je odstop od sodelovalnega načela potreben. Nekdo lahko tako denimo izreče »o tem se ne želim pogovarjati, to je zame občutljiva tema« ali pa »ne morem povedati ničesar, podpisal sem pogodbo o tajnosti« brez kršitve Gricovih principov in maksim. Ob tovrstnem upoštevanju maksim, sama izjava odstopa od sodelovalnega načela ustreza tudi samemu sodelovalnemu načelu. Čeprav se s to izjavo predaja informacij in prepričanj konča, bi bilo težko vzpostaviti, da gre za jezikovno hibo.

Pri trku dveh maksim je ena izmed njiju očitno prekršena v prid druge, zato se je potrebno sklicevati na preostale navedene pogoje. Da ne gre za jezikovno hibo, marveč za zgolj optimalno a še zmeraj vrlo pragmatično ravnanje, je potrebno, da govorec še zmeraj ravna v skladu z osnovami sodelovalnega načela. Prav tako je potrebno, da govorec pravilno presodi, da je prekršiti eno maksimo v prid druge potrebno, ter da prav tako pravilno presodi, katero izmed maksim, ki sta v položaju trka, je primerneje prekršiti (kaj bo ugodneje za konsistenco konverzacije in predajo informacij).

Preziranja in izkoriščanja maksim so kot izpostavlja Grice, najbolj problematična za tako pragmatično obravnavo kot obravnavo skozi prizmo koncepta jezikovnih hib. V ta namen bomo sledili Griceovemu lastnemu naboru primerov, ki jih poda v drugem delu članka.

Najprej obravnava primere, v katerih ni prekršena nobena maksima. Prvi je pogovor med osebama A in B, kjer A pove, da ji je zmanjkalo goriva, B pa odvrne, da je za vogalom avtogaraža. Kljub temu, da ni bila prekršena nobena maksima, izjava še zmeraj vsebuje implikaturo, da lahko A v tej avtogaraži kupi gorivo. Naslednji primer je pogovor, v katerem A izjavi »Zdi se mi, da Smith trenutno nima dekleta,« B pa na to pripomni, da je bil Smith zadnje čase veliko v New Yorku. Smiselna implikatura za tem je, da B predpostavlja, da ima Smith dekle v New Yorku (ter da je to dekle razlog za njegove pogoste obiske mesta). Pomembna razlika med tema primeroma je, da kljub temu, da sta v skladu z našo opredelitvijo oba jezikovno vrla, prvi prenaša znanje, drugi pa zgolj prepričanje, kar ima pomembne implikacije za opredelitev funkcije jezikovnih vrlin in hib.

Pri naslednjem primeru je ena maksima prekršena zaradi trka z drugo. Oseba B na vprašanje, kje živi C, odgovori enostavno, da ta živi na jugu Francije. Gre za kršitev prve maksime količine zato, da bi ugodili maksimi kakovosti – se pravi, B ima omejeno znanje o lokaciji C, natančneje: omejeno na jug Francije. V kolikor bi B imenovala naključno mesto na jugu Francije, češ da tam živi C, brez primernega znanja o tem in torej brez primernih dokazov za svojo izjavo, bi prekršila drugo maksimo kakovosti. Pomembno v tem primeru je tudi, da C s tem, ko se odloči za kršitev maksime zaradi trka, s svojo izjavo ne le izrazi vsebino prepričanja oz. znanja, ampak jasno opredeli tudi obseg svojega znanja.

Sledijo primeri zlorabe maksim- najprej prezir prve maksime količine. Profesor piše priporočilo za študenta filozofije, v katerem hvali njegovo znanje angleščine in delovnost, ne omeni pa znanja filozofije. Implikatura za tem je, da profesor meni, da študent ni dober filozof, kar implicira z izpustitvijo ključnega podatka (tj. o znanju filozofije, kakršnokoli že naj to bo) iz pisma. Ker je skupno znanje profesorja in bralca pisma o pričakovani vsebini pisma brez dvoma konsistentno, je prenos želenega prepričanja učinkovit, v tem primeru ne gre za jezikovno hibo.

Popolnoma drugačni so primeri prezira iste maksime v nekaterih tavtologijah, ki jih navaja Grice, denimo generalizirana tavtološka izjava »Ženske so ženske.« Ta izjava je semantično popolnoma prazna, sklicuje se na predpostavljeno skupno *prepričanje* (in nikakor ne znanje) o stereotipni naravi žensk. Ker je to skupno prepričanje samo predpostavljeno, izjava tudi s pragmatičnega stališča ne doseže drugega, kot da v vsakem potencialnem poslušalcu zbudi njemu samemu lastno generalizirano predstavo ženske.

Poglejmo si še preziranje prve maksime kakovosti, kar se dogaja v ironičnem in metaforičnem govoru. Če oseba A govori o nekem znancu X, ki ga je pred kratkim izdal, pravi Grice, in o njem izjavi, da je dober prijatelj, pove točno to, za kar ve, da ni resnično. Na poznavanju konteksta med udeleženci v konverzaciji pa sloni, ali bo primerno prepričanje, tj. obratno od tega, ki je bilo neposredno izraženo, primerno prenešeno. Pri metafori »ti si [kot] smetana v moji kavi« je podobno potrebno poznati kontekstualne podrobnosti dejanske smetane v kavi (npr. smetana kavo obogati), nato pa potegniti vzporednice na podlagi poznavanja konteksta konverzacije, v kateri je metafora uporabljena, npr. govorec podaja prepričanje o svoji partnerici, da slednja obogati njegovo življenje, kot smetana obogati kavo. Preziranje druge maksime odnosa je nekoliko drugačno. Če oseba A o X-ovi ženi

reče, da ga danes verjetno vara, lahko to temelji na utemeljenih, ampak dokazno nezadostnih kontekstualnih znanjih – denimo da ga je v preteklosti že varala – ali pa je izraženo popolnoma brez dokazov (implicira pa med drugim tudi, da dokazi obstajajo, čeprav ne). Gre za jasen primer neiskrenosti, prav tako pa tudi za najverjetneje najbolj jasen primer jezikovne hibe.

Primeri kršitev maksime odnosa so, trdi Grice, najredkejši. Grice ponudi primer, v katerem med čajanko A reče, da je neka gospa X stara vešča. Po trenutku tišine B komentira, da je vreme zadnje dni zelo prijetno. Implikatura tega je, da je A-jeva opazka neprimerna, cilj pa, da naj ne bo predmet nadaljnjega pogovora. Tovrstna preusmeritev je sorodna s poprej omenjenim odkritim odstopom od maksim ali sodelovalnega načela v celoti – dokler je razlog za kršitev jasno razviden in ugaja opisanim trem kriterijem, ne moremo trditi, da gre za jezikovno hibo.

Ko pridemo do kršitev maksim načina, so dvoumnosti pogosto jezikovno specifične, zaradi česar Griceove primere v slovenskem jeziku težko analiziramo. Vendar pa lahko govorimo o maksimah jasnosti in jedernatosti. Primer nejasnosti, ki ga najdemo v Griceovem delu je primer staršev, ki se v prisotnosti otroka pogovarjata namenoma nejasno, da ta ne bi razumel tematike njunega pogovora. Vprašanje razloga, zakaj natanko starša tako ravnata, je na tej točki nepomembno. Njun način konverzacije onemogoča prenos prepričanj, ki jih njune izjave vsebujejo, na tretjega poslušalca, zaradi česar gre za jezikovno hibo. Upravičenost skrivanja vsebine pred otrokom ni predmet jezikovnega aspekta situacije, marveč kvečjemu moralno ali družbeno vprašanje. Hibnost tovrstnega ravnanja bi bila jasneje razvidna v primeru, kadar bi se na tak način pogovarjala denimo zdravnik in medicinski tehnik v pričo pacienta o stanju slednjega. Nazadnje si poglejmo še Griceov primer kršitve kratkosti in jedrnatosti, kjer nas nagovori, da primerjamo stavka »Gdč. X je zapela 'Home Sweet Home'« in »Gdč. X je proizvedla serijo zvokov, ki se bližnje ujemajo z notnim zapisom za 'Home Sweet Home'« (Grice, 1975, 55). Pri tem se pojavlja implikatura, da cinično razvlečen način zapisa izraža, da je njegov avtor mnenja, da je gdč. X pesem odpela slabo, razglašeno, ali kaj podobnega. Ker pa je tudi to ponovno jasno prepoznavno iz splošnega konteksta in znanja o jeziku in kulturi, ki jezik uporablja, ponovno ne gre za hibo.

Kaj lahko na koncu analize Griceovih maksim povemo o naravi jezikovne vrline in hibe je, da je jezikovna vrlina način konverzacije, ki se sklada s spodnjimi pogoji:

1. Govorec upošteva NAJMANJ sodelovalno načelo.
2. Govorec meni, da je implikat, v kolikor ta obstaja, potreben za interno konsistenco izjave.
3. Govorec upravičeno meni, da je poslušalec zmožen dojeti potrebnost predpostavke IN kaj in zakaj je implicirano, na podlagi znanja jezika, nabora znanj v ozadju, konteksta, itd.

Pod tretji pogoj smo dodali »upravičeno« meni, kar pomeni, da mora imeti govorec zadostne razloge, da predpostavlja poslušalčeve zmožnosti. Denimo, da profesor filozofije jezika pri razlagi Gricea izpusti podatke o pomenu terminov semantika in pragmatika, saj predpostavlja, da so njegovi poslušalci z njima seznanjeni iz začetnih predavanj pri dotičnem predmetu. Zaradi tega pa dijak, ki je prišel poslušati predavanje s starejšim prijateljem, ki je študent filozofije, ne razume določenega dela predavanja. Kljub temu, da ta poslušalec ni bil zmožen dojeti nekaterih stvari, ki jih je povedal profesor, profesor ni prekršil nobenega izmed zgornjih pogojev.

Jezikovna hiba, po drugi strani, je način konverzacije, ki ne zadosti enemu ali več izmed zgornjih pogojev.

Nazadnje naj izrecno opredelimo še to, na kar smo v zgornji analizi nekajkrat že grobo nakazali: kaj pa »počnejo« jezikovne vrline in hibe?

V osnovi jezikovna vrlina počne to, kar počne tudi argumentacijska vrlina, namreč širi znanje, in sicer izven argumentacijskega okolja. Vendar pa funkcija jezikovne vrline ni omejena zgolj na širjenje znanj (pri čemer naj ta dojemamo kot upravičena resnična prepričanja), temveč lahko gre tudi za širjenje enostavnih prepričanj, ki niso nujno resnična, niti nujno upravičena. To pa velja le pod pogojem, da je iz narave implikature oziroma izjave kot celote poslušalcu pod zgornjimi pogoji razvidno, da ne gre za znanje, in da se izražena prepričanja v skladu z maksimama kakovosti, če so neupravičena ali neresnična, ne nanašajo na stvari, o katerih govorec ve karkoli resničnega. Tako je dodatna funkcija jezikovne hibe tudi ta, da lahko širi neupravičena ali neresnična prepričanja na tak način, da jih poslušalec zmotno razume kot upravičena in resnična prepričanja oziroma znanja zaradi zlorabe pragmatičnih maksim in zgornjih pogojev.

Po tej osnovni opredelitvi se lahko obrnemo še na izbor jezikovno pragmatičnih vedenj in pojavov, ki predstavljajo odlične kandidate za poglobljene primere jezikovnih hib, a ležijo izven strogega obsega Griceovih pragmatičnih maksim.

Prvi sklop bodo predstavljali različni načini govorjenja neresnice, kot jih obravnavata Cappelen in Dever (2019). Med temi je prva kategorija preprosto zmoten ali napačen govor, ki se lahko pojavlja na več načinov. Lahko denimo govorimo zmotno enostavno zato, ker se motimo, se pravi, ker ne poznamo resnice o p, imamo pa denimo neupravičene in neutemeljene dokaze na podlagi katerih verjamemo, da p. Nadalje lahko govorimo zmotno zato, ker nismo vložili zadostnega truda v pridobivanje dokazov – najpogosteje gre za govor na podlagi slutenj, intuicije ali čustev. Nazadnje je eden izmed možnih kandidatov za zmoten (ali v tem primeru neresničen) govor fikcija, ki pa od ostalih načinov bistveno odstopa. Kaj pa lahko o teh načinih zmotnega govora povemo v sklopu jezikovnih hib?

Kadar govorimo zmotno na podlagi tega, da verjamemo, da p, imamo pa neupravičene in neutemeljene dokaze za naš govor, kršimo Griceovo drugo maksimo kakovosti, se pravi govorimo nekaj, za kar nimamo ustreznih dokazov – natančneje, kljub temu, da dokaze imamo, so ti neustrezni, tudi če se tega ne zavedamo. Nadalje, prav zaradi našega nezavedanja zmotnosti lastnega govora in pomanjkljivosti dokazov v tovrstni situaciji, nimamo načina, da bi lahko predvidevali, da bo naše prepričanje sogovorec prepoznal kot golo prepričanje ne znanje, saj zaradi svojega zaupanja, v sicer neustrezne dokaze, izraženo prepričanje predstavljamo kot znanje, ko to ni. V skladu z našo opredelitvijo gre v tem primeru brez dvoma za vrsto jezikovne hibe. Nekoliko kontroverzen primer tega bi utegnilo biti Platonovo prepričanje, da je Zemlja krogla. Svoje (sicer resnično) prepričanje je Platon utemeljeval na podlagi tega, da ja krogla najpopolnejša oblika, zaradi česar ima vesolje obliko krogle – kar velja za celoto, pa velja tudi za del, ki je središče vesolja, tj. Zemljo, ki mora imeti zato prav tako obliko krogle. Platonovo sklepanje, da je Zemlja krogla, temelji na dokazih, ki pa niso ustrezni: krogla ni absolutno najpopolnejša oblika (čeprav je najobstojnejša v pogojih, ki vladajo v vesolju – podatek, ki pa Platonu ni bil dostopen), kar velja za celoto ne velja nujno tudi za del, Zemlja pa ni središče vesolja. Platon bi lahko s pomočjo tovrstnega sklepanja na podlagi poljubnih neresničnih premis prišel do poljubnega drugega sklepa o obliki Zemlje ali drugem pojavu. Podobno kot denimo zgodnje novoveško prepričanje, da podgane spontano nastanejo v prisotnosti umazanije, tudi Platonovo prepričanje temelji na zdravorazumsko smiselnih, a neresničnih premisah, le da je Platonov sklep

naključno pravilen – do velike mere je njegovo resnično prepričanje o obliki Zemlje posledica spoznavne sreče. Morda bi lahko Platonovo sklepanje branili na podlagi zgodovinske umeščenosti in omejenih spoznavnih orodij, ki so mu bila na voljo, pa vendar je manj kot stoletje kasneje Eratosten empirično in matematično dokazal okroglost Zemlje in tudi izračunal njen obseg in premer. S tem je prepričanja o obliki Zemlje upravičil. Tudi če Platonu ne pripišemo neposredne spoznavne hibe, pa je njegova argumentacije vendarle jezikovno in logično težavna.

Kadar govorimo zmotno na podlagi tega, da menimo, da p, a nismo vložili zadostnega truda v iskanje dokazov, prav tako prekršimo Griceovo drugo maksimo kakovosti, saj dokazov sploh nimamo, kaj šele, da bi bili ti ustrezni. V tem primeru je zelo verjetno, da se lastnega pomanjkanja dokazov zavedamo. V primeru, ki ga podajata Cappelen in Dever, kjer imamo zlo slutnjo, da je oseba A morilec, a za to nimamo nikakršnih dokazov, to brez dvoma drži. V kolikor se kljub zavedanju pomanjkanja dokazov odločimo predstaviti naše prepričanje kot resnično, poleg maksime kakovosti prekršimo tudi drugi pogoj jezikovne vrline, zaradi česar smo zagrešili jezikovno hibo. V kolikor pa se odločimo predstaviti naše prepričanje kot takšno, kakršno je, se pravi v skladu s tretjim pogojem sogovorcu primerno nakažemo, da naše prepričanje izvira iz intuicije, zanj nimamo dokazov, itd., pa vendarle ni bila zagrešena jezikovna hiba.

Ta dva primera nam postrežeta s taksonomsko lastnostjo jezikovnih hib, ki temelji na Cassamovi (2019) razdelitvi spoznavnih hib na prostovoljne ali neprostovoljne. V primeru nezadostnega truda pri iskanju dokazov je zaradi zavedanja pomanjkanja dokazov zagrešena hiba, če do nje pride, prostovoljna. Po drugi strani se v primeru neustreznih dokazov ne zavedamo neustreznosti dokazov, torej je hiba neprostovoljna (kar izpostavljata tudi Cappelen in Dever).

Fikcija je v pragmatiki, pa tudi v spoznavni teoriji zelo zanimiv pojav, saj je vsebina fikcije brez dvoma neresnična. Ne obstaja namreč šola za čarovnice imenovana Bradavičarka, kot tudi ne obstaja zlobno kraljestvo imenovano Mordor, zato je vse, kar je o njima povedano v dotičnih serijah fikcijskih del, neresnično v kontekstu, v katerem se ideja resničnosti sklicuje na svet, v katerem živimo. Po drugi strani je mogoče reči, da je vse, kar se pojavi v delu fikcije, resnično v svetu, ki ga fikcijsko delo vzpostavlja in opisuje – potrebno pa je izpostaviti, da je delo zares fikcijsko (Lewis 1978), kar ustreza našim pogojem jezikovne vrline.

Naslednja kategorija je laganje, pri čemer Cappelen in Dever izpostavljata, da ga je potrebno ločiti od enostavnega zavajanja. Pri tem je laganje definirano kot »podati izjavo, za katero verjamemo, da je neresnična drugi osebi z namenom, da bi ta druga oseba verjela, da je izjava resnična« (2019, 40). Avtorja poudarjata, da se je potrebno upreti težnji po bolj reduktivni definiciji.

V primeru laganja je jasno, da gre za jezikovno hibo, saj namen laganja po zgornji definiciji vsebuje namerne kršitve prve maksime kakovosti, saj se govorec ne le zaveda, da jezikovno dejanje za konverzacijo ni potrebno, temveč ve, da je škodljivo za konsistenco konverzacije, prav tako pa vsaj predvideva, da sogovornik ni primerno opremljen s predznanji, kontekstom, ipd., da bi se zavedal lažnosti prepričanja, ki ga vsebuje izjava.

Bolj zanimivi so obrobni primeri laganja, kot denimo, ali je mogoče lagati na tak način, da povemo resnico. Faulknerjev (2007) primer predpostavlja scenarij, v katerem A skriva osebo B pred nacisti. Oseba B ob prihodu nacistov pobegne, A pa jim, nevedoč, da B ni več tam, pove, da ne skriva nikogar. V tem primeru namen ostaja enak – A še zmeraj govori nekaj, kar, kot veli (ali bolje prepoveduje) Griceova prva maksima kakovosti, *verjame*, da ni resnično. Če zanemarimo moralne posledice A-jinega ravnanja na stran, je še vedno res, da je A zagrešila jezikovno hibo. Cappelen in Dever sicer trdita, da v tovrstnih primerih A poskusi lagati, a ji lagati ne uspe, pri čemer predpostavljata, da je potrebno za laganje ugoditi tako notranjim kot zunanjim dejavnikom, vendar pa se s tem pogledom ne moremo strinjati. Notranji pogoji, ki jih zahtevata avtorja, vsebujejo namene in mentalna stanja govorca, medtem, ko zunanji pogoji obsegajo dejstva o svetu. Na tej točki si je smotrno zadevo ogledati s stališča govornega dejanja. Austin (1962) v goli osnovi našteje tri vrste: lokucijska, ilokucijska, in perlokucijska. Pri tem ugotavljamo, da gre pri lokuciji, ilokuciji ter perlokuciji za različne aspekte ali dele govornih dejanj in ne za kategorije govornih dejanj (Šetar 2020), pri čemer je lokucija semantična vsebina (kaj je rečeno), ilokucija intencionalna vsebina (kaj je govorčev namen), perlokucija pa konsekvencionalna vsebina (kaj so posledice v zunanjem svetu) vsakega govornega dejanja. Pri tem se lahko ilokucijski in perlokucijski aspekt govornega dejanja razlikujeta, že Austin pa izpostavlja tudi, da je govorno dejanje izvedeno, tudi če ni uspešno. Kaj to pomeni za zgornji primer? Da je ilokucija A-jine izjave nacistom doseči, da slednji verjamejo v resničnost vsebovane propozicije (da B ni tam) in, da jih zavede, saj verjame, da je propozicija neresnična. Po drugi strani je perlokucija te izjave, da nacisti bodisi verjamejo, bodisi ne verjamejo v resničnost propozicije, posledično pa odidejo,

preiščejo hišo, ali karkoli drugega. Če nacisti verjamejo A-jini izjavi, je govorno dejanje uspešno; če ne, je neuspešno. V vsakem primeru pa je bilo izvedeno. A je, z njenega lastnega stališča, nacistom vendarle lagala.

Kljub temu, da je laganje v osnovi najbolj izrazita oblika zavajanja, pa obstajajo tudi primeri, ko laganje ni nujno zavajajoče, denimo sledeče:

> »Študentko, obtoženo goljufanja na izpitu, pokličejo v dekanovo pisarno. Študentka ve, da dekan ve, da je resnično goljufala. A ker je prav tako dobro znano, da dekan ne bo kaznoval nikogar, razen če ta prizna svojo krivdo, študentka torej, četudi ve, da bo dekan vedel, da ne govori resnice, zgolj da bi se izognila kazni reče: /.../ Nisem goljufala« (Carson, 2006, 290).

Dever in Cappelen se v ta primer sicer ne poglabljata, je pa izredno zanimiv v primeru jezikovnih hib. Kljub temu, da študentka ne govori resnice in s tem prekrši Griceovo prvo maksimo kakovosti, pa s tem še ne prekrši pogojev za jezikovno vrlino, saj upošteva najmanj sodelovalno načelo, prav tako pa ve, da je kršitev nujna za pragmatično konsistenco konverzacije, prav tako pa ve tudi, da je sogovorec (tj. dekan) zmožen vedeti oziroma ve, da ni povedala resnice in zakaj.

Manj zanimive, a vendarle vredne omembe so bele laži, kot denimo, če nas prijateljica vpraša, ali ji nova obleka pristoji, mi pa, ji kljub nasprotnemu mnenju, odgovorimo pritrdilno. Namen za tem je moralno dober, saj želimo, da se prijateljica glede svojega nakupa in izgleda počuti dobro, a smo s tem prekršili maksimo kakovosti. Ali gre za jezikovno hibo ali ne, je bržkone odvisno od vsakega posameznega primera – nekatere bele laži, kot denimo zgornja ali pa bela laž zdravnika otroku, da injekcija ne bo bolela, so dovolj pogoste, da ima sogovorec potrebno kontekstualno znanje o rabi tovrstnih laži, da jih prepozna kot implikature, ki ustrezajo pogojem jezikovne vrline. Po drugi strani je lahko bela laž uporabljena v bolj specifičnih okoliščinah, kjer sogovorec nima razpoložljivih sredstev, da jo prepozna kot laž, v tem primeru pa je prvi govorec zagrešil jezikovno hibo.

Nadaljujemo z oblikami zavajanja, ki pa ne vsebujejo neposrednega laganja. Analizirali bomo primere, ki jih ponujata Cappelen in Dever. Prvi je primer obtoženca, ki je na zatožni klopi vprašan, kolikokrat je obiskal oropano trgovino, na kar odvrne, da jo je obiskal petkrat, čeprav jo je obiskal petdesetkrat. Njegova izjava je logično gledano resnična v smislu, da jo je obiskal *vsaj* petkrat, kljub temu pa h

konverzaciji ne prispeva ustrezno. Prekrši namreč prvo maksimo količine, saj ne pove dovolj, na nek način pa tudi eno izmed maksim načina, saj je glede na kontekst namenoma nejasna. Govorec računa na to, da sogovorec, oziroma v tem primeru prej poslušalec, namreč porota, nima zadostnega poznavanja konteksta, da bi dojela vsebovano implikaturo. Zaradi tega, ker namenoma, a tiho odstopa od prispevka v konverzaciji na splošno, pa utegnemo reči celo, da krši sam sodelovalno načelo, torej nedvomno gre za primer jezikovne hibe.

Nadalje si oglejmo primer, v katerem je tobačno podjetje Winston oglaševalo, da njihove cigarete (v nasprotju z ostalimi znamkami) ne vsebujejo aditivov. Implikatura, trdita Cappelen in Dever, je ta, da so zaradi pomanjkanja aditivov njihovi cigareti bolj zdravi (oziroma, saj vendarle govorimo o cigaretih, manj škodljivi) kot cigareti drugih proizvajalcev. Na nek način bi lahko rekli, da je bila prekršena prva maksima kakovosti, saj oglas ne poudarja, da to pomeni, da so cigareti manj zdravi; vendarle pa tovrstnih omejitvenih izjav (v smislu angleškega izraza *disclaimer,* ki pa v slovenščini žal nima lepe ustreznice) v okviru maksim ni pričakovati. Po drugi strani pa so predstavniki podjetja, ki so snovali oglas, predvidevali prav to, da tarčna javnost tega ne bo vedela – nasprotno, predvidevali so, da bodo opazovalci oglasa zmotno prepoznali implikaturo, da pomanjkanje aditivov pomeni, da so cigareti zdravi. Ta implikatura pa sledi iz kršitve druge maksime, tj. maksime načina, ki se navezuje na izogibanje dvoumnosti. Ponovno lahko rečemo, da gre za jezikovno hibo.

Zadnji primer si avtorja izposodita iz dela Jennifer Saul (2012, 73):

> »George namerava ubiti Friedo. Frieda je alergična na arašide in George zanjo pripravi jed, ki vsebuje velike količine arašidovega olja. Frieda je nekoliko zaskrbljena in Georgea vpraša, ali je hrana zanjo varna? Primerjajte sledeča odgovora:
>
> 1. Hrana je zdrava – od nje ne boš zbolela.
> 2. Hrana ne vsebuje arašidov.«

Pri prvi izjavi gre za laganje, saj se soočamo z namerno kršitvijo prve maksime kakovosti – George je namenoma izgovoril neresnico, saj ve, da bo Frieda v najboljšem primeru zelo zbolela. Pri drugi izjavi, po drugi strani, gre za zavajanje, saj je prekršena prva maksima količine hkrati z maksimo odnosa, saj niso podane vse

relevantne informacije. Oboje sta jasna primera hibe – kot izpostavlja Saul, pa tudi Cappelen in Dever, je pomembno sporočilo tega primera, da spoznavne in moralne posledice laganja niso nujno hujše in bolj grajevredne kot posledice zavajanja.

V nadaljevanju bomo analizirali še vlogo resnice v komunikaciji, pri čemer se osredotočamo na teorije posebne resnice in neposebne resnice (ang. *truth-special* in *truth-nonspecial*). Pristop posebne resnice, ki je v filozofiji jezika najpogosteje sprejet, trdi, da je resničnost izjav v komunikaciji nujna za njen uspešen potek. Po drugi strani, Wilson in Sperber (2002) zagovarjata stališče, da je kriterij nujne resničnosti prestrog – komunikacija lahko poteka uspešno tudi tedaj, ko niso vse izjave nujno resnične. Primeri, ki jih avtorja podajata, vsebujejo denimo humor – neka izjava je lahko neresnična, a vseeno učinkovita saj je njen namen proizvesti komični učinek, ne pa podajanje neke resnice o svetu – in posplošene ali morda bolje rečeno poenostavljene izjave: če rečem, da potrebujem viledo, sem povedal, da potrebujem neke vrste vpojno krpo, ne pa nujno tega, da potrebujem vpojno krpo znamke Vileda, kar lahko sogovorec tudi primerno prepozna. Če sprejmemo stališče Wilsona in Sperberja se lažje izognemo trivializaciji jezikovne hibe na vse vrste neidealnega govora. Kot pa opažata tudi Cappelen in Dever, je neposebni pristop nekoliko preveč ohlapen in potrebuje določene dopolnitve: resnica ni posebna v smislu, da ni nujna za uspešno komunikacijo, vendar pa je za uspešno komunikacijo potrebna pragmatična zmožnost občinstva, da prepozna neresničnost govorčeve izjave in namen govorca, da je občinstvo zmožno prepoznati neresničnost njegove izjave.

Laganju soroden, a vendarle od njega ločen pojav, je nakladanje, katerega prvi podrobneje opiše Harry Frankfurt (2005) v eseju *O nakladanju* (ang. *On Bullshit*). Poglavitna razlika med laganjem in nakladanjem leži v tem, da je lažnivec zavezan k resnici v smislu, da se resnice zaveda, a se zavestno odloči govoriti to, za kar ve, da ni resnica. Po drugi strani, nakladač nima nikakršne tovrstne zavezave – nakladaču je popolnoma vseeno, kakšna je resnica. Nakladač izbira ali pa si enostavno izmišljuje izjave ali propozicije o svetu, ki ustrezajo njegovim instrumentalnim ciljem, ne da bi se na kakršenkoli način oziral na dejstva ali dejansko stanje stvari. Cappelen in Dever podajata primer misinformacij, ki jih je konzervativna opozicija v ZDA širila o nekdanjem predsedniku Baracku Obami, češ da ta sploh ni državljan ZDA in je rojen v Keniji. Pripadnikom opozicije, ki so širili to prepričanje, je bilo docela nepomembno, ali je to res ali ne, marveč je bilo pomembno le, da jim koristi pri širjenju nezaupanja v predsednika Obamo.

Po Frankfurtovi definiciji nakladač zagreši jezikovno hibo, saj prekrši drugo maksimo kakovosti, tj. govori za kar nima dokazov – na slednje se namreč požvižga – ob tem pa ne zadosti drugemu in tretjemu pogoju za jezikovno vrlino, tj. kršitev ni potrebna in poslušalec nima podlage razbrati, da gre za kršitev.

Cappelen in Dever pa se s tako široko definicijo ne strinjata, saj trdita, da ta zaobjema tudi zgodbičarje, ki se prav tako požvižgajo na dejstva in resnice o svetu, a s popolnoma drugačnim namenom. Zgodbičar prezira dejstva in resnico, a ob tem občinstvo jasno ve, na kakšen način in zakaj to počne. Avtorja trdita, da Frankfurtovi definiciji manjka element zavajanja, saj se nakladač skrivaj požvižga na resnico, pred poslušalci pa izgleda kot normalen konverzacijski partner. Prav tako ima zgodbičar primeren odnos do resnice, saj spoštuje resnico v zunanjem svetu, pa tudi interno resnico znotraj fikcijskega sveta. Nakladač ne vzpostavlja fiktivnega sveta, v katerem bi bila njegova izjava resnična, saj se sploh ne zanima za resnico.

Poleg navadnega nakladanja pa obstaja še fenomen globokega nakladanja, pri katerem se nakladač ne le požvižga na resnico, marveč se prav tako ne ozira na to, ali so njegove izjave sploh smiselne, konsistentne, ali celo ali sploh imajo kakršenkoli pripisljiv pomen (Cohen 2002). Cappelen in Dever v prvi vrsti ilustrirata govor brez pomena na primeru pesmi Lewisa Carrola z naslovom *Jabberwocky*, kjer že sam naslov ne pomeni prav nič, prav tako pa tudi večina popolnoma izmišljenih besed in izrazov v tem otroškem delu. Seveda pa gre znova za primer zgodbičarstva, kjer avtor poskrbi za to, da se občinstvo zaveda, kaj je prekršeno in kako.

Cappelen in Dever omenita tudi globoko nakladanje v kontekstu Carnapove verifikacionistične teorije pomena. Res je, da se večina sodobnih filozofov jezika z njo ne strinja, vendar tudi v sodobnih razpravah najdemo zanimive primere. Vzorčen primer je tekst Alana Sokala, ki je bil objavljen v reviji *Social Text* kljub temu, da je bil spisan v celoti brezpomensko. Sokal je namenoma uporabljal izražanje konsistentno z globokim nakladanjem, kar so kasneje opredelili kot Sokalovo šalo. Sokal in Jean Bricmont (1999) sta nadgradila to intenco v knjigi, v kateri poudarjata vlogo globokega nakladanja v psevdo-znanstvenem šarlatanizmu, pri čemer dodatno poudarjata primere v delu Jacquesa Lacana, češ da slednji uporablja nedefinirane izraze in besedne zveze, ki ustrezajo prej globokemu nakladanju kot pa akademskemu govoru oz. filozofski terminologiji. Cappelen in Dever opozorita še na primer koncepta rase, ki se v sodobnem javnem, akademskem in političnem dialogu rabi bodisi kot biološko določena skupina, družbena vrsta, ali pa pravzaprav

ne pomeni nič. Avtorja povzemata po Appiahu (2002), da je sodobna biologija dokazala, da rasa ni biološka, saj ima kakršna koli biološka vrsta, na katero naj bi se posamezna rasa nanašala, preveliko število izjem v smislu posameznikov, ki naj bi ji na videz pripadali, a ji ne, ali pa teh, ki ji naj ne bi, a (npr. genetsko) ji. Prav tako težko trdimo da gre za družbeno vrsto, saj za nobeno raso, kot te tradicionalno dojemamo, ni moč trditi, da imajo vsi pripadniki te rase iste družbeno določljive lastnosti (npr. običaje, vere, kulinariko, itd.). Edina smiselna razlaga je torej, da izraz rasa pravzaprav ne označuje ničesar opredeljivega in je torej brezpomenski – gre za obliko globokega nakladanja.

Oba primera, tako Lacanovi izrazi kot izrazi rase, sta po zgornji opredelitvi jezikovni hibi. A za razliko od navadnega nakladanja, pri primerih globokega nakladanja ne gre za kršitev druge maksime kakovosti, marveč gre za kršitve maksime načina. Globoko nakladanje je namreč nejasno, dvoumno ipd. in zavajajoče na tak način, da se poslušalcu predstavlja kot jasno, enoznačno in opredeljeno, pri čemer pa poslušalec nima pogojev (če pa jih ima, pa govorec računa na to, da nima pogojev), da jih prepozna, da gre za dvoumne, nejasne, ipd. figure govora.

Naslednji kandidat za jezikovne hibe so leksikalni učinki – oziroma, natančneje, zloraba leksikalnih učinkov. Cappelen in Dever leksikalne učinke opredeljujeta kot učinke določenih vrst izražanja, ki v svoji naravi niso kognitivni, marveč afektivni ali čustveni, mednje pa sodijo metafore, poimenovanja, kodirane besede (ang. *code words, dogwhistles*) in slabšalnice.

Metafora, trdi Donald Davidson (1978), v nasprotju s splošnim prepričanjem, ni večpomenska, marveč ima samo en, dobeseden pomen – vse ostalo so nepomenski (se pravi, ne-semantični) učinki metafore, ki se izražajo skozi njeno interpretacijo. Dobra metafora se tako zanaša na to, da lahko občinstvo prepozna, zakaj je metafora – pri katerih gre večinoma za namerne kršitve maksime načina, natančneje jasnosti in nedvoumnosti – na mestu, ter na podlagi splošnega konteksta besedila in primernih predznanj (vsaj do zadostne mere) prepozna primerno interpretacijo glede na avtorjev namen. Zdi pa se, da lahko obstajajo primeri, v katerih je avtorjev namen interpretacija metafore v skladu z njegovim namenom, ki pa se zanaša na nepoznavanje določenih kontekstov in znanj. Cappelen in Dever podajata primer, v katerem je Donald Trump ameriško demokratsko stranko oklical za »ladjo brez krmila,« s čemer je želel doseči, da občinstvo prepozna njegovo namero, da demokratsko stranko opiše kot neorganizirano, brezciljno, ipd. V primeru, da je

demokratska stranka popolnoma sprejemljivo organizirana in ima popolnoma sprejemljivo zarisane politične cilje, je implikatura te metafore napačna, poslušalec, kateremu manjka predznanje oz. kontekstualno znanje o političnem delovanju te stranke, pa bo implikaturo sprejel za resnično. Medtem, ko je poslušalec ob tem, ko je implikaturo sprejel za resnično, kriv neke vrste spoznavne hibe, pa ostaja vprašanje, ali je predsednik Trump kriv zagrešitve jezikovne hibe.

Brez dvoma se je pri podajanju svoje izjave držal minimalno sodelovalnega načela, saj je rdeča nit konverzacije bržkone podajanje političnega mnenja. Je menil tudi, da je implikatura nujna za interno konsistenco izjave? Tukaj se zadeva zaplete – v luči tega, da je metafora prepoznavna občinstvu in da interna konsistenca izjave vzdrži samo metaforo, ne gre za jezikovno hibo. Vendar pa je morda vredno pogledati, *kaj* natanko je Trump *želel* povedati. V skladu z Davidsonom je Trump povedal, da je demokratska stranka dobesedno ladja, ki nima krmila, želel pa je povedati, da je neorganizirana in brezciljna. Če to privzamemo kot ilokucijo metafore kot govornega dejanja, nenadoma jezikovna hiba leži v tem, da je Trump prekršil eno izmed maksim kakovosti – povedal je namreč nekaj, za kar je bodisi vedel, da ni resnično, bodisi ni imel zadostnih dokazov za to, da bi lahko trdil, da je resnično. Ta primer pa nam pove, da metafora sama po sebi ni jezikovna hiba, marveč je to izjava, ki se skriva v implikaturi za metaforo – pri tem gre prej za zmotno govorno dejanje.

Poimenovanja, o katerih govorita Cappelen in Dever, so predvsem poimenovanja blagovnih znamk, pa tudi poimenovanja ljudi, tj. osebna imena. Avtorja opisujeta denimo leksikalne učinke imen blagovnih znamk kot je učinek imena Coca-Cola v smislu asociacije tega imena z Božičem, prijetnimi občutki ipd., ki ga podjetje »izdeluje« in promovira že skoraj stoletje, ali pa v negativnem primeru ime podjetja Malaysian Airlines, ki zaradi dveh smrtonosnih nesreč leta 2014 še danes nosi slabo konotacijo. Ker je slednje pridobilo svoj negativen leksikalni učinek zaradi naključnega spleta okoliščin, seveda podjetju kot sporočevalcu ni mogoče pripisati nikakršnega namenskega ali lingvističnega delovanja, ki bi lahko proizvedlo jezikovno hibo. V primeru Coca-Cole pa prav tako ni mogoče določiti kakršnekoli implikature v oglaševanju podjetja. Da Božiček pije dotično pijačo je fiktiven scenarij, da utegnejo člani neke družine ob Božiču uživati v pitju Coca-Cole je enostavno uprizoritev možne situacije, itd. Četudi je namen podjetja, da potrošniki kot poslušalci njegovih konverzacijskih naprezanj njihov proizvod povežejo z ugodnimi občutki v obdobju praznikov, ne gre za jezikovne hibe marveč za neko ne-jezikovno obliko psihološke manipulacije.

Leksikalne učinke lahko sprožijo tudi slabšalnice in psovke. Pri tem gre za naslednje pristope: deskriptivni, ekspresivni, presupozicijski in prohibicionistični. Deskriptivni trdi, da so slabšalnice in psovke le konotirani opisi konceptov: v kolikor nekomu rečem, da je čefur, to pomeni le, da gre za priseljenca iz drugih držav bivše Jugoslavije, pri čemer osnovno koncepcijo tega priseljenca opremim z dodatnimi (negativnimi) atributi, tj. impliciram, da je len, pokvarjen, nasilen. Ekspresivni pogled na slabšalnice in psovke trdi, da slednje izražajo neko osebno negativno stališče govorca do nevtralnega jedrnega koncepta, v primeru čefurja torej, da sem do priseljencev iz bivše Jugoslavije sam negativno nastrojen, ne pa da nujno pridam nek specifičen slabšalni pomen – ta je lahko sam po sebi objektivno neopisljiv, denimo v primeru psovke »jebemti,« ki ne izraža ničesar natančno določenega. Presupozicijski pogled smatra, da ko nekoga okličem za čefurja, ne le opišem, kakšen menim da je, marveč predpostavljam, da takšen dejansko je, in da takšni dejansko so vsi pripadniki populacije, ki je opisljiva z jedrom tega izraza (tj. priseljenec iz držav bivše Jugoslavije). Nazadnje, prohibicionistični pristop opredeljuje psovke in slabšalnice kot besede, ki so psovke in slabšalnice zaradi tega, ker jih kot slabšalne ali neprimerne dojemajo osebe, na katere se nanašajo. Če torej popolnemu neznancu rečem, da je čefur, ta pa je na ta račun upravičeno užaljen, gre za slabšalnico; podobno velja, če uporabljam oznako »čefur« v splošnem govoru, večina oseb, ki jo ta označuje, pa se s to rabo ne strinja. V kolikor pa, po drugi strani, v šali isto izrečem dobremu prijatelju, v tovrstnem kontekstu ne gre za slabšalnico.

Dejstvo je, da slabšalnice poskušajo nekaj ali nekoga prikazati kot slabšega, kot v resnici je. Kot olepšave, tudi slabšalnice, pa najsi bodo deskriptivne, ekspresivne ali presupozicijske, pomenijo kršitev maksime načina, natančneje zahteve po jasnosti, saj so namenoma oblikovane tako, da je dejanski pomen izraza nejasen oziroma prikrit; prav tako kot metafore kršijo drugo maksimo kakovosti, saj povedo nekaj, za kar govorec nima zadostnih dokazov. Vendar pa obstaja pomembna razlika – olepševalnice osnovni koncept opišejo na tak način, da se zanašajo na pomanjkanje znanja občinstva o dejanski naravi osnovnega koncepta. Slabšalnice, po drugi strani, osnovni koncept opišejo tako, da se zanašajo na to, da bo občinstvo prepoznalo tako osnovnega referenta kot negativne pomenske konotacije, ki jih raba slabšalnice prenaša. Zaradi tega v jezikovnem smislu ne gre za hibo. Prav tako raba slabšalnice ne predvideva, da občinstvo ni zmožno razpoznati, da govorec nima dokazov za implikature, vsebovane v negativnih pomenskih elementih rabljenega izraza – zanaša se na to, da je govorcu v celoti vseeno za dokaze. V bistvu se zanaša se govorčevo spoznavno brezbrižnost, kljub temu, da je govorec lahko zmožen prepoznati

pomanjkanje dokazov in s tem tudi to implikaturo. Slabšalnic tako, kljub njihovi moralni spornosti, ne moremo opredeliti kot jezikovne hibe – gre za način pomensko označenega izražanja, ki je neposredna posledica spoznavne ali intelektualne hibe in se zanaša na to, da občinstvo poseduje taiste ali podobne hibe.

Ostaneta nam še dve možnosti, ki bi lahko generirali jezikovne hibe: zatiranje in jezikovno utišanje. Cappelen in Dever (2019) orišeta moralne težave zatiranja in utišanja ter podata dobro argumentacijsko shemo, zakaj sta jezikovni različici teh pojavov težavni ne le v perlokuciji (izven-jezikovnem učinku govornega dejanja) marveč tudi v ilokuciji (jezikovnem učinku, namenu govorca).

Jezikovno zatiranje predstavita na primeru rasističnega lastnika restavracije, ki zaposlenim ukaže, da je odslej prepovedano streči temnopoltim. Razvidno je, da tudi, če delavci tega ukaza ne bi ubogali (torej ni perlokutivnega učinka), sam po sebi ni nič manj sporen, saj je namen lastnika, da s svojim govornim dejanjem zatira temnopolte, prav tako pa obstaja še skrit perlokutivni učinek normalizacije tovrstnega zatiranja.

Cappelen in Dever podajata neverjetno zapleten primer jezikovnega utiševanja: »Alex se je odločila ubiti Beth. Ker je jezikovno nadarjena, se je odločila to storiti na precej zvit način. Alex postane režiserka igre, Beth pa ena izmed igralk. Alex uvede strogo pravilo, da ob začetku vaje vedno reče »akcija« /.../ Alex nato Beth nastavi kozarec zastrupljene vode. Beth, ki začuti učinek strupa, želi reči »Pomagajte, zastrupili so me. Pokličite zdravnika.« A preden lahko to izreče, Alex reče »Akcija.« (Cappelen in Dever, 2019, 173 –174).

V tem primeru je Alex povzročila, da Beth ni uspela storiti govornega dejanja, tudi če besede, ki jih je nameravala izreči, izreče, saj se te besede od trenutka ko Alex reče »Akcija« naprej smatrajo kot del fiktivnega govora v gledališki igri. S tem jo je Alex torej utišala. Izpostavimo naj, da če bi Alex namesto tega razširila govorice, da je Beth hipohonder, zaradi česar ljudje ne bi verjeli njenim trditvam o zastrupitvi, ne bi šlo za utišanje, marveč ignoriranje, saj bi Beth še zmeraj uspešno izvedla govorno dejanje, ki pa enostavno ne bi bilo sprejeto

V primeru jezikovnega utišanja se zdi, da gre za nekakšen odstop od rdeče niti pogovora: Beth govori o dejanski zastrupitvi, Alex pa pogovor preusmeri na interni kontekst gledališke igre; neimenovana ženska v drugem primeru želi, v realnem

kontekstu, tj. dejanski situaciji, v kateri se nahaja, povedati, da ne želi odnosa. Ne gre zgolj za menjavo teme, kar bi pomenilo kršitev maksime odnosa, marveč gre za celovit odstop od sodelovalnega načela – utiševalec odstopi od jezikovnega sodelovanja z utišanim – in s tem za jezikovno hibo kot posledico prekršitve prvega pogoja jezikovne vrlosti.

Enako se primeri v jezikovnem zatiranju, le da zatiralec odstopi od sodelovanja z družbenim normativom v celoti, in šele posledično od sodelovalnega načela. Rasistični lastnik restavracije tako noče sodelovati z družbenim normativom, da so temnopolte osebe enakopravne in upravičene postrežbe v restavraciji, njegova izjava »temnopoltim ne strežemo« pa tako predstavlja odstop od sodelovalnega načela v smislu odstopa od kooperacije z ostalimi govorci na splošno.

Našo analizo lahko sklenemo z naslednjimi ugotovitvami:

V namene vrle konverzacije, se pravi jezikovne vrline, smo vzpostavili tri nujne pogoje, ki zahtevajo minimalno upoštevanje:

(i) sodelovalnega načela,
(ii) potrebnost implikature za konsistenco izjave ali konverzacije, ter
(iii) upravičeno predpostavko govorca, da ima občinstvo zmožnost prepoznati pomen in potrebnost implikature.
(iv) V kolikor je najmanj enemu izmed teh pogojev nezadoščeno, je to zadosten pogoj za jezikovno hibo.

Na podlagi te definicije in analize primerov »slabega jezika« v Cappelen in Dever (2019) smo identificirali sledeče jezikovne hibe:

(i) Zmoten govor iz nezadostnih dokazov: Govorčeva izjava A je podkrepljena z naborom dokazov n, pri čemer so dokazi, vsebovani v n, sami po sebi zmotni.
(ii) Zmoten govor iz pomanjkanja dokazov: Govorčeva izjava A ni podkrepljena z dokazi, saj govorec ni vložil truda v iskanje dokazov.
(iii) Laganje: Govorec verjame, da njegova izjava A ni resnična.
(iv) Nepopoln zavajajoč govor: Govorec poda izjavo A na način, da izjava A vsebuje minimalno nujno, ne pa tudi zadostno količino podatkov.

(v) Psevdo-implikatura: Govorec poda izjavo A na način, da izjava vsebuje nejasnost, zaradi katere občinstvo predpostavi implikaturo, ki jo govorec namerava sporočiti.

(vi) Nakladanje: Govorec poda izjavo A brez ozira na resničnostno vrednost izjave A.

(vii) Globoko nakladanje: Govorec poda izjavo A na način, da izjava A ni neresnična, marveč v celoti brezpomenska.

(viii) Zavajajoča parafraza (pogosto olepševanje): Govorec poda izjavo A z rabo izrazoslovja, ki prikrije podrobnosti pomena izjave A.

(ix) Implikatura v metafori: Govorec poda izjavo A, pri čemer bodisi verjame, da ni resnična, bodisi zanjo nima zadostnih dokazov, na metaforičen način z namenom, da prikrije neresničnost ali nedokazanost A.

(x) Jezikovna posplošitev: Govorec poda izjavo A o P tako, da sporoča, da imajo vsi P lastnost F, pri čemer najmanj nima zadostnih dokazov za vsebino A.

(xi) Neupravičeno zastopništvo: (Posamezni) govorec poda izjavo A tako, da implicira, da je dejanski (skupinski, razpršeni) govorec izjave neka skupina X, ki ji pripada.

(xii) Jezikovno utišanje: Govorec 1 govorcu 2 prepreči podajanje izjave X tako, da vpelje kontekst ali izjavo A, ki izniči pomen in/ali namen izjave X.

(xiii) Jezikovno zatiranje: Govorec poda izjavo A, katere vsebina v celoti odstopa od družbenega normativa in je posledično nesprejemljiva v konverzaciji na splošno.

Ta seznam ni popoln in dopušča dodajanje drugih jezikovnih figur, ki utegnejo biti primerni kandidati za jezikovne hibe. Jezikovna hiba, kot jo tukaj definiramo, prav tako ni izoliran in popolnoma svojstven pojav, marveč gre za manifestacijo spoznavne hibe, ki ne more obstajati v odsotnosti povezane spoznavne hibe. Spoznavna hiba pa sama po sebi ni avtomatsko tudi jezikovna hiba – jezikovna hiba se kot manifestacija spoznavne pojavi takrat, kadar akter prepričanje, ki nastane kot posledica spoznavne hibe, širi s pomanjkljivo ali zavajajočo komunikacijo. Z drugimi besedami: kakor hitro sporočimo tovrstno prepričanje na način, ki je poleg spoznavne pomanjkljivosti tudi jezikovno zavajajoč, gre za jezikovno hibo kot manifestacijo spoznavne. Na koncu bi poudaril še, da jezikovne hibe praviloma niso karakterne dispozicije, kot so denimo značajske lastnosti z visoko zvestobo pri

Cassamovi opredelitvi spoznavnih hib, marveč gre za trenutni način komunikacije, podobno kot gre pri hibah z nizko zvestobo za zdrs načina mišljenja. Možni izjemi dopuščam v primerih lažnivcev in nakladačev, ki lahko ti dve jezikovni hibi prikazujejo bolj konsistentno (tj. z višjo stopnjo zvestobe), kadar sta neposredno povezani s spoznavnimi hibami zlonamernosti in brezbrižnosti.

**Literatura**


Aberdein, A. (2010). Virtue in Argument. *Argumentation, 24*, 165–179.

Aberdein, A. (2013). Fallacy and Argumentational Vice. V Mohammed, D. in Lewinski, M. (ur.), *Virtues of Argumentation: Proceedings of the 10th International Conference of the Ontario Society for Study of Argumentation (OSSA)*. Ontario Society for the Study of Argumentation.

Aberdein, A. (2014). In Defence of Virtue: The Legitimacy of Agent-Based Argument Appraisal. *Informal Logic,* 34(1), 77–93.

Appiah, A. (2002). *As If: Idealizations and Ideals*. Harvard University Press.

Aristotel. *Nikomahova Etika*.

Austin, J. L. (1962). *How to Do Things with Words*. Oxford University Press.

Cappelen, H. in Dever, J. (2019). *Bad Language*. Oxford University Press.

Carson, T. L. (2006). The Definition of Lying. *Nous,* 40(2), 284–306.

Cassam, Q. (2019). *Vices of the Mind*. Oxford University Press.

Cohen, D. H. (2005). Arguments that Backfire. V Hitchcock, D. in Farr, D. (ur.), *The uses of argument* (str. 58–65). OSSA.

Cohen, D. H. (2007). Virtue Epistemology and Argumentation Theory. V Hitchcock, D. (ur.), *Dissensus and the search for common ground*. OSSA.

Cohen, G. A. (2002). Deeper into Bullshit. V Buss, S. in Overton, L. (ur.), *Contours of Agency: Themes from the Philosophy of Harry Frankfurt* (str. 321–339). MIT Press.

Davidson, D. (1978). What Metaphors Mean. *Critical Inquiry, 5*(1), 31–47.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

Faulkner, P. (2007). What is Wrong with Lying? *Philosophy and Phenomenological Research, 75*(3), 535–557.

Frankfurt, H. (2005). *On Bullshit*. Princeton University Press.

Greco, J. (1999). Agent Reliabilism. V Tomberlin, J. (ur.), *Philosophical Perspectives 13: Epistemology*. Ridgeview.

Grice, H. P. (1975). Logic and Conversation. V P. Cole in J. L. Morgan, *Syntax and semantics 3: Speech acts* (str. 41–58). Academic Press.

Lewis, D. (1978). Truth in Fiction. *American Philosophical Quarterly, 15*(1), 37–46.

Montmarquet, J. A. (1993). *Epistemic Virtue and Doxastic Responsibility*. Rowman and Littlefield.

Saul, J. M. (2013). *Lying, misleading, and what is said: An exploration in philosophy of language and in ethics*. Oxford University Press.

Sokal, A. in Bricmont, J. (2003). *Intellectual Impostures*. Profile Books.

Sosa, E. (1991). *Knowledge in Perspective*. Cambridge University Press.

Sosa, E. (2007). *Apt Belief and Reflective Knowledge, Volume 1: A Virtue Epistemology*. Oxford University Press.

Šetar, N. (2020). *A Monosemic Account of Modality in Speech Act Theory* [Magistrsko delo, Univerza v Mariboru, Filozofska fakulteta]. Digitalna knjižnica Univerze v Mariboru. https://dk.um.si/IzpisGradiva.php?id=77264&lang=eng

Šetar, N. (2024). *Epistemologija škodljivih prepričanj: med spoznavno hibo, argumentacijo in zavajajočo retoriko* [Doktorska disertacija, Univerza v Mariboru, Filozofska fakulteta]. Digitalna knjižnica Univerze v Mariboru. https://dk.um.si/IzpisGradiva.php?id=88440

Wilson, D. in Sperber, D. (2002). Truthfulness and Relevance. *Mind and Language, 111*(443), 583–632.

Zagzebski, L. (1996). *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge University Press.

Univerzitetna založba
Univerze v Mariboru