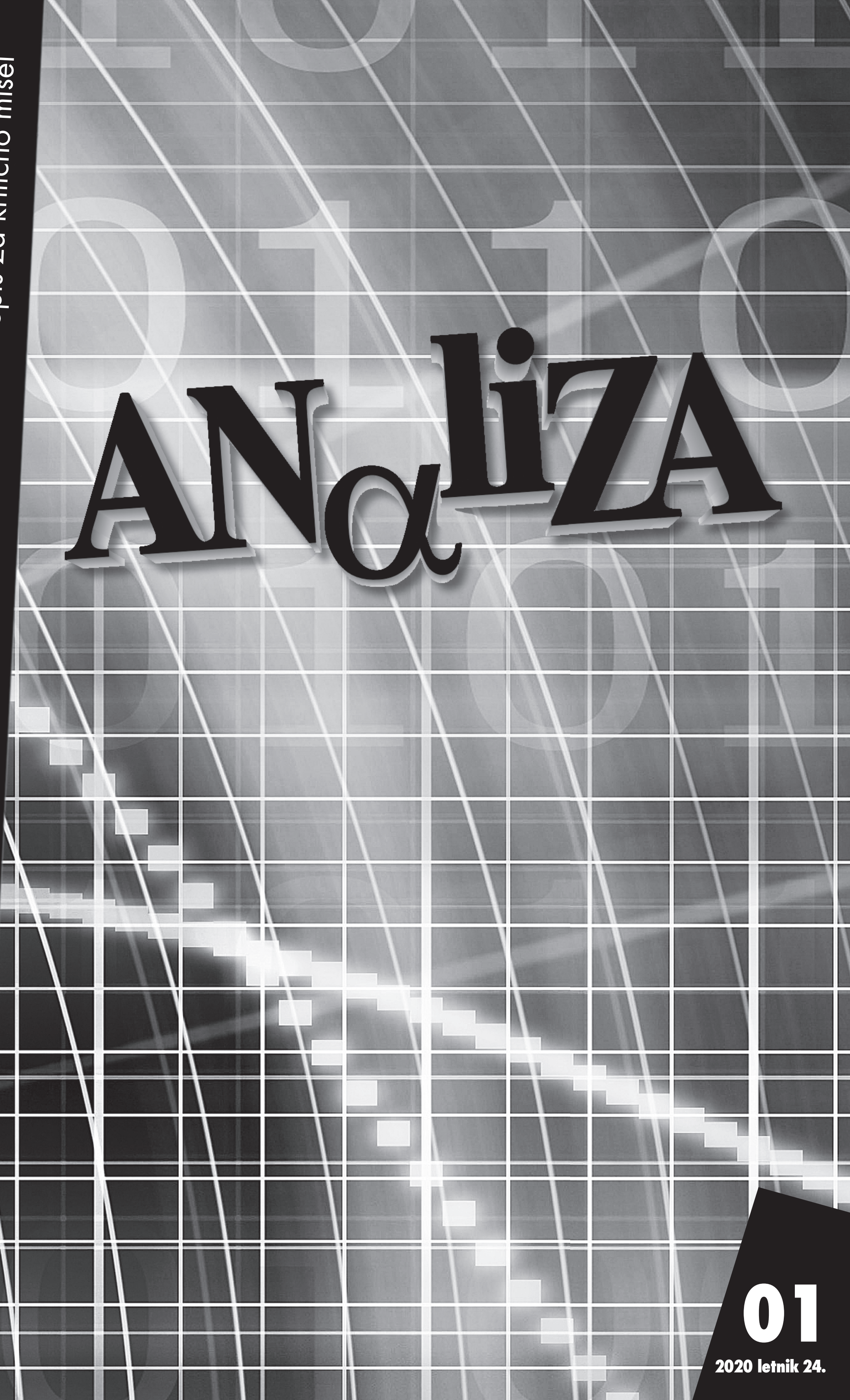
časopis za kritično misel

ANαliZA



Društvo za analitično filozofijo in filozofijo znanosti

01

2020 letnik 24.

ANαliZA, časopis za kritično misel, številka 1, leto XXIV., 2020

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

Naročila sprejemamo na zgornji naslov.

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

Lektor: Borut Cerkovnik

Oblikovanje: K.reg@

Tisk: Design studio d.o.o., Obrežna ulica 33, 2000 Maribor

Naklada: 200 izvodov

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

Revija Analiza je indeksirana v THE PHILOSOPHER'S INDEX.

ISSN 1408-2969

Kazalo

Etika in AI, etika obžalovanja

Logika, filozofija jezika in filozofija znanosti

Zgodovina filozofije

Etika in AI, etika obžalovanja

Janez Bregant
Filozofska fakulteta Univerze v Mariboru

Umetna inteligenca v praksi (2. del): nekaj etičnih pomislekov

Nagel razvoj umetne inteligence (UI) v zadnjem obdobju in njena samoumevna uporaba v našem vsakodnevnem življenju sta povečala zanimanje za vse vrste umetno izdelanih avtonomnih sistemov tako na strani javnosti kot stroke. Ker UI ni več razumljena zgolj kot priročno orodje, ampak tudi kot samostojni akter, ne moremo mimo vprašanja, kakšen vpliv ima njena uporaba na naše življenje v moralnem smislu. Odgovori na vprašanja, kot so »Ali je UI odgovorna za svoja dejanja?«, »Kaj so družbene, pravne in kulturne posledice njenih odločitev?«, »Ali naj bi bili avtonomni sistemi prosto dostopni na trgu?« itn., so tisto, kar v veliki meri določa, ali bomo UI v prihodnosti zaupali ali ne. V članku predstavimo pet nevarnosti, ki nam ob nebrzdani uporabi UI pretijo in s pomočjo analize trenutnega dogajanja na področjih, s katerimi so povezane, zaključimo, da družba vse bolj postaja talec industrije inteligentnih sistemov, ki v nobeni fazi svojega razvoja (oblikovanje, izgradnja, namestitev in ocenjevanje) niso narejeni tako, da bi v interakciji z nami spoštovali človekove pravice in ravnali skladno s sprejetimi družbenimi vrednotami.

Ključne besede: umetna inteligenca, strojno učenje, obdelava naravnega jezika, etika, človekove pravice.

0. Uvod

Cilj ustvarjalcev modelov UI je bil že od nekdaj narediti takšen umetni sistem, ki bo sposoben samostojno prepoznati, obdelati in rešiti naloge ter probleme, na katere bo naletel v svojem okolju, ne da bi pri tem človek moral natančno določiti vsak njegov korak. V idealnem primeru bi se morali biti modeli UI sposobni prilagajati spremembam v okolju, kar bi od njih zahtevalo tudi sprejemanje odločitev. Kot smo videli v prejšnjem članku (Bregant, 2019), se prednost umetnih sistemov kaže tam, kjer je potrebno iz velike in neurejene količine podatkov razbrati ustrezne oziroma zahtevane vzorce. V korist strojnega odločanja se običajno navaja argument, da se s tem iz postopka sprejemanja odločitev izloči predsodke, ki pri človeškem odločanju pomembno vplivajo na rezultat. To pa ne drži, saj tehnologija niti slučajno ni nevtralna. Algoritem je dober le toliko, koliko so dobri podatki, ki jih obdeluje, pri čemer lahko predsodke, ki se zrcalijo v zbranih informacijah posvoji do te mere, da postane obstoječi problem še večji. Še svež je primer rasistične UI, ki ni sposobna prepoznati obrazov temnopoltih ljudi, zaradi česar jih označuje kot gorile. Ti v naši družbi globoko zakoreninjeni predsodki, se v slovarju UI imenujejo *predsodki strojnega učenja*,[1] naloga razvijalcev umetnih sistemov pa je, da jih prepoznajo in njihov vpliv čim bolj omejijo (Ogola, 2019).

Tako je očitno, da so odločitve modelov UI zgolj navidezno objektivne, iz česar izhaja kup praktičnih moralnih problemov. Kdo je odgovoren za odločitve, ki jih sprejmejo avtomatizirani umetni sistemi?, Ali se dajo moralne odločitve, ki jih človek sprejema intuitivno, sprogramirati?, Ali sme UI, razvita v azijskem okolju, sprejemati moralne odločitve v evropskem okolju? itn. Vprašanje je, katere moralne kriterije mora UI izpolniti, da bo pri sprejemanju svojih odločitev, ki se dotikajo naših življenj (npr. prijava za službo, odobritev kredita, odločitev o krivdi toženega v sodnem rocesu itd.), pravična? Politične smernice glede tega, kako uskladiti strojno odločanje s pravom in moralo, je na ravni Evropske unije pripravila tako imenovana *Strokovna skupina za UI*,[2] ki jo je leta 2018 imenovala Evropska komisija. Njen cilj je bil določitev etičnih standardov, ki naj bi pomagali pri tem, da bi odslej na UI gledali z zaupanjem. Poročilo vključuje naslednje zahteve:

(i) nadzor (umetni sistemi morajo človeku pomagati pri sprejemanje argumentiranih odločitev, skladnih z njegovimi pravicami in svoboščinami; poleg tega morajo vključevati tudi primerne nadzorne mehanizme, ki se pri svojem delu zgledujejo po človeškem ravnanju);

(ii) varnost (umetni sistemi morajo biti robustni in varni, vključevati morajo rezervni načrt v primeru, da gre kaj narobe, pa tudi natančni in zanesljivi);

(iii) zaščita podatkov (umetni sistemi morajo zagotavljati in spoštovati pravico do zasebnosti; poleg tega morajo vključevati tudi mehanizme upravljanja s podatki, ki zagotavljajo kakovost njihove obdelave in omogočajo dostop do njih vsem, ki imajo do tega pravico);

(iv) transparentnost (umetni sistemi morajo vključevati mehanizme za sledljivost sprejetih odločitev, človek pa mora vedeti, kdaj je v interakciji s takšnim sistemom in česa je ta zmožen);

(v) nediskriminacija (umetni sistemi se morajo izogibati nepravičnim odločitvam na osnovi predsodkov, ker to povečuje marginalizacijo že tako odrinjenih skupin, dostopni pa morajo biti vsem skozi svoje celotno življenjsko obdobje);

[1] Angl. Machine-Learning-Biases.

[2] Angl. European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG).

(vi) trajnostna orientacija (umetni sistemi morajo biti prijazni do okolja, od njih naj bi v končni fazi imelo korist več generacij, njihov vpliv na vsa, v naravi živeča bitja, pa mora biti skrbno preverjen in premišljen);

(vii) odgovornost (razviti je treba mehanizme, s katerimi se zagotavlja in preverja moralna in/ali pravna odgovornost umetnih sistemov ter njihovih dejanj; poleg tega je treba priskrbeti ustrezna pravna sredstva, ki se uporabljajo v primeru morebitnih kršitev) (Ogola, 2019).

V članku predstavimo pet področij človekovega življenja, na katerih se UI nepremišljeno in nebrzdano uporablja, pri čemer ni o spoštovanju omenjenih smernic za to, da bi dočakali zaupanja vredno UI, ne duha ne sluha. Prvo področje je (če z njihovimi opisi malo dramatiziramo) *zbiranje podatkov*, kjer umetni sistemi vedo o nas več, kot mi sami, drugo *odpoved zasebnosti*, kjer nekdo vedno ve, kaj počnemo, tretje *razvoj orožja*, kjer bodo bodoče (vedno bolj pogoste) vojne potekale med stroji, četrto *slovo od služb*, kjer smo vsi nadomestljivi z UI in peto *uničevanje okolja*, kjer pri proizvodnji elektrike, nujne za potešitev naših spletnih potreb, uporabi fosilnih goriv ni videti konca. V članku zaključimo, da trenutna nepreudarna uporaba umetnih sistemov na omenjenih področjih, predstavlja resno grožnjo obstoječim moralnim načelom, na katerih sloni družba.

1. Kopičenje in koncentracija podatkov

Očitno je, da si danes življenja brez družbenih omrežij, spletnega nakupovanja in elektronske pošte nihče več ne predstavlja. Preko platform, kot so *Facebook*, *YouTube*, *Twitter*, *Instagram* itn. uporabniki delijo svoje izkušnje z drugimi. Na različne načine, to je z besedilom, glasbo, sliko ali všečkom, izražajo svoja mnenja o na primer znamkah, filmih, izdelkih, vladah itn. Ni pomembno, ali kritizirajo, zavračajo, sprejemajo, polemizirajo itn., s tem odkrito kažejo, kaj so njihove preference, pričakovanja in prioritete, s čimer vplivajo na svoje zasebno in javno življenje. Tukaj nam je UI v pomoč, saj lahko z njeno uporabo mnenja ljudi, izražena na takšne načine, odkrivamo, povezujemo in interpretiramo, kar v splošnem imenujemo *rudarjenje podatkov*.[3]

Rudarjenje podatkov v osnovi ne pomeni, da iz količine, ki je na voljo, vzamemo tiste, ki so za nas iz različnih razlogov uporabni, ampak predvsem odkrivanje skritih vzorcev in odstopanj v njih ter ocenjevanje tega, ali so med seboj povezani: na primer združevanje kupcev z istim okusom, opozarjanje na anomalije v proizvodnem postopku, ki kažejo na napake v delovanju sistema ali iskanje zveze med količino pridelka in vremenom. Kot vir se uporabljajo besedila, slike, filmi, glasba itn., iz katerih nevronske mreže s pomočjo strojnega učenja izluščijo želje-

[3] Angl. *Data-Mining*.

ne informacije, največkrat to, kako se ljudje pri posredovanju različnih vsebin počutimo, kar omogoča UI tvorjenje emocionalnih razredov, v katere se potem vsebina razvršča (Dengel, 2019c).

Problem nastane, ko se rudarjenje podatkov izrodi v njihovo zbiranje, uporabnik sodobnih UI modelov pa postane izdelek. Pri tem prednjači pet velikih tehnoloških podjetij: *Google*, *Apple*, *Facebook*, *Amazon* in *Microsoft* (GAFAM). *Google* je s svojimi brezplačnimi storitvami počasi, a zanesljivo, postal ljubljenec občinstva št. 1, ki je v zameno za svoje izdelke želel le eno, predajo naših osebnih podatkov, glede česar smo mu rade volje ustregli in mu tako omogočili neprestano črpanje informacij. *Gmail*, *Chrome*, *YouTube*, *Maps*, *Nest* itd., mu omogočajo, da ve, kaj zanima večino uporabnikov spleta. Njegov iskalnik je najbolj obiskana stran na spletu, opravi 90% vseh iskanj na spletu, operacijski sistem Android, ki je nameščen na 75% vseh mobilnih telefonov, pa trenutno največji sovražnik zasebnosti na sploh. Aplikacije so narejene tako, da naše podatke posredujejo naprej, na primer o lokaciji, razvoju nerojenega otroka, tlorisu stanovanja, nakupovalnih navadah itn., z nakupom *Google Asistenta* pa dobimo upravljalca naprav, ki ves čas posluša, spremlja in beleži, kaj počnemo ter tako pridno polni *Ggoglove* podatkovno zbirke. Ker *Mastercard* še ni njegov, se je z njim dogovoril, da bo zbiral podatke tudi o nakupih v klasičnih trgovinah, opravljenih s to kreditno kartico, ki vsebujejo ime in priimek, čas, lokacijo, kupljeno blago, količino itn. Skratka, *Google* ve, kje smo, kdaj vstanemo, kdaj gremo v službo, kaj in kje smo nakupovali, kako dolgo smo bili na malici, kdaj smo kolesarili, kakšen je naš srčni utrip in drugo[4] (Masten, 2019a).

Za boljši občutek glede tega, o kakšni količini podatkov, ki je na voljo omenjenim družbam, govorimo, si oglejmo, koliko informacij je na svetu znotraj različnih omrežij v obtoku v eni minuti: 18 milijonov poslanih SMS sporočil, 4.3 milijona ogledov video vsebin na *YouTubu*, 481.000 poslanih čivkov, 187 milijonov poslanih elektronskih sporočil, 3.7 milijona iskanj z *Googlom*, 973.000 vpisov v *Facebook*, za 862.823 ameriških dolarjev opravljenih nakupov, 375.000 naloženih aplikacij, 67 nameščenih virtualnih asistentov itn. (Dengel, 2019c). Povečana računska moč ter lahek in poceni dostop do zmogljivih računalnikov, pa niso omogočili zgolj pridobivanja, obdelave in uporabe podatkov, ampak tudi možnost zlorabe, še posebej, ker so informacije razporejene izredno neenakomerno. Vprašanje je, ali lahko kdo zagotovi, da se zbrani podatki, ki so bili tako ali drugače (zakonito ali nezakonito) pridobljeni, ne bodo uporabljali za izsiljevanje?

[4] Tudi ostali tehnološki velikani niso nobena izjema: *Microsoft* se ukvarja s ciljnimi oglasi, tehnologijo prepoznavanja obraza, virtualno resničnostjo itn., *Facebook* preko oglasov in všečkov zbira informacije o imenih in naslovih ljudi, odnosih med njimi, njihovih družinah, lokaciji in pogovorih itn. ter je tako lastnik največje podatkovne baze o nas, *Amazon* pa preko spletne trgovine, v kateri je mogoče kupiti tako rekoč vse in *Alexi*, ki govori iz zvočnika *Echo*, podatke pa pošilja v oblak, pozna naše želje, potrebe in interese, da o naslovu dostave, kreditnih karticah in mestu nakupa niti ne govorimo (Masten, 2019a).

Skoncentriranost podatkov, ki so, kot se rado reče, nova nafta, v rokah posameznikov, saj na primer pri *Googlu*, *Amazonu* in *Facebooku*, drugje pa ni dosti drugače, odloča en sam človek, omenjenim podjetjem omogoča prevlado na informacijskem trgu, saj upravljajo, če malo pretiravamo, s skoraj celotno količino podatkov, ki je na voljo. To pa lahko ima negativne posledice na družbo, ker peščica ljudi s filtriranjem podatkov o svetu, vpliva na njegovo dojemanje in dogajanje v njem, zaradi česar že lahko govorimo o upravljanju družbe s pomočjo UI. Družba postane potem takšna kot avto, podobno kot avto upravljamo z vzvodi, z vzvodi upravljamo tudi družbo. Tako kot lahko avto zavije levo ali desno, lahko »levo« ali »desno« zavijejo tudi družbe, odvisno od želje, potreb ali interesov upravljalca, to je *Googla*, *Facebooka* ali *Amazona* (Grobelnik, 2018). V praksi se je to pokazalo pri zlorabi osebnih podatkov podjetja *Cambridge Analytica*, ki je z aplikacijo, ponujeno *Facebooku*, s prošnjo sodelovanja v akademski raziskavi, nezakonito pridobilo osebne podatke 50 milijonov uporabnikov Facebooka (njegov dizajn je omogočal tudi to, ne samo zbiranje odgovorov) in iz njih izdelalo njihove psihografične profile, s katerimi so ugotavljali, kakšno propagando je treba uporabiti, da bodo glasovali po naročnikovih željah.

Nekateri raziskovalci menijo, da smo ljudje v dobi digitalnih tehnologij, ki jih s pridom uporabljamo, postali surovina, iz katere tehnološka podjetja naredijo končni izdelek, to je napovedi o nas: kaj bomo kupili, kam bomo šli, s kom se bomo družili, kaj bomo delali, koga bomo volili itn. Ti podatki se prodajajo naprej tudi z namenom spreminjanja naših navad, običajev in želja, s čimer izgubimo pravico do prihodnosti. S tem GAFAM glede življenjskih navad ustvarjajo monopol ne zgolj nad sedanjostjo tretjine človeštva, ampak tudi nad njihovo prihodnostjo, saj lahko s ciljnimi izdelki poljubno vplivajo na njihovo obnašanje in ga usmerjajo skladno z lastnimi interesi, kar lahko imenujemo *nadzorovalni kapitalizem*. (Zuboff, 2019)

Vprašanje je, kako to preprečiti? Ena možnost je podružabljanje korporacij, kar pa seveda ni realno, druga možnost pa je regulacija, ki bi razbila monopol tehnoloških gigantov in družbo tako zaščitila pred njimi.[5]

2. Izguba zasebnosti

Mediji so pred kratkim poročali, da se bo državam, ki pod pretvezo večje varnosti, uporabljajo nadzorne kamere z vgrajenim sistemom za prepoznavanje obraza, pridružila tudi Indija, katere glavno mesto New Delhi že brez tega velja za eno izmed najbolj nadzorovanih mest na svetu. UI, ki represivnim organom na ta način pomaga pri identifikaciji ljudi, ni novost, že lep čas je v uporabi v številnih

[5] Kitajska je regulacijo izvedla tako, da je pred časom postavila velik požarni zid, ki je omenjenim podjetjem preprečil zbiranje podatkov s področja njihove industrije (Grobelnik, 2018).

evropskih državah: Veliki Britaniji, Nemčiji, Franciji, Danski, Švedski, Nizozemski, Italiji, Srbiji in še kje. Očitno postajajo algoritmi, ki pomagajo policiji pri preprečevanju kriminala in iskanju krivcev ter tako povečujejo njeno učinkovitost, del našega vsakdanjika ne glede na to, da namestitev nadzornih kamer, prepoznavanje obrazov s slik, ki jih naredijo, in hitra identifikacija oseb, ki so na njih, mobilni telefoni, ki stalno sporočajo lokacijo imetnika itn., bistveno zmanjšujejo našo zasebnost in odpirajo neskončne možnosti zlorab. Na primer iz vsega tega je mogoče rekonstruirati dnevno rutino neke osebe in odstopanje od nje: kje je doma, kje je v službi, kaj so njeni hobiji, kaj so njeni interesi, kaj je njena priljubljena hrana itd. Zdi se, da bi se lahko v bližnji prihodnosti uresničila nočna mora vseh tistih, ki že dolgo opozarjajo na to, da bo razvoj novih tehnologij v slogu 1984, kjer te veliki brat opazuje, ljudi povsem prikrajšal za svobodo.

Opozoriti velja vsaj na tri probleme, ki se pojavljajo pri uporabi sistemov za prepoznavanje obraza. Prvič, v praksi so se izkazali za neučinkovite, saj je odstotek napačnih identifikacij še vedno visok: 0.3 % zveni malo, vendar na mestu, kjer je pretok ljudi velik, na primer letališča, to lahko pomeni več sto krivičnih zaslišanj ali aretacij (Rajšek, 2019). Drugič, takšni algoritmi imajo težave s prepoznavanjem ljudi, ki so temnopolti, pripadajo etničnim manjšinam ali pa so spremenili spol, saj so v fazi urjenja v svojo podatkovno bazo shranili največ obrazov belopoltih ljudi,[6] pri čemer gre krivdo pripisati njihovim učiteljem, to je programerjem. Tretjič, kako je mogoče, da so se po Evropi tako razpasli, ko pa je zbiranje biometričnih podatkov, med katere poleg značilnosti obrazov sodijo tudi druge fiziološke in vedenjske značilnosti posameznika, v Evropski uniji prepovedano s Splošno uredbo o varstvu osebnih podatkov (SUVP)[7] (Rajšek, 2019).

Težava je v tem, da država, ki želi namestiti takšno tehnologijo v svojih mestih, s ponudnikom takšnih storitev, sklene pogodbo, ki ostane tajna. Oba seveda trdita, da biometrične podatke izbrišeta, ne vemo pa, kaj se z njimi resnično dogaja, saj niti enim niti drugim ne gre zaupati. Razen njune besede nimamo nobenega drugega zagotovila, da jih ne shranjujeta v le njima znanih podatkovnih bazah. Če to sedaj povežemo z UI, ki je do te mere sposobna posnemati različne glasove (tudi slavnih osebnosti), da jih ne ločimo od pravih in ki je pri obdelavi naravnega jezika že tako napredovala, da se ne razlikuje več od človeškega osebnega asistenta, potem imamo resen problem. Ko pokliče mamo in ji čestita za rojstni dan, čeprav ga nima, je to še najmanj, če pa ji ukažemo, da naj z glasom Primoža Rogliča pokliče sto oseb in jih prosi za donacijo 50 evrov v dobrodelne namene, nakaže pa naj jih na naš transakcijski račun, stvar ni več tako nedolžna (Kremp, 2018). Ali se ne bi morali takšni sistemi UI, ko nekoga pokličejo, najprej predstaviti, da ljudje ne bi bili zavedeni? Nekateri avtorji (Welsh, 2016) so to predlagali že pred leti,

[6] To aktivisti za človekove pravice imenujejo *tehnološki rasizem*.

[7] Angl. General Data Protection Regulation (GDPR).

tako bi se namreč zlorabam v primeru, ko virtualni pomočnik pri telefoniranju posnema glas nekoga, ki ga poznamo, izognili. Samo upamo lahko, da bo nova zakonodaja za etično uporabo UI, ki jo pripravlja Evropska komisija tako dovršena, da države, ki bi rade postavile nadzorne kamere z vgrajenim sistemom za prepoznavanje obraza, ne bodo našle lukenj, ki jim jo bodo omogočale zaobiti.

Vse to pa ni nič proti temu, kar pod krinko zagotavljanja varnosti državljanov z različnimi pilotnimi projekti nadzora, katerih bistvo je nenehno zbiranje vseh dostopnih informacij o posameznikih, počne Kitajska. Najstrožji je tisti v provinci Sinkiang,[8] gre pa za eno izmed različic prihajajočega sistema spremljanja in ocenjevanja državljanov, ki ljudi glede na vse zbrane podatke, to je, kje delajo, kaj govorijo, kam hodijo, s kom se družijo, kaj objavljajo itn., razvršča v tri skupine: zaupanja vredne, povprečne in zaupanja nevredne. Skladno s tem so eni nagrajeni, drugi pa kaznovani, to pa je mogoče le, če informacijska tehnologija razdrobljene baze podatkov, pridobljene s sistemi za prepoznavanje obraza, zgradbe telesa in načina hoje (iz tega dobi oblast tudi podatke o spolu, starosti, rasi itn.), poveže v enovite profile fizičnih in pravnih oseb. Na primer, če uporabnik mobilnega telefona prečka cesto poleg prehoda za pešce, ga aplikacija, ki jo mora imeti naloženo na telefonu, da država vidi, kaj dela, o tem prekršku obvesti in mu s transakcijskega računa odtegne denar za plačilo kazni: algoritem zazna, obsodi, razsodi in kaznuje. Ljudje s povezavo na splet v bistvu državi sami izdajajo vse skrivnosti, ki so jih včasih morale mukoma odkrivati agenti obveščevalnih služb. Seveda pa ni cilj takšnega množičnega zbiranja informacij zgolj nadzor početja, ampak tudi njegovo napovedovanje. UI namreč omogoča prepoznavanje in razvrščanje vzorcev, iz njih pa je mogoče predvideti kup stvari: posameznikovo mnenje, s kom se bo družil, koga bo volil, kaj bo kupil ali kje bo ob določeni uri dneva. Očitno je, da gre v državi z več kot 800 milijonov uporabnikov spleta, kar je daleč največ na svetu, splet je sicer v državni lasti, saj daje pasovno širino v najem ona, s čimer lahko gleda, kaj se po žicah in valovih pretaka, za državni projekt, njegov cilj pa ustvariti enotno mrežo različnih povezanih baz podatkov o ljudeh in podjetjih, da bi lahko nadzorovala in usmerjala njihovo obnašanje. Omenjen sistem bo po letu 2020 obvezen za vse Kitajce, trenutno jih je 1.4 milijarde.[9] Kitajska je tako dober primer države, ki je UI zlorabila za postavitev vseprisotnega, največjega in najučinkovitejšega nadzorovalnega aparata v človeški zgodovini (Masten, 2019b).

[8] Gre za največjo kitajsko provinco, ki meri kar 1.6 milijona km^2 in leži na severozahodu države, znana pa je po zatiranju muslimanske manjšine Ujgurov.

[9] Kitajska pa uvaja tudi tako imenovane bralnike možganskih valov. Gre za opremo, ki meri človekovo možgansko aktivnost, iz nje pa je mogoče razbrati, v kakšnem čustvenem stanju se človek nahaja. Nositi jo morajo številni pripadniki vojske, nekateri delavci v proizvodnih enotah in v transportnem sektorju ter še kdo, omogoča pa, da je tudi naša notranjost pod neprestanim nadzorom delodajalca in s tem države (Masten, 2019a, 2019b).

3. Inteligentno orožje

Velika nevarnost, ki zahteva moralni preudarek, je izdelava UI v vojaške oziroma vojne namene. Medtem ko so deloma avtonomni sistemi orožja že dolgo v uporabi, na primer. troti, s katerimi je mogoče na daljavo in brez lastnih žrtev uničevati sovražnikove cilje, pa povsem avtomatizirano orožje šele prihaja.[10] V sicer zgodnji fazi izdelave so robotski sistemi za samostojno delovanje na bojišču, nekaj takega, kar smo lahko do sedaj opazovali zgolj v takšnih in drugačnih znanstveno-fantastičnih filmih. Razvoj UI gre v smeri ustvarjanja sistemov, ki bodo sposobni brez neposrednega človekovega ukaza ali nadzora ubijati ljudi. Njihova prednost v primerjavi z na primer jedrskim orožjem je v tem, da zanje ne rabimo nobenih redkih elementov, izdelati jih je možno zelo enostavno in to iz komponent, ki jih lahko deloma že danes kupimo preko spleta (Stöcker, 2015). To je kot naročeno tudi za teroriste, ki lahko takšno UI uporabljajo za podporo pri svojih napadih, v splošnem pa se lahko tako razvijejo modeli, ki majhnim skupinam pomagajo, da z nizkimi stroški povzročijo ogromno škodo. Ker bodo hitro zreli za masovno proizvodnjo, je samo vprašanje časa, kdaj se bodo pojavili tudi na črnem trgu in v rokah diktatorjev, oligarhov in generalov, z namenom boljšega nadziranja prebivalstva, zatiranja miroljubnih protestov ali etničnega čiščenja. Dejstvo je, da UI, če se bo še dalje razvijala brez organizacijskega, varnostnega in tehničnega nadzora, predstavlja grožnjo.

Zato je leta 2015 skupina vodilnih znanstvenikov s področja UI in robotike napisala odprto pismo z naslovom *Autonomous weapons: an open letter from ai & robotics researchers*, v katerem opozarjajo na nevarnost razvoja avtonomnega orožja, s svojimi podpisi pa so se izrekli tudi za njegovo prepoved izdelave po vzoru prepovedi uporabe kemičnega in biološkega orožja s strani kemikov ter v vesolju postavljenega jedrskega in laserskega orožja s strani fizikov. Poudarjajo, da je nadomeščanje vojakov z roboti sicer dobro z vidika zmanjševanja žrtev, a slabo z vidika določanja praga, kdaj se za vojno odločimo, ker se ta zniža. S tem, ko avtonomno orožje imenujejo »jutrišnje kalašnikovke«, slikovito ponazarjajo, kam nas lahko globalno povpraševanje po njem pripelje. Opozarjajo, da je idealno za atentate, destabilizacijo držav, podrejanje prebivalstva in izbrisanje posameznih etničnih skupin. Zaradi tega menijo, da z razvojem sistemov UI v vojaške namene človeštvo ne bi ničesar pridobilo, ampak izgubilo.

Resnici na ljubo to ni prvi poskus vplivati na regulacijo področja razvoja inteligentnega orožja, poročilo *Human Rights Watch*-a je že leta 2012 opozarjalo na njegovo nevarnost, res pa je, da je seznam podpisnikov nove peticije dolg, imena pa ugledna, zaradi česar lahko samo upamo, da bodo njihovi pozivi tokrat padli na plodna tla.

[10] Govori se o tretji revoluciji v vojskovanju, takoj za iznajdbo smodnika in jedrskega orožja.

4. Izguba služb

C. Mims, reporter *Tech-Blogs-Quartz*, je imel svoje prvo srečanje z (delovnim) robotom na zabavi MIT-ja. Stal je v kotu in baje izgledal osamljeno, dokler se mu Mims ni približal, vzel njegovo roko in jo potopil v škatlo z majhnimi predmeti. *Baxter*, kot mu je bilo ime, je enega izmed njih zagrabil, potem pa ga je Mims odpeljal do mize, kjer ga je robot odložil. Nadaljeval je *Baxter* sam, vneto je hodil do škatle in iz nje vlekel predmete ter jih nosil na mizo. Ni ga motilo, da so bili različnih oblik in velikosti, niti da so bili gibi, ki mu jih je pokazal Mims in iz katerih se je v nekaj sekundah naučil, kako to narediti, vse prej kot nazorni. Zbrano je opravljal svoje delo, delo, ki ga sicer opravlja Janez Novak v bližnjem Amazonovem paketno distribucijskem centru (Schulz, S., 2016).

Pred nečim podobnim je svaril že Weizenbaum (1976), to je pred nadomeščanjem ljudi z UI, kar na dolgi rok vpliva na spremembo področja zaposlovanja. Tehnološki giganti kot *Apple*, *Microsoft*, *Google* itd., si prizadevajo za to, da bi v prihodnosti čim več delovnih postopkov postalo avtomatiziranih, to je, da bi opravljanje dela, ki je sedaj v domeni ljudi, postopoma v celoti in na vseh področjih gospodarstva prevzeli stroji. Ker bo vedno več opravil samodejnih, ljudje ne bodo več imeli služb, zaradi česar bo za njihovo preživetje potrebno poskrbeti drugače (prekvalifikacija, UTD, nadomestila ipd.). Na primer, avtonomna vozila bodo odvzela službe taksistom, šoferjem avtobusov in tovornjakov, strojevodjem itn. S tem bo človek, vsaj tako se zdi, postal odvečna oziroma presežna delovna sila. Povečalo se bo število ljudi, ki živijo pod pragom revščine, kar bo vodilo v socialne nemire po vzoru, recimo, protesta »rumenih jopičev«, ki so se leta 2018 začeli v Parizu. Skratka, družba bo postala še bolj nezadovoljna in razslojena, kot je sedaj. Ali je ta strah upravičen?

Dejstvo je, da bo naslednji val avtomatizacije in digitalizacije trg dela pretresel do temeljev in to prej, kot si mislimo. Ni sicer prvič, da se bo kaj takega zgodilo, res pa je, da so vse spremembe do sedaj sledile nekemu naravnemu zakonu: stroji so prevzeli umazane in zdravju škodljive službe, s tem pa ustvarili nove, bolj zdrave in kreativne za ljudi, ki so na začetku običajno pomenile nadzor, vzdrževanje in popravilo avtomatiziranih sistemov. Tokrat naj bi bilo drugače, ne bo šlo zgolj za zamenjavo fizičnega z intelektualnim delom, ampak tudi za nadomeščanje ljudi s stroji v službah, ki zahtevajo ustvarjalnost in načrtovanje. Zdi se, da so na tapeti skoraj vsi poklici, študija Univerze v Oxfordu pa kaže, da je samo v ZDA v nevarnosti 47% služb in to na vseh področjih: od kmetijstva, preko industrije do storitev (Benedikt, Osborne, 2013). Deutsche Bank, na primer, je v okviru svojega varčevalnega programa napovedala, da naj bi do leta 2022 službo izgubilo 18.000 zaposlenih. Zasluga za to gre uporabi UI, ki je v določenih sektorjih poslovanja izjemno povečala njeno produktivnost. Deutsche Bank sploh ne skriva, da bo v prihodnosti še odpuščala, njen cilj je zmanjšati stroške, povečati prihodke in, zanimivo, s pomočjo uporabe UI celo izboljšati uporabniško izkušnjo strank.

Kakorkoli, vsi strokovnjaki ne mislijo, da je prihodnost področja zaposlovanja zaradi uporabe UI tako črna. Nastali naj bi novi poklici, pri tistih, ki se omenjajo kot primeri, pa fantazija očitno ne pozna meja: strokovnjak proti staranju, pisatelj v virtualni realnosti, urbani kmet, psiholog domačih živali, da o vseh vrstah računalničarjev in programerjev ter vzdrževalcev in čistilcev robotov niti ne govorimo. Cilj naj bi bil ljudi osvoboditi rutine in jim tako pomagati, da se lahko posvetijo bolj kreativnim in produktivnim nalogam (Schulz, S., 2016).

Kljub temu vse ni tako rožnato. Število zaposlitev za nedoločen čas naj bi se zmanjšalo, skoraj nihče več ne bo istega dela opravljal celo življenje, služba pa bo odvisna od projektov, na katerih bodo delale vedno različne skupine ljudi. Pospešena avtomatizacija za delavca pomeni, da se mora spremeniti, postati mora bolj ustvarjalen, prilagodljiv in poln idej, z drugimi besedami, neprestano mora dokazovati, da še ni prišel čas za njegovo zamenjavo z UI (Schulz, T., 2016). Drugi izvedenci so še bolj pesimistični, menijo, da bo število ustvarjenih delovnih mest manjše od števila izgubljenih, prekvalifikacije pa drage in da se sčasoma delavcev ne bo več izplačalo izobraževati, da bi lahko držali korak s stroji, ampak jih bo smotrnejše odpustiti.

Ima pa seveda revolucija trga zaposlovanja, ki je pred nami, če ironiziramo, eno prednost: ker bodo sčasoma nevronske mreže postale tako močne, da sploh ne bomo več rabili delati, bomo imeli več časa zase. Država pa bo z univerzalnim temeljnim dohodkom (UTD) poskrbela za to, da bo vsak izmed nas imel dostojno življenje in se lahko posvetil razmišljanju o tem, zakaj je sploh na svetu, ali pa si bo za to, da se bo dokopal do tega odgovora, sestavil računalnik (Schulz, S., 2016).

Kakorkoli, že Joseph Weizenbaum je verjetno zaradi opisane groze, ki jo je doživel z *Elizo*, leta 1976 navedel nekaj poklicev, v katerih ljudi v nobenem primeru ne bi smeli nadomestiti računalniki: policaj, sodnik, vojak, socialni delavec, terapevt in delavec v podporni službi strankam (Weizenbaum, 1976). Še več, takšna uporaba UI naj bi ogrozila človeško dostojanstvo, saj omenjene službe od zaposlenih zahtevajo izkazovanje in posedovanje verodostojnega občutka empatije, to je sočutja. Njegovo opozorilo je bilo žal preslišano, hiter razvoj UI z vedno bolj sofisticiranimi proizvodi, je na moralna vprašanja glede njihove izdelave in uporabe gladko pozabil.

Kljub temu se zdi, da smo še vedno v fazi, ko stroj človeka sicer lahko preseže pri marsikaterem opravilu, ga tudi pretenta ali prevzame del njegovega posla, nikakor pa ga še ne more nadomestiti. Upamo lahko samo, da bo tudi nadaljnji razvoj UI potekal postopoma, in sicer preko vedno tesnejšega sodelovanja človeka in stroja. To nam bi pomagalo, da postanemo učinkovitejši, bolj prilagodljivi in iznajdljivi ter se tako bolje pripravimo na to, kaj nam bo prihodnost z vidika služb resnično prinesla.

5. Obremenitev okolja

Za svoj skokovit razvoj v zadnjih letih se ima UI zahvaliti predvsem eni stvari: tako imenovanemu globokemu učenju, ki smo ga že omenili, ne pomeni pa nič drugega kot to, da se stroji sami učijo, kar jim omogoča, da postajajo vse pametnejši in zmogljivejši. Ideja je, poenostavljeno rečeno, ustvarjanje večplastnih mrež iz umetnih nevronov in kopiranje delovanja človeških možganov z njimi. Problem je, da živimo v času, ko je skrb za okolje postala osrednja politična tema, zaščita podnebja pa že zdaj naša glavna naloga, saj trenutni hiter razvoj UI niti slučajno ni skladen z zeleno okoljsko politiko.

Zadnja študija Univerze v Amherst Masschusetssu je pokazala, da je urjenje nevronskih mrež drago z dveh vidikov: prvič, finančno, zaradi cene elektrike, ki je za to potrebna, in drugič, okoljsko, zaradi ogljičnega odtisa, ki ga proizvede delovanje osrednje procesne enote (Strubell, Ganesh, McCallum, 2019). V njej so avtorji preučevali štiri modele UI, ki temeljijo na nevronskih mrežah in so sposobni obdelovati naravni jezik. Gre za sisteme, ki na takšen ali drugačen način analizirajo jezik, natančneje, algoritme, ki se uporabljajo kot virtualni asistenti, prevajalniki, pisci zgodb itn. Da bi izmerili porabo energije njihovih procesorjev in grafike, so štiri takšne modele izmenično urili en dan. Iz količine energije, ki so jo modeli porabili za to, da so opravili posamezne dele naloge, so izračunali, koliko elektrike bodo porabili do konca treninga, potem pa to pretvorili v količino CO_2, ki ga bodo pri tem izpustili v ozračje. Rezultati so pomenljivi: urjenje zgolj enega takšnega sistema proizvede 313 ton CO_2, kar je približno petkrat toliko kot avto v svoji celotni življenjski dobi. Potreba po ogromni količini energije je posledica velike računske moči stroja, brez katere ne bi mogel obdelati ogromne količine podatkov, kar pa je za uspešen trening, kot smo že omenili, ključnega pomena.

Da so računalniki energetsko potratni, ni nič novega, že dolgo je znano, da eno iskanje na na primer Googlu proizvede toliko CO_2 kot 2 grelnika vode za čaj, to je 15 gramov. Kriva je elektrika, ki jo rabimo za napajanje računalnika in za pošiljanje zahteve na strežnike po svetu. Pred leti je IBM-ov superračunalnik v ugankarskem kvizu sicer premagal človeška tekmeca, vendar je pri tem porabil 85 kW energije; za primerjavo, človeški možgani pri sodelovanju v takšnem šovu porabijo zgolj 20 W. In še en podatek, že danes porabijo računski/podatkovni centri 2 % vse elektrike na svetu (cel IT sektor pa 7 %) in proizvedejo toliko CO_2 kot cela letalska industrija skupaj. Vprašanje je torej, kako trajnostno sploh je strojno učenje? (Lobe, 2019)

Obstaja pa še en pomemben vidik, ki je povezan z ogromno porabo energije: od kod sploh izvira elektrika, ki jo spletni giganti porabijo za svoje delovanje? Glede na zadnje poročilo Greenpeaca se pri tem med znanimi najslabše odrežejo *Amazon*, *Microsoft* in *Netflix*, ki več kot tretjino svoje energije pridobijo iz premoga, najbolje pa *Apple*, *Facebook* in *Google*, ki je več kot polovico pridobijo iz obnovljivih virov[11] (Cook et al, 2017). Slednji so tudi na vrhu pri vključevanju zavez o uporabi obnovljivih virov energije v svojo politiko, sledljivosti njene nabave in energetski učinkovitosti. O kitajskih ponudnikih spletnih storitev, kot so *Alibaba*, *Baidu*, *Tecent* in ostali, nima smisla razpravljati, bolj ali manj vsi pokrijejo dve tretjini svojih potreb po energiji z elektriko pridobljeno iz premoga, prav tako pa capljajo daleč zadaj glede posredovanja podatkov o svojem energetskem odtisu, sledljivosti uporabljene zelene energije in vključevanju trajnostnih zavez v politiko svojega delovanja.

Glede na to, da »je internet centralni živčni sistem sodobnega globalnega gospodarstva« (Cook et al, 2017: 5), ni čudno, da je tudi izredno velik porabnik energije. Izdelava in napajanje naših domačih naprav, podatkovnih centrov in strežnikov vseh vrst, če naštejemo zgolj najbolj očitne, terjata svoj davek. In prihodnost ni svetla: promet na spletu naj bi se do leta 2020 povečal za trikrat, število njegovih uporabnikov pa naj bi naraslo na 4 milijarde. Očitno je, kaj to pomeni. Zato je ključno naslednje vprašanje: ali bomo pri pokrivanju energetskih potreb IT sektorja (in na sploh) sposobni narediti prehod iz fosilnih goriv na obnovljive vire in se tako izogniti podnebnim spremembam?

V nekem smislu bi lahko naša odvisnost od spleta to celo pospešila, ampak zgolj v primeru, (i) če bi kot njegovi uporabniki tehnološka podjetja znali prisiliti v to, na primer z izogibanjem nakupovanju njihovih »umazanih« izdelkov, da bi njihovo digitalno infrastrukturo poganjala zgolj čista elektrika, (ii) če bi tudi ponudniki energije to prepoznali kot priložnost in zagotovili ugodne fiksne cene za elektriko, pridobljeno iz obnovljivih virov in (iii) če bi od tehnoloških družb zahtevali, da gradijo svoje ime na zavezanosti k uporabi obnovljivih virov energije in transparentnem posredovanju informacij o njeni nabavi ter zavedanju, da bodo tako prispevali k zaščiti podnebja (Cook et al, 2019). Omenjena študija Univerze Amherst v Masschusetssu pa vidi rešitev glede izjemne potratnosti digitalnih tehnologij v optimizaciji modelov UI, to je v razvoju energetsko bolj učinkovitega hardvera (računalnikov/strojev) kot tudi softvera (algoritmov/programov) (Strubell, Ganesh, McCallum, 2019). To naj bi prispevalo k bistveno varčnejšemu urjenju nevronskih mrež, ki bo zaradi kasnejše masovne uporabe takšnih modelov UI, na primer različni asistenčni programi v avtomobilih s sposobnostjo preprečevanja nesreč, v resnici še cenejše.

[11] *Apple* pridobi iz obnovljivih virov celo 83 % vse potrebne energije.

6. Sklep

Ali se razvijalci, ponudniki in uporabniki UI danes držijo vsaj minimalnih etičnih standardov, ki jih je kot smernice za zaupanja vredno UI predlagala Evropska komisija? Res je, da gre pri tem za političen projekt, pa vendarle. Pri njihovi pripravi so sodelovali posamezniki iz akademske sfere, industrije in civilne družbe, zaradi česar imajo kljub vsemu določeno težo. V članku s pomočjo analize petih področij človekovega življenja ugotovimo, da je nebrzdana in nepremišljena uporaba UI v konfliktu s človekovimi pravicami in družbenimi vrednotami.

Prvo nevarnost predstavlja skoncentriranost podatkov, ki največjim petim tehnološkim podjetjem na svetu, *Googlu*, *Applu*, *Facebooku*, *Amazonu* in *Microsoftu* (GAFAM), omogoča prevlado na informacijskem trgu, zaradi česar je strah pred negativnimi posledicami, ki bi jih njihovo filtriranje podatkov o svetu lahko imelo na družbo, kot kaže primer *Cambridge Analytice*, povsem na mestu. Druga nevarnost je nameščanje kamer s sposobnostjo takšnega in drugačnega prepoznavanja osebnih značilnosti človeka, ki bistveno zmanjšujejo našo zasebnost in svobodo, saj je mogoče iz tako zbranih podatkov do potankosti spoznati rutino človeka in odstopanja od nje ter skladno s tem prilagoditi ciljno ponudbo zanj na spletu celo do te mere, da se bo spremenil glede na naše želje, potrebe in interese. Tretja nevarnost je razvoj inteligentnih vojaških sistemov, ki so sposobni brez neposrednega človekovega ukaza ali nadzora ubijati ljudi brez lastnih žrtev, kar vstop v vojno samo še olajša. Njihova izdelava ne predstavlja praktično nobenega finančnega bremena in je v rokah avtokratov kot naročena za discipliniranje državljanov, zatiranje demonstracij ali etnično čiščenje. Četrto nevarnost predstavlja izguba služb, kar je posledica tega, da stroji že zdaj v vedno večji meri opravljajo delo, ki je bilo še pred kratkim povsem v domeni ljudi. Človek vse bolj postaja presežna delovna sila, ki lahko preživi le, če se neprestano prilagaja, dela pod prekarnimi pogoji in dokazuje, da še ni zrel za zamenjavo. Vprašanje je, kako dolgo še? Peta nevarnost pa je izjemno velika poraba energija za delovanje UI na globalni ravni in temu primeren slab ogljični odtis, kar ni skladno z današnjimi trendi na področju varovanja okolja. Zaskrbljujoče je predvsem to, da v večini primerov še vedno preveč elektrike za izdelavo in pogon naših pametih naprav pridobimo iz umazanih fosilnih goriv, namesto da bi pri tem prevladovala energija proizvedena iz obnovljivih virov.

Kaj lahko glede tega storimo? Glede skoncentriranosti podatkov smo v težkem položaju, podružabljanje korporacij ni realna možnost, regulacija digitalnih gigantov pa se lahko kot v primeru Kitajske hitro sprevrže v digitalno cenzuro in diktaturo države. Pri izgubi zasebnosti bi lahko pomagala domišljena in dovršena zakonodaja brez pravnih praznin, ki bi onemogočale, da bi jo države zaobšle. Seveda bi morala veljati za vse, kar pa je slejkoprej iluzija, če jo bomo dobili v EU, bo to že velik dosežek. Tudi v zvezi z izdelavo avtomatiziranega orožja so naši izgledi slabi. Težava je, da države financirajo njegov razvoj, podjetja, ki se s tem ukvarjajo, pa denar velikodušno in brez zadržkov sprejemajo. Ko je izdelek narejen, so financerji tudi njegovi kupci, zaradi česar je uspešnost prizadevanj za prepoved izdelave in uporabe samodejnega orožja pod velikim vprašajem. Glede revolucije na trgu delovne sile med nekaterimi strokovnjaki prevladuje prepričanje, da bo nastalo več služb, kot pa jih bo izginilo. Pa ne samo to, novi poklici bodo ljudem omogočili večjo ustvarjalnost in učinkovitost ter jih tako osvobodili rutine in zasičenosti. Ker so stroji sposobni prevzeti tudi takšne službe, je opisan idealen scenarij po našem mnenju možen zgolj v enem primeru: nevronske mreže so tako močne, da ljudem ni več treba delati, država pa vsakemu izmed nas nakazuje UTD kot plačilo za dostojno življenje. Ali je to uresničljivo, bo pokazal čas. Pri manjši porabi energije pa bi lahko pomagal pritisk na tehnološka podjetja, da mora njihovo digitalno infrastrukturo poganjati zgolj energija, pridobljena iz obnovljivih virov. To pa zahteva dovolj veliko število ozaveščenih uporabnikov, ki si lahko na takšen način narejene izdelke, običajno dražje od konkurence, ki jaha na premogu, tudi privoščijo. Verjetno je bolj realna rešitev za zelen IT sektor na globalni ravni v tem trenutku optimizacija urjenja nevronskih mrež s pomočjo varčnejšega hardvera in softvera.

Kakorkoli, svet pametnih strojev ni več literarna fikcija. UI naše življenje spreminja, mi pa temu, tako se zdi, v moralnem smislu, ne posvečamo dovolj pozornosti. Razvoj inteligentnih strojev je hiter, glede tega ne moremo storiti nič. Mogoče ga lahko usmerjamo, ampak potem moramo s tem začeti zdaj.

Artificial Intelligence at Work: Some Ethical Concerns

A recent development of artificial intelligence (AI) and its self-evident use in our everyday life have increased interest of the public and experts for all kinds of artificially made autonomous systems. Since AI is not anymore understood merely as a handy tool but also as an independent agent, the question of its impact on our life in terms of morality emerges naturally. Answers to questions such as »Is AI responsible for its actions?«, »What are social, legal and cultural consequences of its decisions?«, »Should its autonomous models be freely accessible in the market?« etc., determine to a great extent whether AI will be trustworthy in the future or not. The paper introduces five threats associated with our unbridled use of AI and concludes through the analyses of current developments in the fields related to those threats that the society is increasingly becoming a hostage of the AI industry: intelligent systems are not made in a way that they would, while interacting with people, in any of their evolution stages (design, construction, installation and evaluation) respect human rights nor act according to accepted social values.

Keywords: artificial intelligence, machine learning, natural language processing, ethics, human rights.

Literatura

Bregant, J. (2019). »Umetna inteligenca v praksi (1. del): razvoj, obnašanje in učenje strojev«. *Analiza*, 2, str. 39–55.

Autonomous weapons: an open letter from ai & robotics researchers (2015). Dostopno na: https://futureoflife.org/open-letter-autonomous-weapons/?cn-reloaded=1 [29.11.2019].

Benedikt, C., Osborne, M. A. (2013). »The future of employment: how susceptible are jobs to computerisation?«. Dostopno na: https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf [30.11.2019].

Cook et al (2017). »CliCking Clean: Who is Winning the RaCe to Build a gReen inteRnet?«.

Dengel, A. (2019c). »Multimedia-Data-Mining: Trends und Emotionen in Big Data erkennen«. V Dengel, A., Socher, R., Kirchner E. A., Ogolla, S., Künstliche Intelligenz: Die Zukunft von Mensch und Maschine. Hamburg: ZEIT Akademie Gmbh, str. 74–84.

European Commission's High-Level Expert Group on Artificial Intelligence (2018). *Ethics guidelines for trustworthy AI.*

Grobelnik, M. (2018). »Podatki so nova nafta. Kdor ima dostop do podatkov, lahko rešuje probleme«. *Mladina*, 39.

Human Rights Watch (2012). Losing humanity: The Case against Killer Robots. Dostopno na: https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf [29.11.2019].

Kremp, M. (2018). »Google Duplex ist gruselig gut«. Spiegel Online. Dostopno na: https://www.spiegel.de/netzwelt/web/google-duplex-auf-der-i-o-gruselig-gute-kuenstliche-intelligenz-a-1206938.html [28.11.2019].

Lobe, A. (2019). »KI ist alles andere als grün«. Spektrum.de. Dostopno na: https://www.spektrum.de/news/kuenstliche-intelligenz-verbraucht-fuer-den-lernprozess-unvorstellbar-viel-energie/1660246 [01.12.2019].

Masten, A. (2019a). »Kaj vse pomeni klik na 'Strinjam se': O ekonomiji podatkov in monopolih«. MMC RTV SLO. Dostopno na: https://www.rtvslo.si/mmc-podrobno/na-pragu-digitalne-diktature-brez-zasebnosti-in-brez-svobode/489378 [02.12.2019].

Masten, A. (2019b). »Ko nadzor prevlada nad svobodo: Veliki kitajski brat: prva digitalna diktatura«. MMC RTV SLO. Dostopno na: https://www.rtvslo.si/mmc-podrobno/na-pragu-digitalne-diktature-brez-zasebnosti-in-brez-svobode/489378 [02.12.2019].

Ogola, S. (2019). »Verantwortung, Erklärbarkeit und Transparenzalgorithmischer Entscheidungen«. V Dengel, A., Socher, R., Kirchner E. A., Ogolla, S., *Künstliche Intelligenz: Die Zukunft von Mensch und Maschine*. Hamburg: ZEIT Akademie Gmbh, str. 93–101.

Rajšek, J. (2019). »Veliki evropski brat se prebuja – 'Obraza ne moremo zamenjati'«. *MMC RTV SLO*. Dostopno na: https://www.rtvslo.si/evropska-unija/veliki-evropski-brat-se-prebuja-obraza-ne-moremo-zamenjati/502913 [02.12.2019].

Schulz, S. (2016). »Die Jobfresser kommen«. Spiegel Online. Dostopno na: https://www.spiegel.de/wirtschaft/soziales/arbeitsmarkt-der-zukunft-die-jobfresser-kommen-a-1105032.html [30.11.2019].

Schulz, T. (2016). »Wie Kollege Computer Ihren Job überflüssig macht«. Spiegel Online. Dostopno na: https://www.spiegel.de/wirtschaft/kuenstliche-intelligenz-wie-kollege-computer-ihren-job-ueberfluessig-macht-a-1110661.html [30.11.2019].

Stöcker, C. (2015). »Künstliche-Intelligenz-Forscher warnen vor künstlicher Intelligenz«. Spiegel Online. Dostopno na: https://www.spiegel.de/netzwelt/netzpolitik/elon-musk-und-stephen-hawking-warnen-vor-autonomen-waffen-a-1045615.html [28.11.2019].

Strubell, E., Ganesh, A., McCallum, A. (2019). »Energy and Policy Considerations for Deep Learning in NLP«. *Študija Univerze v Massachusetssu Amherst*. Dostopno na: https://arxiv.org/pdf/1906.02243.pdf [01.12.2019].

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. New York: PublicAffairs.

Welsh, T. (2016). »Turing's Red Flag«. *Communications of the ACM*, 7, str. 34–37. Dostopno na: https://www.cse.unsw.edu.au/~tw/wcacm15.pdf [28.11.2019].

Weizenbaum, J. (1976). *Computer Power and Human Reason*. San Francisco: W. H. Freeman.

Smiljana Gartner
Filozofska fakulteta Univerze v Mariboru

Hurkov koncept racionalnega obžalovanja v perspektivi Dancyjevega moralnega partikularizma

Vprašanje racionalnega obžalovanja lahko osvetlimo s pomočjo monističnih in pluralističnih teorij dobrega. V članku je predstavljeno, kdaj je, po mnenju Hurke, obžalovanje racionalno za nebenthamovskega monista in kdaj v pluralističnih teorijah dobrega. Prav tako je prikazano, kakšno stališče zavzame Dancy, ki je predstavnik partikularizma, glede monističnih in pluralističnih razlag obžalovanja. Vendar, če je sprejet koncept obžalovanja v teoriji razlogov in če je holizem upravičena teorija, potem trdimo, da mora biti ideja opredeljenega racionalnega obžalovanja zavrnjena, kar pa je v nasprotju z Dancyjevim stališčem.

Ključne besede: monizem, pluralizem, obžalovanje, partikularizem, teorija razlogov.

Uvod

Pričnimo z zgodbo o mravlji in škržadu oziroma murnu. Slednji je med poletjem pel, medtem ko je mravlja delala in zbirala pridelek za zimo. Ko je prišla zima, je bil škržad lačen in žalosten ter je odšel prosit mravlje, da mu dajo za jesti. Mravlje ga zavrnejo, saj želijo, da obžaluje svoja dejanja v poletnem času. Vzemimo spremenjeno bajko, ki jo je navedla M. Nussbaum, kjer mravlja vpraša škržada: »Ti je sedaj žal, da si vse poletje zgolj pel?« Škržad ji odvrne: »Zelo žal. Tako, kot sem vedel, da mi bo. Toda sedaj je sedaj. Takrat sem ravnal racionalno.« (Hollis, 1996: 60). Ali je škržad občutil obžalovanje in ali ima za obžalovanje tehtne razloge, sta vprašanji, ki ju lahko umestimo na polje monističnih in pluralističnih teorij dobrega in na kateri bomo odgovorili na koncu članka.[1] Pred tem bomo predstavili: (i) kdaj je, po mnenju Hurke, obžalovanje racionalno za nebenthamovskega monista in kdaj v pluralističnih teorijah dobrega; (ii) kakšno stališče zavzame Dancy, ki je predstavnik partikularizma, glede monističnih in pluralističnih razlag obžalovanja, saj zagovarja trditev, da nam monizem ne more ponuditi dobre razlage obžalovanja, pluralizem, natančneje partikularizem, pa lahko. V tretjem delu bomo zavrnili Dancyjevo stališče. Če namreč sprejmemo koncept obžalovanja v teoriji razlogov in če je holizem upravičena teorija, potem moramo idejo racionalnega obžalovanja zavrniti. Iz česar pa še ne izhaja, da je potrebno zavrniti tudi partikularizem.

I.

Koncept obžalovanja lahko razložimo na različne načine. Prva in enostavna opredelitev je, da se pri obžalovanju soočamo z občutki v sedanjosti, ki so povezani z dejanji, ki smo jih naredili, ali prepričanji, ki smo jih imeli v preteklosti, ter njihovimi posledicami. Obžalovanje lahko tako povežemo z značilnostmi, ki smo jim dajali prednost v preteklosti (s preferencami v preteklosti), in tistimi značilnostmi, ki jih dajemo prednost sedaj. Ali praktično prikazano, če bi nekdo lahko potoval v preteklost, bi v drugo sprejel drugačno odločitev kot prvič. V zgodbi s škržadom bi to pomenilo, da če bi lahko zavrtel čas nazaj, bi se odločil drugače, to je, ne bi pel in igral, temveč bi se raje pripravljal na zimo. Njegove preference, če bi se lahko vrnil v preteklost, bi bile drugačne, razlogi za prepričanje in dejanje prav tako. V primeru, da bi zavrtel čas nazaj in ne bi sprejel drugačne odločitve, torej bi odregiral enako, bi imel enake razloge za dejanja, potem lahko, kot v spremenjeni zgodbi M. Nussbaum, izključimo obžalovanje. Pa vendar se vsi ne strinjajo s tem. B. Williams trdi, da četudi bi sprejeli enako odločitev, storili enako dejanje in je le-to najboljše možno, so situacije, kjer obžalovanja ne moremo izločiti že zaradi same narave dejanja (Williams, 1985). Kahn temu pritrjuje in trdi, da četudi smo sprejeli najboljšo možno odločitev, še vedno obstaja »neizločljiv konflikt in bi morali do neke stopnje obžalovati dejanje, ne glede na to, kakšno dejanje je« (Kahn, 2011: 15). R. Sorensen je nasprotnega mnenja − da ni razumno obžalovati odločitve, ki si jo presodil kot najboljšo, zato je bolečina obžalovanja nepotrebna, saj nas sili v nesmiselno ukvarjanje s preteklostjo (Sorensen, 1998: 528–533). Ali bi torej morali vedno imeti določeno stopnjo obžalovanja in četudi je obžalovanje odveč, se mu ne moremo izogniti (Kahn) in je tako

[1] Temelji na Ezopovi bajki, ko je škržad v antični Grčiji predstavljal cenjeno bitje. Tudi Platon uporabi mit in metaforo o škržadu v Fajdrosu. V članku uporabljamo besedo škržad, četudi je v SSKJ zapisana kot škržat, vendar nas je sklop člankov in pesmi o teh bitjih (Matičetov, 2000: 587–597) prepričal, da je škržad z d-jem pristnejša beseda.

psihološka nuja ali pa ga ne bi smeli imeti? Morda lahko odgovor poiščemo s pomočjo monističnih in pluralističnih teorij dobrega.

Za etiškega pluralista so etiško pomembne lastnosti, razlogi oziroma oblike »dobrega« bolj ali manj enakovredne, to je, ne obstaja ena (vrednota, vrednost, lastnost, razlog), ki je nad vsemi, zato jih tudi ne moremo razvrstiti po pomembnosti. V tem članku jih uporabljamo v kontekstu »kar je dobro po sebi oz. za kar si je vredno prizadevati zaradi njega samega.« (Klampfer, 2002: 90). Če ima neka stvar intrinzično vrednost, si prizadevamo zanjo že zaradi nje same, ne glede na to, ali ima zaželene posledice ali ne. Nosilci intrinzične vrednosti pa so, tako Audi, realizacije stanj stvari in te realizacije, ti konkretni elementi so intrinzično dobri v luči njihovih intrinzičnih lastnosti (Audi, 2004: 123) oziroma nekaj je intrinzično dobro, kadar je takšno zaradi svoje intrinzične narave. To pa za Moora pomeni, da ima takšno vrednost v vsaki situaciji, v kateri se pojavi (Moore, 2000: 303).

Za etiškega monista obstaja ena sama vrsta stvari, ki so dobre po sebi. Obstaja zgolj ena etiško pomembna lastnost dejanj oziroma ena oblika »dobrega«, ki je intrinzično taka in po kateri se meri in presoja vrednost vsega drugega. Na primer, četudi prepoznamo prijateljstvo kot nekaj dobrega, nekaj zaželenega oziroma kot odnos oz. stanje stvari s pozitivno vrednostjo, je lahko za hedonista, paradigmatičnega monista ta zgolj instrumentalna, kolikor prijateljema zagotavlja edino dragoceno stvar, to je užitek.

Isaiah Berlin je monizem definiral, če povzamemo njegovo stališče v treh točkah, na naslednji način: a) vsa izvorna vprašanja morajo imeti pravilen odgovor, in sicer zgolj enega; b) vsak pravilen in resničen odgovor je mogoče odkriti, pri čemer pa mora obstajati zanesljiva pot, da ga odkrijemo; c) ko odkrijemo odgovore, so le-ti kompatibilni drug z drugim (lahko so razvrščeni hierarhično ali kako drugače), saj ena resnica ne more biti nekompatibilna z drugo, in posledično tvorijo celoto. To temelji na predpostavki, da je univerzum, četudi sprejmemo idejo o velikem številu ciljev in vrednot, še vedno harmoničen in koherenten. Berlin nasprotuje vsem trem točkam, saj trdi, prvič, da obstaja veliko število raznovrstnih ciljev in vrednot človeka, in drugič, da vsi niso kompatibilni med seboj. Iz tega izhaja, tretjič, da konflikta vrednot (in tragedije, ki jo konflikt povzroča) ne moremo nikoli popolnoma izločiti iz (osebnega ali družbenega) življenja. Natanko to pa je tisto, kar daje, četrtič, svobodi vrednost (Berlin, 1958: 29–32). Ne le da so dobrine in vrednote nekompatibilne, tudi neprimerljive so, kar v skladu z Berlinovim razumevanjem neprimerljivosti pomeni, da ni enega merila oziroma načela, po katerem bi lahko lastnosti med seboj primerjali in ovrednotili.

W. D. Ross in J. Rawls prav tako zavračata monizem. V okviru svojega ugovora monističnemu konsekvencializmu zagovarjata etiški pluralizem, a trdita, da ljudje obljub ne izpolnjujejo zato, ker bi bile posledice izpolnjevanja obljube boljše kot posledice njihovega prelamljanja, temveč preprosto zato, ker so to obljubili. »Ko posameznik izpolni obljubo, ker meni, da je dolžan to storiti, je jasno, da pri tem ne razmišlja o vseh posledicah svojega (ne)dejanja, še manj o tem, da je to najboljša možnost. V resnici veliko več razmišlja o preteklosti kot o prihodnosti. Vzrok, zaradi česar misli, da je njegovo dejanje pravilno, je dejstvo samo, da je obljubil, da bo tako storil – razmišlja zgolj to in običajno nič drugega« (Ross, 2002: 17).[2]

Razmišljanje o preteklosti pa je značilno tudi za obžalovanje in vrednostni monizem, kot ga razume B. Hurka. Njegovo razlago oziroma definicijo tako imenovanega racionalnega obžalovanja bi lahko predstavili v dveh korakih:

a) intenzivnost obžalovanja je sorazmerna z izgubo stvari, ki jo obžalujemo;

b) ravno zaradi tega proporcionalnega pogleda na obžalovanje, ki ga vključimo v definicijo racionalnega obžalovanja, lahko o obžalovanju govorimo tako v etiškem monizmu kot v etiškem pluralizmu. Za boljše razumevanje in upravičitev racionalnega obžalovanja bomo uporabili (c) tudi Dancyjevo stališče o razmerju del – celota, ki bi ga lahko sprejel Hurka. V nadaljevanju si poglejmo predstavitev vseh treh točk.

a. Izhodiščni argument o izgubi nečesa intrinzično dobrega

> A1. Če je nekaj intrinzično dobro, potem je takšno samo po sebi.
>
> A2. Če je nekaj dobro samo po sebi, je primerno, racionalno in dobro, da to nekaj ljubimo zaradi njega samega.
>
> Torej,
>
> A3. Če izgubimo nekaj, kar je dobro samo po sebi, je primerno, racionalno in dobro, da občutimo obžalovanje.

[2] Vredno(s)tni pluralist ni relativist. Za prvega ima konflikt lastnosti, vrednosti, pravil, vrednot, sistemov itn., razmerje med dobrim in slabim, pravilnim in nepravilnim objektivno vrednost in racionalno razlago ter upravičitev. Vredno(s)tni relativist pa trdi, da ni nobeno merilo vrednotenja pravilno ali nepravilno, niti bolj ali manj pravilno, utemeljeno ali razumsko sprejemljivo od drugega. Vredno(s)tni relativist »zanika, da bi bilo karkoli na svetu takšno, da je za *vsakogar* […] razumno, da si to prizadeva s svojim ravnanjem uresničiti« (Klampfer, 2002: 90). Nemerljivost in neprimerljivost se največkrat zagovarjata pri vrednostnem političnem relativizmu, ki je povezan s političnim liberalizmom, izhaja pa, vsaj verzija, ki jo zagovarjata Berlin in Galston, iz moralnega vrednostnega pluralizma. Neprimerljivost, ki je tako osnova pluralista, pa pomeni, da ne moremo govoriti o boljših, slabših ali enakih elementih, saj neprimerljivost predpostavlja, da ni skupne enote, po kateri bi lahko vse elemente vrednotili, ni vrstnega reda vrednot, tudi ni »vrhovnega dobrega« ali prve vrline družbenih institucij (npr. pravičnosti), po kateri bi vrednotili vse ostalo, ali kateri bi vse ostalo sledilo (Bellamy, 2001: 4–8; Galston, 2004: 5–6).

V monističnih teorijah bi Hurka z idejo o racionalnosti obžalovanja oziroma s tem, da je intenzivnost obžalovanja proporcionalna izgubi stvari, ki jo obžalujemo, ne imel težav. Namreč, samo če obstaja natanko eno intrinzično dobro (ena dobra vrsta stvari (predmet, dejavnost, priložnost, užitek, dosežek) in če le-to izgubiš oz. ji ne slediš, je primerno, dobro, pravilno in racionalno, da občutiš obžalovanje. V primeru več intrinzično dobrih stvari pa se lahko situacija zaplete oziroma jo lahko razpletemo z načelom proporcionalnosti.

b. Sorazmernost obžalovanja

Vzemimo primer, kjer imamo več možnih dejanj glede na izhodiščne pogoje oziroma imamo nekaj možnih posledic dejanja. Pri tem imajo ta različna dejanja, te različne posledice, različno vrednost ali drugače, lahko govorimo o različnih količinah dobrega, ki izhajajo iz storjenega dejanja. Akter izbira med različnimi možnostmi, pri tem pa si zastavimo vprašanje, kakšno je razmerje med intrinzično dobrimi stvarmi in obžalovanjem. Odgovor smo poiskali v Hurkovem stališču o odnosu do intrinzično dobrih stvari. Predstavimo ga lahko z naslednjim argumentom:

> B1. Če je nekaj intrinzično dobro, potem je primerno, racionalno in dobro, da to nekaj ljubimo zaradi njega samega.
>
> B2. Pozitivna naravnanost oziroma neke vrste ljubezen do *katerekoli* intrinzično dobre stvari zaradi nje same je racionalna stvar.
>
> B3. Stvari, ki so intrinzično dobre, so lahko različno intrinzično dobre.
>
> Torej,
>
> B4. Naše občutke in čustva moramo porazdeliti, in sicer nekaj večjemu dobremu, nekaj manjšemu dobremu, še več, intrinzično boljšim stvarem več, intrinzično slabšim (a še vedno dobrim) manj.

Zgoraj navedeni sklep sledi načelu proporcionalnosti. Občutke in čustva porazdelimo, kar pomeni, da je tudi psihološka navezanost na katerokoli intrinzično dobro stvar porazdeljena, a vedno prisotna. Razlog za prisotnost psihološke navezanosti je v sami intrinzični naravi stvari. Njihova vrednost se, tako Moore, naj ne bi spreminjala ali celo izginila.

Če akter izbere dejanje, ki proizvede večje dobro, s tem »izgubi« manjše dobro, se zastavi vprašanje, ali bo imel psihološko reakcijo na izbiro. Z drugimi besedami, sklep prvega argumenta, to je A3 je bil, da če izgubimo nekaj, kar je dobro samo po sebi, je primerno, racionalno in dobro, da občutimo obžalovanje. V primeru, ko imamo več možnih intrinzično dobrih stvari in izberemo zgolj eno, je, če sedaj uporabimo proporcionalni pogled na obžalovanje, intenzivnost obžalovanja sorazmerna izgubi stvari, ki jo obžalujemo oz. je sorazmerna opuščeni intrinzično dobri stvari. Tako lahko, če nadaljujemo z zgornjim argumentom, obstoj obžalovanja razložimo tudi v etiškem pluralizmu:

B5. Racionalnost obžalovanja se veže na izgubo stvari, ki ima intrinzično vrednost.

B6. Vsaka izguba nečesa intrinzično dobrega je proporcionalna izgubi stvari, ki jo obžalujemo.

Torej,

B7. Intenzivnost obžalovanja je proporcionalno porazdeljena.

S tema argumentoma smo predstavili Hurkovo stališče. Ravno ta proporcionalni pogled na obžalovanje, ki ga vključimo v definicijo razumnega obžalovanja, nam omogoča govoriti o obžalovanju tako v etiškem monizmu kot v etiškem pluralizmu. In četudi bi se lahko vrnili v preteklost in bi naredili enako izbiro, npr. ubili eno osebo in s tem rešili deset drugih, saj bi bilo tako dejanje v celoti gledano najboljše oz. najmanj slabo, lahko sedaj razumemo, zakaj je lahko kljub temu, da smo izbrali večje dobro, racionalno obžalovati svojo odločitev. Tudi Williams, Kuhn in Dancy trdijo, da bi tako konsekvencialist kot deontolog še vedno čutila obžalovanje. So namreč situacije, t.i. tragične dileme, kjer se ne moremo izogniti temu, da storimo napačno dejanje, in karkoli storimo, plačati bomo morali moralno ceno (Dancy, 1993: 219). A to so izjeme.

Taki rešitvi bi lahko ugovarjali tudi rekoč, da lahko o pravem etiškem konfliktu govorimo zgolj v primeru, ko izbiramo med dvema stvarema z enako težo, npr. ko imata obe alternativi vrednost +1. Vendar, če ima Hurka prav in je primerno (proporcionalno, seveda) čutiti obžalovanje tudi v primeru izbire večjega dobrega, potem bi lahko dokazovali, da je obžalovanje primerno tudi takrat, ko izbiramo med enakimi količinami dobrega. Sorensen (1998: 531) takšno razmišljanje zavrne. Pri tragičnih dilemah, kjer imata obe alternativi enako vrednost, recimo +1, imamo razlog, da se odločimo za eno od njih (namesto za nobeno), čeprav nimamo razloga, da bi se odločili raje za eno kot za drugo. V takem primeru je zato primerno in pravilno obžalovati, da smo postavljeni pred izbiro, ne pa tudi, da, da smo izbrali eno in ne drugo.

c. Obžalovanje in relacija del–celota

Vzemimo naslednji primer, ki postavi zadnje zapisani argument pod vprašaj. Izbrati moramo med petimi počitniškimi cilji, ki smo jih že izbrali iz kataloga stotih. Ti cilji imajo, glede na idejo sorazmernosti, za nas enako vrednost, kar pomeni, da si vsakega izmed petih želimo enako močno oziroma da nam vsak izmed njih predstavlja enako količino dobrega. Do vseh imamo torej enako naravnanost (ljubezen), vendar pa moramo na koncu izbrati zgolj en počitniški cilj. Iz Hurkove ideje, da je intenzivnost obžalovanja sorazmerna vrednosti dobljene/izbrane in izgubljene/opuščene stvari, lahko izpeljemo, da je obžalovanje primerno za vsak počitniški cilj, saj je do vsakega enaka naravnanost. Če pa je obžalovanje primerno za vsak počitniški cilj, ki ni bil izbran, in seštejemo obžalovanja neizbranih ciljev, je rezultat tega, da negativni občutki presežejo pozitivne. Podobno bi se nam zgodilo, če ima ena izmed petih destinacij manjšo prednost. Izguba štirih dobrih destinacij pretehta dobitek ene, ki je boljša od vsake posamične. S tem bi bilo razumljivejše, zakaj bi nekdo imel negativno psihološko reakcijo (npr. obžalovanje) v primeru izbire večjega dobrega.

Hurka se je tega zavedal, zato je dejal, da moramo nekako zamejiti intenziteto racionalnega obžalovanja. Če tega ne storimo, se lahko zgodi, da bodo izbirek, v primeru da bi neizbrane značilnosti ovrednotili in sešteli, pretehtali negativni občutki, kar bi bilo iracionalno.

Vendar, prvič, a lahko govorimo o seštevanju in drugič, če ne, kako potem zamejiti intenziteto racionalnega obžalovanja? Po Hurki je vsaka značilnost (vsak od petih počitniških ciljev) del racionalnega odziva, ki pa mora biti oziroma imeti, če želi biti del tega odziva, intrinzično vrednost. Torej, vsak element disjunkcije (vseh pet ciljev) bi moral imeti intrinzično vrednost. Vzemimo, da so elementi disjunkcije deli celote, ki zajema vseh pet počitniških ciljev. Zaradi Zmote seštevanja[3] ne moremo uporabiti pri nadaljnji razlagi modela kuhinjske tehtnice, vendar pa lahko, glede na to, da vsak element, vsi počitniški cilji, ostajajo ves čas v igri, to pomeni, da so prisotni še po izbiri, uporabimo model organskih celot. Natančneje, Dancyjev model organskih celot, pri katerem vrednost celote ni identična seštevku vrednosti njenih delov, temveč je identična s seštevkom vrednosti prispevajočih delov (kar pomeni, da obstajajo v celoti deli, ki ne prispevajo k vrednosti celote). To ne pomeni, da neprispevajoči elementi nimajo nobene vloge, pomeni lahko celo nasprotno, da so v določeni situaciji nujni deli celote (v našem primeru elementi disjunkcije), a ne nujni v smislu intrinzičnosti, temveč v smislu omogočanja. Takšno stališče bi lahko sprejel tudi Hurka, saj bi to pomenilo, da smo zamejili intenziteto racionalnega obžalovanja, saj ne deluje po principu seštevanja oziroma tehtanja celotne teže oziroma vrednosti vseh prisotnih elementov (vseh pet destinacij z njihovimi vrednostmi), temveč s tehtanjem oziroma s seštevkom prispevajočih delov, tistih, ki pridonesejo h končni odločitvi, in tistih, ki omogočijo končno odločitev.

[3] Zmota seštevanja se nanaša na tisto moralno razmišljanje, v katerem naj bi imeli opravka s tako imenovanim učinkom kuhinjske tehtnice. Pri slednjem gre za to, da ima snov, ki jo tehtamo, vedno enako težo, ne glede na to, ali jo tehtamo samo ali pa skupaj s čim drugim (kakšno drugo snovjo). Analogno tehtanju naj bi bilo moralno razmišljanje. Tisto, kar delamo pri njem, je namreč to, da damo na eno stran razloge za dejanje, na drugo pa razloge proti dejanju. Potem vidimo, kaj je težje oziroma, katero dejanje je bolj pravilno. Npr. »govoriti resnico« ima vrednost + 2, »držati obljubo« pa + 1. Če damo vsako na svojo stran, bo »govoriti resnico« pretehtalo »držati obljubo«. Toda ne vedno. Vzemimo, da imamo na strani obljube še tretjo lastnost z vrednostjo + 1,5. Ker naj bi bila vsota razlogov proti dejanju, če jih »tehtamo skupaj«, enaka njihovi vsoti, če jih »tehtamo posamično«, stran, kjer je »držati obljubo«, pretehta stran, kjer je »govoriti resnico«.

V nasprotju s povezavo, ki smo jo naredili med Hurko in Dancyjem sami ne vidimo razloga za to, zakaj bi neizbrani počitniški cilji nujno predstavljali dele celote, še posebej, ker ne predstavljajo niti omogočajočih niti neomogočajočih pogojev pri vrednosti izbranega počitniškega cilja. V primeru da je vrednost petih ciljev enaka, da noben počitniški cilj ne pretehta drugih, je izbira lahko ali posledica naključne izbire (morda žreba) ali uvida. Če to drži in neizbrani deli nimajo nobene vloge, ko smo opravili izbiro, potem obžalovanje ni racionalno.

Kar je še bolj čudno, je stališče Hurke, ki obžalovanju (ki izhaja iz neizbranih počitniških ciljev) pripisuje časovno komponento. Hurka pravi: »Obžalovanje, četudi racionalno, bi moralo biti omejeno, ne zgolj zaradi razlogov, ki so povezani z užitkom in bolečino, temveč tudi zaradi tega, ker s časom njihova intrinzična primernost izgine« in »Obžalovanje je tako kot instanca proporcionalne ljubezni racionalno, toda tako kot vse takšne ljubezni, tudi obžalovanje sčasoma postane zaradi oddaljene možnosti manj racionalno« (Hurka, 1996: 560). Če je to res tako, se nam zastavita dve vprašanji: prvo, ali pomeni *intrinzična primernost* intrinzično značilnost ali pa način, po katerem nekaj vrednotimo; in drugo, kdo nam lahko zagotovi, da tako imenovana intrinzična primernost ne izgine. Odgovora na obe vprašanji sta v medsebojni odvisnosti, saj če lahko intrinzična primernost izgine, potem ne moremo govoriti o intrinzični značilnosti, kot jo običajno razumemo (in o kateri smo govorili v zgornjih odstavkih), in izgubimo racionalno obžalovanje. Toliko o kritiki proporcionalnega stališča do obžalovanja in z njim povezanim dojemanjem intrinzičnosti.

II.

Stališče Hurke, da njegova ideja racionalnega obžalovanja podpira tako etiški pluralizem kot monizem, je v nasprotju s stališčem Dancyja, ki trdi, da je pojem racionalnega obžalovanja element pluralizma in ne monizma. V nadaljevanju bomo poskušali zagovarjati (vsaj za zdaj) upravičenost Dancyjevega stališča. Preden predstavimo zagovor, pa še nekaj besed o samem monizmu, to je o vrsti monizma, ki ga imamo v mislih in ki mu odvzemamo možnost racionalnega obžalovanja. Vzemimo monistično etiško teorijo, pri kateri zgolj užitek predstavlja vrednost (ena vrsta utilitarizma), po kateri vrednotimo posamezna dejanja. Če imamo enake vrste užitka, to je govorimo o enaki lastnosti, na primer petnajstminutno ali tridesetminutno plavanje, potem obžalovanje, ki ga čutimo, ni racionalno. Če pa nimamo enakih vrst užitka, to je imamo različni značilnosti, na primer petnajstminutni užitek plavanja ali petnajstminutni užitek plesanja, lahko zavzamemo stališče nebenthamovskega utilitarizma ali benthamovskega utilitarizma (v obeh primerih govorimo o etiškem monizmu). Slednji zagovarja trditev, da je edina intrinzična lastnost užitka prijetnost. To pomeni, da lahko v primeru, ko primerjamo plavanje in plesanje, govorimo zgolj o različni količini užitka in ne o kakšnih nadaljnjih različnih notranjih lastnostih. Užitki torej, ki izhajajo iz različnih virov, se tako ne razlikujejo po intrinzičnih lastnostih, iz tega pa izpeljemo, da nimamo nikakršnega vzroka oziroma razloga za obžalovanje v primerih, ko je količina enaka.

V nasprotju s tem nebenthamovski monizem trdi, da se primera intrinzično razlikujeta. Če namreč primerjamo petnajst minut plavanja in petnajst minut plesa, lahko govorimo o lastnostih, ki so prisotne v obeh primerih, na primer »prijetnosti«, in pa o lastnostih, ki se ne pojavijo v obeh primerih, a so v konjunkciji s tistimi, ki si jih delita. Hurka trdi, da lahko še vedno govorimo o monizmu, saj obstaja zgolj ena intrinzična lastnost, ki se veže na dobro oziroma ki dejanje naredi »dobro«, in to je prijetnost. Pa vendar je obžalovanje zaradi drugih konjunkcijskih lastnosti racionalno. Zdi se, da je omenjeno stališče prej pluralizem kot monizem (pri čemer se zastavlja vprašanje, ali lahko govorimo o različnih intrinzičnih lastnostih), saj je monizem, na primer klasični utilitarizem, teorija, kjer je užitek edini nosilec intrinzične vrednosti. Tisto, kar utilitarista zanima, je, koliko užitka bo prineslo prvo dejanje in koliko drugo. Izbira je dejanje z večjo količino užitka. Kaj pa če sta oba enaka užitka? V omenjenem primeru plavanja in plesa, če smo izbrali tisto, ki nam bo prineslo več užitka, bi bilo obžalovanje za utilitarista popolnoma nekonsistentno in neracionalno psihološko stanje. To pomeni, da o monizmu in obžalovanju ne moremo govoriti v eni sapi.

Tako smo ponovno obtičali s pluralizmom. Dancy trdi: »Ko si izbral večje dobro napram manjšemu dobremu, pri čemer to manjše dobro ni vključeno v izbrano dobro, je občutek obžalovanja za manjšim dobrim, ki ga ni več, racionalen« (Dancy, 1993: 123). Za zagovor tega stališča znotraj teorije razlogov Dancy vpelje dva koncepta: prvič, koncept manka in pa drugič, koncept premaganih razlogov. Poglejmo to na že omenjenem primeru. Sprejeti moramo racionalno odločitev med dejanjem *A* in dejanjem *B*. Po natančnem in podrobnem pregledu obeh primerov smo racionalno izpeljali, da predstavlja *B* večje dobro v primerjavi z *A*-jem. Tako izberemo *B*. Ker pa ima boljša možnost lahko neki manko (ki smo ga pri vrednotenju *B*-ja upoštevali), ki ga manjše dobro oziroma slabša možnost nima, je racionalno občutiti obžalovanje (oziroma, povedano drugače, boljša možnost ima določene »nevrednosti«, lastnosti, ki ji zmanjšujejo vrednost, katerih možnost *A* (slabša možnost) nima), a je še vedno boljša izbira. Koncept manka je torej povezan s konceptom obžalovanja oziroma z njim lahko pojasnimo razumnost obžalovanja zaradi neizbire *A*-ja. Glede na to, da pri monizmu tega manka nimamo (govorimo namreč zgolj o enem nosilcu intrinzične vrednosti), tudi obžalovanja ni. Takšno stališče temelji na vlogi, ki jo je Dancy pripisal premaganim intrinzičnim vrednostim. Le-te niso prispevajoče vrednosti, vendar so še vedno prisotne z vso svojo močjo, ki jih premorejo. Torej jih ne moremo ignorirati in izbrisati oziroma zavreči. Če pa drži, da jih ne moremo neupoštevati oziroma da je njihova moč še vedno prisotna, potem morajo imeti neko vlogo. Po Dancyjevi razlagi je ravno obžalovanje pokazatelj, da še vedno imajo neko vlogo in moč.

III.

Ob poskusu odgovoriti na vprašanje, zakaj se pojem obžalovanja izpostavlja kot pomemben element v razpravah, bomo prešli iz vrednosti na razloge, saj menimo, da se odgovor na vprašanje, ali je obžalovanje upravičeno, nahaja v teoriji razlogov. Partikularizem in generalizem nam bosta dala odgovor na začetno vprašanje, to je zakaj je pojem obžalovanja pomemben element filozofskih razprav. V želji po razlikovanju med generalistom in partikularistom si zastavimo naslednje vprašanje: Ali obstajajo kakšne opisne lastnosti (ali vsaj ena), ki so vedno etiško relevantne in so etiško relevantne vedno na enak način? Če odgovorimo pritrdilno, smo generalisti, če nikalno pa partikularisti. Vzemimo Kantov primer skrivanja prijatelja pred morilcem, ki je pravkar pozvonil na naša vrata. Lahko se zlažemo in rešimo prijatelja, lahko izrečemo resnico in ga obsodimo na smrt. V primeru da po odločitvi, da rešimo prijatelja, trdimo, da je lastnost »govoriti resnico« še vedno etiško relevantna (in etiško relevantna na isti način), prav tako pa tudi njena moč oziroma vrednost, smo generalisti. Vzrok se skriva v invariantni polarnosti. Generalist, kot je Williams, sicer opisuje in razlaga primere, kjer se zdi, da opisna lastnost ni več etiško relevantna, ker je premagana, pa vendar, če obdrži isto vrednost in je prisotna, potem je tudi, kot razlog, etiško relevantna. Moralna obligacija je tako kategorična. Premagani razlogi, tako J. Dancy, niso prispevajoči razlogi, vendar so še vedno prisotni z vso svojo močjo in vrednostmi, ki jih premorejo. Torej jih ne moremo ignorirati in izbrisati oziroma zavreči. Če pa drži, da jih ne moremo neupoštevati oziroma da je njihova moč še vedno prisotna, potem morajo imeti neko vlogo. Po Dancyjevi razlagi je ravno obžalovanje pokazatelj, da še vedno imajo neko vlogo in moč. Partikularist, v nasprotju z generalistom, trdi, da je razlog utišan takrat, kadar je dejansko brez moči, brez vrednosti. Utišani razlog ni več etiško relevanten oziroma ni etiško relevanten na enak način.[4] Zato ni racionalnega obžalovanja.

Lahko pa pogledamo še iz druge strani. Mark Timmons trdi, da nekateri filozofi (predvsem partikularisti), s čimer se sam ne strinja, zagovarjajo način, po katerem lahko ločimo med utišanimi in premaganimi razlogi. Kriterij za presojanje je oblika določene psihološke reakcije, ki jo ima akter. Kriterij za prepoznavo, ali imamo opravka z utišanimi ali s premaganimi značilnostmi, pa je racionalno obžalovanje (Timmons, 2002: 261). Četudi Timmonsu to stališče ni pogodu, menimo, da je omenjeni kriterij dober, saj ima utišana značilnost naslednji lastnosti: (i) lastnost, ki je utišana, nima več (po utišanosti) nobene vloge (odločili smo se že, da bomo, na primer, storili dejanje *A* in ne dejanja *B*), in (ii) ker je izgubila vso svojo moč, obžalovanje ni racionalno. Torej, obžalovanje ima smisel zgolj, če so razlogi lahko premagani, ne da bi bili obenem utišani. V Kantovem primeru s pri-

[4] Vendar, če je izkupiček pri obeh primerih enak (ne glede na to, ali smo generalisti ali partikularisti), se pravi, da bi morali lagati, potem ni razlike med premaganimi in utišanimi razlogi, prav tako pa tudi ni razlike med atomizmom in holizmom.

jateljem in z morilcem bi to pomenilo: če obžalujemo, je lastnost »izreči laž« premagana, če ne, je utišana. Takšno stališče bi lahko imenovali partikularizem brez premaganih razlogov, kar pa ni po Dancyjevi volji. Je pa po volji Aristotela: Človek z vrlinami »dela z veseljem ali vsaj brez obžalovanja, kajti dejavnost, ki je v skladu z vrlino, mora biti prijetna in neboleča, o obžalovanju pa ne sme biti govora« (Aristotel, EN, 1120a). O obžalovanju bi tako lahko govorili zgolj v primeru, da nismo delovali v skladu z vrlino, ali pa da razlogi, ki so delovali za dejanje, ki ni bilo izbrano, niso bili po izbiri utišani.

Glede na vse dosedanje ugotovitve se torej zdi, kot trdimo v tem članku, da racionalno obžalovanje ni kompatibilno s takšnim partikularizmom, kot ga zagovarja Dancy, razen če bi sprejel idejo o premaganih razlogih, kar pa postavlja pod vprašaj sam partikularizem. Če namreč sprejmemo razlago koncepta racionalnega obžalovanja v teoriji razlogov in hkrati menimo, da je holizem upravičena teorija, potem moramo racionalno obžalovanje zavrniti. Tisto, kar imamo, je tako partikularizem brez obžalovanja.

Pa se vrnimo k La Fontainovi bajki Mravlja in škržad: škržad mravlji, ki ni bila odprtih rok, in se ji je to, kar je počel škržad čez poletje, zdelo greh, odgovori, da se zaveda, da je noč in dan pel, a doda, »saj nimate nič proti?« (Fontaine, 2000: 591). Še več, pred tem ji tudi pove, da bo plačal, kot po navadi, v mesecu juliju. To pa je natanko tisti mesec, ko škržadi »pojejo in igrajo«. Škržad torej:

– ne želi živeti na tuj račun in se neupravičeno okoristiti z delom drugih;

– tudi ni ravnal lahkomiselno, sedaj pa ne bi vedel, kaj bi;

– besede, da bo plačal, kot je v navadi, nakazujejo na njegovo zavedanje glede igranja in nenabiranja hrane, kar pomeni, da pristopa racionalno do obojega;

– glede na to, da naj bi plačal z lastnim igranjem, to pomeni, da ima igranje vrednost zanj in za druge, če pa običajno s tem plača v mesecu juliju;

– škržad ni beračil niti se pokesal ali videl kakršnokoli težavo v tem, škržad je namreč mravljo prosil zgolj za posojilo, ali kot reče »Plačal bom, kot je v navadi, v juliju pri moji vesti, glavnico in vse obresti« (Fontaine, 2000: 591).

Tako je interpretacija, da škržad ne čuti racionalnega obžalovanja, popolnoma upravičena. Še več, želi si, da bi v mrzli zimi imel hrano (a je zaradi poletnih odločitev seveda nima), pa vendar, če bi lahko potoval v preteklost, si ne bi želel, da bi se moral odločiti drugače oziroma se (ob vseh enakih elementih konteksta) ne bi odločil drugače. Škržad tako ravna po premisleku, je togo racionalen, konsistenten in brez racionalnega obžalovanja (Bliss, 2002: 427).

Zaključek

Obžalovanje je kot značilnost ali kot psihološka reakcija, ki je prisotna po izbiri med večimi možnostmi, po trditvah Hurke racionalno, pri čemer je to omogočeno, vsaj po njegovem mnenju, v nebenthamovskih teorijah in v nekaterih pluralističnih teorijah dobrega. Razlago utemeljuje na konceptu ljubezni in intrinzičnosti, kar smo zavrnili kot neupravičeno oziroma neprimerno.

Prav tako smo zavrnili Dancyjevo pozicijo, ki se nam je zdela, ker je zavrnil zgoraj omenjeno, prepričljivejša, saj meni, da je mogoče racionalno obžalovanje ustrezno razložiti v pluralizmu in partikularizmu razlogov. Lahko bi imeli željo takšnemu partikularizmu ugovarjati, saj bi menili, da je, četudi smo racionalno rešili konflikt, obžalovanje še vedno prisotno. Tisto, kar pa ni prisotno, je konsistentna teorija, ki bi razložila, zakaj je to tako. Edina upravičena razlaga, ki jo imamo, je, da občutimo obžalovanje, ker sta oba elementa, tako element *A*, ki smo ga izbrali, kot element *B*, ki ga nismo, še vedno v igri.

Glede na to, da so razlogi, ko smo odločitev sprejeli, utišani in ne premagani, ni nobenega upravičenega razloga občutiti racionalno obžalovanje. Prav tako kot ni nobenega upravičenega razloga za trditev, da sta oba elementa še vedno v igri. V primeru namreč, da sta, potem naša odločitev ni bila dobra oziroma racionalna. Če se vrnemo k bajki, ki smo jo omenili na začetku tega poglavja, pri interpretaciji le-te, ki jo je podala M. Nussbaum, škržad odgovori mravlji na vprašanje, če mu je sedaj kaj žal, da je zgolj pel vse poletje. »Zelo žal,« ji odgovori škržad, »tako, kot sem vedel, da mi bo. Toda sedaj je sedaj, takrat sem ravnal racionalno. Bitje, ki je neracionalno, si ti, saj se upiraš svoji sedanji želji, da bi mi pomagala« (Hollis, 1996: 60). Četudi je razvidno iz odgovora, da sprejme idejo obžalovanja, lahko izpeljemo, da mravlja predstavlja generalista in škržad partikularista (glede na obžalovanje Dancyjevega partikularista). Kar imamo sedaj, sta partikularistični škržad, ki se predaja življenju (in sprejme vso odgovornost za lastna (ne)dejanja), in pa generalistična mravlja, ki sprejema racionalne odločitve in ima nenehno slab občutek, to je, ji je žal za stvari, ki jih ni storila oziroma, v primeru škržada, da mora sprejemati odločitve, ki jo mučijo. Tako smo lahko generalisti z obžalovanjem ali partikularisti brez obžalovanja. V tem članku smo zagovarjali partikularizem. Brez obžalovanja, seveda.

The Hurka's Concept of Rational Regret through Perspective of Dancy's Moral Particularism

The notion of rational regret has been mostly discussed in the Theory of Value. The article presents B. Hurka's position on when regret is rational in non-Benthamite monistic as well as moderate pluralistic theories of the good. It is shown that J. Dancy, on the other hand, rejects the idea that monism provides us with an adequate explanation of regret, but is convinced that pluralism and consecutive particularism can provide it. In the end, an argument is made against this claim of Dancy's, both in terms of value and in terms of reasons. Namely, if the concept of regret is accepted in the theory of reasons, and if holism is plausible, we will reject the possibility of rational regret.

Keywords: monism, pluralism, regret, particularism, theory of reasons.

Literatura

Aristotel (2002). *Nikomahova etika*. Ljubljana: Slovenska matica.

Audi, R. (2004). *The Good in the Right*. Princeton, Oxford: Princeton University Press.

Bellamy, R. (1999). Liberalism and Pluralism: Towards a politics of compromise. Taylor and Frances e-Library (2001).

Berlin, I. (1958). »Two Concepts of Liberty«. V Berlin, I., *Four Essays on Liberty,* Oxford: Oxford University Press.

Bliss, C. (2002). »Life-Style and the Standard of Living«. V Nussbaum, M. in Sen, A. (ur.), *The Quality of Life*, Oxford: Oxford University Press.

Dancy, J. (1993). *Moral reasons.* Oxford: Blackwell.

Galston, W. A. (2004). Liberal Pluralism: The Implications of Value Pluralism for Political Theory and Practice. Cambridge: Cambridge University Press.

Hollis, M. (1996). Reason and Action. Essays in the Philosophy of social science. Cambridge: Cambridge University Press.

Hurka, T. (1996). »Monism, Pluralism, and Rational Regret«. *Ethics*, 106, str. 555–575.

Kahn, L. (2011). *Conflict, Regret, and Modern Moral Philosophy.* Pridobljeno 1. 9. 2016, s http://www.academia.edu/226939/Conflict_Regret_and_Modern_Moral_Philosophy.

Klampfer, F. (2002). »Vrednotenje in morala«. V Miščević, N., Kante, B., Klampfer, F. in Vezjak, B., *Filozofija za gimnazije*, Ljubljana: Cankarjeva založba.

La Fontaine, J. (2000). »Škržad in mravlja«. *Primorska srečanja: revija za družboslovje in kulturo*, 24, 231/232, str. 591.

Matičetov (2000), Primorska srečanja: revija za družboslovje in kulturo, 24, 231/232, str. 587–597.

Moore, G. E. (2000). *Principia Ethica.* Ljubljana: Študentska založba.

Ross, W. D. (2002). *The Right and the Good*. New York: Oxford University Press.

Sorensen, R. (1998). »Rewarding Regret«. *Ethics,* 108, str. 528–537.

Timmons, M. (2002). *Moral Theory: An Introduction.* Maryland: Rowman & Littlefield Publishers, Inc.

Williams, B. (1985, 2011). *Ethics and the Limits of Philosophy*. London & New York: Routledge.

Logika, filozofija jezika in filozofija znanosti

Danilo Šuster
Filozofska fakulteta Univerze v Mariboru

Begging the Question – Proper Justification or Proper Conversation?

Petitio - kršitev pravil upravičenja ali pravil dialoga?

Aristotel ponuja dve veliki pojasnili zmote *petitio* (krožno sklepanje): napačna poteza v argumentativnem dialogu ali pa epistemska napaka (premisa je manj "utemeljena" od sklepa). V drugem primeru védenja o premisi ne moremo imeti neodvisno od védenja o sklepu. Dialektični pristop uporablja pojem sprejetja (angl. *committment*) in se sklicuje na normativnost pravil dialoga. Predlagam hibridni model, ki temelji na teoriji F. Jacksona: razlog za sprejetje obvez v dialogu je epistemski.

Ključne besede: argument, *petitio*, epistemski model, dialektični model, sprejetje.

> »I think, therefore I am.« Descartes has begged the question here, because when he said »I think,« he'd already implied »I am« (or how else could he think?). Yet his fallacy continues to persuade people, over three hundred years later.[1]

1.

An influential paper starts with a remark (Sinnott-Armstrong, 1999: 174): »No topic in informal logic is more important than begging the question. Also, none is more subtle or complex.« Not just informal logic, reasoning and argumentation, the circles of justification and inference in general – all these topics have to address this fallacy, if the fallacy it is. The issues are complex and the literature is large and growing, but there remains something deeply puzzling when arguments are discredited as being circular, or question-begging, or being a *petitio principii* – the labels are different, but I will assume in this paper that they describe more or less the same phenomenon.

[1] *Mission Critical* – an old web page, probably no longer active, though still accessable: http://missioncritical.royalwebhosting.net/part2/circular.html, January 8th, 2021.

Let me start with a quick textbook definition: to beg the question is to assume the truth of what one seeks to prove, in the effort to prove it, according to Copi and Cohen (1991: 102). In a later edition *petitio* is described as an informal fallacy in which the conclusion of an argument is stated or assumed in any one of the premises (Copi et al., 2013: 140). Textbooks will often say that an argument begs the question if any doubt about the conclusion would equally infect the premises. A question immediately arises: is this a *logical* fallacy at all? Mill reminds us that any valid argument is such that if one doubts the conclusion, one ought to doubt the premises. So every valid argument begs the question? Some tricks are needed to avoid this generally unwanted conclusion. Hurley states that the fallacy of begging the question is committed whenever the arguer creates the *illusion* that inadequate premises provide adequate support for the conclusion, for instance: »*Clearly*, terminally ill patients have a right to doctor-assisted suicide. After all, many of these people are unable to commit suicide by themselves.« But the argument »No dogs are cats. Therefore, no cats are dogs« commits no fallacy because no *illusion* is created to make inadequate premises appear as adequate (Hurley, 2013: 163). No illusions, no disappointments according to the Japanese proverb?! I nevertheless think that the »no dogs« argument remains a disappointment if the argument is used to remove anyone's doubt about the conclusion. And the first argument will remain fallacious even after omitting seducing words »clearly« and »after all« .

Since perfectly valid arguments can beg the question, the mistake involved cannot be of a logical kind. A very radical diagnosis was proposed by Robinson (1971). If the fallacy is not one of deductive logic then, as matter of fact, no arguments beg the question. »The prohibition of begging the question is not a law of logic, nor a maxim of good scientific method. It is merely a rule of an old fashioned competitive game.« And he adds: »There are only two proper ways of condemning an argument. One is to say that the conclusion does not follow from the premises. The other is to say that you do not accept the premises as true … Begging the question appears to be neither of these. So it is not a proper accusation.« (Robinson, 1971: 114)

Let me first note that the accusation of begging the question is really often misused so as to cover any assumption found problematic in the argumentative exchange. You disagree with your opponent? You find her conclusion implausible? There must be something wrong with her premises, most likely she already assumes the conclusion in one form or another. Augustus De Morgan has already remarked that:

> There is an opponent fallacy to the *petitio principii* which, I suspect, is of more frequent occurrence: it is the habit of many to treat an advanced proposition as a begging of the question the moment they see that, if established, it would establish the question. [...] Are there not persons who think that to prove any previous propositions, which necessarily leads to the conclusion adverse to them, is taking an unfair advantage? (quoted by Walton, 1991: 257)

I have actually heard an academician argue in the following way: »So you argue for, say, decriminalization of (soft) drugs. I disagree strongly! Let me see, hmm …, your conclusion follows from the results that show that the strict policy of prohibition is unsuccessful. But we all know that you are usually biased in your interpretations of data. Your premise is thus unacceptable, *petitio*!«

Following Robinson one could say that this accusation merely marks a disagreement with one of the premises. But how about *circularity* proper as the criterion? It is often said that »begging the question« occurs when the same proposition is asserted twice, both as a premise and as a conclusion. But it would be too restrictive to limit the defect to cases when the conclusion appears *explicitly* as a premise for that would allow to avoid the charge by a mere reformulation of premises. Surely, the following argument begs the question: »Since firefighters must be strong men willing to face danger every day, it follows that no woman can be a firefighter.« One could try to amend the criterion by saying that the defect resides in the fact that the conclusion appears *implicitly* in the premises. And what does *that* mean? Do one's premises include implicitly one's conclusion if they cannot all be true unless the conclusion is? Well, then all *valid* arguments beg the question, »so it is not a proper accusation,« just as Robinson has said.

The *equivalence* conception of the fallacy is that some premise of the argument is equivalent to the conclusion (explicitly or implicitly). I think that a more lax, *dependency* conception is the right way to go (cf. Walton 2006). We should start with a *probative* function of an argument – the premises provide evidence of a kind that gives the respondent a reason to accept the conclusion. The conclusion depends, justificatorily, on the premise, the »flow of inference« is the »flow of evidence« and it has to go from the premise to the conclusion. The premises are used to remove the respondent's doubt about the conclusion. But where it is also required that (some kind of) inference be made in the other direction, from the conclusion to the premise, the argument begs the question. In a *petitio* the premise depends on the conclusion in a way that undermines the probative purpose of arguing. Justification is an epistemic notion and I am inclined to accept broadly epistemic criteria of begging the question and arguments in general. Within a broadly epistemic theory the principal goal of argumentation is, roughly, to induce belief or elicit a reasoned change in view. In the case of *petitio* arguing is *epistemologically* unsuitable for the purpose of proving the conclusion in that particular discussion. But we should also acknowledge that wrongness of *petitio* stems from a pragmatic and contextual notion of how an argument is *used* in an argumentative dialogue. Can the two approaches be reconciled?

The split between »dialectical« and »epistemic« approaches to begging the question goes back to Aristotle. In the *Prior Analytics* (64b 33) *petitio* is characterized as the attempt to prove what is not self-evident by means of itself. But demonstration proceeds from what is more certain or better known: if a man tries to prove what is not self-evident by means of itself, he begs the original question (64b 37). To beg the question in this sense is to violate the epistemic principle of the priority in knowledge of the premises over the conclusion in a demonstration. Aristotle uses the same terminology in the framework of the *dialectical* or conversational account of the fallacy. In the *Topics* the account is set in terms of contentious disputation between two or more parties. Begging the question is said to occur when the party who is supposed to be arguing for a certain thesis asks to be granted the thesis as a premise to be conceded by his opponent. In *Sophistical Refutations* Aristotle discusses *petitio* in the context of »arguments used in competitions and contests« where one party has the task of proving a proposition (the »question« to be proven) to another party (165b 12). To carry out his task, the first party will have to ask the second party to grant or concede premises that the first party can use in his argument. However, if the first party were to ask (beg for) the very conclusion as a concession, without doing the required work of proving it, then he would be »begging for the question at issue.« This would not be allowed, because the prover would be avoiding the task of proving the proposition at issue.

2.

I have sketched some main dilemmas about *petitio*: there are deflationary views (begging the question is *not* a fallacy at all) and inflationary views (*any* argument that contains an unwarranted presupposition or inadequately supported premise is a *petitio*). Moreover, how to avoid the verdict that every valid argument begs the question? And which is better, the epistemic or the dialectical approach?

Let me start with *The Bank Manager Example* (»a staple of many textbooks«):

> Manager: Can you give me a credit reference?
>
> Smith: My friend Jones will vouch for me.
>
> Manager: How do we know he can be trusted?
>
> Smith: Oh, I assure you he can.

In this dialogue one person is supposed to vouch for the reliability of the other. The reliability of the vouchee is in doubt and some secure source is needed to reassure this doubt. But if the reliability of the voucher is questioned, the reliability of the vouchee cannot be used to reassure this doubt, because it is itself in doubt, in the first place (cf. Walton, 2006: 248). One could say that Smith falsely presents *his* reliability as an accepted *starting point* in a dialogue. Circular arguments, in general, are fallacious because they violate normative rules of dialogue which demand consensual starting points according to modern defenders of *pragma-dialectical* approach (van Eemeren and Grootendorst, 2004).

Argumentation, according to this approach, is understood as an effective means of resolving a difference of opinion in accordance with discussion rules acceptable to the parties involved. Fallacies are illegitimate moves in a given discourse context, they violate rules of a critical discussion. There are eight rules or commandments and the sixth rule (the starting point rule) states (van Eemeren and Grootendorst, 2004: 193):

> Discussants may not falsely present something as an accepted starting point or falsely deny that something is an accepted starting point.

By falsely presenting something as a common starting point, the protagonist tries to evade the burden of proof; the techniques used for this purpose include advancing argumentation that amounts to the same thing as the standpoint (*petitio principii*). Circular arguments are fallacious because they violate normative rules of dialogue which demand consensual starting points.

Begging the question according to this account is a dialectical mistake that depends on the violation of some general rule of dialogue. Let me use another evergreen (two sentences from this passage are given as an exercise by Copi and Cohen, 1991: 110):

> To know or tell the origin of the other divinities is beyond us, and we must accept the traditions of the men of old time who affirm themselves to be the offspring of the gods – that is what they say – and they must surely have known their own ancestors. How can we doubt the word of the children of the gods? Although they give no probable or certain proofs, still, as they declare that they are speaking of what took place in their own family, we must conform to custom and believe them. [Plato, *Tymaeus* 40d–e]

Like *the Bank Manager Example* this case is on the borderline with some other fallacies. It has to do with trustworthiness, so it relates to *ad verecundiam* and *ad hominem* that also have to do with matters of the reputation of an agent and the trustworthiness of a source. On the face of it we have inductive reasoning »from tradition«:

> We must conform to custom and believe the men of old time.
>
> The men of old time declare that they are speaking of what took place in their own family.
>
> The men of old time must surely have known their own ancestors.
>
> The men of old time affirm themselves to be the offspring of the gods.

So,

> The men of the old time were the children of the gods.

Not blatantly fallacious (the power of custom carries at least some probative force, one usually knows one's ancestors), but the problem is created by the very *inductive* nature of the argument: is it not possible for the premises to be true and the conclusion false? Could men of old times err? But then we get the rhetorical question: How can we doubt the word of the children of the gods? The possibility of error is removed, but the question is begged. Here is a reconstruction proposed by Iacona and Marconi (2005, 33):

> 1 We cannot doubt the word of the children of the gods. P
>
> 2 [The men of the old time were the children of the gods.] P (?)
>
> 3 We cannot doubt the word of the men of the old time. 1, 2
>
> 4 The men of the old time affirmed themselves to be the children of the gods. P
>
> 5 The men of the old time were the children of the gods. 3, 4

The intended conclusion (5) is just the missing premise (2) which the arguer inserts in order to make the argument valid. Iacona and Marconi describe the case as a rhetorically efficient *covert petitio*, an invalid argument which hints at reasoning that can only be carried out once a premise is added, but the integrated argument is used to make *patent petitio* (in this case one of the premises being equivalent to the conclusion). But Marconi and Iacona ignore the larger inductive setting. To make the example even more vivid I will consider a dialogue between Achilles (supposed to be the son of nymph *Thetis*) and the Tortoise inspired by the classic Lewis Carroll (1895).

> Tortoise: Can you, Achilles, state some credible evidence for your being the son of the sea nymph *Thetis*?
>
> Achilles: I am speaking of what took place in my own family and I surely know my immediate ancestors.
>
> Tortoise: How do we know *you* can be trusted?

> Achilles: Oh, I can assure you. How can you doubt the word of a child of the gods?

In *The Bank Manager Example* when Smith gives Jones as a reference who can endorse him, the argument presumes that Jones is trustworthy because of his reputation, or because he is a member of a profession that is trustworthy, etc. But if one person endorses another, then the second cannot be used as a reference to endorse the first. Similarly, Achilles is trustworthy because he is a member of the club where membership requires special qualities. And how do we know that he really is a member of this club? Because members of the club do not lie. Membership in the club guarantees trustworthiness, but we have to accept his trustworthiness in order to count him as a member of the club.

On the conversational approach in the event of a difference of opinion one discussant puts forward a *standpoint* (»I am a child of gods«) and the other discussant calls that standpoint into question (»A son of gods, I really doubt that!«). The discussants are not in agreement on the acceptability of a certain standpoint. If any attempt to resolve this difference of opinion by means of a regulated discussion is to have any chance of success, it is necessary for the discussants to adopt a number of propositions accepted by both parties as their starting points. The best method for judging cases of begging the question is to keep track of the commitments in the dialogue in relation to the theses to be proved by both sides. So let us include this dimension:

> Achilles (the proponent): I really am a child of gods.
>
> Tortoise (the respondent): Force me, argumentatively, to accept the conclusion as true. Pick up your note-book again and kindly enter mutually accepted propositions.
>
> Achilles: Proceed! And no tricks this time – I will be on watch!
>
> Tortoise: Tell me whether you agree with the following proposition: There are people.
>
> Achilles: How trivial!
>
> Tortoise: So true, write it down.
>
> Achilles: Done.
>
> Tortoise: There are gods.
>
> Achilles: How trivial!
>
> Tortoise: So true, write it down.
>
> Achilles: Done. And you should also add: Gods are trustworthy.

> Tortoise: Well, they have their weaknesses, but I am, for the sake of discussion, prepared to grant that gods do not lie.
>
> Achilles: I will also write down: We must conform to custom.
>
> Tortoise: Granted, provisionally, for the sake of discussion.
>
> Achilles: According to the custom the word of the children of the gods should not be doubted.
>
> Tortoise: Accepted.
>
> Achilles: One knows what took place in his childhood in his own family.
>
> Tortoise: Well, not always, in my case there were those hundreds of eggs and ….
>
> Achilles: All right, all right, I just want to rationally convince you, the Euclidian iron logic (hmm …) is not really required. So let us say: A person (human or god-like) usually knows the nature of his mother and father. And I certainly know them!
>
> Tortoise: *Usually* is not always. Maybe *you* err?
>
> Achilles: You can trust my word.
>
> Tortoise: Hmm, let me have a look in your note-book. Sorry, trust is not enlisted.
>
> Achilles: But I told you I know my parents and Thetis, a divine being, is my mother!
>
> Tortoise: Let me have a look …, sorry, not on the list! We only have: a person (human or god-like, and why not a turtle-like?) usually knows her mother and father. Maybe you do know your mother, but you wrongly think she is a goddess?
>
> Achilles: But look, you agreed that gods are trustworthy.
>
> Tortoise: Indeed.
>
> Achilles: And children of gods are speaking of what took place in their own family.
>
> Tortoise: Well, I believe that children usually are honestly speaking of what they think took place in their own family, but … .
>
> Achilles: So how could you possibly doubt my word, a word of a child of the gods?

Achilles has taken on the burden to fulfill a probative function by putting forward an argument to the Tortoise. The initial standpoint (»Achilles is a son of goddess«) cannot, of course, form any part of the list of propositions that are acceptable to both parties, otherwise there would be no difference of opinion, nothing to argue for. In the case of begging the question, the error that is made is that the protagonist (intentionally or unintentionally) makes use of a proposition that, as he can know beforehand, is not to be found in the list of propositions that are acceptable to both parties. As I developed the case, the proponent (Achilles) is at least implicitly aware of this dialectical rule and tries to present the contested proposition as a conclusion (»so, …«). But we get the feeling that the dialogue could go on and on. The Tortoise is never presented with the premise he can commit to, independently of his doubts about the conclusion.

Instead of an accepted starting point one might speak, more generally, about *acceptances* in a conversation (cf. Hazlett 2006). A proposition *p* is accepted in conversation *c* when all the speakers of *c* are allowing utterances that express *p* into *c* (i. e. they are prepared to meet utterances expressing *p* with positive behavior such as saying 'Yes' and nodding, and are not prepared to meet utterances expressing p with negative behavior such as saying 'No' and shaking their heads). A proposition *p* is in question in *c* when neither *p* nor not-*p* is accepted in *c*. On this Gricean analysis an argument then begs the question to the extent that it violates the so called Submaxim of Relation (Hazlett, 2006: 356):

> You are allowed to use as premises only what you can reasonably expect your audience to accept based on their contribution to the conversation and whatever else you know about them *qua* speaker.

Walton (2006), inspired by Hamblin, uses the notion of a *commitment* in dialogue. What an arguer is committed to is what she has gone on record as saying in a dialogue, judging from the textual evidence in the case. Through arguing we commit ourselves to propositions and the commitments incurred are what we are concerned with when evaluating the argument. The fallacy of begging the question is then a failure that relates to how the respondent's commitments are used by the proponent's argument attempt. To rationally convince the respondent to come to accept the conclusion that she doubts, the proponent needs to use an argument with premises that consist only of propositions that the respondent is committed to, or is prepared to accept, independently of the proposition to be proved by the proponent.

Standpoints, commitments or acceptances – begging the question, according to this approach, is a matter of what your audience has indicated a willingness to accept in a dialogue. But where do commitments come from? Why are some of them included on the initial list and the others are not? What is the *rationale* for the relevant rule? Let us join Achilles and the Tortoise once again:

> Tortoise: Why do I doubt your word? In words of a famous philosopher yet to be born, we have to consider, whether it be more probable, that you should either deceive or be deceived, or that the fact of your divine origin should really have happened. And I think that the falsehood of your testimony is less miraculous.
>
> Achilles: So you do not believe me? You think I am not trustworthy?
>
> Tortoise: I just considered our list and since trust was not enlisted, I …
>
> Achilles: You call me a *liar*?
>
> Tortoise: I never said that! Look, let us inspect our list …
>
> Achilles (shouts): Beware the rage of Achilles!
>
> Tortoise (hides her head under the shell): But we agreed to use rational argumentation as an effective means of resolving a difference of opinion in accordance with discussion rules acceptable to both parties involved.
>
> (Suddenly a noise is heard, knocking on the door and shouting): Achilles, son of Peleus! Agamemnon, son of Atreus, the king of Argos is asking you to join Achaeans in the just war to effectively and ultimately resolve a difference of opinion with the Trojans.
>
> Tortoise (quietly): I knew they were still slightly primitive and do not understand the subtleties of logic in spite of their philosophers. (Sighs and adds). As a matter of fact, for the next two thousand years the argument of power will prevail over the power of argument.
>
> Achilles (overhears her): I really wonder how could somebody who has doubts about *modus ponens* complain about *my* logical abilities.

The Tortoise gives *epistemic* reasons for his doubt in the contested commitment. Divine origin would count as miraculous for him, but, according to Hume (1748/2000, 86–87): »a miracle is a violation of the laws of nature; and as a firm and unalterable experience has established these laws, the proof against a miracle, from the very nature of the fact, is as entire as any argument from experience can possibly be imagined.« Given that he doubts the conclusion, the evidence provided by Achilles for the contested premise is no evidence for him. In general, I think that given the probative purpose of the argument commitments and acceptances are not arbitrary but subject to *epistemic* constraints.

3.

Why is a certain proposition on the *index prohibitorum* as a commitment in an argumentative dialogue? Well, because it does not advance the issue. Arguments have different functions but I will assume that the main use of an argument is *probative* – an argument should »prove« its conclusion (make it knowable, rationally believable or acceptable, remove doubts …, – Walton 2006 makes this clear). A probative function of argument uses the premises to provide evidence that gives the respondent a reason to accept the conclusion. The question-begging commitment does not remove the respondent's doubt about the conclusion: the reasons offered by the proponent for the contested premise are ineffective against this respondent. We can agree that when one begs the question in a dialogue one breaks the rules of *conversation*, but the *rationale* for these rule is a prohibition of an epistemic deadlock. At least when fulfilling the probative function is required by the type of conversation the participants are engaged in.

I would thus argue for a *hybrid* model of begging the question, integrating epistemic dimension into a dialectical setting. The most promising approach which also provides a plausible solution to the problem of deduction (all valid arguments are *petitio*) was offered by Jackson (1984). On Jackson's view arguing has two basic functions. One is the 'teasing out' function, which basically amounts to proving something to those who already had the resources to do it themselves in the first place – showing them how propositions that they already believe entail other propositions which they had not previously recognized as consequences of what they already believed:

> The act of propounding an argument may have brought a half-buried piece of information to the surface, may have alerted me to the relevance of certain facts to my final concern, or may simply have enabled me to see how to get it altogether, so as to make transparent what I want to know (Jackson, 1984: 27).

I think that the »teasing out« function explains the initial puzzlement about the usefulness of *Cogito*. How about *petitio* proper? Here we have to consider the second function of arguing, the evidence-borrowing. In presenting an argument, speakers advertise themselves as having a certain sort of evidence for the premises which the audience may not possess. Hearing the argument enables the audience to 'borrow' that evidence, thereby coming to have justification for the premises, and a fortiori for the conclusion. This purpose of arguing has to do with the new information conveyed by the selection of the premises. Consider the valid argument: »Fred is an expert tax lawyer. Expert tax lawyers are wealthy. Therefore, Fred is wealthy«. Anyone who doubts that Fred is wealthy should doubt that Fred is an expert tax lawyer. Suppose the respondent has doubts about the conclusion. Still, the evidence backing the premise that Fred is an expert tax lawyer might be something »new« for her, so to speak (say, she bases her initial doubts on his modest life-style) and not undermined by the evidence offered in the argument. The argument is not question-begging for *this* respondent (Jackson, 1984: 34). This, in brief, offers a solution to the problem of all valid arguments being a *petitio*.

In order to understand the structure of *petitio* we have to go beyond the explicit premise/conclusion structure and consider the evidential support for the premises. It is just too simple to say that the very conclusion of an argument is fallaciously assumed in one of the premises or that the premises presume, openly or covertly, the very conclusion that is to be demonstrated. The argument begs the question for the addressee S if one of the premises, P, which is supported (according to the speaker/proponent) by evidence E, S who antecedently doubted the conclusion would adopt assumptions against the background of which the evidence E would not support P. And an argument is »begging the question proper« when one propounds »an argument such that any (sane) audience which was in doubt about the conclusion would have background beliefs relative to which the evidence provided by propounding the argument has no impact« (Jackson, 1984: 35). This might be the case with »the children of the gods« argument, since any rational audience who antecedently doubted the conclusion (»Achilles is the son of gods«) would have background beliefs relative to which the evidence provided (Achilles is trustworthy because gods are trustworthy) has no impact.

Or perhaps not? The notion of »sane« audience is notoriously elusive. There is a difference between saying that the evidence on display is no evidence for that *particular* audience (say the Tortoise) and saying that the evidence is no evidence for *any* (sane, rational, normal) audience. Would anyone who sanely doubted the conclusion have background beliefs relative to which (s)he would not regard all of the evidence implicitly offered for the premises (scripture? clergy?) as evidence? The way the dialogue was presented one might say that some new evidence does »pour« in, after all (the first-person testimony – »I am speaking of what took place in my own family«). This evidence is ineffective against the Tortoise and her Humean scepticism, but it might still carry some evidential weight for a different, not necessarily insane audience.

In any case, what makes an argument question-begging depends on argumentative features of the context in which it is proposed. There are rules against including certain premises (commitments) in the argumentative dialogue. But the *rationale* for such rules is ultimately epistemic. Let me test this proposal against what looks to be a very clear exposure of the conversational model. Late David Lewis in his recently published correspondence discusses van Inwagen's version of the consequence argument for the incompatibility of free will and determinism. Lewis adopts the compatibilist's position and objects to one of the premises in the argument which is »awfully close to an explicit denial of compatibilism.« And then he writes in a letter to van Inwagen (Lewis, 2020: 90):

> To beg the question isn't to argue from premises; to argue from premises that fail to be neutral with respect to the conclusion; to argue from premises that your opponent is free to deny; or to argue from premises that your opponent probably would deny. (It isn't a fallacy to argue, or to argue validly, or to argue with a free man, or to argue with a stubborn man.) The most promising account of begging the question supposes that there's a dialogue going on and doesn't directly apply to a monologue (or -graph): there's a status of being under challenge, there are ways for a participant to give something that status, and there's a rule against using challenged premises, the point of which rule presumably is to avoid going on and on in a dead-locked condition.

Two parties are involved in dialogue and one is trying to prove something to the other according to some procedural rules. The proponent presents a premise which would automatically come under challenge as soon as the conclusion was in dispute. There is a rule against using such a premise. What might the rule be? Do not use a contested premise if you want to avoid going on and on in a dead-locked condition. But why would the discussion go on and on? Well, the contested premise is supported (according to the proponent) by the evidence such that the respondent who doubts the conclusion has background beliefs relative to which this evidence has no impact. Perhaps a premise under challenge is very close to being just a reformulation of the conclusion or the premise is connected to the conclusion in an *epistemically* illegitimate way. The rule prohibits endlessly restating the old and ineffective evidence in new clothing.

Epistemic advancement is the name of the game: an argument uses the premises to provide evidence of a kind that gives the respondent a reason to accept the conclusion. I thus agree with Walton (2006: 238) that the failure to provide such reasons, a probative failure, is at the root of the fallacy of begging the question. But I think that a hybrid model, based on Jackson's combination of pragmatic and epistemic components provides the most plausible approach in analyzing the phenomena of begging the question. Not every valid argument begs the question, because we have to consider the role of potentially new evidence revealed by the proponent’s selection of premises. In most general terms in the case of *petitio* a statement (P) is made that presupposes or depends upon the point at issue (Q) (cf. Fogelin, 1987: 95). We can agree that reasoning is embedded in a conversation that represents a goal-directed dialogue (to establish whether Q is the case) that the questioner (proponent) and respondent are taking part in. The proponent should offer an argument with premises that provide evidence that supports the conclusion that the respondent doubts or disagrees with. And there is a rule against using certain premises. But the *rationale* for the rule should be spelled out in epistemic terms: when Q is in dispute a certain sort of evidence is discredited. Some new or extra evidence has to be introduced in the argumentation. The respondent must be presented with premises (and thereby the sources of evidence) she can accept, independently of her doubts about the conclusion (such that her doubts do not block the »flow« of evidence). One *could* say that begging the question is a violation of some general rule of dialogue. Yet the *rationale* for this rule is epistemic – sorry, this commitment is unacceptable because of its epistemic »corruption«, your premise depends on the conclusion in an epistemically illegitimate way. Proper (argumentative) conversation requires proper (epistemic) justification.

Begging the Question – Proper Justification or Proper Conversation?

Since Aristotle there are two main approaches in the explanation of begging the question (*petitio*): a dialectical mistake (an improper move in an argumentative dialogue) and an epistemic mistake. According to the latter begging the question is committed when the premises of an argument cannot be known independently of knowing the conclusion of the argument. Dialectical approaches use the notion of a commitment (acceptance, standpoint) and rules of dialogue as their basis. I propose a hybrid model, inspired by Jackson: the *rationale* for introducing commitments and rules is epistemic.

Keywords: argument, begging the question, epistemic model, dialectical model, commitments.

Literature

Carroll, L. 1895. »What the Tortoise said to Achilles.« *Mind* 4, 278–280.

Copi, I. M. and Cohen, C. 1991. *Introduction to Logic* 8th. New York: McMillan.

Copi, I. M., Cohen, C. and McMahon K. 2013. *Introduction to Logic 14*th. Harlow: Routledge.

Fogelin, R. 1987. *Understanding Arguments*. New York: Harcourt Brace Jovanovich.

Hazlett, A. 2006. »Epistemic Conceptions of Begging the Question.« *Erkenntnis* 65, 343–363

Hume, D. 1748 *et seq. An Enquiry Concerning Human Understanding*, Ed. Tom L. Beauchamp. New York: Oxford University Press, 2000.

Hurley, P. J. 2013. *A Concise Introduction to Logic* 12th. Cengage Learning.

Iacona, A. and Marconi, D. 2005. »*Petitio principii*: What's wrong?« *Facta Philosophica* 7, 19–34.

Jackson, F. 1984. »Petitio and the Purpose of Arguing.« *Pacific Philosophical Quarterly* 65, no. 1 (January 1984): 26–36.

Lewis, D. K., Beebee, H. and A. R. J. Fisher, eds. 2020. *Philosophical Letters of David K. Lewis: Volume 1: Causation, Modality, Ontology*. Oxford, New York: Oxford University Press.

Robinson, R. 1971. »Begging the Question.« *Analysis* 31, 113-117.

Sinnott-Armstrong, W. 1999. »Begging the Question.« *Australasian Journal of Philosophy* 77, 174–91.

van Eemeren, F. H. and Grootendorst, R. 2004. *A Systematic Theory of Argumentation. The pragma-dialectical approach*. Cambridge: Cambridge University Press.

Walton, D. N. 1991. Begging the Question: Circular Reasoning as a Tactic of Argumentation. New York, Greenwood Press.

Walton, D. N. 2006. »Epistemic and Dialectical Models of Begging the Question.« *Synthese* 152, 237–284.

Ilya Y. Bulov
Institute of Philosophy, Russian Academy of Sciences
e-mail: bulovilya@gmail.com

Concepts as Representations, as Senses, and as Abilities

Pojmi kot reprezentacije, kot smisli in kot sposobnosti

Obstajajo trije glavni pristopi k ontologiji konceptov: reprezentacionalizem, fregejanstvo in mešani pogled, ki združuje reprezentacionalizem in fregejanstvo. Reprezentacionalizem na koncepte gleda kot na mentalne reprezentacije, medtem ko fregejanstvo trdi, da so koncepti fregejevski smisli. Vsaka od teh razlag ima več prednosti. Kljub temu imajo ti pogledi poleg svojih prednosti tudi pomembne slabosti, ki bodo poudarjene v tem članku. V tem kontekstu predlagam alternativno stališče, ki ga imenujem »abilitizem«. Abilitizem je pristop k ontologiji konceptov, ki trdi, da so koncepti zmožnosti za usklajevanje skupin sposobnosti. Ta pristop rešuje vse probleme reprezentacionalizma, fregejanstva in mešanega pogleda. Abilitizem je tudi zelo primeren za razlago pluralizma konceptov in delnega lastništva konceptov. Vendar pa obstaja nekaj možnih ugovorov glede abilitizma, na katere bom odgovoril v zadnjih razdelkih.

Ključne besede: koncepti, mentalne reprezentacije, sposobnosti, ontologija konceptov, reprezentacionalizem.

Introduction

The ontology of concepts is an ancient topic. We have known about these discussions since the time of Plato, or perhaps even earlier. Yet there exist many contemporary philosophers who discuss this problem (Fodor, 1987; Laurence, Margolis, 2007; Peacocke, 1992; Zalta, 2001).

The question I want to begin with is »What are concepts?« Fortunately, philosophers and cognitive scientists have a conventional answer to this question. The conventional description of concepts was originally proposed by John Locke in his *Essay Concerning Human Understanding*. Locke characterizes concepts (or »ideas« in his terminology) as »materials of reason and knowledge« (Locke, 1690/1979: 104). In the contemporary literature on the subject, we find similar descriptions: »building blocks of thought« (Solomon, Medin, Lynch, 1999: 99), »constituents of thoughts« (Prinz, 2002: 2), »units of thought« (Carey, 2009: 5), etc. Unfortunately, this description is not very informative. We cannot deduce anything about the ontological aspect of concepts from it. Therefore, we cannot use this characterization to answer questions like »How do concepts exist?«, »Where do they exist?«, »What kind of existence is it?«, »Are they psychological entities or abstract entities? Or both?« etc. There are three main approaches to the ontology of concepts: representationalism, Fregeanism, and the mixed view. In this article, I will point out the drawbacks of these approaches. Then I will propose a less popular view, which I call »abilitism«, and show how it is better than the alternatives. Then I will answer some possible objections to the abilitist approach.

Representationalism

The first way to approach the problem of the ontology of concepts is to adopt representationalism. Representationalism (sometimes called the »psychological view«) is based on the Representational Theory of Mind, whose main theorist is Jerry Fodor (Fodor, 1975; Fodor, 1987). According to representationalism, concepts are mental representations (or mental symbols) in the language of thought. For example, the concept DOG[1] is a mental symbol that refers to a dog.[2] Representationalism is well suited to explain the compositional properties of concepts or the productivity of thought. Representationalists can easily extend their approach with the Language of Thought Hypothesis (LOTH)[3] (Fodor, 1975), which they can use to explain compositionality. If LOTH is true and the relation between a concept and a thought is similar to the relation between a word and a sentence, then we can explain how our minds are able to construct an infinite number of meaningful thoughts using combinatorial rules. In other words: If LOTH is true, we can explain the productivity of thought. For example, using just the three basic concepts DOG, LOOK, and CAT, we can construct an infinite number of thoughts:

[1] Concepts are usually capitalized in academic literature.

[2] However, the concrete details of this reference depend on the theory of reference we accept (e.g., the causal theory of reference describes it through the causal relations).

[3] The Language of Thought Hypothesis (LOTH) is the hypothesis that mental representation has a linguistic structure. According to LOTH, the process of combining mental representations into thought is similar to the process of building up a sentence from words. Additionally, the produced thought will be more or less the sum of its parts (concepts or other thoughts).

THE DOG IS LOOKING AT THE CAT.

THE CAT IS LOOKING AT THE DOG.

THE CAT IS LOOKING AT THE DOG LOOKING AT THE CAT.

THE CAT IS LOOKING AT THE DOG LOOKING AT THE CAT LOOKING AT THE DOG.

Etc.

However, in addition to the advantages, this view has significant disadvantages. Its main shortcoming is the idea of mental representation. Mental representation is a term that in theory refers to a natural phenomenon. However, this concept is so vague that we cannot really say what this natural phenomenon is. The only thing we can say for sure about mental representations is that they are *in the mind*. The plethora of different definitions of mental representation is a good illustration of this problem. For example, L. Roitblat describes mental representations as all internal changes caused by experience (Roitblat, 1982). At the same time, A. Newell claimed that mental representations are mental entities that designate facts of the world (Newell, 1980). As for the designation, Newell describes it as follows: »An entity *X* designate an entity *Y* relative to a process *P* if, when *P* takes *X* as input, its behavior depends on *Y*« (Newell, 1980: 156). Another definition of mental representations comes from S. Laurence and E. Margolis, who describe them as components of propositional attitudes (Laurence, Margolis, 2007: 563). Moreover, all of these definitions themselves have problems. Roitblat's definition is too broad, resulting in the term »mental representation« including entities such as psychological trauma. Newell's definition does not take into account that the behavior of concepts does not always depend on the facts of the world (the transformation of a concept may be caused by mental processes). The definition proposed by S. Laurence and E. Margolis is tautological. Thoughts, according to the mainstream definition, are propositional attitudes (Crane, 2016: 19). If so, then mental representations are constituents of thoughts. However, if concepts are constituents of thoughts, as we noted earlier, and mental representations are constituents of thoughts, then the statement »concepts are mental representations« is simply a tautology.[4]

As we can see, the term »mental representation« itself needs justification. Therefore, we should not use it, because describing an unclear term (»concept«) by another unclear term (»mental representation«) would be a bad strategy.

[4] This issue with the term »mental representation« was also noticed by Robert Cummins who characterize it as an umbrella term.

The second problem with representationalism is that it does not give us a good explanation of representations of certain kinds of concepts. Namely, for non-actual and very abstract concepts. For example, most of us certainly possess the concept ALLIGATOR WITH BIG EYES AND A SMALL INJURED TAIL.[5] We own it because we know what that concept means. When we hear the phrase »alligator with big eyes and a small injured tail« we know exactly what it means. We are also prepared to act on that knowledge. For example, we know that alligators are dangerous, and we should be careful if we want to help that alligator. But do we really have representations for all very specific concepts like the concept ALLIGATOR WITH BIG EYES AND A SMALL INJURED TAIL? I think that is very unlikely. Laurence and Margolis objected to this criticism. They had the objection that representationalism only captures actual concepts and therefore we do not actually possess these sophisticated concepts until we hear them (Laurence, Margolis, 2007: 577). Yet we usually know (or at least are familiar with) these concepts and we are able to act on this knowledge. If this is so, how should we account for it other than that we possess these concepts? Laurence and Margolis have not given us a satisfactory explanation. Therefore, I think that possessing these concepts is the only acceptable option.

Similarly, there is no good way to describe very abstract concepts like MATHEMATICAL ADDITION in terms of mental representations. What kind of representation should that be? Someone might guess that it could be the concrete example of a mathematical addition or a set of them (e.g. 3+4=7). However, we can easily imagine a person who remembers some examples of addition, but does not understand the general principle behind this mathematical operation.

I have mentioned the main problems with representationalism. Nevertheless, this view also faces other difficulties that I have not addressed here (e.g., the regress argument (Blackburn, 1984; Laurence, Margolis, 1997)).

Fregeanism

A different approach to the ontology of concepts, which I call Fregeanism (after Frege), was proposed by Christopher Peacocke (Peacocke, 1992). According to Fregeanism, concepts are Fregean senses, which means that they are abstract entities (like mathematical objects) whose existence is independent of the mind. According to Fregeanism, a concept can be described as a mode of representation of a referent (Peacocke, 1992: 3). For example, the concept DOG is a (purely abstract) mode of presentation of a dog. Peacocke also noted that we are capable of distinguishing concepts from one another on the basis of different possession

[5] This objection is similar to Dennet's argument against mental representations. The difference is that Dennet uses belief (not a concept) as an example (»Zebras don't wear overcoats«) (Dennett 1977: 104).

conditions. The possession condition is a set of logical queries that we have to cope with in order to possess the concept (Peacocke, 1992: 9). For example, we possess the concept CONJUNCTION by virtue of finding these logical queries persuasive: *p, q —> pCq; pCq —> p; pCq —>q.*

As for the perceptually based concepts, their possession conditions are only related to the perceptually based premises. For example, the possession condition for the concept RED would be the fact that we find the inquiry »this ball is red« persuasive, provided the ball is in a red region of our vision and we are in regular circumstances.

In addition to describing the conditions of possession, Peacocke also proposes the theory of determination. This theory, according to Peacocke, explains how possession conditions and the external world determine the semantic content of concepts (Peacock, 1992: 17). For example, the semantic content of the concept CONJUNCTION would be a determinate truth function. Thus, the determination theory would say that this truth function »makes transitions of the forms mentioned in the possession conditions truth-preserving under all assignments to their constituents *p* and *q*« (Peacock, 1992: 18).

As we can see, Fregeanism is a very sophisticated approach to the ontology of concepts. The Fregean approach avoids the main errors of representationalism. It does not use vague terms like »mental representation«. In contrast, Fregean sense is a fairly clear concept. Fregeanism can also explain the ontology of non-actual terms and concepts like MATHEMATICAL ADDITION. According to Fregeanism, they are simply abstract objects that are independent of our minds, and the mind grasps them only when they can be useful.

In spite of the fact that Fregeanism solves many of the problems of representationism, it also has its own drawbacks. The first disadvantage of Fregeanism is that it does not have enough explanatory power to deal with psychological aspects of concepts: categorization, concept acquisition (Carey, 2009), prototype effect (Rosch, 1973; 1978), etc. Possession theory and determination theory cannot explain such phenomena.

The second problem is that Fregeanism cannot explain how the senses, which are abstract objects, are related to the human mind and brain, which are part of the physical world. There are two aspects to this problem. The first and more general one is that abstract objects cannot have causal relations with physical processes if we accept physical causal closure (Kim, 1993). If this is so, we cannot explain the process of »grasping« concepts and other interactions between our psyche and concepts. The second and more specific aspect of this problem is that Fregeanism invokes the problem of modes of representation (Schiffer, 1990). This problem can be illustrated in the following way. Suppose that Peter Parker and Spider-Man are the names of the same person. Suppose further that Gwen Stacy knows her classmate Peter Parker and she has also heard from her friends about the exploits of Spider-Man. However, she does not know that Peter Parker is Spider-Man. If she finds out that Peter Parker is Spider-Man, that would be new information for her. In contrast, the information that Peter Parker is Peter Parker will not be new to her. The reason these situations are so different is that Gwen Stacy uses two modes of representation when thinking of the same person. Hence, we are tempted to think that presentation modes are Fregean senses (or at least related to them). But, as noted above, Fregean senses are abstract entities. Therefore, they cannot be mental processes or causally connected to mental processes. Moreover, as abstract objects themselves, Fregean senses can have multiple modes of presentation in a human mind.

We can try to solve this problem by saying that language provides us with modes of presentation. However, it is easy to object to the fact that even a creature that cannot acquire language can have multiple modes of presentation for the same object. For example, we can imagine the situation where a cat does not know that the person on whom that cat is sitting and the thing touching the cat's tail are the same object. However, if the cat turns around, it would possibly understand that.

The mixed view

There is also the mixed view proposed by Steven Laurence and Eric Margolis (Laurence, Margolis, 2007) and by Wayne Davis (Davis, 2008), which combines representationalism and Fregeanism. On the mixed view, types of concepts are senses, and tokens of concepts are mental representations. For example, the concept type DOG is an abstract mode of representation for a dog, and the token is a mental symbol corresponding to a dog. According to Laurence and Margolis, the mixed view is supposed to solve the problem of presentation modes. They see the solution in what they call »mental orthography«, which is a mental copy of the formal properties of concept types (i.e. Fregean senses). Laurens and Margolis suggest that mental orthography can be understood in two ways. The first interpretation states that mental orthography is a property of concepts that enables them to function as representational forms and express Fregean concepts. According to the second interpretation, which takes up Fodor's conception (Fodor, 1998), mental orthography is a fundamental property of the human mind that allows it to use concepts as logical entities. According to this interpretation, mental orthography is itself a mode of presentation. Moreover, this way of thinking about mental orthography suggests that concepts are the basic units of thought and lack internal structure.

The first problem with this solution is that the mental orthography still sheds no light on the nature of the relation between Fregean senses and modes of presentation. Is it a causal relation? If so, then the physical causal closure is violated. If not, then why is there an isomorphism between them? The second problem is that the mental orthography hypothesis ignores the type/token relation between Fregean senses and mental representations proposed by Laurence and Margolis. If modes of representation are copies of Fregean senses in our minds, then the former are tokens of the latter. However, the term »mode of presentation« is not the equivalent of the term »mental representation« (mode of presentation is just an aspect or function of a mental representation). If this is so, we cannot say that there is a type/token relation between mental representation and Fregean sense, and the mixed view is just one version of Fregeanism. The third problem is that the mixed view uses the term »mental representation« and therefore inherits some of the problems of representationalism (especially the problem with the vagueness of the term »mental representation«).

Abilitism

There is also a more Aristotelian way of thinking about concepts, which I will defend here. I will call this view »abilitism«. In *De Anima*, Aristotle writes, »It follows that the soul is analogous to the hand«. This way of thinking, as I will show, can be effectively adapted to modern understandings of concepts. Another source of inspiration for this view is the work of Michael Dummett (Dummet, 1993) and Anthony Kenny (Kenny, 2010), in which they propose ideas similar to the abilitist approach.

According to abilitism, concepts are cognitive abilities that generally coordinate groups of other abilities (cognitive and non-cognitive). First, we need to clarify what a cognitive ability is. Cognitive ability is the ability that our psyche uses to accomplish certain tasks such as categorization, object recognition, etc. (Benjafield et al., 2010). I believe that the term »ability« itself is quite intuitive, but if necessary, as Kenny noted, we can describe »ability« using the Aristotelian distinction of *potential* and *actual*, where an ability is a potential state and a performance of that ability is an actual state (Kenny, 2010). A theoretical model of the ability to master a particular set of sufficient abilities can be called a type of concept. The token of a concept would be the concrete ability that exists in the mind of a particular person. Thus, the bearer of a token of the concept would be a human or non-human being who possesses a sufficient group of abilities and has the ability to manage them. The owner of the ability (i.e., the bearer of a token of a concept) is distinct from the ability. The ability is also different from its execution. For example, the ability to distinguish the red color is different from the concrete action in which a person uses this ability to select the red napkin. We should also distinguish between the ability and the vehicle of the ability. The vehicle of the ability is a concrete physical implementation of the ability, which is usually the neural networks and their interactions.

Now we can illustrate the abilitist approach using the concept DOG as a toy example. The concept DOG is the ability to coordinate abilities such as 1) the ability to distinguish dogs from other objects; 2) the ability to use the word »dog«; 3) the ability to compare and combine the image of a dog with other images; 4) the ability to compare and combine the language unit »dog« with other language units; 5) the ability to discover relationships between DOG and other concepts (e.g. PUPPY or HOUND). The bearer of a token of the concept DOG in this example may be a person who has at least the abilities (1), (3), and (5) from the list above and the ability to coordinate these abilities with each other. The performance of the concept DOG would be the use of some of the listed skills. Finally, the vehicle of the concepts DOG would be the neural mechanisms that the person uses to perform these abilities.

Anthony Kenny defended a different version of the abilitist approach in his article. He wrote that the concept of *X* is just the ability to use the word »*X*« (Kenny, 2010). This view oversimplifies the cognitive processes involved in possessing and acquiring a concept, because possessing the ability to use the word »*X*« is far from sufficient for possessing a concept *X*. The ability to use the word »*X*« is not enough to possess a concept. It is only *one of the abilities* a person must master in order to acquire the concept. For example, imagine a person *Z* who can use the word »tomtit«. She knows how to pronounce this word, and she uses it effectively when *Z* is talking to her friends about their favorite animals. During these discussions, *Z* can say that she admires tomtits and that she especially likes their black heads and the yellow band across their chests (real tomtits usually have these features) and never say anything else about them. However, *Z* thinks that tomtits are mammals and not birds. Based on this knowledge, should we conclude that *Z* owns the concept TOMTIT? I don't think so. Even though *Z* may actually use the word »tomtit« during the conversation, she does not actually possess the concept. In order to possess it, she must be able to adequately categorize tomtits (in our situation, she obviously cannot). This example shows that possessing the ability to use the word »*X*« is not nearly enough to possess concept *X*.

One of the main advantages of abilitism is that it can effectively explain concept pluralism, which saves the term »concept«. Based on some experiments in cognitive science, we can conclude that different concepts can have different structures. In addition, we can conclude that different concept carriers (or even the same concept carrier in different ages) can have the same concept but with different structures. I call »concept pluralism« the theory that states that the structure of the same concept can vary from person to person and can change over time in the person's mind. One of the examples of experiments I will mention is the study by Susan Gellman, in which she points out differences between categorization mechanisms we use when we process concepts of artifacts and categorization mechanisms we use when we process concepts of animal kinds (Gellman, 1988; Gellman, Markman, 1986). According to abilitism, different concepts tend to have different sets of abilities depending on the bearer of a concept, the type of concept (a concept-ideal, a concept of an abstract thing, a concept of a natural kind, etc.), and other variables. Therefore, there is no difficulty in explaining these phenomena. For example, in the experiment mentioned above, we can assume that people tend to have different sets of abilities for concepts of artifacts and for concepts of animal kinds.

Another example is the experiment where Susan Carey found that children tend to associate the concept ANIMAL with the concept PERSON, while adults do not associate these concepts (Carey, 1985; Carey, Johnson, 2000). Abilitism can explain these observations. According to abilitism, a concept carrier may sometimes lose a skill associated with that concept or acquire a new skill during the course of life. Of course, we can ask whether these two or more states of the set of abilities separated by time are the same concept. This question has the same structure as the problem of personal identity. Therefore, we can adapt some of the answers to this problem. For example, we can adapt narrative theory (Schechtman, 2014), which means that each concept has a historical narrative that maintains its unity.

Abilitism accounts for conceptual pluralism while preserving the term »concept«. The abilitist approach unifies various phenomena under the term »concept« and thus does not make this term too broad. In contrast, representationalism does not have this advantage. According to the representationalist view, we can explain the conceptual change discussed above in two ways. First, we can assume that the change of concept ANIMAL requires the replacement of the mental representation *animal#1* by another mental representation – *animal#2*. Therefore, we must assume that there are at least two concepts of an animal: ANIMAL #1 and ANIMAL #2. The problem with this explanation is that we must assume the existence of a very large number (potentially infinite) of similar concepts (ANIMAL #1, ANIMAL #2, ANIMAL #3, ..., ANIMAL #N etc.). Another way to explain this conceptual shift with representationalism is to expand the notion of »mental representation«, which is already too broad, as I have shown before. Therefore, this is also a poor explanation.

The second main advantage of abilitism is that »cognitive ability« is a clear concept, unlike the concept »mental representation«. We showed earlier how it can be clarified using the dictionary definition and the Aristotelian distinction between potential and actual.

The third advantage is that, unlike Fregeanism, the abilitist approach has no commitment to the problem of abstract entities. Abilitism simply does not use such entities in its framework.

The fourth point in favor of abilitism is that it can explain partial ownership of a concept better than the alternatives. For some concepts, we can say that we partially own them (e.g., the example with *Z* and the concept TOMTIT). Abilitism explains the gradation of possession of a concept by assuming that a person can initially acquire some abilities as for a concept and does not acquire other abilities with regard to that concept. In contrast, representationalism and Fregeanism do not have as good an explanation for this phenomenon.

Finally, abilitism has a good explanation for concepts such as ALLIGATOR WITH BIG EYES AND A SMALL INJURY TAIL and MATHEMATICAL ADDITION. According to abilitism, these are simply cognitive abilities that manage sets of other cognitive abilities (e.g., the ability to distinguish the species of an alligator, the ability to add numbers, the ability to talk about both, and so on).

As we have just seen, abilitism has many strong points compared to other views. Therefore, I believe that abilitism is the best explanatory model for the problem of the ontology of concepts. However, there are some possible objections to abilitism. I will discuss them below.

Objection#1: Abilitism does not explain compositionality

The first possible objection to abilitism is that it does not explain compositionality and therefore cannot explain the productivity of thought. At first sight, this objection seems justified, since faculties do not have the same combinatorial properties as representations and Fregean senses. Indeed, how do we combine concepts according to abilitism? I think there are two possible explanations, which are not necessarily mutually exclusive. The first possibility is to say that the combination of terms is the combination of different capacities. For example, imagine the situation where person *X* has two concepts: *A* and *D*. Concept *A* has the set of abilities {*a,b,c*} and *D* has the other set {*d,e,f*}. According to abilitism, when *X* combines concepts *A* and *D* to *AD*, she actually combines the sets of abilities {*a,b,c*} and {*d,e,f*} to {*a,b,c,d,e,f*}. However, the combination process is usually not so straightforward. In a standard situation, the context also affects the process of combination. Depending on the context and properties of the concepts, some capabilities may be dropped or/and some additional capabilities may be added. For example, {*a,b,c*} and {*d,e,f*} may be combined to form {*a,d,f,g*}. This context dependency was highlighted by Zachary Estes and Sam Glucksberg when they showed that one concept usually assigns its properties to another during the combination process (Estes, Glucksberg, 2000). For example, when we combine the concept SHARK and the concept LAWYER, we attribute the most salient properties (or the abilities to distinguish them, according to the abilitist view) of the concept SHARK (»predatory«, »aggressive«, »vicious«) to the concept LAWYER, which has some relevant dimensions for attribution (»temperament«, »competence«, »cost«). It is obvious that these traits and dimensions strongly depend on the context in which a person is located.

Another solution to the combination problem is to propose that we have a distinct ability to combine different mental entities (e.g., visual images, linguistic entities, etc.). This solution does not necessarily rule out the first one. We can have the general ability to combine concepts, and the set of specific abilities to combine particular aspects of concepts. I think the question of which solution is better and whether they are compatible is a matter of future research.

Objection#2: Abilitism does not explain knowing-that

Another possible objection to abilitism is that it does not explain knowledge-that. Knowledge-that is knowledge about a fact (e.g., knowledge about a weather report). In contrast, knowledge-how presents itself in our minds in the form of an ability (e.g., the ability to swim) (Ryle, 1945). G. Ryle uses this distinction between knowing-that and knowing-how to show that a person normally has two different faculties of the mind (which can nevertheless be used simultaneously). Obviously, abilitism suggests that all concepts should be classified as knowing-how. The point of the objection is that if some thoughts can be classified as knowledge-that, and all thoughts consist of concepts, then we need a good explanation of how knowledge-how converts to knowledge-that at the level of thought.

There are two possible solutions. The first is to accept a radical anti-intellectualism according to which all knowledge can be expressed in (and eventually reduced to) concepts of knowing-how (Hetherington, 2011). If all knowledge can be explained in terms of knowledge-how, then there is no need to explain the compatibility between abilitism and knowledge-that. However, radical anti-intellectualism is a rather controversial approach (Adams, 2009), so it needs justification itself. An analysis of all the advantages and disadvantages of this position is beyond the scope of this article. Here, I would only like to highlight that radical anti-intellectualism can possibly be a solution to the aforementioned problem.

Another solution is to say that knowledge-that is mere modal imagery (visual imagery, auditory imagery, etc.) that is normally handled with specific capacities associated with those images. If so, these modal images can be described as knowledge-that, and we need not abandon abilitism, since modal images are not concepts. This solution seems preferable. However, the first solution can also be used if it turns out that radical anti-intellectualism is correct.

Conclusion

As has been shown, there are three main approaches to the ontology of concepts: representationalism, Fregeanism, and the mixed view. All of them have significant drawbacks: the problem of describing mental representations, the problem of abstract entities, the problem of non-actual concepts, etc. Abilitism, which proposes that concepts are cognitive abilities, is a viable alternative to these approaches. This view solves all the problems mentioned above. Abilitism is also well suited to explaining concept pluralism and partial ownership of concepts. There are some possible objections to abilitism, but as has been shown, they are easily answered.

Concepts as Representations, as Senses, and as Abilities

There are three main approaches to the ontology of concepts: representationalism, Fregeanism, and the mixed view combining representationalism and Fregeanism. The representationalist view views concepts as mental representations, while Fregeanism states that concepts are Fregean senses. Each of these accounts has several advantages. Nevertheless, along with their advantages, these views also have significant disadvantages, which will be highlighted in this article. In this context, I propose an alternative view, which I call »abilitism«. Abilitism is an approach to the ontology of concepts that holds that concepts are capabilities for coordinating groups of capabilities. This approach solves all the problems of representationalism, Fregeanism, and the mixed view. Abilitism is also well suited to explain concept pluralism and partial ownership of concepts. However, there are some possible objections to abilitism, which I will answer in the final sections.

Keywords: concepts, mental representations, abilities, ontology of concepts, representationalism.

References

Adams, M (2009). »Empirical evidence and the knowledge-that/knowledge-how distinction«. *Synthese,* 2009, 1, pp. 97–114.

Benjafield, J., Smilek, D., Kingstone, A. (2010). *Cognition.* 4 ed. New York: Oxford University Press.

Blackburn, S. (1984). *Spreading the Word*. Oxford: Oxford University Press.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carey, S. (2009). *The Origin of Concepts.* Oxford: Oxford University Press.

Carey, S., Johnson, S. (2000) »Metarepresentations and conceptual change: Evidence from Williams syndrome«. In Sperber D. (ed.), *Metarepresentation: A multidisciplinary perspective,* New York: Oxford University Press, pp. 225–264.

Cummins, R. (1989). *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Davis, W. (2003). *Meaning, Expression, and Thought*. Cambridge, MA: Cambridge University Press.

Dennett, D. (1977). »A Cure for the Common Code«. In Dennett, D. (ed.), *Brainstorms*, Cambridge, MA: MIT Press, pp. 90–108.

Dummett, M. (1993). »What Do I Know When I Know a Language?«. In Dummett, M. (ed.), *Seas of Language*, pp. 94–105.

Estes, Z., Glucksberg, S. (2000). »Interactive property attribution in concept combination«. *Memory & Cognition*, 28, pp. 28–34.

Fodor, J. (1987). Psychosemantics: The Problem of Meaning in the Philosophy of Mind. Cambridge, MA: MIT Press.

Fodor, J. (1998) *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford University Press.

Frege, G. (1948) »Sense and Reference«. *The Philosophical Review,* 3, pp. 209–230.

Gelman, S. (1988). »The development of induction within natural kinds and artifacts categories«. *Cognitive Psychology*, 1, pp. 65–95.

Gelman, S., Markman, E. (1986). »Categories and induction in young children«. *Cognition,* 3, pp. 183–209.

Hetherington, S. (2011). *How to Know: A Practicalist Conception of Knowledge*. Oxford: Wiley-Blackwell.

Kenny, A. (2010) »Concepts, Brains, and Behaviour«. *Grazer Philosophische Studien,* 1, pp. 105–113.

Kim J. (1993). Supervenience and Mind: Selected Philosophical Essays. Cambridge University Press.

Laurence, S., Margolis, E. (1997). »Regress Arguments Against the Language of Thought«. *Analysis,* 1, pp. 60–66.

Laurence, S., Margolis, E. (1999). »Concepts and Cognitive Science«. In E. Margolis, S. Laurence (ed.), *Concepts: Core Readings,* Cambridge, MA: MIT Press, pp. 3–81.

Laurence, S., Margolis, E. (2007): »The Ontology of Concepts – Abstract Objects or Mental Representations ?«. *Noûs*, 4, pp. 561–593.

Locke, J. (1690). *An Essay Concerning Human Understanding*. P. H. Nidditch, ed. Oxford: Oxford University Press, 1979.

Newell, A. (1980). »Physical symbol systems«. *Cognitive Science*, 2, pp. 135–183.

Peacocke, C. (1992). *A Study of Concepts.* Cambridge. MA: MIT Press.

Prinz, J. (2002) Furnishing the Mind: Concepts and Their Perceptual Basis. Cambridge. MA.: MIT Press.

Roitblat, L. (1982). »The meaning of representation in animal memory«. *Behavioral and Brain Sciences,* 3, pp. 353–406.

Rosch, E. (1973). »Natural categories«. *Cognitive Psychology,* 3, pp. 328–350.

Rosch E. (1978). »Principles of Categorization«. In Rosch, E., Lloyd, B. (ed.), *Cognition and Categorization*, Hillsdale, NJ: Lawrence Erlbaum Associates, strr. 27–48.

Ryle, G. (1945). »Knowing How and Knowing That«. *Proceedings of the Aristotelian Society*, 1, pp. 1–16.

Schechtman, M. (2014). Staying alive: personal identity, practical concerns, and the unity of a life. Oxford: Oxford University Press.

Schiffer S. (1990). »The Mode-of-Presentation Problem«. In Anderson, C. and Owens, J. (ed.), *Propositional Attitudes: The Role of Content in Logic, Language, and Mind,* Stanford: CSLI Publications, pp. 249–268.

Maria Beatrice Buonaguidi
King's College London, Velika Britanija

The nature of ψ: Epistemic and ontic views of the wave function

Narava ψ-ja: epistemični in ontični pogled na valovno funkcijo

Dilema o pomenu valovne funkcije je morda ena najbolj kontroverznih in spregledanih v sodobni fiziki. V znanstveni skupnosti obstaja široko nesoglasje glede tega, ali je treba na valovno funkcijo gledati kot na resnični fizični objekt ali pa ustreza stanju nepopolnega znanja o osnovnem sistemu, na primer gostoti verjetnosti. Ti pogledi se imenujejo ψ-ontični in ψ-epistemični. Cilj tega članka je dati jasen vpogled v problem in značilnosti obeh pogledov. Po nekaj razjasnitvenih uvodnih predstavah bom podala pregled, kako so to vprašanje zaznali in interpretirali pionirji kvantne teorije, kot sta Bohr in Einstein, nato pa nadaljujem z modernejšimi časi in imeni, kot so Bell, Kochen in Specker. Nato se bom osredotočila na sodobno literaturo in poskušala izčrpno predstaviti prednosti in slabosti vsakega pogleda. Na koncu bom razpravljala o nekaterih vprašanjih, ki nam jih postavlja vprašanje valovne funkcije, na primer o našem odnosu do resničnosti.

Ključne besede: kvantna mehanika, valovna funkcija, znanstveni realizem, ontologija.

1.Introduction

What *is* the wave function? We all know that it forms the core of the quantum mechanical formalism, that its dynamics are specified by the Schroedinger equation, and that Born's rule relates it to the probability of measuring a certain state or observable quantity. But should we regard them as something real, »part of the furniture of the world« (Callender, 2015), or rather as a representation of an incomplete state of knowledge about an underlying entity? The question of the ontological status of the wave function arises partly from wave-particle duality (Callender, 2015; Norsen, 2010), and partly from the measurement problem (Gao, 2017).

The wave function is the equation for »matter waves« proposed by the de Broglie hypothesis, but it lives in 3N-dimensional configuration space, while particles live in 3D space: are the latter emergent from the former, or are they two ontologically distinct entities? Is there a physical entity that corresponds to the wave, or is it just a mathematical tool? What do entanglement and collapse in projective measurement tell us about the structure of reality? The wave function appears to be intrinsically different from the wave equations for electromagnetic radiation because of the space in which it propagates; this leaves open whether or not the wave behavior of matter should be considered as emanating from a physical entity such as a field.

These questions have been part of the foundations of quantum mechanics since its birth; they are essential for us to understand what does and does not belong to the ontology of the world. The theory seems to suggest the existence of an underlying physical reality (Gao, 2017), but what is it? And most importantly, what does the wave function mean in all this?

1.1 Mathematical formalism and physical entities

There are some caveats to be made when exploring this topic, as Maudlin points out in (Albert and Ney, 2013). We should keep in mind that the wave function itself is a mathematical object, let us call it ψ, and we should distinguish it from the quantum state, i.e., the entity it represents, λ. Our efforts to interpret ψ should not reify it and confuse it with λ, but try to investigate what is the relation between the two.

It is an open question whether the quantum state is uniquely determined by the wave function or whether there is something else in the theory, and whether the wave function directly and literally represents the quantum state or our state of knowledge about it. Does ψ supervene over the quantum state or not? The answers to these questions will determine our stance on the ontology and status of ψ. Unfortunately, the task of constructing an ontology starting from a theory so heavily laden with mathematical formalism complicates the distinction between what is part of the mathematical description and what is a physical entity, and exposes scientists to the risk of naively reading the metaphysics of a theory from its formal structure (Ney, 2015). (Norsen, 2010) notes that »mathematically equivalent formulations can have radically different implications in regard to ontology and causality«, and points to this risk. In recent years, however, the scientific community has embraced the idea that the question of the status of the wave function should be approached rigorously, using both mathematical and heuristic arguments, i.e. a thorough engagement with the formalism to find out what lies behind it (Leifer, 2014a).

Maudlin (Albert and Ney, 2013) paints a clear picture, albeit inevitably biased by his view, as Ney points out in (Ney, 2019), of what we should be looking for when determining the ontology of a theory. I think it is worth considering it as a kind of terminological clarification for what follows.

1. Primary observables: these are the phenomena, that is, what we directly observe as measurement outcomes. The choice of primary observables is itself a

theoretical act, and thus a stance, because they obviously do not correspond to the ontology of quantum mechanics. If we are realists, there are objective facts about the state of the world that are independent of observation; they may be different from phenomena, i.e. what we perceive about reality in measurements may not be a complete picture of it and we may not be able to grasp the underlying reality, but that underlying reality exists independently of our observations.

2. Ontology: these are the things we take to be physically real, which are supposed to explain the primary observations, following the laws. An epistemic distinction has been made between primary and secondary ontology: The primary observables supervene on the state of the primary ontology, but not on that of the secondary ontology, i.e. if we change the latter, the data remain the same, because the secondary ontology does not directly affect the behavior of the phenomena, but is rather theoretical in nature, whereas a change in the primary ontology affects a change in the observables. The secondary ontology is something that we postulate as real, but whose existence is not directly related to the outcome of our measurements. The wave function, if we consider it to be part of the ontology, will most likely be part of the secondary ontology, while other entities, such as particles, form the primary ontology. It is desirable that the elements that form the ontology of a theory are local, since we assume that the constituents of reality obey locality (Norsen, 2010). This follows from the fact that we require a theory to obey causality. Nonlocal causality would imply influence from a distance, thus generating nondeterminism.

3. Laws: The behavior of primary and secondary ontology obeys dynamical laws. It is unclear whether we should conceive of laws as governing this behavior – this view is called *nomological primitivism* (Ney, 2019) – or rather as a mathematical description that scientists have devised to fit the spontaneous behavior of ontology (*Humean view*). Another interpretation of the status of laws is that they are spontaneous dispositions of ontology: In this case, they are not prescriptive, but they are related to properties of ontology (*dispositionalism*; see (Dorato, 2014; Ney, 2019; Dorato, 2015)).

1.2 Epistemic and ontic

The answers that physicists and philosophers have given to these questions initially suggest two views: ψ complete views and theories of hidden variables. The first view assumes that the wave function is »all there is«, i.e. the only element of our ontology (Goldstein, 2011). The best known interpretation of this kind is the Many-Worlds or Everettian view.

In this review, however, I will focus on the second view, according to which the ontology is not exhausted in the wave function, but is complemented by other elements, mainly due to the results of the 1935 EPR paper (Einstein, Podolsky and Rosen, 1935), discussed in section 2.2. Within this framework, the wave function can be considered as part of the ontology or not; in other words, it can be considered as indicating a real physical entity or as a simple tool that allows us to derive some information about the underlying reality (Harrigan and Spekkens, 2010).

The first view is called ψ-ontic, the second ψ-epistemic. As (Harrigan and Spekkens, 2010) put it, »In ψ-ontic models, distinct quantum states correspond to disjoint probability distributions over the space of ontic states, whereas in ψ-epistemic models, there exist distinct quantum states that correspond to overlapping probability distributions«. An epistemic model allows a quantum state, λ_x, to be in support of two distinct probability distributions that overlap in the region occupied by it, $|\psi_\alpha\rangle$ and $|\psi_\beta\rangle$. In contrast, there is no such state in an ontic model: disjoint probability distributions correspond to different ontic states.

Formally, a hidden-variable theory is ψ-ontic if, for arbitrary preparation procedures P_ϕ, P_ψ, associated with different wave functions φ and ψ, the probabilities of obtaining a quantum state λ,

$$p(\lambda \mid P_\phi) p(\lambda \mid P_\psi) = 0$$

and epistemic otherwise (Harrigan and Spekkens, 2010). This means that in an ontic model the quantum state supervenes on the wave function, while in an epistemic model this is not the case.

Moreover, we can distinguish between realist and antirealist epistemic views: The former claim that although there is an underlying ψ, the wave function does not describe it directly, but rather our state of knowledge about it, which is considered incomplete and probabilistic. The latter, on the other hand, states that we cannot assume that there is a ψ-state that is different from ψ, blurring the distinction between the state and the wave function. This article will focus more on the realist views than on the antirealist or instrumentalist views.

2. Historical overview

In the last chapter I outlined the problem at hand and clarified the distinction between epistemic and ontic views of the wave function. Now it is worth looking at the debate from a historical perspective, since the question of the nature of the wave function has long been overlooked. This was probably because the orthodox Copenhagen interpretation was fundamentally instrumentalist and antirealist about ψ, i.e. a »shut up and calculate!« attitude – to quote David Mermin – prevailed (Dorato, 2015). The wave function was perceived as epistemic and subjective, only as an operational tool, and its meaning was not explored. However, this canonical picture is partly due to an oversimplification of Bohr's views (Halvorson, 2019; Ladyman, 2019), and since Einstein there have been attempts to investigate the topic further, to find limitations and difficulties for epistemic models, and to make new ontic proposals.

2.1 Bohr and the Copenhagen interpretation

The Copenhagen interpretation, led in particular by Bohr et al[1], is often simplistically presented as alternatively antirealistic, operationalist, or instrumentalist. It is commonly seen as an interpretation of the wave function as an accounting device, a computational tool representing probabilistic information about the results of measurements, according to Born's rule. Although this is consistent with some of Bohr's views, the popular picture has retained only its most extreme features.

(Bohr, 1948) highlights the difficulty of assigning physical properties to the states represented by the quantum mechanical formalism; for this reason he gives ψ a symbolic interpretation and uses it exclusively as a tool for deriving predictions. The motivation for this is not an antrealist or skeptical view of physical reality, but derives from the highly abstract nature of the wave function, which makes it difficult to reconduce it to any physical concept.

As noted in particular in (Ladyman, 2019), Bohr explicitly acknowledged the existence of particles, e.g. in his Nobel Prize lecture, and therefore cannot be seen as rejecting the idea of an objective, observer-independent physical reality. However, the quantum mechanical formalism is so different from our classical perception that it would be risky to interpret it realistically: We can only observe phenomena and make probabilistic predictions about them.

Therefore, the concept of causality, which is crucial for scientific explanation, has to be replaced by the concept of complementarity, i.e. by attributing classical features to quantum description, in order to try to cope with the fact that we are unable to deal with quantum mechanical concepts because of the way science and our perception of the world have developed. As Bohr puts it, this problem of interpretation has nothing to do with metaphysics, but rather with language: our language as scientists is so biased with classical notions that it cannot picture ontology at the quantum level and therefore must use the concept of complementarity to try to make sense of it.

In any case, we are denied a full understanding or even observation of what is real at the quantum mechanical level, and the wave function is just a tool cast in a language that mimics the classical wave equation, allowing us to derive predictions. As (MacKinnon, 2012) summarizes, the Bohr interpretation is minimal rather than antirealistic, sufficient to make predictions about experiments without in-

[1] Described in MacKinnon, 2012.

consistency, and thus it holds that the wave function is an epistemically complete description of reality. ψ is then not part of the ontology, but knowledge of the ontology is closed to us, so the wave function provides us with all the information we need and care about in the system.

2.2 Einstein and the EPR paper

Einstein was a staunch critic of Copenhagen Quantum Mechanics, as he hoped for a new theory that could restore realism and determinism and that would make the probabilistic treatment of ψ only a matter of epistemic imprecision rather than an intrinsic limit. In other words, he hoped to establish a realist ontology given over to the minimal interpretation of Bohr. He was convinced that the wave function was an ontologically incomplete description, since it could only describe ensembles of systems (Fine, 1993).

Fine points out that Einstein's position on quantum mechanics was never very clearly stated and that we should take into account that his view is skewed by the desire to push the scientific community towards the new theory mentioned above. However, we can characterize his view as instrumentalist, subjective, and a proponent of hidden variables. He holds that ψ is an ensemble description because it describes only the statistics for measurement outcomes, according to Born's rule: it is thus subjective because it describes only our knowledge of the state of the system, and collapse corresponds to information updating. Thus, his epistemic view of the nature of the wave function must be considered negative.

His proposal for a model with hidden variables, together with an argument for the incompleteness of ψ, can be seen in the EPR paper of 1935 (Einstein, Podolsky and Rosen, 1935). Even though the paper was written almost entirely by Podolsky and somewhat pushes and modifies Einstein's views, as noted in (Harrigan and Spekkens, 2010; Fine, 1993), I will still consider it as the first example of a formal proof to investigate the meaning of the wave function.

The EPR paper argues for the ontological incompleteness of ψ by invoking a kind of »reality criterion« (Fine, 1993 which states that »If, without in any way disturbing a system, we can predict with certainty [...] the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity« (Einstein, Podolsky and Rosen, 1935).

It follows that if we consider observables described by two non-commuting operators, either the wave function description is not complete, or two non-commuting quantities cannot have simultaneous reality. To show that the first possibility holds, the paper considers two systems that are separated after a certain time of interaction; the state of one system cannot be determined exactly because the two wave functions are superimposed. Different measurements can be made on the first system: Thus, the wave function of the combined system can be expressed in different terms.

$$\Psi(x_1, x_2) = \sum_{n=1}^{\infty} (\psi_n(x_1) u_n(x_2))$$

$$\Psi(x_1, x_2) = \sum_{n=1}^{\infty} (\phi_n(x_1) v_n(x_2))$$

The paper then states that »As a consequence of two different measurements performed upon the first system, the second system may be left in states with two different wave functions«, although no objective change can occur on the second system since the systems no longer interact. This is just to establish the many-to-one correspondence of wave functions to reality. The two states that the second system then still has can be eigenfunctions of non-commuting operators. When we make a measurement on the second system, we can only determine one of the two eigenvalues at a time. However, when measuring on the first system, both quantities seem to be reality at the same time. Therefore, the description by the state function cannot be complete.

The paper then explicitly calls for theories of hidden variables as the only way to give the wave function a realistic ontological status. Therefore, the wave description of matter alone is insufficient to provide a complete prediction of reality.

2.3 No-go theorems

After the EPR paper, many other papers appeared in the context of hidden variable models that attempted to determine the meaning of the wave function in a formal way. These use quantum mechanical toy models, called *ontological models*, to derive necessary results about the state of ψ using formal and rigorous mathematical arguments, following the method used in the EPR paper (Leifer, 2014a). As explained by (Leifer, 2014b), ontological models are »attempts at explaining quantum predictions in terms of real physical properties (λ-ontic states) existing independently of the experimenter«.

The first notable no-go theorem for hidden variables is Bell's theorem (Bell, 1966), which challenges the refutation of the existence of hidden variables in quantum mechanics by Von Neumann. Bell goes further than EPR and highlights the fact that hidden variables, even if they exist, cannot appear in a simple measurement, as this would render quantum mechanics observationally inadequate, thus emphasizing the sound but incomplete character of the state function description. For Bell, »The question at issue is whether the quantum mechanical states can be regarded as ensembles of states further specified by additional variables, such that given values of these variables together with the state vector determine precisely the results of individual measurements« (Bell, 1966). The paper refutes the assumptions in the proof of Von Neumann and thus establishes the necessary existence of hidden variables and, most importantly, shows that models with hidden variables represent non-local and non-separable trajectories, which resolves the EPR paradox by showing a spacelike interaction between systems and decisively challenges the concept of space-time and causality in quantum mechanics. Systems are thus instantaneously affected by distant systems, producing an undesirable non-local causality. The results of Bell's work can then be summarized by a paradox: Either hidden variables must exist, and therefore ψ must be supplemented by an additional ontology, but the wave function is shown to behave nonlocally, or the wave function is complete, which is not the case by the EPR work, but local.

As (Harrigan and Spekkens, 2010) show, Bell's theorem is only needed for ontic models with hidden variables to rule out locality, since epistemic models are intrinsically nonlocal. In other words, as noted by (Norsen, 2010), for any theory in which the wave function has beable status, it is necessarily a non-local beable.

So far, then, the no-go theorems suggest that the most accurate interpretation of the wave function should be a ψ-incomplete ontic view in which the ontology is instantiated by some hidden variables that could be local beables and a nonlocally behaving wave function that has real physical existence in some way. We will see that hidden variables and nonlocality are not the only constraints.

The 1967 Kochen-Specker paper, which deals with the proof of the existence of hidden variables in quantum mechanics, shows, by providing an epistemic ontological model in 2 dimensions, that for any Hilbert space with a number of dimensions d≥3, deterministic ψ-epistemic theories are necessarily measurement contextual (Kochen and Specker, 1967). (Aaronson et al. 2013) further show that epistemic theories in 3-dimensional Hilbert space and above cannot satisfy symmetry under unitary transformations and maximal nontriviality, i.e., overlap of probability distributions over non-orthogonal states.

Finally, (Pusey, Barrett, and Rudolph, 2012) shows the inconsistency of nontrivial ψ-epistemic theories: If we assume a separability condition, the wave function must be ontic.

Thus, it seems that all no-go theorems point in the direction of a ψ-ontic theory of hidden variables, which may be mathematically consistent and correctly reproduce the quantum mechanical predictions, restricting the possibility of finding nontrivial and accurate ψ-epistemic models. However, we will see in the next section that there are problems with the ontic views that are best solved by adopting an epistemic perspective. The formal strategy of no-go theorems, though it has certainly shed some light, is far from resolving the question of the ontological status of the wave function.

3. The epistemic-ontic debate

3.1 Configuration space realism

According to the so-called ontic views of the wave function, ψ constitutes a real physical entity that is part of the ontology, even if it is not directly observable. Thus, when we talk about it, we refer to a function that describes a kind of field, just as the electromagnetic wave equation describes the electromagnetic field. This is exactly the view held by the supporters of wave function realism (Albert and Ney, 2013; Gao, 2017; Ney, 2015, 2019).

Wave function realism takes ψ to represent a field in configuration space, i.e., the multidimensional space containing all possible configurations of any system in the universe, modeled by a Hilbert vector space with potentially infinitely many dimensions. In fact, this is the space in which the wave function »lives in« as a mathematical object: wave function realism assumes that this space has a literal corresponding physical counterpart, and is not purely abstract.[2]

This configuration space is then conceptually very different from the phase space we encounter in thermodynamics and statistical mechanics as the space of all possible microstates, because it does not represent a space of probabilities or an ensemble state space, but a concrete alternative to our familiar 4-dimensional spacetime, as (Ney, 2015) suggests. Configuration space, as the space in which the wave function operates, would then be the fundamental physical space of the world, while 3-dimensional space or 4-dimensional spacetime would be contingently emergent. Albert (Albert and Ney, 2013) explains the generation of geometric appearances as a matter of dynamics: an adequate coordinate transformation in n-dimensional space would generate the apparent 3-dimensional one. In other words, while it may appear to us that we live in a 4-dimensional manifold of 3-D space and time, the real structure of »space« is that of a multidimensional one, over which there is a field called »ψ« that describes the behavior of matter.

However, if we consider quantum field theory, the configuration space itself is defined from the field operators, which in turn have a connection to spacetime, so it may just be a matter of convention which one we consider fundamental (Myrvold, 2015).

The wave function realist view is compatible with both a ψ-complete and a hidden variable ontology: in the former case, the wave function would determine the emergence of particles and all our observables; in the latter case, it would be paired with other local beables for which we can only partially determine properties (Albert and Ney, 2013; Ney, 2015).

[2] See (Halvorson, 2019) on literal interpretation on the quantum state.

Configuration space realism would provide an explanation for the phenomenon of entanglement and restore locality; as pointed out by (Norsen et al., 2015), quantum phenomena such as the violation of the Bell inequality and quantum teleportation, for example, would require configuration space realism.

Following (Ney, 2015), suppose we have a system with two electrons that are in an indeterminate *x*-spin state. Their state before the measurement can be represented in two ways:

$$\psi_{singlet} = \frac{1}{\sqrt{2}}|x_{\uparrow}\rangle_1|x_{\downarrow}\rangle_2 - \frac{1}{\sqrt{2}}|x_{\downarrow}\rangle_1|x_{\uparrow}\rangle_2$$

$$\psi_{independent} = \frac{1}{2}|x_{\uparrow}\rangle_1|x_{\uparrow}\rangle_2 + \frac{1}{2}|x_{\uparrow}\rangle_1|x_{\downarrow}\rangle_2 + \frac{1}{2}|x_{\downarrow}\rangle_1|x_{\uparrow}\rangle_2 + \frac{1}{2}|x_{\downarrow}\rangle_1|x_{\downarrow}\rangle_2$$

If we send the electrons in opposite directions to two Stern-Gerlach devices and deflect them to detect their *x*-spin, we assume that the first electron has 2 possible positions, *A* and *B*, while the second one can be in positions *C* or *D*. *A* or *C* results when the electrons are in a spin-up state, *B* or *D* when they are in a spin-down state. The state of the system can be represented as:

$$\psi_1 = \frac{1}{\sqrt{2}}|A\rangle_1|D\rangle_2 - \frac{1}{\sqrt{2}}|B\rangle_1|C\rangle_2$$

$$\psi_2 = \frac{1}{2}|A\rangle_1|D\rangle_2 + \frac{1}{2}|A\rangle_1|C\rangle_2 + \frac{1}{2}|B\rangle_1|C\rangle_2 + \frac{1}{2}|B\rangle_1|D\rangle_2$$

The wave function realism allows us to draw a distinction between these two states, since each combination occupies a different point in the configuration space. Then the wave function will have different locations in the two cases, while if we consider only the probability, we will get the same information about the system in both cases. Then this interpretation does not take the trajectory of the system as non-locally distributed between different possibilities, because the wave function results local in the configuration space. The configuration space allows us to satisfy the separability requirement, i.e., that the physical state supervene on the specification of the local beables at any point in spacetime (Myrvold, 2015).

By making the configuration space physical and fundamental, we would solve the paradoxes raised by the no-go theorems. However, we should point out that although locality is recovered in configuration space, the beables in 3-D space remain nonlocal (Norsen, 2015). This raises the problem of the interaction between high-dimensional wave function and low-dimensional peas, i.e., particle-like beables, which will be further explored in the next section.

(Gao, 2014, 2017) solves this problem while maintaining the fundamental character of our familiar spacetime by explaining entangled states described by the wave function with random discontinuous motion of particles. He also proposes a particle ontology by stating that since the wave function of an N-body system corresponding to N particles has 3N coordinates in configuration space, these must correspond to 3-dimensional positions.

3.2 Problems with ontic views

As stated by Heisenberg in 1955, cited in (Bacciagaluppi and Valentini, 2006),

»For [de Broglie and] Bohm, particles are 'objectively real' structures, like the point masses of classical mechanics. The waves in configuration space are also objectively real fields, like electric fields... [But] what does it mean to call waves in configuration space 'real'? This space is a very abstract space. The word 'real' goes back to the Latin word 'res' meaning 'thing'; but things are in ordinary three-dimensional space, not in an abstract configuration space«.

By restoring locality and separability, configuration space realism offers a viable solution to the problems posed by no-go theorems about hidden variable theories. However, it has significant problems that could be arguments for an epistemic view (Gao, 2017).

The first and most obvious is how to recover the 3-dimensional space of our experience and observations from a multidimensional ontology. As pointed out by Maudlin in (Albert and Ney, 2013), the observables should be directly determined by the ontology: The wave function should then be able to specify from the configuration space for each N-body system the positions of the particles that we detect in 3-D space. However, there are many ways in which N particles can move in 3-D space, all compatible with only one evolution in 3N-D space, and nowhere can we find a specification within ψ of which dimensions correspond to which particles (Monton, 2002). There is no »natural« correspondence in the Hamiltonian that specifies how objects in 3N-D space correspond to three-dimensional ones, and we cannot even find one in some physical facts. (Myrvold, 2015) claims that the problem is specifically finding 3-D objects in the wave function, not in the relation of the wave function to ordinary space. This is because in this picture the fundamental space of the world is configuration space, therefore the wave function would have virtually no relationship to it. The problem of emergence arises when we need to identify a correspondence between object positions in configuration space and in 3-D space, and it is not an obvious one, since such a correspondence would be required to preserve locality and separability, so the physical state supervene on a specification of local conditions at each point. Note that the physical state in 3-D does not supervene on the state of the wave function (Monton, 2002).

The second problem with ontic views is how to treat the collapse of the wave function in measurement (Gao, 2017; Ney, 2015). According to an epistemic view that does not treat the wave function as a representative of something real, the collapse corresponds merely to an update of the information we have about a system. In contrast, we must take collapse seriously if we consider the wave function as a direct representation of a state. It then becomes an undefined but real physical process that instantaneously changes the wave function and thus the state, which in turn introduces indeterminism and non-locality into the description of physical processes. Indeed, as proved by (Harrigan and Spekkens, 2010), locality is also ruled out for ontic views, and this is far from unproblematic. If collapse is a physical process that cannot be explained, the only ψ ontic theory that could avoid this problem is the Many Worlds interpretation, but believing in an infinite branching of reality is, after Occam's razor, less likely to get us to the truth than postulating that collapse is either an epistemic or a nonlocal physical process (although allowing for nonlocality is still pretty bold). In a configuration space realist view, collapse could be viewed as the movement of the wave function to a precise position in configuration space, but this too is instantaneous.

The third problem is the indistinguishability of nonorthogonal states (Leifer, 2014b). From an ontic point of view, two nonorthogonal states represent different arrangements of physical reality, but the distinction cannot be detected in any way. This could in principle be explained by an epistemic view according to which, since nonorthogonal quantum states correspond to overlapping probability measures about the true ontic states, indistinguishability results from the fact that the ontic state is in the overlap region. However, Leifer shows that even this epistemic explanation becomes imprecise and implausible for higher Hilbert space dimensions.

The last point is the eigenvalue-eigenstate half link, i.e., the fact that »when a physical system is in an eigenstate of an observable, the system has an observation-independent property, with value being the eigenvalue corresponding to the eigenstate« (Gao, 2017). The wave function of a system, assuming that it corresponds to something real, would then be determined by the set of observables of which it is an eigenstate. But is this only the case if the wave function is an eigenstate of some observables? In other words, does the quantum state have properties only if it is specified by the wave function as eigenvalues of observables? If the wave function is real, then shouldn't the properties always be present, and if that is the case, why aren't they always detectable? This is related to the crucial fact that there is a many-to-one correspondence of wave functions to quantum states, which seems to undermine the onticity of the wave function.

All of these problems relate to a configuration space realist view as we consider it the main proponent of ontic models; if we were to consider other views according to which the wave function is real but not in configuration space, entanglement would become a physical process that could not be explained, hence such views would be far more problematic than configuration space realism (Harrigan and Spekkens, 2010; Ney, 2019). Therefore, I will not consider them as viable options.

The problems we have just considered can be solved by an epistemic view; however, we must note that the concept of epistemic view is very broad and can account for various degrees of antirealism (Gao, 2017). The exact formulation of our wave function ontology depends on the preferred version of QM and is therefore a stance rather than a discovery, allowing us to interpret the theory differently (Monton, 2002). Certainly, to arrive at a more accurate ontology, we need to consider different formulations when the only source from which we can derive it is a mathematical formalism (Albert and Ney, 2013; Callender, 2015). As we have already seen, ψ-epistemic views are strongly undermined by no-go theorems that prove their limited nature. However, they might solve some of the difficulties raised by ontic views. I think epistemic views are worth considering as operational theories that leave aside the question of the status of the wave function and adopt a classical particle ontology, treating the wave function in its abstract and nonlocal nature as a mathematical tool that is useful but mysterious. For example, (Bartlett et al., 2012) have attempted to develop an epistemic version of quantum mechanics, Gaussian Quantum Mechanics, by implementing classical Liouville mechanics with the uncertainty principle, but, as the authors themselves point out, the model is limited and is proposed as an exercise in axiomatization.

In the final section, I will explore the philosophical implications of our answers in relation to the ontological status of the wave function, focusing on the notion of law, mathematical object and causality, and how formalism and ontology can be related.

4. Interpretive perspectives

It seems, then, that reducing the debate about the ontological status of the wave function to its core raises the question of what the fundamental physical space of reality is and what exists in that space. Configurational space realism presents us with what (Solé, 2013) calls the problem of perception and the problem of missing invariants; its main difficulty is explaining the 3-dimensional character of our experience, which should be determined in some way by ontology, as indicated in chapter 1, and the fact that the dynamical symmetries of non-relativistic quantum mechanics also seem to suggest a 3-D space. The emergence of 3-D beables is then still not accounted for. In addition, a mixed ontology with both high-dimensional ψ- and low-dimensional particles presents us with the problem of communication, i.e., how to establish a connection that causally explains the influence of the former part of the ontology on the latter.

Since ontic views face these obstacles and realistic epistemic views are ruled out by no-go theorems, it is questionable whether there really is a well-defined and knowable quantum state λ, since its relation to ψ and to local beables is puzzling for both epistemological and ontological reasons. We are then tempted to think of λ as related to the Kantian noumenon: the state that exemplifies reality but is intrinsically impossible to know, and unlike the phenomenon that is our only possible epistemic object (Kant, 1934).

Thus, as (Callender, 2015) suggests, the problem can be viewed in part as a two-space problem, borrowing a Cartesian notion: How can we connect the epistemically inaccessible fundamental space of reality – the space in which λ exists – with the space we map into our knowledge with models and theories, the one in which we construct our laws and ontology? Which is the 3-D space and which is the configuration space? Can we consider our observations and measurements as part of epistemic space or ontic space? Although this is a very broad question and I cannot possibly explore the answer in this review, I will focus on the views expressed in the literature. The main solutions to this »separation in spaces« have been of opposite natures: either enlarging the ontic space, i.e. the space where particles arguably live, into configuration space, as Albert suggests, or, following (Norsen, 2010), reducing the space where the wave function lives to a 3D space.

Rather than putting the two parts of our ontology in the same space, according to (Callender, 2015), the problem of communication can also be solved by eliminating part of the ontology: either we can consider local beables as emerging from the wave function, or we can eliminate the wave function in favor of a particle ontology and treat ψ as abstract.

This seems pretty much an interpretive stance: How can we know exactly whether the wave function is ontological or merely operational? Why do we tend to reify the wave function by associating it with a direct representation of the quantum state of reality, even though it cannot fully describe ontology? (Callender, 2015) suggests that we look at the mathematical formalism to solve this puzzle. We have no problem accepting that the Hamiltonian function in the Hamilton-Jacobi formulation of classical mechanics lives in multidimensional phase space because we consider it a mathematical object and a lawlike entity: Why can't we consider the wave function as related to the Hamiltonian function (Norsen, 2015)? Considering ψ as a lawlike entity would resolve the debate between epistemic and ontological status by definitively removing ψ from ontology, although its role in determining quantum state remains.

However, this shifts the same debate about the nature of a law: are the laws of physics primitive, i.e. they generate and govern the world, or dispositional or

Humean, in that they arise as descriptions of intrinsic tendencies of matter (Callender, 2015; Dorato, 2015)? The notion of a nomological entity is also unclear and should be treated with caution (Ney, 2019; Dorato and Laudisa, 2015). Ney criticizes the primitivist view of laws, but there still seems to be a sense in which the laws of nature seem to be causative and not merely descriptive. This would be the case of ψ if we view it as nomological, since it does not supervene on particle positions.

Indeed, laws are not necessarily tied to one mathematical formalism: Classical mechanics can be cast in a quantum Koopman-Von Neumann framework, and quantum mechanics can be cast in a classical Hamilton-Jacobi formalism (Callender, 2015). The way laws are formulated depends only on the set paradigm and can therefore be considered internalistic (Kuhn, 1996). However, the quantum wave function differs from the classical wave function in that it is linear and single-valued and acts as a causal agent that generates the motion of objects (Callender, 2015).

This fact raises some difficulties in considering ψ as nomological (Goldstein, 2011; Dorato and Laudisa, 2014): laws are not supposed to be dynamical entities, i.e. to have explicit time dependence; moreover, ψ is contingent and can be prepared as an initial condition: May we consider a law of nature as contingent? Moreover, the many-to-one correspondence between wave function and quantum state seems to reject the nomological nature of ψ[3] (Ney, 2019). For these reasons, (Dorato and Laudisa, 2014) argue that nomological realism is inconsistent because if we consider the wave function as a mathematical object, as a law is supposed to be, it must be interpreted instrumentally or minimally, and not as fundamentally causally active. In other words, a nomological view of the wave function can only require an abstract ontology, since the mathematical nature of laws implies that they cannot be primitive, but only Humean. Therefore, a nomological view can only be instrumental and laws cannot govern beables without being instantiated as physical entities.

If we consider the wave function as part of the laws of nature, then it is still unclear whether it is a description of the evolution of a system or the information we have about it, or whether it governs it as another entity, for example, the equivalent of a field. If the latter is the case, there would have to be an entity in spacetime equivalent to the wave function that governs the beables. So we are left with our original question, and as in Bohr's day, the safest answer is to suspend our judgment about the wave function and treat it only operationally.

[3] Note that Bohmian mechanics proponents, such as (Norsen, 2015), tend to consider the universal wave function, Ψ, instead of the conditional wave function, ψ, as lawlike. However, even Ψ is a solution to a dynamical equation, namely the Wheeler-deWitt equation, and it is dubious how a law of nature can be the solution to an equation of motion.

5. Conclusion

Finally, I would like to make a brief remark about what it means to be a realist with respect to the wave function. We have seen that, no matter how formal may be our methods, no mathematical proof has yet been able to clarify the ontological status of the wave function; so it seems that realism with respect to it must be an interpretive stance. What then does it mean to judge the best interpretation of the status of the wave function? Are we looking for empirical coherence and empirical accuracy, or rather metaphysical truthfulness? Will there ever be a way to know with certainty what the wave function is and how it relates to the quantum state when our observations seem inherently incomplete? Would we have to postulate a partial reality in which properties are instantiated only when the corresponding operators commute in measurement? In other words: Is realism about the wave function and the quantum state even desirable, or should we settle for a phenomenological metaphysics? As (Dorato and Laudisa, 2014) argue, we can be realists about quantum mechanics, the quantum state, and therefore about the reality of the world we see, even if we choose not to be realists about the wave function. What is certain is that a realist attitude will be beneficial in the search for the truth about ψ (Ladyman, 2019; Halvorson, 2019). For now, the assumption that »there is an underlying reality« is surely the most comforting to uphold (Gao, 2017): no matter how much quantum mechanics hints at skeptical scenarios, the existence of reality will always be a cornerstone of our beliefs.

The nature of ψ: Epistemic and ontic views of the wave function

The dilemma concerning the meaning of the wave function is perhaps one of the most controversial and overlooked in contemporary physics. There is widespread disagreement in the scientific community as to whether the wave function should be viewed as a representative of a real physical object or whether it corresponds to a state of incomplete knowledge about the underlying system, such as a probability density. These views are respectively called ψ-ontic and ψ-epistemic. This article aims to give a clear insight into the problem and features of both views. After some clarifying preliminaries, I will give an overview of how the question was perceived and interpreted by the pioneers of quantum theory, such as Bohr and Einstein, and then move on to more modern days with the likes of Bell, Kochen, and Specker. Then I will focus on the contemporary literature and try to give a comprehensive account of the pros and cons of each view. Finally, I will discuss some of the questions that the question of the wave function raises for us, such as our relationship to reality.

Keywords: quantum mechanics, wave function, scientific realism, ontology.

References

Aaronson, S. et al. (2013). »Psi-Epistemic Theories: The Role of Symmetry”. arXiv.org 88.3.

Albert, D. and Ney, A. (2013). *The wave function: Essays on the metaphysics of quantum mechanics*. New York; Oxford: Oxford University Press.

Bacciagaluppi, G. and Valentini, A. (2006). »Quantum Theory at the Crossroads: Reconsidering the 1927 Solvay Conference«.

Bartlett, S., Rudolph, T., and Spekkens, R. (2012). »Reconstruction of Gaussian quantum mechanics from Liouville mechanics with an epistemic restriction«. arXiv.org 86.1.

Bell, J. S. (1966). »On the Problem of Hidden Variables in Quantum Mechanics«. *Reviews of Modern Physics* 38.3, pp. 447–452.

Bohr, N. (1948). »On The Notions of Causality and Complementarity«. *Dialectica* 2.3-4, pp. 312–319.

Callender, C. (2015). »One world, one beable«. *Synthese* 192.10, pp. 3153–3177.

Dorato, M. (2015). »Laws of nature and the reality of the wave function«. *Synthese* 192.10, pp. 3179–3201.

Dorato, M. and Laudisa, F. (2014). »Realism and instrumentalism about the wave function. How should we choose?«

Einstein A., Podolsky B., and Rosen N. (1935). »Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?« *Physical Review* 47.10, pp. 777–780.

Fine, A. (1993). »Einstein’s Interpretations of the Quantum Theory«. *Science in Context* 6.1, pp. 257–273.

Gao, S. (2017). The Meaning of the Wave Function: In Search of the Ontology of Quantum Mechanics. Cambridge: Cambridge University Press.

Gao, S. (2014). »The Wave Function and Particle Ontology«.

Goldstein, S. and Zanghì, N. (2011). »Reality and the Role of the Wavefunction in Quantum Theory«. arXiv: 1101.4575.

Halvorson, H. (2019). »To Be a Realist about Quantum Theory«. In Olimpia Lombardi et al., ed., *Quantum Worlds: Perspectives on the Ontology of Quantum Mechanics*. Cambridge University Press, pp. 133– 63.

Harrigan, N. and Spekkens, R. (2010). »Einstein, Incompleteness, and the Epistemic View of Quantum States«. *Foundations of Physics* 40.2, pp. 125–157.

Kant, I. (1934). *Critique of pure reason*. Abridged ed. London: Macmillan.

Kochen, S. and Specker, E. (1967). »The Problem of Hidden Variables in Quantum Mechanics«. *Indiana University Mathematics Journal* 17.1, pp. 59–87.

Kuhn, T. S. (1996). *The structure of scientific revolutions*. 3. ed.. Chicago: University of Chicago Press.

Ladyman. J. (2019). »What Is the Quantum Face of Realism?« In Olimpia Lombardi et al., ed., *Quantum Worlds: Perspectives on the Ontology of Quantum Mechanics*. Cambridge University Press, pp. 121–132.

Leifer, M. (2014a). »Is the quantum state real? An extended review of ψ-ontology theorems«. arXiv.org 3.1.

Leifer, M. (2014b). »ψ -Epistemic Models are Exponentially Bad at Explaining the Distinguishability of Quantum States«. *Physical Review Letters* 112.16.

MacKinnon, E. (2012). »Interpreting Physics Language and the Classical/Quantum Divide«. *Boston Studies in the Philosophy of Science*, 289. Dordrecht: Springer Netherlands.

Monton, B. (2002). »Wave Function Ontology«. *Synthese* 130.2, pp. 265–277.

Myrvold, W. (2015). »What is a wavefunction?« *Synthese* 192.10, pp. 3247–3274.

Ney, A. (2015). »Fundamental physical ontologies and the constraint of empirical coherence: a defense of wave function realism«. *Synthese* 192.10, pp. 3105–3124.

Ney, A. (2019). »Locality and Wave Function Realism«. In Olimpia Lombardi et al., ed., *Quantum Worlds: Perspectives on the Ontology of Quantum Mechanics*. Cambridge University Press, pp. 164–182.

Norsen, T. (2010). »The Theory of (Exclusively) Local Beables«. *Foundations of Physics* 40.12, pp. 1858–1884.

Norsen, T., Marian, D., and Oriols, X. (2015). »Can the wave function in configuration space be replaced by single-particle wave functions in physical space?« *Synthese* 192.10, pp. 3125–3151.

Pusey, M., Barrett, J., and Rudolph, T. (2012). »On the reality of the quantum state«. *Nature Physics* 8.6, pp. 475.

Solé, A. (2013). »Bohmian mechanics without wave function ontology«. *Studies in History and Philosophy of Modern Physics* 44.4, pp. 365–378.

Sam Dickson
University of York

Mathematical Causation

Matematična vzročnost

Nekateri trdijo, da nas intuitivne sodbe o matematičnih trditvah vodijo do tega, da verjamemo v matematični platonizem. Toda matematični predmeti platonističnih teorij naj ne bi bili prostorsko časovni in ločeni od sveta, natančneje so nevzročni. To stališče ima težave; če so matematični predmeti ločeni od sveta, se zdi, kot da ne vplivajo na svet. Svet bi bil tak, kot je, tudi če matematični predmeti ne bi obstajali. Poleg tega ljudje, kot je Benacerraf (1973), trdijo, da glede na vzročno teorijo vednosti in glede na njihovo nevzročno naravo nikoli ne bi mogli vedeti o matematičnih predmetih in jih zato ne bi smeli postulirati. Drugje v literaturi ljudje ponujajo neogibne argumente *za* matematične predmete (npr. Baker, 2009). Zdi se, da se morajo naše najboljše znanstvene teorije sklicevati na razlage, ki kvantificirajo matematične predmete, zato bi jim morali verjeti ravno tako, kot verjamemo v neopazljive predmete znanosti. Potem lahko rečemo, da vemo, da matematični predmeti obstajajo s sklepanjem na najboljšo razlago, ki temelji na naših najboljših znanstvenih teorijah. Toda to nam ne pove, *kakšni* so matematični predmeti, še pomembneje pa je, da še vedno ne odgovori na argumente »ne naredi nobene razlike«. Morda bomo prisiljeni verjeti vanje, vendar nam to ne pove, ali dejansko kaj *naredijo* ali ne. Želim razpravljati o enem od načinov, kako lahko matematični predmeti nekaj *naredijo*. Mislim, da obstaja koristen in informativen način, kako lahko o matematičnih objektih govorimo kot o vzročnih. To storim tako, da razpravljam o primeru matematične omejitve, kot je predlagal Marc Lange (2017). Izpopolnil bom pojem matematične omejitve in govoril o relaciji omejitve na splošno. Nato bom nadaljeval z razpravo o modelih strukturnih enačb in o tem, kako jih lahko uporabimo za predstavitev vzročnih relacij. To še posebej ustreza intervencionističnemu pogledu na vzročno zvezo, ki ga bom opisal, in kako ga lahko uporabimo kot test za ugotavljanje, katere relacije so vzročne. Mislim, da matematična omejitev uspešno prestane ta test. V tem okviru bom razpravljal o posebnem modelu strukturne enačbe, ki predstavlja relacijo omejitve. Ta posebna relacija omejitve je tudi neposredno vzročna. Mislim, da se struktura te relacije naravno preslika na strukturo arhetipskega primera matematične omejitve. Ne samo, da si delita strukturo, ampak se obe relaciji obnašata na enake načine v intervencionističnih obravnavah. To bi nam moralo dati razlog za trditev, da je matematična omejitev vzročna. Za tiste, ki jih zanima epistemologija matematike, lahko to omogoči izogibanje ugovorom v Benacerrafovem slogu; skrivnostno *bi* bilo, kako vemo o nevzročnih matematičnih objektih, toda glede na to, da so dani matematični objekti vzročni, lahko razložimo svojo vednost. Za tiste, ki so naklonjeni ali se zavzemajo za argumente, da to ne naredi nobene razlike, imamo spet odgovor, da so matematični predmeti vzročni, zato *bi* to, če ne bi obstajali, naredilo razliko v svetu.

Ključne besede: intervencionizem, protidejstvenik, model strukturnih enačb, vzročnost, omejitev.

1. Introduction

It is widely accepted in the philosophy of mathematics that mathematical objects, if they existed, would be acausal. Indeed, this acausal nature poses the strongest challenge to mathematical Platonists, since it makes it difficult to explain how we can have knowledge about mathematical objects. Contrary to this orthodoxy, this paper discusses a sense in which mathematical objects might be causal. This is intended as an attempt to tell us not only what sort of things mathematical objects are, but also to explain how it is that they can make a difference and we can gain knowledge about them. This is a controversial view of mathematical objects, but I think there is precedent in the literature for this claim. To try to justify this claim, I will briefly describe the mathematical constraint relation proposed by Marc Lange (2017) and what it is supposed to *do*. To prove that the mathematical constraint can be a causal relation, I will try to show that it fits a typical causal pattern in a structural equation model and that under an interventionist interpretation we can consider the mathematical constraint to behave like a causal relation. In this paper, we build a notion of causality in the tradition of interventionism and structural equation models. After establishing this notion, it will then be applied to a case in mathematics to show that we can view the mathematical constraint relation as a causal relation. I will then discuss some issues that are likely to be raised in response to this view, and respond to them first.

2. Motivations

Given the way we talk about numbers, and certain intuitions, mathematical objects seem to be non-spatiotemporal, acausal entities. Benacerraf and others have raised problems in this regard (Benacerraf 1973 & Liggins, 2010). Given their non-spatiotemporal and acausal nature, it is inexplicable how we can know about mathematical objects. Since knowledge about them seems impossible, our theories postulating them face a serious problem. This argument need not take exactly this form, and instead of referring to knowledge, it may refer to the ability to form justified beliefs about mathematical objects, or it may be granted a warrant to refer to them, and sometimes it is not even expressed in causal terms. The essence of any form of this objection is that without a connection to mathematical objects, the postulating of those objects is in some sense suspect. If this connection is not causal, then it is puzzling exactly what it is. I would respond that there is a connection between mathematical objects and us, and furthermore that connection is a causal one, so it is not puzzling or *ad hoc*. I agree that mathematical objects seem to be non-spatiotemporal, but I don't think that rules out their being causal. It seems plausible that mathematical objects ground the truth of certain statements and facts in the world. I will argue that this justification is a kind of causation, even though mathematical objects are not spatiotemporal; for they act on the world by constraint. To use a classic example: We cannot divide 23 strawberries (of equal size) equally among 3 people (without them cutting them up) *because* 23 is not divisible by 3 (into whole natural numbers)[1]. The mathematical fact constrains the way the world can be, it limits the possible actions mother can take in dividing the strawberries (Lange 2017).[2] It seems reasonable to say that this is a grounding claim. In a sense, mother cannot divide the strawberries in this way *because of the fact that* 23 is indivisible by 3. I think this is a perfectly natural reading of constraint. Those tempted by Jessica Wilson's (2014) discussion of grounding might want to say that constraint is not identical to big-G grounding, but is instead a kind of small-g grounding relation. Alternatively, those who follow Karen Bennett (2017) might say that constraint, like grounding, is one of the construction relations. However you want to read it, I think it is appropriate and correct to place constraint in this region. I prefer a Bennett-like picture myself. Grounding and causation are two kinds of broad dependency relations, and constraint represents a small area where they overlap. But I think that as long as one accepts grounding in the first place, constraint can be read as neutral across all interpretations, and one need not accept Bennett's picture of building. Returning to the topic of discussion, I argue that grounding by constraint, which is present in mathematical cases like the one above, is a kind of causality because it shares a structure with straightforward causal relations. In structural equation models (SEMs), constrainers in constraint relations occupy the same place as causes in random relations, in interventionist treatments of causality, and so we should be prepared to accept them as causal. I will discuss a basic SEM before moving on to more complicated ones, and then move on to a final example that I believe demonstrates a structure shared by directly causal constraint relations and mathematical relations.

3. Structural Equation Models & Interventionism

SEMs are models designed to represent causal relationships in a clear and simple way, and they allow us to infer causal relationships through statistical data. They have been used with great success in statistical modeling in fields such as physics and economics, and are considered a very useful heuristic. They also have many applications in philosophy, particularly in counterfactual reasoning and the

[1] Of course, 23 is divisible by 3, it just does not yield a whole natural number. For the purposes of this example and simplicity we restrict oursemselves to division into natural numbers. This is also the reason we imagine the strawberries cannot be cut up in this scenario. Perhaps the children are unusually picky eaters and will only eat strawberries if they are whole, but still demand an even distribution. Hereafter all these caveats will be assumed and »23 is not divisible by 3« will be referring purely to division into whole natural numbers.

[2] This is a simple and restricted example but I think it serves to show the principles at play here and demonstrate how one might come to the conclusion that mathematical objects could be causal. There are many other examples of mathematical constraint, and many more complicated ones, these will obviously require slightly different and more complex treatments but those treatments will still be in the spirit of the account proposed in this paper.

analysis of causality. In a SEM, one assigns values to variables and specifies some rules for how the variables interact with each other to model the particular relationship.[3] In the following illustrative causal SEMs, *C* and *E* generally stand for cause and effect, respectively. In the following simple case of a stone being thrown against a window, we can see how SEMs look and work:

Variables

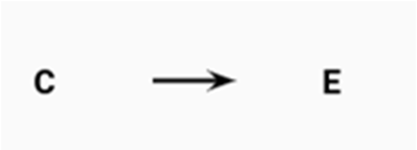
C: Whether Suzy throws the stone

E: Whether the window shatters

Structural equations *E*=*C*

Assignment *C*=1; *E*=1

Graphical representation



(Wilson 2018: 741)

C causes *E* if interventions that change the value of *C* affect the value of *E* in a particular way. This is the interventionist account of causality and is developed more clearly in some places, e.g. Woodward (2003). Specifically in relation to the causality/grounding unity, this idea is taken up by Alastair Wilson (2018). Wilson suggests that the way interventionism is applied to show that a relationship is causal also works in cases of grounding. For reasons of theoretical unity, among others, he suggests that we simply treat grounding as a kind of causality, a genus of the same kind, so to speak (Wilson, 2001). I am quite sympathetic to this idea, although I am not entirely sure that it works in all cases. For my purposes, I merely want to show that a certain kind of grounding (by constraint) is identical to a certain kind of causation (again, by constraint). I will nevertheless draw on Wilson's examples intended to show that grounding and causation are the same thing in order to build an understanding of interventionism and SEMs. If we look at the above model under an interventionist treatment of causation, we can see how this relationship is causal. We determine whether *C* causes *E* by making interventions on *C*, e.g. preventing its occurrence, and see how this affects *E*. As Woodward

[3] A more thorough exploration of SEMs, their use and applications, and their relevance to the topic of causation in philosophy can be found in Hitchcock (2018).

notes, »we can explain what is for a relation between *X* and *Y* to be causal by appealing to facts about other causal relations involving *I, X*, and *Y*, and counterfactual claims involving the behaviour of *Y* under interventions on *X*« (Woodward 2003: 105).

Interventionist explanations of causality are a type of counterfactual explanation that tend to assert that a relationship between *X* and *Y* is causal, based on counterfactuals such as »if *X* had not happened, *Y* would not have happened«. Specifically, interventionism states whether a relationship between *X* and *Y* is causal based on what would happen if an intervention were to occur on *X* with respect to *Y*. An intervention is a technical term generally defined by four criteria:

> *I* is an intervention on *X* iff:
>
> *I* causes *X*.
>
> *I* isolates *X* from previous causes of *X*, so that the value of *X* is fixed by *I* alone.
>
> Any path from *I* to an effect *Y* goes through *X*.
>
> *I* is independent of any variable *Z* which causes *Y* and is on a path which does not go through *X*. (Woodward 2003: 105)

If these criteria are met, then an action can appropriately be called an intervention and the causal relationship can be tested. If an intervention on *X* in relation to *Y* causes *X* and goes on to cause *Y*, then we can confirm that *X* causes *Y*.

We can see how this would work in the case of Suzy. We see Suzy throw the stone, and we see the window break, but she denies that she caused the window to break. Let's imagine that Greg was also standing next to the window at this point, claiming to cast a window breaking spell (Hitchcock, 2018). Now let's imagine that what actually happened is that Greg did not actually cast a spell and the breaking of the window was actually Suzy's fault. We can see how we could recreate the situation and perform interventions to determine this. We could hold fixed Suzy's throwing of the stone while varying Greg's casting of the spell to break the window, and vice versa. Soon we will see dependency patterns emerge between Suzy's throwing and breaking the window. If we are still unsure, we could make further interventions by varying the speed of Suzy's stone throwing, or perhaps varying the obstacles between Suzy and the window. Ultimately, however, we will be able to show that Suzy's relationship to shattering the window is a true dependency relationship, but Greg's casting of the spell is not. Although this is a simple case and in such a situation we would not need to apply the interventionist test to determine the causal relationship, this method can be adapted to more complex scenarios where it is not so clear. In particular, we will see this in the case of mathematics, described later. Interventions and SEMs allow us to determine what the effect depends on, what the cause is in a given situation. In the case of mathematics, we will see how it appears that this process leads to the result that certain situations are causally dependent on mathematics.

4. Structural Equation Models & Grounding

Now that we have considered interventionism in relation to direct causality and SEMs, we can see how it can be used in relation to grounding relationships. By constructing an appropriate model, we can show what examples of a true dependency relationship exist. I will not define appropriateness rigidly here, as it will vary from case to case; but as an example: If we assume smoking to be the only relevant causal factor for lung cancer, then according to the model above, smoking would not cause lung cancer in interventionist treatments of causality. This is because in some cases people smoke and do not develop lung cancer due to other relevant factors, such as genetic predisposition. When you factor all of these things into a SEM you get the right result, this is where appropriateness comes into play. You might ask why I am modeling grounding relationships with SEMs. What I want to show is that using an appropriate SEM and combining it with interventionism reveals a true dependency relationship. I think this is sufficient to call something causal, but if one is reluctant to do so for various reasons, one might at least be tempted to concede that constraint is a genuine dependency relation. I think such a concession would still allow us to answer the epistemic and »makes no difference« challenges because the constraint relation would be so similar to causality. So I think we would have permission to use it to determine what exists, just as we use causality in the eleatic criterion.

As mentioned earlier, SEMs also model different cases of grounding. Models with the structure of the Suzy case also describe simple grounding:

> Simple: Singleton
>
> Variables
>
> *C*: Whether Socrates exists
>
> *E*: Whether the singleton set {Socrates} exists.
>
> (Wilson 2018: 741).

Again, we can see how interventionism in this case will expose the dependency relation as genuine by counterfactuals like »if Socrates did not exist, {Socrates} would not have existed«, which are clearly true. In worlds where there is no Socrates, intuition tells us that there would be no singular set containing Socrates. Indeed, it seems difficult to imagine a world in which there is a singular set containing Socrates but no Socrates. This clearly shows that there is some sort of dependency relation at play.

It is worth considering a potential problem here, since answering it will help clarify why interventionism was chosen over other counterfactual explanations of causality. As has been pointed out, interventionism is a kind of counterfactual explanation. If it is true to say »if not-*X* then not-*Y*« then it is the case that *X* caused *Y*, but what about cases where it is also true to say that if not-*Y* then not-*X*? For example, in the Socrates case above. Finally, it is true to say that if {Socrates} did not exist, then Socrates would not have existed, but we do not think that this is a causal or grounding relation. We want to say that there is a dependence relation at play in the first case, but not in the second. In a sense, the second case tracks the wrong relationship; Wilson (2018: 736) calls such cases wrong-trackers. Interventionism can allow us to avoid wrong-trackers, as Wilson notes: »The distinction between right-tracking and wrong-tracking counterfactuals is then derived in the interventionist framework from a distinction between appropriate and inappropriate causal models. Right tracking counterfactuals are those with antecedents specifying some combination of interventions on model variables in some appropriate model, and with consequents specifying some values for other model variables in that model.« (Wilson 2018: 738). I think this kind of response is related to what we mean by an intervention, and if we look at the criteria that define an intervention above, we can see why. There would be no method of causing (or producing) the nonexistence of {Socrates} that does not go via eliminating Socrates from existence. We cannot satisfy the third criterion, since in such a case the intervention on *X* with respect to *Y* would not go via *X*, there would be a simple causal dependence between the intervention and *Y*. Perhaps one way to put this idea is in terms of conceivability. It is conceivable that sets do not exist, we can imagine a world without sets. So there are worlds in which Socrates exists, but {Socrates} does not exist. But it could be argued that it is inconceivable for a set to exist without its members, so we cannot imagine an intervention on {Socrates} that leaves Socrates unchanged. It is important to remember that interventionism is not a reductive account of causation, but only an account of when we can call a relation causal or test for it, and in this case no such test can be made. Ultimately, interventionism, in combination with SEMs, can detect genuine dependency relations and distinguish them from false dependency relations.

As mentioned earlier, counterfactual explanations tend to call a relationship causal when it is the case that if *X* had not happened, *Y* would not have happened. However, sometimes the relationship between *X* and *Y* is causal, even if this counterfactual is false. For example, if there is a »back-up« or preemptive cause, *Z*, that would have caused *Y* if *X* had not happened. Cases like these tend to defeat counterfactual explanations of causation, but this can be accommodated by interventionism and SEMs. This is where the notion of an appropriate model comes in, because including all the relevant factors allows us to see the true dependency relationships in the model. A simple example of this is the case of an *Assassination*.

Two assassins, *A* and *C*, aim their guns at a target, *B*. They both always hit their target, and if they hit their target, it definitely results in death. And when *A* doesn't shoot, for whatever reason, *C* shoots instead. When it happens, *A* fires, and *B* dies. It's clear that *A* caused *B*'s death, but it's not correct to say that if *A* hadn't fired, *B* wouldn't have died because *C* would have fired instead. It seems that a simple counterfactual fails at this point. But the interventionist approach can deal with this. Once you establish the notion of an appropriate model, we can see that simply taking *A* and *B* into account will lead to false causal reports. Interventions in an appropriate model that includes *C* would reveal the genuine causal dependency relationships that are present.

Another such case is presented below:

Early Pre-emption: Marsupials

Variables

C: Whether wombat bites the plant

P: Whether wombat swallows the plant

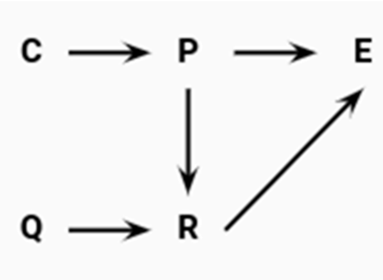
Q: Whether kangaroo sees the plant

R: Whether kangaroo eats the plant

E: Whether the plant is digested

Structural Equations $P=C$ $R=\max(Q-C, 0)$ $E=\max(P, R)$

Assignment $C=1$; $P=1$; $Q=1$; $R=0$; $E=1$



Graphical representation

Wilson (2018: 744)

It is not the case that if the wombat had not bitten into the plant, the plant would not have been digested because the kangaroo was there to eat the plant instead. Let's imagine this model without considering the variables *Q* and *R*. We could imagine an intervention being applied to this model, such as muzzling the wombat to see if the plant was digested. As it turns out, the wombat would be muzzled, so unable to bite the plant and therefore unable to swallow the plant, but the plant would be digested because the kangaroo would eat it. It seems that we have shown that the wombat biting the plant does not cause the plant to be digested, but that is the wrong result. What has happened is that we have modeled this situation with an inappropriate model. The model is inappropriate because the kangaroo is also involved. That's why it's so important to include the variables *Q* and *R*. We could imagine someone doing the above intervention and noticing the kangaroo's involvement. Given this, we could imagine someone muzzling both the kangaroo and the wombat and observing whether the plant is digested, or releasing the wombat while the kangaroo is restrained and observing what happens to the digest variable. Performing such interventions allows us to determine the dependence relationships and thus establish that the relationships are causal.

At this stage, we have seen how SEMs and the interventionist approach can be used to model and test relationships, allowing us to determine which ones are causal (or at least true dependence relationships). Both are useful tools for assessing causality and, as we will see, also useful for talking about constraint as a relationship.

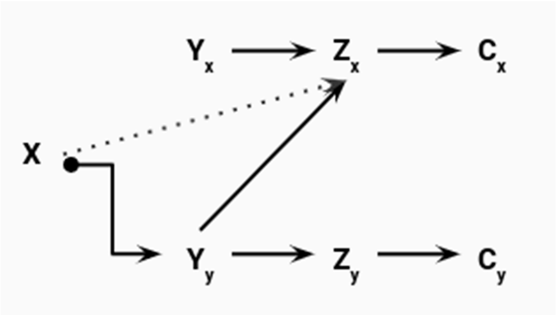
5. Constraint Relationships

With this understanding of SEMs, we can move on to discuss constraint relations. I will explain a constraint relation, which I hope tells us intuitively that it is causal, before showing that the same structure applies to cases of mathematical constraint. The example we will use is a river. The river flows down a hill and comes to some sort of plateau. There are two potential courses (*A* and *B*) that the river could then take from this plateau to flow further downhill. As it happens, the river only flows down one of them, *B* (perhaps because it is closer or lower). One day a tree falls and blocks the path of river *B*, causing the river to have to make a new path through another course (*A*). It seems fairly obvious that the fallen tree has caused the river to flow down through course *A*. It also seems obvious that the fallen tree has made it impossible for the water to flow down river channel *B*. This seems then, to be a constraining relationship. The tree has restricted the range of possible »actions« of the water, it has made certain »options« impossible. In the example with the strawberries, the mathematical fact restricts the range of possible distributions, it causes certain distributions of the strawberries to be impossible. The mathematical fact restricts the physical world. The cases are parallel and both seem equally constraining. It is also worth noting that the case is causal and restrictive, due to the same elements. The cause is the fact that the tree fell (or the event) and the effect is the fact that the water flows as it does (or the event). The constraint is the fact that the tree has fallen, what results from that, the constrained, is the fact that the water flows as it does. Now, of course, you could say that these relationships are not identical, but that instead there are two in play, but I don't see a strong case for that. As I will argue, these relationships behave in the same way under similar interventions, and the patterns of dependence remain the same. Given that they exist between the same elements, we should conclude, on intuition and parsimony grounds, that there is in fact only one relation at play here, perhaps described in different ways, and that this relation is causal and constraining.

Constraint: river

Structural equations: Y_x=Y_y+1, Z_x=Y_x, C_x=Z_x, Y_y=X-1, Z_y=Y_y, C_y=Z_y.

Assignment: X=1; Y_x=1; Z_x=1; C_x=1; Y_y=0; Z_y=0; C_y=0.



Graphical representation

The dashed line from X to Z_x represents the transitive dependence of Z_x on X, because X ultimately excludes certain other possibilities (Y_y, ... Y_L); it is the exclusion of these other possibilities (Y_y, ... Y_L); it is the exclusion of these other possibilities that causes $Z_{x.}$ We can see how this abstract structure applies to the river case by assigning variables.

Variables:

X: Tree falls

Y_x: Course A open

Z_x: Water flows through course A

C_x: Consequences

Y_y: Course B open

Z_y: Water flows through course B

C_y: Consequences

In the river case let us suppose that we perform an intervention on X and change its value to 0, i.e., the tree does not fall. We will see that the value of the variable Z_y changes to 1, i.e., the water flows through B and not through A. In this case, it is conceivable that we build a scale model of the tree/river case and make interventions on this model so that we can apply these conclusions to the »real« case. But let us also consider this case with the mathematical variables plugged in:

X: 23 is indivisible by 3.

Y_x: 23 objects divisible between 3 people non-evenly

Z_x: Strawberries are divisible in a particular way

C_x: Consequences

Y_y: 23 objects evenly divisible among 3 groups

Z_y: 23 strawberries evenly divisible among 3 people

C_y: Consequences

Suppose 23 had been divisible by 3, then it would have been possible to divide 23 objects evenly into 3 groups, and furthermore mother would have divided her 23 strawberries evenly among her 3 children. The variable Z_y would have changed in exactly the same way. Likewise, if mother would have divided her strawberries evenly, then she would not have divided them unevenly, i.e., the variable Z_x will also change in the same way as in the case of the tree. Thus, in an interventionist treatment, the fact that 23 is indivisible by 3 caused the 23 strawberries to be indivisible between 3 children. That the course is blocked is the cause of the water flowing down the open one, just as the indivisibility of 23 objects into 3 groups is the cause of the mother distributing the strawberries as she eventually does. An important point to make at this point is that you might want to resist this by placing the cause in this case outside of mathematics. Maybe the cause in this situation is what led to the mother only buying 23 strawberries and not one more (or that she only has 3 children), maybe the cause is some sort of social conditions. I think this is too hasty, because the reason we point to mathematics as *the* explanation (and I want to say cause) is that it explains why such a distribution is impossible in any case, whereas such social conditions would not. Moreover, the social facts may well form an important part of the background conditions that led to the distributional scenario, but it should matter what the cause is within the scenario, to put it that way. We can compare this to the assassination case discussed earlier: The background conditions that led to the assassins being assigned a target are relevant in building the scenario, and we may want to include them as part of the full causes of the scenario. But these background conditions do not thereby prevent the assassin's shooting from being a cause of the target's death. I suggest that we should treat the mathematical case in a similar way. We might want to include the background conditions as part of the relevant causes of the situation, but *in the* *situation* the cause we can point to argumentatively is still the mathematical fact, the social facts that matter do not prevent that causal attribution.

6. Possible Problems

6.1 Counterarguments

6.1.1 Intuitive Counterarguments

Before concluding, I would like to discuss some possible problematic differences. In the case of the tree, we are dealing with simple contingent facts that could not easily be true. In the case of mathematics, we are dealing with necessary truths, things that it seems could not have been otherwise. In this case, it is less clear that we are dealing with causality and not some other relation. The answer is that we can use counterpossible statements to take the place of contingent statements in the tree case, such as »if the tree had not fallen«. Counterpossibles like »if 23 had been divisible by 3« can be used in the math case, and we can see if certain situations would have held true in that case. Many people consider counterpossibles to be simply trivially true because of their impossible antecedents, and I think it's worth making some intuitive arguments for why that's premature. We can do this by comparing two counterpossibles:

> *E*. If 23 had been divisible by 3, then 3 would have been a factor of 23.
>
> *F*. If 23 had been divisible by 3, then 3 would not have been a factor of 23.

It seems clear that *E*. must appear to be true and *F*. must appear to be false. These are not merely trivial statements. Now this non-triviality may admittedly stem from the fact that in *E*. the consequent is merely a reformulation of the antecedent, whereas in *F*. it is a plain contradiction. That may be a good point, but we can still construct non-trivial counterpossibles:

> *G*. If[4] 23 were divisible by 3, then a calculator would be able to perform the division operation on 23 and 3.
>
> *H*. If 23 were divisible by 3, then a calculator would not have been able to perform the division operation on 23 and 3.

Again, *G.* should intuitively strike us as true, while *H*. should intuitively strike us as false. Also, compare these two with a further counterpossible:

> *J*. If 23 were divisible by 3, then Paris would be in France.

[4] Perhaps the antecedent should be lengthened to include »and had mathematicians known this fact« to avoid makes-no-difference style complaints about the inefficacy of mathematical truths on our practice. This would still produce a counterpossible.

A counterpossible such as *J.* seems empty in an important way that *E-H* are not. This intuitive judgement should hopefully show that there is room for the notion that counterpossibilities can have non-trivial truth values. Now it is of course the case that in mathematical cases, as opposed to physical constraints cases, the modality we are dealing with is different. But the abstract structure of these relations is the same, and how they behave when intervened upon is unchanged, which seems to be more important in determining a relation as causal; the modality may be regarded merely as a difference in degree rather than in kind.

6.1.2 Counterpossibles in science

Some people may not be persuaded by so-called »intuitive« counterpossibles like those we discussed above. Simply put, one might simply deny that these have different truth values, they are one and all true, and any appearance to the contrary is mere appearance. However, there are counterpossibles that are not so easily dismissed. Many scientific theories and models appeal to counterpossibles in their explanations and predictions. If these were all and only trivially true, then we might worry that the scientific endeavour is under threat. Moreover, the judgement that such scientific counterpossibles are non-trivial is not at all pre-theoretical or based only on intuition. Instead, such a judgement is based on scientific reasoning. In other words, such examples provide a good amount of positive support for the non-triviality of counterpossibles. Not only do we need to think through and figure out with more than mere intuition that such counterpossibles are non-trivial, but we also need to judge them as non-trivial in order to make progress and predictions.

Numerous examples of such counterpossibilities are offered and discussed by Tan (2019). Not only are there numerous examples of counterpossibles used in science, but they are also used in different ways for different purposes. One area where Tan (2019) focuses on their use is in scientific explanation. He offers an archetypal example of a counterpossibility and discusses why viewing it as a counterpossibility and as non-trivial is the correct judgement . The case offered is:

> »If diamond were not covalently bonded, then it would be a better electrical conductor« (Tan, 2019: 40).

Tan claims that this is a scientific explanation for the fact that diamond cannot conduct electricity, while solid carbon in some other forms can. The reason why covalent bonding explains this fact is because covalent bonds do not leave free electrons because they »consume« all the electrons that form the strong bond. In other substances, free electrons enable electrical conductivity (Tan, 2019: 40). The property of poor conductivity that diamond has is caused by these bonds and thus by the microphysical structure. So this counterfactual offers an explanation by highlighting this dependence relationship. But one might wonder if this is indeed a counterpossible, one might wonder if diamond could have been bound differently, and so whether this is just a simple counterfactual. There are two ways to approach this, we can consider whether something is called diamond because of its microphysical structure or because of its theoretical role in science (Tan, 2019). If we go the first way, we can easily see that this is a counterpossible, because if something is diamond only because of its microphysical structure, then something that had a different microphysical structure would not be diamond. It is a metaphysical necessity that diamond has the structure that it has. So it is metaphysically impossible for diamond to be bound in any other way.

If you go the second way, you might think that we define diamond by its theoretical role, the diamond substance is the substance that makes *x*, *y*, and *z*. But the reason diamond is different from other substances and the reason it does the things it does is because of its microphysical structure. In other words, nothing else could do the things that diamond does without its microphysical structure. Nothing could fill the role of diamond without actually being diamond. So again, in other words, it is metaphysically impossible for diamond to be bound in any other way than it actually is. So it seems that »If diamond had not been covalently bonded...« is a counterpossible. Tan (2019) goes further than this, insisting that this is also counterpossible which is true, and not-vacuously so. For it describes an empirical fact that the poor conductivity of diamond depends physically on its microphysical structure. Thus, science relies on non-vacuous counterpossibles for scientific explanation (Tan, 2019: 42). It is easy to see that this is not an isolated case, as many explanations of why substances have the properties they do will rely on a similar explanatory structure in a scientific context. So it looks like we will need to invoke non-trivial counterpossibles in some cases, so hopefully their use in this paper will not seem so controversial, keeping that in mind.

6.2 Intelligibility of interventions

Another possible problem is the intelligibility of interventions: We can understand what it means to stop the tree from falling, but we can't understand what it means for 23 to be divisible by 3. I think this objection is pretty closely related to the last one. It's easy enough to understand counterfactuals like »if the tree hadn't fallen« because we know what this world would be like, and we can imagine ourselves holding everything fixed in this world except for the tree falling. But with counterpossibles like »if 23 had been divisible by 3«[5] is less clear, and what that world would look like is puzzling. If we change this mathematical fact, then of course people will think we need to change more of the math, there will be a massive wave of changes throughout the world, and the objection will be that there is too much to make sense of to be able to make any statement about the truth or falsity of things in such a world non-trivial. So we could make any arbitrary statements about imagined interventions in mathematical facts, but they will not make sense, and in any case would not allow us to conclude what would lead from

[5] Or more precisely, »if 23 had been divisible by 3 into whole, natural numbers«.

them because we do not know what else would be true or false in that world. To respond to this kind of worry, I would like to make use of a response offered by Marc Lange (2019) in response to these questions concerning another theory. Lange sees the crux of the problem as being that counterpossibles »spill out« into the world to such an extent that we cannot make sense of what is true and false in this world. He claims that this is the case with many common counterfactuals that people would use. For example, we tend to think of counterfactuals like »if Julius Caesar had lived today . . .« and are happy to say that we can hold this fixed without changing anything. But that's not really the case, because we can ask how come Julius Caesar is alive today, did he time travel? That would require a lot of differences, was it a time travel machine built by humans, did he fall through a wormhole in a sci-fi-esque accident? Maybe instead of time travel it's just that the baby that was born and grew up to be Julius Caesar was actually born in 2007. But if he didn't have the same upbringing (which is probably impossible, considering how much society has changed), then he really isn't Julius Caesar. Lange's point is that although it may seem that ordinary counterfactuals hold when »everything else is fixed«, this is not in fact the case (Lange, 2019). All counterfactuals undulate because the world is fundamentally contiguous. So criticising counterpossibles for this is a weaker criticism than intended, since it also applies to counterfactuals. We may have to do some work to figure out what exactly gets fixed in counterpossible worlds and what remains unchanged, but the fact that this work has to be done is not a problem. Alternatively, it seems that in ordinary rippling counterfactual cases we can simply stipulate that such counterfactuals are non-trivially true or false, so this should be available to us in the counterpossible case as well (Lange, 2019). I think Lange got this right, just because a counterfactual seems easy to understand does not mean it is, and just because a counterpossible seems impossible to understand does not mean it is. Moreover, the literature on impossible worlds is constantly expanding, and much work has been done in this area to show that impossible worlds have non-trivial truth conditions. It may well turn out that all of these accounts are false, but while the debate is still lively I don't think we should jump to that conclusion, and so in the meantime we can make do with such accounts to justify the kinds of counterpossibles and impossible interventions we want.

7. Conclusion

In conclusion, the constraining relation seems to be a real one that operates on the world. Moreover, it is at least arguable that this relation shares a structure with relations that we would normally want to describe as causal. Given this, it seems as if we should conclude that constraint is a causal relation. This means that because mathematical objects are involved in constraint relationships, they are involved in causal relationships. Given this, there seems to be an open causal path available to us that could allow us to avoid the Benacerraf problem and gain knowledge of mathematical objects, and also to avoid »makes no difference« arguments, because mathematical objects do make a difference, thus allowing us to postulate them. In this contribution we have seen one application of this thinking and how it can give us knowledge of the natural numbers. In a similar vein we can see how one might construct a mathematical version of the pre-emption case we discussed earlier. The fact that 17 does not have 2 as a factor is *because* it is a prime number, but equally even if it had not been prime, it is an odd number, and this also explains why it does not have 2 as a factor. This seems like some pre-emption/overdetermination is at play here (perhaps the explanation could be considered to be in the other direction, with the fact the number is odd being more important, but clearly there will be some sort of pre-emption/overdetermination). Perhaps such an explanation could be of use in a constraint explanation of a variation of the classic cicadas case (Baker, 2005). The reason that 17-year life cycle cicadas do not intersect with predators every 2 years is *because* 17 is prime, but even if 17 had not been prime, it is an odd number. So one might wish to say this non-intersection is overdetermined/pre-empted by mathematical facts. Of course so far we have only considered how natural numbers can be involved in constraint relations, and that is a small part of mathematics. Thankfully, the epistemic advantage of appealing to constraint could give us much more. Certain percolation phase transitions seem to exhibit constraint and this would ultimately grant us knowledge of the real numbers. Lange (2017:7–8) argues that the example of the Königsberg bridges is a constraint explanation. It is a constraint explanation because the arrangement of bridges and landmasses makes it impossible to traverse all the bridges exactly once in a single attempt, called an Eulerian path. To perform an Eulerian path, it must be the case that, considered as a network, either every vertex, or every vertex but two, is touched by an even number of edges, I will refer to this as the Eularian fact. We can see how we could plug this into the above model to produce a similar constraint explanation.

Variables:

X: Eulerian fact

Y_x: The Königsberg bridge arrangement can be crossed exactly once each in a single trip

Z_x: A given person, P, performs an Eulerian path over the bridges

C_x: Consequences

Y_y: The Königsberg bridge arrangement cannot be crossed exactly once each in a single trip

Z_y: A given person, P, performs a non-Eulerian path over the bridges

C_y: Consequences

What I hope this shows is that the constraint approach to mathematical explanation, treated as a causal relation can give us knowledge of, and a connection to, further mathematical objects. This provides us with a project and a process going forward, treating constraint as causal, or even as an important dependence relation in its own right, has the potential to give us a path to knowledge about many different areas of mathematics and go a large portion of the way to solving the epistemic problem.[6]

Mathematical Causation

Some argue that intuitive judgments about mathematical statements lead us to believe in Mathematical Platonism. But the mathematical objects of platonistic theories are supposed to be non-spatiotemporal and detached from the world; more precisely, they are acausal. This is problematic because if mathematical objects are detached from the world, then it would seem that they make no difference to the world. The world would be the way it is even if there were no mathematical objects. Moreover, people like Benacerraf (1973) argue that given a causal theory of knowledge, and given its acausal nature, we could never know about mathematical objects and therefore should not postulate them. Elsewhere in the literature, indispensability arguments are made *for* mathematical objects (e.g., Baker, 2009). Our best scientific theories seem to rely on explanations that quantify about mathematical objects, so we should believe in them just as we believe in the unobservable objects of science. We can then say that we know mathematical objects exist by inference to the best explanation based on our best scientific theories. But that doesn't tell us what mathematical objects are *like*, and most importantly, it still doesn't answer the »makes no difference« argument. We may be forced to believe in them, but that doesn't tell us whether or not they actually *do* anything. I'd like to discuss one way that mathematical objects might *do* something. I think there is a useful and informative way we can talk about mathematical objects as causal. I do this by discussing a case of mathematical constraint as proposed by Marc Lange (2017). I will elaborate on the notion of mathematical constraint and talk about the constraint relation in general. I then discuss structural equation models and how they can be used to represent causal relationships. This fits particularly well with the interventionist view of causality that I will describe, and how it can be used as a test to determine which relationships are causal. I think that mathematical constraint passes this test. Using this framework, I will discuss a specific structural equation model that represents a constraint relationship. This specific constraint relationship is also clearly causal. I think that the structure of this relation naturally maps on to the structure of an archetypal example of a mathematical constraint. Not only do they share a common structure, but both relations behave in the same way in interventionist treatments. This should give us reason to say that mathematical constraint is causal. For those interested in the epistemology of mathematics, this may provide a workaround to Benacerraf-type objections; it *would* be puzzling how we know about acausal mathematical objects, but since mathematical objects are causal, we can explain our knowledge. For those who sympathize with or are concerned about the »makes no difference« argument, we again have an answer: mathematical objects are causal, so it would make a difference to the world if they did not exist.

Keywords: interventionism, counterfactual, structural equation model, causality, constraint.

[6] An earlier version of this article appeared in *Dialectic*. Available at: https://uoydialecticjournal.wordpress.com/

Bibliography:

Baker, A., 2005. »Are there Genuine Mathematical Explanations of Physical Phenomena?«. *Mind* 114(454), pp. 223–238. Retrieved (2016 Oct 13) from <https://academic.oup.com/mind/article/114/454/223/992237?login=true>

Benacerraf, P. (1973). »Mathematical Truth«. *The Journal of Philosophy* 70(19), pp. 661-679. Retrieved (2016 Nov 17) from <http://www.jstor.org/stable/2025075>

Bennett, K. (2017). *Making Things Up*. Oxford: Oxford University Press.

Hitchcock, C. (2018). *Causal Models*. The Stanford Encyclopedia of Philosophy. Retrieved (2021, February 1) from <https://plato.stanford.edu/archives/sum2020/entries/causal-models/>.

Lange, M. (2017). Because Without Cause: Non-Causal Explanations in Science and Mathematics. New York: Oxford University Press.

Lange, M. (2019). *What could mathematics be for it to function in distinctively mathematical scientific explanations?*. Mathematics and its applications: Philosophical Issues. 23-24th September. University of Leeds.

Liggins, D. (2010). »Epistemological Objections to Platonism«. *Philosophy Compass*. 5 (1), pp. 67-77. Retrieved (2017 April 1) from <http://onlinelibrary.wiley.com/wol1/doi/10.1111/j.1747- 9991.2009.00259.x/full>.

Wilson, A. (2018). »Metaphysical Causation«. *Noûs* 52(4), pp. 723-751. Retrieved (2020 February 5) from <https://doi.org/10.1111/nous.12190>.

Wilson, J. (2014). »No work for a theory of grounding«. *Inquiry: An Interdisciplinary Journal of Philosophy* 57(5-6), pp. 535-79. Retrieved (2020 May 16) from <https://doi.org/10.1080/0020174X.2014.907542>.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

Zgodovina filozofije

John Locke

Prevedel Božidar Kante

Ni nobenih prirojenih praktičnih načel[i]

Tretje poglavje

§ 1

ČE te spekulativne maksime, o katerih smo govorili v zgornjem poglavju, nimajo dejanskega univerzalnega soglasja v vsem človeštvu, kakor smo dokazali tam, je še veliko bolj očitno glede *praktičnih načel*, da jim *umanjka univerzalni sprejem*; in mislim, da bo težko najti primer kakšnega moralnega pravila, ki bi se lahko potegovalo za tako splošno in pripravljeno soglasje kot *Kar je, je* ali bi bilo tako očitna resnica kot *Nemogoče je za isto stvar, da je in da ni.* S čimer je razvidno, da so še bolj oddaljena od naziva prirojena kot druga in je dvom, da so prirojeni vtisi v duh, o teh moralnih načelih močnejši kot o drugih. S tem sploh ni dvoma o njihovi resnici. So enako resnična, čeprav ne enako razvidna. Te spekulativne maksime nosijo s sebnoj svojo lastno razvidnost, toda moralna načela zahtevajo razmišljanje in razpravo in nekaj vaje duha, da bi odkrili gotovost njihove resnice. Niso odprte kot naravne črke/znaki, vtisnjeni v duh, ki bi morali biti, če kaj takega obstaja, nujno vidni sami od sebe in biti nedvomni s svojo lastno lučjo in znani vsakemu človeku. Vendar to ni okrnitev njihove resnice in gotovosti, nič bolj kot resnice ali gotovosti, da so trije koti trikotnika enaki dvema pravima kotoma, ker ni tako razvidna kot *Celota je večja od dela* niti tako primerna, da z njo soglašamo na prvi mah. Utegne zadostovati, da je ta moralna pravila mogoče dokazati in je torej naša lastna zmota, če ne pridemo do nekega njihovega poznavanja. Toda nevednost, ki jo veliko ljudi ima o njih, in počasnost soglasja, s čimer jih sprejemajo drugi, sta jasna dokaza, da niso prirojene in take, da se ponujajo pogledom ljudi brez raziskovanja.

[i] Prevod dela Johna Locka, *An Essay Concerning Human Understanding*, Peter H. Nidditch (ed.), Oxford: Clarendon Press, 1975, str. 65–103.

§ 2

Ali obstajajo kakšna taka moralna načela, o katerih vsi ljudje soglašajo, se sklicujem na vsakega, ki je bil zgolj zmerno podkovan v zgodovini človeštva in je pogledal v dalj onstran dima svojega lastnega dimnika. Kje je tista praktična resnica, ki je univerzalno sprejeta brez dvoma ali vprašanja, kot bi moralo biti, če bi bila prirojena? *Pravičnost* in spoštovanje pogodb, to je tisto, o čemer *se zdi, da večina ljudi soglaša*. To je načelo, za katerega se misli, da sega k razbojniškim gnezdom in k združbam podležev; in tisti, ki so šli dlje k temu, da so se znebili same humanosti, obdržijo vero in pravila pravičnosti drug do drugega. Priznavam, da sami izobčenci to počnejo med sabo, vendar ne da bi te sprejemali kot prirojene zakone narave. Udejanjajo jih v praksi kot pravila primernosti znotraj svojih lastnih skupnosti, vendar ni mogoče dojeti, da se oklepa pravice kot praktičnega načela tisti, ki dela pošteno s svojimi tovariši cestnimi razbojniki in hkrati ropa ali ubije prvega častnega človeka, ki mu pride na pot. Pravica in resnica sta skupna vez družbe in zato morajo celo izobčenci in roparji, ki prekinejo z vsem svetom poleg sebe, obdržati vero in pravila enakosti med seboj, sicer ne morejo držati skupaj. Toda ali bi kdorkoli rekel, da imajo tisti, ki živijo s sleparjenjem in ropanjem, prirojena načela resnice in pravice, ki jih dopuščajo in z njimi soglašajo?

§ 3

Nemara bodo opozarjali, da *tiha privolitev njihovega duha soglaša s tistim, čemur oporeka njihova praksa*. Odgovarjam, *prvič*, da sem vselej mislil, da so dejanja ljudi najboljši interpreti njihovih misli. Ker pa je gotovo, da sta praksa večine ljudi in nekateri odprti poklici ljudi bodisi podvomili ali zanikali ta načela, ni mogoče vzpostaviti univerzalne privolitve (čeprav bi jo moali iskati zgolj med odraslimi ljudmi), brez katere ni mogoče sklepati, da so prirojena. *Drugič*, zelo nenavadno in nerazumno je predpostaviti prirojena praktična načela, ki se končajo zgolj v kontemplaciji. Praktična načela, izpeljana iz narave, so tu zaradi delovanja in morajo ustvariti ustreznost dejanja, ne zgolj spekulativno soglasje z njihovo resnico, sicer jih je zaman razlikovati od spekulativnih maksim. Narava je, priznam, postavila v človeka željo po sreči in odpor do bede: to sta res prirojeni praktični načeli, ki (kot morajo praktična načela) še naprej nenehno delujeta in vplivata na vsa naša dejanja, ne da bi nehali: ta načela je mogoče opaziti pri vseh osebah in v vseh dobah, stalno in univerzalno, vendar so to nagnjenja poželenja po dobrem, ne pa vtisi resnice v razum. Ne zanikam, da obstajajo naravna stremljenja, vtisnjena v duh ljudi in da od samih prvih primerov čutov in zaznavanja obstajajo nekatere stvari, ki so jim prijetne, in druge, ki so ji nedobrodošle, nekatere stvari, h katerim so nagnjeni, in druge, ki se jim izogibajo, vendar to ne naredi nič za prirojene znake v duhu, ki naj bi bili načela vednosti, uravnavajoč našo prakso. Taki naravni vtisi v razum so tako daleč od tega, da bi bili s tem potrjeni, da je to argument proti njim, kajti če bi obstajali kaki znaki, ki ji je narava vtisnila v razum kot načela vednosti, si ne bi mogli kaj, da jih ne bi zaznavali kot nenehno delujoče v nas in da vplivajo na našo vednost, kot zaznavamo tiste druge na voljo in poželenje, ki nikoli ne nehajo biti stalni izviri in motivi vseh naših dejanj, h katerim nas, to nenehno čutimo, močno silijo.

§ 4

Drug razlog, ki me spravlja v dvom glede kakršnegakoli prirojenega praktičnega načela, je, da mislim, *da ni mogoče predlagati nobenega moralnega pravila, od katerega človek ne bi mogel upravičeno zahtevati neki razlog*, kar bi bilo popolnoma smešno in nesmiselno, če bi bila prirojena ali toliko kot samorazvidna, kar bi vsako prirojeno načelo moralo nujno biti in ne bi potrebovalo nobenega dokaza za potrjevanje njegove resnice niti želelo kakršnegakoli razloga, da bi dobilo svojo odobritev. Za tistega, ki bi po eni strani spraševal ali po drugi strani skrbel za to, da bi dal neki razlog, *zakaj ni mogoče, da ista stvar je in je ni*, bi se mislilo, da nima zdravega razuma. Ta propozicija nosi s seboj svojo lastno luč in razvidnost in ne rabi nobenega drugega dokaza. Tisti, ki razume termine, soglaša z njo zaradi nje same ali pa ga sicer nikoli ne bo nič zmožno prepričati, da bo to storil. Toda ali bi morali tisto najbolj neomajno pravilo morale in temelj vse družbene vrline *Drugim bi morali delati, kar si sami želimo, da bi drugi delali nam* predlagati nekomu, ki ga prej nikoli ni slišal, vendar je zmožen razumeti njegov pomen? Ali ne bi mogel brez kakršnegakoli nesmisla vprašati zakaj? In ali ne bi bil tisti, ki ga je predlagal, zavezan, da mu prikaže njegovo resnico in razumnost? Kar jasno kaže, da ni prirojeno, kajti če bi bilo, si ne bi niti želelo niti sprejelo nobenega do-kaza, vendar potrebuje (vsaj takoj, ko je slišano in razumljeno), da je sprejeto in da se z njim soglaša kot nepobitno resnico, o kateri človek nikakor ne more dvo-miti. Tako da je resnica vseh teh moralnih pravil očitno odvisna od nekega druge-ga njihovega antecedenta, iz katerega morajo biti izpeljana, kar se ne bi moglo zgoditi, če bi bila bodisi prirojena bodisi toliko kot samorazvidna.

§ 5

To, da bi se ljudje morali držati svojih dogovorov, je gotovo veliko in nezanikljivo pravilo v morali, vendarle pa, če bi kristjana, ki ima sodbo o sreči in bedi v drugem življenju, vprašali, zakaj mora človek držati svojo besedo, bo kot *razlog* dal tole: ker Bog, ki ima moč večnega življenja in smrti, to od nas zahteva. Toda če bi vprašali *Hobbsovega* privrženca, bi odgovoril: ker to zahteva javnost in te bo *Leviathan* kaznoval, če tega ne storiš. Če pa bi vprašali kakega od starih *poganskih* filozofov, bi odgovoril takole: ker bi bilo nečastno, pod človekovim dostojanstvom in v nasprotju z vrlino, z najvišjo popolnostjo človeške narave, če bi storili drugače.

§ 6

Od tod izvira velika raznolikost mnenj glede moralnih pravil, ki ji gre najti med ljudmi, v skladu z različnimi vrstami sreče, ki jih pričakujejo ali pa jih predlagajo samemu sebi, da bi jih dosegli, kar se ne bi moglo zgoditi, če bi bila praktična načela prirojena in neposredno vtisnjena v našega duha z božjo roko. Priznavam, da je eksistenca Boga na tako mnogo načinov očitna in poslušnost, ki mu jo dolgujemo, tako skladna z lučjo razuma, da velik del človeštva priča o zakonu narave, vendar mislim, da mora biti dopuščeno, da utegne večje ševilo moralnih pravil prejeti od človeštva zelo splošno privolitev, ne da bi bodisi poznalo bodisi dopuščalo resnični temelj moralnosti, ki je lahko zgolj volja in zakon Boga, ki vidi ljudi v temi, ima v svojih rokah nagrade in kazni in dovolj moči, da pokliče na zagovor najbolj domišljavegas kršitelja. Kajti Bog je z neločljivo vezjo povezal skupaj *vrlino* in javno srečo in naredil njuno udejanjanje nujno za ohranitev družbe in vidno *dobrodejno* za vse, s katerimi ima opraviti vrli človek; ne gre se čuditi, da bi moral vsakdo ne zgolj dopustiti, temveč priporočiti in poveličevati ta pravila drugim, iz izpolnjevanja katerih je gotov, da bo požel prednost zase. On utegne zaradi interesa kot tudi iz prepričanja vzklikati, da je sveto tisto, ki če je enkrat poteptano in profanirano, on sam ne more biti niti srečen niti varen. To, čeprav ničesar ne odvzema od moralne in večne obveznosti, ki jo ta pravila očitno imajo, vendarle kaže, da zunanje priznanje, ki jim ga ljudje dajejo s svojimi besedami, ne dokazujejo, da so prirojena načela: ne, ne dokazuje niti tega, da ljudje privolijo vanje notranje v svojem lastnem duhu kot neprekršljiva pravila svoje lastne prakse, ker ugotavljamo, da lastni interes in prijetnosti tega življenja povzročijo, da jih mnogi ljudje zunanje priznavajo in odobravajo celo tedaj, ko njihova dejanja zadostno dokazujejo, da zelo malo upoštevajo zakonodajalca, ki je predpisal ta pravila, niti pekla, ki ga je predpisal za kaznovanje tistih, ki jih kršijo.

§ 7

Kajti če ne bomo v javnosti dopuščali preveč iskrenosti izpovedim večine *ljudi*, temveč bomo sodili, da so njihova dejanja interpreti njihovih misli, bomo ugotovili, da ljudje nimajo *nobenega* takega notranjega čaščenja teh pravil, niti tako *popolnega prepričanja o njihovi gotovosti* in obveznosti. Veliko načelo morale *Drugim delaj, kar terjaš, da delajo tebi* je bolj priporočeno kot udejanjeno. Toda kršitev tega pravila ne more biti večji greh/hiba od norosti učiti druge, da ni nobenega moralnega pravila niti obveznosti; to bi bilo v nasprotju s tistim interesom, ki se mu ljudje žrtvujejo, ko ga kršijo sami. Morda bo *vest* tista, ki nas bo silila zadrževati take kršitve in bo tako ohranjena notranja obveza in veljava pravila.

§ 8

Čemur odgovarjam, da ne dvomim, da utegne mnogo ljudi, ne da bi bilo to zapisano v njihovih srcih, na enak način, kot pridejo do vednosti drugih stvari, priti do privolitve v večje število moralnih pravil in biti prepričani o njihovi obveznosti. Drugi lahko tudi pridejo do enakega mnenja s pomočjo svoje izobrazbe, družbe in običajev svojih dežel, ki bo, *kakorkoli že pridobimo prepričanje*, *rabilo za to*, *da* *bo pognalo v delovanje vest*, ki ni nič drugega, kot zgolj naše lastno mnenje ali sodba o moralni pravilnosti ali izprijenosti naših lastnih dejanj. Če pa je vest dokaz prirojenih načel, utegnejo biti prirojena načela tudi nasprotja teh prirojenih načel, saj nekateri ljudje z enakim nagnjenjem vesti opravljajo tisto, čemur se drugi izogibajo.

§ 9

Vendar ne morem razumeti, kako bi katerikoli *človek* mogel kdajkoli *prekršiti* ta *moralna pravila* z *zaupanjem* in *mirnostjo*, če bi bila prirojena in vtisnjena v naš duh. Poglejte zgolj vojsko, ki ropa neko mesto, in kakšno je spoštovanje ali občutenje moralnih načel ali kakšen je očitek vesti za vse nasilje, ki ga povzročajo. *Ropanje, umori, posilstva* so razvedrila ljudi, osvobojenih kazni in cenzure. Ali niso obstajali celotni narodi in narodi najbolj civiliziranih ljudi, med katerimi je bilo izpostavljanje njihovih otrok in to, da so jih puščali na poljih, da so umirali zaradi pomanjkanja ali divjih zveri, praksa, ki so jo tako malo obsojali ali o njej imeli pomisleke, kot njihovo spočetje. Ali jih ne v nekaterih deželah še vedno dajejo v isti grob z svojimi materami, če umrejo ob porodu ali pa se jih znebijo, če domnevni astrolog razglasi, da so rojeni pod nesrečno zvezdo? In ali ne obstajajo kraji, ko ob določeni starosti ubijejo ali izpostavijo svoje starše sploh brez kakršnegakoli kesanja? V delu Azije bolnega, ko za njegov primer menijo, da je obupen, nesejo ven in ga položijo na zemljo, preden je mrtev in ga pustijo tam, izpostavljenega vetru in vremenu, da pogine brez pomoči ali *usmiljenja. (α) Med Mingrelijanci, ljudmi, ki izpovedujejo krščanstvo, je običaj, da* svoje otroke pokopljejo žive brez pomislekov, (β) obstajajo kraji, kjer pojedo svoje lastne otroke, (γ) na *Karibih* ponavadi skopijo svoje dečke, da bi se zredili in jih potem pojedli. (δ) In *Garcilasso de la Vega* nam pripoveduje o ljudstvu v *Peruju*, kjer so običajno redili in jedli otroke, ki so jih dobili od svojih ženskih ujetnic, ki so jih v ta namen imeli kot konkubine in ko je bilo obdobje plodnosti mimo, so ubili tudi same matere in jih pojedli, (ε) vrline, s katerimi so *Tououpinambosi* verjeli, da zaslužijo raj, sta bili maščevanje in pojesti obilje svojih sovražnikov, (ζ) sploh niso imeli imena zas boga, *Lery, str.* 216. Nobenega priznavanja boga, nobene religije, nobenega čaščenja, *str.* 231. Svetniki, ki so kanonizirani med *Turki*, preživljajo življenje, ki ga ne moremo povezovati z zmernostjo. Pomemben odlomek v tej zadevi, iz *Voyage of Baumgarten*, ki je knjiga, na katero ne naletimo vsak dan, bom navedel obširno, v jeziku, v katerem je napisana: *Ibi* (*sc. prope* Belbes *in* Ægypto) *vidimus sanctum unum Saracenicun inter arenarum cumulos, ita ut ex utero matris prodiit nudum sedentem. Mos est, ut didicimus* Mahometistis, *ut eos, qui amentes et sine ratione sunt, pro Sanctis colant et venerentur. Insuper et eos qui cum diu vitam egerint inquinatissimam,voluntariam demum pœnitentiam et paupertatem, sanctitate venerandos deputant. Ejusmodi verò genus hominum libertatem quandam effrœnem habent, domos quas volunt intrandi, edendi, bibendi, et quod majus est, concumbendi; ex quo concubitu, si proles secuta fuerit, sancta similiter habetur. His ergo hominibus, dum vivunt, magnos exhibent honores;*

mortuis verò vel templa vel monumenta extruunt amplissima, eosque contingere ac sepelire maximæ fortunæ ducunt loco. Audivimus hæc dicta et dicenda per Interpretem à Mucrelo nostro. Insuper sanctum ilium, quem eo loci vidimus, publicitus apprimè commendari, eum esse Hominem sanctum, divinum ac integritate præcipuum; eo quod, nec fœminarum unquam esset, nec puerorum, sed tantummodo asellarum concubitor atque mularum. Peregr. Baumgarten, I. 2. c. 1. str. 73. Več istovrstnega, ki zadeva te plemenite svetnike med *Turki*, lahko vidimo pri *Pietru de la Valli* v njegovem pismu z dnem 25. januarja 1616. Kje so potem tista prirojena načela pravice, usmiljenja, hvaležnosti, enakosti, čistosti? Ali, kje je tisto univerzalno soglasje, ki nam zagotavlja, da obstajajo taka prirojena pravila? Umori v dvobojih, ko jih je moda naredila častne, so storjeni brez grižnje vesti: ne, v mnogih krajih je nedolžnost v tem primeru največja sramota. In če se ozremo po tujini, da bi videli, kakšni so ljudje, bomo odkrili, da se kesajo na enem mestu, da so storili ali opustili, za kar so drugi na drugem mestu menili, da zaslužijo.

§ 10

Tisti, ki bo skrbno pregledal zgodovino človeštva in se ozrl daleč naokrog na razna človeška plemena in ravnodušno pregledal njihova dejanja, bo zmožen zadovoljiti samega sebe s tem, da je z muko imenovati kako načelo morale ali misliti *pravilo vrline* (izključena so zgolj tista, ki so absolutno nujna, da držijo skupaj družbo, ki so običajno tudi zanemarjena med različnimi družbami), ki ni ne vem kje *prezrto* in ki ga na splošen način obsojajo *cele družbe* ljudi, ki jih vodijo praktična mnenja in pravila življenja, nasprotna drugim.

§ 11

Tu bodo nemara ugovarjali, da ni nikakršen argument, da *pravilo ni znano, ker je prekršeno*. Priznavam, da je ugovor dober, kjer ga ljudje, čeprav zakon kršijo, vendar ne zavračajo, kjer strah pred sramoto, cenzuro ali kaznijo nosi znak nekakšnega spoštovanja, ki ga ima zanj. Vendar ni mogoče pojmiti, da bi *celoten narod* ljudi moral *javno zavrniti* in se odreči tistemu, za kar vsakdo od njih, gotovo in nezmotljivo, ve, da je zakon, ker se tako mora zgoditi tistim, ki ga imajo naravno vtisnjenega v duha. Možno je, da lahko imajo ljudje *pravila morale*, za katera v svojih zasebnih mislih ne verjamejo, da so resnična zgolj zato, da bi obdržali reputacijo in cenjenje med tistimi, ki so prepričani o njihovi obveznosti. Vendar si ni mogoče zamišljati, da bi celotna družba ljudi morala javno in po lastni izjavi zavračati in zavreči pravilo, ki v svojem lastnem duhu ne bi mogli, da ne bi bili nezmotljivo gotovi, da je zakon, niti biti nevedni, da vsi ljudje, s katerimi imajo opraviti, vedo, da je tak. Zato pa mora vsakdo od njih pričakovati od drugih vse zaničevanje in odvratnost, ki gre nekomu, ki izpoveduje, da nima humanosti; na nekoga pa, ki zamenjuje znane in naravne mere pravilnega in napačnega, ni mogoče gledati kot na izrecnega sovražnika njihovega miru in sreče. Katerokoli praktično načelo, ki je prirojeno, si ne more kaj, da ne bi bilo znano vsakomur, da je pravično in dobro. Zato je malo manj kot protislovje predpostavljati, da bi morale celotne nacije ljudi tako v svojih izpovedih kot praksi enodušno in univerzalno lagati o tem, o čemer z najbolj nepremagljivo evidenco vsakdo od njih ve, da je resnično, pravilno in dobro. To je dovolj, da nam zadosti, da za nobeno praktično pravilo, ki je kjerkoli univerzalno in z javno privolitvijo ali dopuščanjem prekršeno, ne moremo predpostavljati, da je prirojeno. Vendar lahko k odgovoru na ta ugovor dodam še nekaj.

§ 12

Prekršitev pravila, pravite, ni nikakršen argument, da je pravilo neznano. Priznam, toda *splošno dopuščeno kršenje pravila kjerkoli*, pravim jaz, *je dokaz, da ni prirojeno*. Na primer, vzemimo katerokoli od teh pravil, ki so najbolj očitne izpeljave iz človeškega razuma in prilagodljiva naravnim nagnjenjem največjega dela ljudi in se jih zgolj najmanjše število ljudi drzne zanikati ali nepremišljeno dvomiti o njih. Če lahko o kateremkoli razmišljamo, da je naravno vtisnjeno, mislim, da nobeno ne more imeti primernejše pravice, da je prirojeno, kot tole: *Starši, varujte in ljubite svoje otroke*. Ko torej rečete, da je to prirojeno pravilo, kaj s tem mislite? Bodisi da je to prirojeno pravilo, ki ob vsaki priliki vzbudi in usmerja dejanja ljudi, bodisi da je resnica, ki jo imajo vsi ljudje vtisnjeno v svoj duh in ga zato poznajo in z njim soglašajo. Vendar v nobenem od teh pomenov, ni prirojeno. *Prvič*, da ni načelo, ki vpliva na dejanja vseh ljudi, sem dokazal s prej navedenimi primeri, niti nam ni treba iskati tako daleč, kot sta *Mingrelija* ali *Peru*, da bi odkrili primerke takega zanemarjanja, zlorabe, odklanjanja in uničevanja svojih otrok, niti gledati na to, kot kaj več kot na brutalnost nekaterih divjih in barbarskih narodov, ko se spomnimo, da je bila običajna in neobsojena praksa med *Grki* in *Rimljani*, da so brez usmiljenja in kesanja izpostavljali svoje nedolžne otroke. *Drugič*, da je to prirojena resnica, ki jo poznajo vsi ljudje, je tudi zmotno, kajti *Starši, varujte svoje otroke*, je tako daleč od prirojene resnice, da sploh ni nobena resnica, temveč je zapoved, ne pa propozicija in torej ni zmožna resnice ali neresnice. Da bi omogočili privolitev vanjo kot resnično, jo je treba preoblikovati v nekako tako propozicijo, kakršna je naslednja: *Dolžnost staršev je varovati svoje otroke*. Toda to, kaj je dolžnost, ni mogoče razumeti brez zakona, niti ni mogoče poznati ali predpostavljati zakona brez zakonodajalca ali brez nagrade in kazni, tako da ni mogoče, da bi moralo biti to ali katerokoli praktično načelo prirojeno, *to je*, vtisnjeno v duha kot dolžnost, ne da bi predpostavljali, da so *ideje* Boga, zakona, obveznosti, kazni, onstranskega življenja, prirojene. Kajti to, da kazen v tem življenju ne sledi kršitvi tega pravila in zato nima moč zakona v deželah, kjer mu splošno dovoljena praksa nasprotuje, je samo na sebi razvidno. Vendar so te *ideje* (ki morajo biti vse prirojene, če naj bo tako karkoli kot dolžnost) tako daleč od tega, da bi bile prirojene, da ni nobenega študioznega in premišljujočega človeka, še manj vsakega, ki se rodi, v katerem naj bi jih našli jasne in razločne. In to, da ena od njih, za katero se od vseh drugih zdi najbolj verjetno, da je prirojena, to ni (mislim na *idejo* Boga), mislim, da se bo v naslednjem poglavju izkazalo zelo razvidno vsakemu premišljujočemu človeku.

§ 13

Iz tistega, kar je bilo rečeno, tako mislim, lahko varno sklenemo, da za *katerokoli praktično pravilo, ki je na kateremkoli kraju splošno in z dopuščanjem kršeno, ne moremo domnevati, da je prirojeno*, kajti ni mogoče, da bi morali ljudje brez sramu ali strahu zaupljivo in mirno kršiti pravilo, za katerega si ne morejo kaj, da očitno ne bi vedeli, da ga je postavil Bog in bi gotovo kaznovali njegovo kršitev (kar morajo, če bi bilo prirojeno) v meri, ki bi omogočila zelo slabo kupčijo kršitelju. Brez take vednosti človek nikoli ne bi mogel biti gotov, da je karkoli njegova dolžnost. Nevednost ali dvom o zakonu, upanja, da se bomo izognili vednosti o ali moči zakonodajalca ali podobno, utegnejo povzročiti, da ljudje popustijo trenutnemu poželenju; vendar naj vsakdo vidi napako in šibo ob njej in s kršitvijo ogenj, ki ga je pripravljen kaznovati, skušajoče zadovoljstvo in roko vsemogočnega, vidno dvignjeno in pripravljeno, da se maščuje (kajti to se mora zgoditi, kjer je katerakoli dolžnost vtisnjena v duha), in potem mi povejte, ali je za ljudi s tako perspektivo, tako gotovo vednostjo, kot je ta, mogoče nesmotrno in brez predsodka kršiti zakon, ki ga nosijo s seboj v neizbrisnih značajih in ki strmi v njihov obraz, medtem ko ga kršijo? Ali lahko ljudje, hkrati ko čutijo v sebi vtisnjene edikte vsemogočnega zakonodajalca, z zaupanjem in veseljem prezirajo in s svojimi nogami gazijo svoje najbolj svete prepovedi? In naposled, ali bi bilo možno, da medtem ko neki človek odkrito kljubuje temu prirojenemu zakonu in najvišjemu zakonodajalcu, bi morale vse priče, da, celo voditelji in vladarji ljudi, ki polno v enakem smislu čutijo tako zakon kot zakonodajalca, molče gledati skozi prste, ne da bi izpričevali svoj odpor ali pa na to metali najmanjšo krivdo? Načela dejanj so res nastanjena v poželenjih ljudi, vendar so ta tako daleč od prirojenih moralnih načel, da če bi jim pustili, da se popolnoma razmahnejo, bi ljudi privedla do sprevračanja celotne morale. Moralni zakoni so postavljeni kot uzda in omejitev teh pretiranih želja, in tega ne morejo izpolniti drugače kot z nagradami in kaznimi, ki bodo pretehtale zadovoljstvo, ki bi ga kdorkoli predlagal zase pri kršenju zakona. Če bi bilo torej karkoli vtisnjeno v duh vseh ljudi kot zakon, bi morali vsi ljudje imeti neko gotovo in neizbežno vednost, da bo neka gotova in neizbežna kazen spremljala njegovo kršitev. Kajti če bi ljudje lahko bili nevedni o ali pa bi dvomili o tem, kaj je prirojeno, potem bi vztrajanje in poudarjanje prirojenih načelih ne imelo nobenega smisla; ta sploh ne zagotavljajo resnice in gotovosti (za kateri gre), toda ljudje so v enakem negotovem, plavajočem stanju tako z njimi kot brez njih. Prirojeni zakon mora spremljati očitna nedvomna vednost o neizbežni kazni, dovolj veliki, da naredi kršitev zelo nezaželeno, razen če ne predpostavljajo tudi prirojenega evangelija. Tukaj se ne bom motil, kot da zato, ker zavračam prirojeni zakon, mislim, da ni nobenih drugih kot pozitivnih zakonov. Obstaja velika razlika med prirojenim zakonom in zakonom narave; med nečem, vtisnjenim v našega duha pri samem nastanku, in nečem, kar ne poznamo, vendar lahko pridemo do vednosti o tem z uporabo in zaradi vprege naravnih zmožnosti. In menim, da enako puščajo na cedilu resnico tisti, ki naletijo na nasprotne skrajnosti, in bodisi potrjujejo prirojeni zakon bodisi zanikajo, da obstaja zakon, ki ga je mogoče spoznati z lučjo narave, to je, brez pomoči pozitivnega razodetja.

§ 14

Razlika, ki obstaja med ljudmi v njihovih praktičnih načelih, je tako očitna, da mi, tako mislim, ni treba reči ni več, da bi pokazal, da bo nemogoče najti kakršnakoli moralna pravila s tem znakom splošnega soglasja. Dovolj je, da kdo sumi, da je domneva o takih prirojenih načelih zgolj mnenje, sprejeto po mili volji, kajti tisti, ki tako zaupljivo govorijo o njih, so tako varčni, da *bi nam povedali, katera so ta načela*. To bi bilo lahko upravičeno pričakovati od tistih ljudi, ki močno poudarjajo to mnenje in daje priložnost, da ne zaupamo bodisi njihovi vednosti bodisi dobrohotnosti, ki so razglašajoč, da je Bog vtisnil v duh ljudi temelje vednosti in pravila življenja, vendarle zelo malo naklonjeni informiranju njihovih sosedov ali miru človeštva, saj jim ne pokažejo, katera med raznimi mnenji so tista, ki odvračajo ljudi. Toda če bi res obstajala kaka taka prirojena načela, ne bi bilo nobene potrebe po njihovem učenju. Če bi ljudje odkrili, da so take prirojene propozicije vtisnjene v njihov duh, bi jih bili zlahka zmožni razlikovati od drugih resnic, ki so se jih potem naučili in jih iz njih izpeljali in ne bi bilo nič manj lahko kot vedeti, kaj in koliko jih je. Ne bi moglo biti večjega dvoma o njihovem številu, kakršen je o številu naših prstov in potem je verjetno, da bi jih bil vsakršen sistem pripravljen našteti. Toda ker si nihče, kakor vem, doslej ni drznil dati njihovega seznama, ne moremo kriviti tistih, ki dvomijo o teh prirojenih načelih, ker nam celo oni, ki od ljudi zahtevajo, da verjamejo, da take prirojene propozicije obstajajo, ne povedo, kaj so. Zlahka je predvideti, da če bi morali različni ljudje različnih sekt skrbeti, da bi nam dali seznam tistih prirojenih praktičnih načel, bi zapisali zgolj taka, ki so primerna njihovim posebnim hipotezam in bi ustrezale podpori naukom njihovih partikularnih šol ali cerkev: očiten dokaz, da ni nobenih takih prirojenih resnic. Ne, velik del ljudi je daleč od tega, da bi v sebi samih našli taka prirojena moralna načela; z zanikanjem svobode človeštvu in s tem, da delajo ljudi za nič drugega kot zgolj stroje, ne odvzemajo zgolj prirojenih, temveč vsakršna moralna pravila in ne puščajo možnosti, da bi verjeli v kako tako pravilo, tistim, ki ne morejo razumeti, kako je lahko kakršnakoli stvar, ki ni svobodni agent, zmožna zakona. In na tej podlagi morajo nujno zavrniti vsa načela vrline tisti, ki ne morejo združiti *morale* in *mehanizma*, ki ju ni mogoče zlahka pomiriti ali uskladiti.

§ 15

Ko sem to napisal, sem bil obveščen, da je moj lord *Herbert* v svoji knjigi *De Veritate* nakazal ta prirojena načela; vprašal sem ga za svet in upal, da bom v človeku s tako veliko duhovitostjo našel kaj, kar bi me na tej točki utegnilo zadovoljiti in zaključiti mojo raziskavo. V njegovem poglavju *De instinctu naturali*, str. 76, izd. 1656, sem naletel na teh šest znakov njegovih *notitiæ Communes*, 1. *Prioritas*. 2. *Independientia.* 3. *Universalitas.* 4. *Certitudo.* 5. *Necessitas*, to je, kot jo razlaga, *faciunt ad hominis conservationem. 6. Modus conformationis*, to je *Assensus nullâ interpositâ morâ.* In na poznejšem koncu svoje male razprave *De Religione Laici* o teh prirojenih načelih pove tole: *Adeo ut non uniuscujusvis Religionis confinio arctentur quœ ubique vigent veritates. Sunt enim in ipsâ mente coelitàs descriptiœ nullisque traditionibus, sive scriptis, sive non scriptis, obnoxiœ*, str. 3. In, *Veritates nostrœ Catholicœ, quœ' tanquam indubia Dei effata in foro interiori descripta.* Ko je tako dal znake teh prirojenih načel ali skupnih pojmov in zatrdil, da jih je v duh ljudi vtisnila božja roka, nadaljuje, da bi jih zapisal, ti pa so naslednji: 1. *Esse aliquod supremum numen.* 2. *Numen illud coli debere.* 3. *Virtutem cum pietate conjunctam optimam esse rationem cultûs divini.* 4. *Resipiscendum esse à peccatis.* 5. *Dari prœmium vel poenam post hanc vitam transactam.* Čeprav dopuščam, da so to jasne resnice in take, če so pravilno razložene, da se racionalne bitje le stežka izogne temu, da z njimi soglaša, vendar mislim, da je on daleč od tega, da bi dokazal, da so to prirojeni vtisi *in Foro interiori descriptœ*, kajti moram si dovoliti pripomniti:

§ 16

Prvič, da teh pet propozicij ali niso vse ali pa jih je več kot tistih skupnih pojmov, ki jih je v našega duha zapisal božji prst, če bi bilo razumno verjeti, da je bila sploh kakšna tako zapisana. Obstajajo še druge propozicije, ki imajo celo po njegovih lastnih pravilih ravno tako pravico biti tako prvotne in ravno tako sprejete za prirojena načela kot vsaj nekatere od teh petih, ki jih našteva, *namreč*, *Delajte, kar si želite, da bi delali vam* in morda nekaj sto drugih ob dobrem premisleku.

§ 17

Drugič, da vseh njegovih znakov ni najti v vsaki od njegovih petih propozicij, *namreč* njegov prvi, drugi in tretji znak se sploh ne ujema z nobeno od njih, prvi, drugi, tretji, četrti in šesti se slabo ujemajo z njegovo tretjo, četrto in peto propozicijo. Kajti poleg tega, da nas zgodovina mnogih ljudi, ne, celotnih narodov, uči, da so dvomili ali niso verjeli v nekatere ali vse, sam ne morem uvideti, kako je lahko tretja, *namreč, Da je vrlina, združena z usmiljenjem, najboljše čaščenje Boga*, prirojeno načelo, ko je tako težko razumeti ime ali zvok besede *vrlina* in je podvržena tako veliki negotovosti v svojem pomenu ter se o stvari, za katero stoji, tako močno razpravlja in jo je težko poznati. In zato to ne more biti nič drugega kot zelo negotovo pravilo človeške prakse in rabi zgolj zelo malo za vodenje naših življenj in ga je zato zelo neprimerno določiti za prirojeno praktično pravilo.

§ 18

Obravnavajmo to propozicijo v njenem pomenu (saj je smisel, ne pa zvok tisti, ki je in mora biti načelo ali skupen pojem), *to je, Vrlina je najboljše čaščenje Boga,* je namreč zanj najbolj sprejemljiva, ki bo – če za *vrlino* štejemo, kot je to najbolj običajno – tista dejanja, ki jih v skladu z različnimi mnenji več dežel imamo za hvalevredne, propozicija, tako daleč od tega, da bi bila gotova, da ne bo resnična. Če bi *vrlino* imeli za dejanja, skladna z božjo voljo ali s pravilom, ki ga predpisuje Bog, ki je resnično in edino merilo vrline, ko je vrlina uporabljena, da pomeni tisto, kar je po svoji lastni naravi prav in dobro, potem bo ta propozicija *Da je vrlina najboljše čaščenje Boga* najbolj resnična in gotova, vendar zelo malo uporabna v človekovem življenju, saj ne bo pomenila nič več, kot *namreč, Da je Bog zadovoljen z delovanjem, ki ga zapoveduje*; človek lahko gotovo ve, da je resnična, ne da bi vedel, kaj je tisto, kar Bog dejansko zapoveduje in tako je tako daleč od vsakršnega pravila ali načela svojih dejanj, kot je bil prej. Mislim pa, da bodo le redki imeli propozicijo, ki ne pomeni nič več od tega, *namreč,* Da je Bog zadovoljen z delovanjem, ki ga zapoveduje, za prirojeno moralno načelo, zapisano v duh vseh ljudi (kakorkoli resnično in gotovo že utegne biti), ker nauči tako malo. Kdorkoli dela tako, bo imel razlog za razmišljanje, da je na stotine propozicij, prirojenih načel, ker jih je mnogo, ki imajo ravno tako dober naziv kot ta, da bi bile sprejete kot take, ki jih še nihče nikoli ni postavil v rang prirojenih načel.

§ 19

Niti ni veliko bolj poučna četrta propozicija, (*namreč*) *Ljudje se morajo pokesati svojih grehov*, dokler se ne določi, katera so tista dejanja, ki so mišljena kot grehi, kajti beseda *peccata* ali grehi, je postavljena, ko je to običajno, da pomeni na splošno slaba dejanja, ki bodo pritegnila kazen na tiste, ki jih delajo; kakšno veliko načelo morale je lahko to, ki nam pove, da nam bi moralo biti žal in da bi morali s tem nehati, kar nam bo prineslo nesrečo, ne da bi vedeli, katera so tista posebna dejanja, ki bodo to storila? Res je, to je zelo resnična propozicija in je ustrezna, da bi si jo vbili v glavo in sprejeli tisti, za katere se domneva, da so jih naučili, katera dejanja vseh vrst *so grehi*; vendar si je mogoče zamisliti, da niti ta niti prejšnja nista prirojeni načeli niti nista kakorkoli uporabni, če bi bili prirojeni, razen če ne bi bili v duh ljudi vtisnjena partikularna merila in meje vseh vrlin in grehov in bi bila prirojena načela tudi tista, o katerih, tako mislim, gre zelo dvomiti. In zato menim, da bi se zdelo stežka mogoče, da bi Bog moral vtisniti načela v duh ljudi z besedami, ki imajo negotov pomen, kakršni sta *vrline* in *grehi*, ki med različnimi ljudmi stojijo za različne stvari. Ne, ne more se domnevati, da se to sploh dogaja s pomočjo besed, ki so v teh načelih zelo splošna imena in jih ni mogoče razumeti, če ne poznamo partikularij, ki jih obsegajo. In v praktičnih primerih je treba merila vzeti iz vednosti o samih dejanjih in iz njihovih pravil, povzetih iz besed in pred poznavanjem imen, ta pravila pa mora človek poznati ne glede na to, katerega jezika se je kakorkoli po naključju naučil, bodisi angleščine bodisi japonščine, ali pa če se sploh ni naučil nobenega ali če ni nikoli razumel rabe besed, kot se zgodi v primeru gluhih in nemih ljudi. Ko bo prepoznano, da ljudje, ki ne poznajo besed ali jih niso naučili zakonov in običajev njihove dežele, vedo, da je del čaščenja Boga, da ne ubiješ drugega človeka, da nimaš stikov z več kot eno žensko, da ne povzročaš splava, da ne izpostavljaš otrok, da od drugega ne jemlješ, kar je njegovega, čeprav si to želiš zase, temveč nasprotno, rešiš in zadovoljiš njegove potrebe in kadarkoli si storil nasprotno, se moraš pokesati, ti mora biti žal, in moraš skleniti, da tega ne boš več počel. Ko pravim, da bo dokazano, da vsi ljudje dejansko poznajo in dopuščajo vsa ta in tisoč drugih takih pravil, od katerih vsa sodijo pod ti dve splošni besedi, ki sta bili uporabljeni zgoraj, *namreč pod virtutes et peccata, pod vrline in grehe*, bo več razlogov za dopuščanje, da so ta in podobna pravila splošni/skupni pojmi in praktična načela, saj bo naposled univerzalno soglasje (če bi kaj takega bilo glede moralnih načel) o resnicah, o katerih je poznavanje mogoče pridobiti drugače, le stežka dokazalo, da so prirojene, kar je vse, za kar se borim.

§ 20

Tu niti ne bo kaj pomembno, če ponudimo tisti zelo pripravljen, vendar ne zelo bistven odgovor, (*namreč*) *da utegnejo prirojena načela zatemniti* in naposled *povsem ponositi duh ljudi z izobrazbo in običaji* ter s splošnim mnenjem tistih, med katerimi živimo. Zatrjevanje teh načel, če je resnično, povsem odstranjuje argument univerzalnega soglasja, s katerim se poskuša dokazati to mnenje o prirojenih načelih, razen če ne bi tisti ljudje mislili, da je razumno, da bi se njihova zasebna prepričanja ali prepričanja njihove stranke morala izdajati za univerzalno soglasje, stvar, ki se je ne počne redko, ko ljudje, ki domnevajo, da so oni sami edini mojstri pravega razuma, zavržejo glasove in mnenja ostalega človeštva kot nevredna upoštevanja. In potem je njihov argument naslednji: Načela, ki jih vse človeštvo dopušča za resnične, so prirojena; tista, ki jih dovoljujejo ljudje pravega razuma, so načela, ki jih dopušča vse človeštvo; mi in tisti, ki so naših misli, smo ljudje razuma, zato soglašamo, da so naša načela prirojena, kar je zelo lep način argumentiranja in bližnjica do nezmotljivosti. Kajti sicer bi bilo zelo težko razumeti, kako obstajajo nekatera načela, ki jih vsi ljudje dejansko priznavajo in z njimi soglašajo, pa vendarle ni nobenega od teh *načel,* ki jih iz duha mnogih ljudi *izbriše pokvarjen običaj in slaba izobrazba*, kar pomeni reči, da jih vsi ljudje dopuščajo, vendar pa jih veliko ljudi zanika in se z njimi ne strinja. In res nam bo predpostavka takih prvih načel le malo rabila za naš namen, mi pa bomo ravno toliko v zadregi z njimi ali pa brez njih, če bi jih lahko kaka človeška sila, kakršna so volja naših učiteljev ali mnenja naših tovarišev, v nas spremenila ali povzročila njihovo izgubo. In kljub vsemu temu bahanju s prvimi načeli in naravno lučjo, bomo ravno tako v temi in negotovosti, kot če sploh ne bi bilo nobene take stvari, saj je enako, če nimamo nobenega pravila in takega, ki bi se od vsepovsod skrivilo, ali pa med raznimi in nasprotnimi pravili ne bi vedeli, katero je pravo. Vendar kar zadeva prirojena načela, si želim, da bi ti ljudje povedali, ali jih je mogoče ali pa ne z izobrazbo in običajem zamegliti ali izbrisati: če jih ne moremo, bi jih morali enako najti v vsem človeštvu in morajo biti jasno v vsakem človeku, če pa jih lahko spremenijo naključna pojmovanja, jih moramo potem najti najbolj jasne in razločne najbliže izviru, v otrocih in nepismenih ljudeh, ki so prejeli najmanj vtisov tujih mnenj. Ti naj zavzamejo tisto stran, ki jo želijo; gotovo bodo odkrili, da ni združljiva z vidnim stanjem stvari in vsakdanjim opazovanjem.

§ 21

Zlahka priznavam, da obstaja veliko število *mnenj*, ki jih ljudje različnih dežel, izobrazbe in temperamentov sprejemajo in se jih oklepajo *kot prvih in nedvomnih načel, za mnoga od katerih*, tako zaradi njihove nesmiselnosti kot tudi nasprotovanja drug drugemu, *ni mogoče, da bi morala biti resnična.* Vendar pa so vse te propozicije, kakorkoli že oddaljene od razuma, ponekod tako svete, da se bodo ljudje, ki imajo dober razum v drugih pogledih, prej ločili od svojega življenja in česarkoli njim najljubšega, kot pa si dovolili dvomiti ali drugim preizpraševati njihovo resnico.

§ 22

To je tisto, kakorkoli čudno že utegne biti, kar vsak dan potrjuje izkustvo in se morda ne bo zdelo tako čudovito, če pogledamo *načine* in korake, *kateri* so to povzročili, in kako se lahko dejansko zgodi, da *nauki*, ki so bili izpeljani iz nič boljšega izvira, kot je praznoverje dojilje ali avtoritete starke, lahko sčasoma in s soglasjem sosedov *zrastejo do dostojanstva načel* v religiji ali morali. Kajti tisti, ki pazijo (kot pravijo), da bi bila načela dobra za otroke (in malo je tistih, ki nimajo zanje nabor takih načel, v katere verjamejo), vcepljajo v brezskrben, vendar še vedno v razum brez predsodkov (kajti prazen papir sprejema vsakršne črke) tiste nauke, ki bi jih otroci obdržali in izpovedovali. Te nauke jih učijo takoj, ko so zmožni kakršnegakoli dojemanja in še vedno takrat, ko rastejo, ki jim jih potrjujejo bodisi z očitnim izpovedovanjem bodisi tihim soglasjem vsi, s katerimi imajo opraviti, ali vsaj tisti, o katerih modrosti, vednosti in usmiljenju si ustvarijo neko mnenje, tisti, ki nikoli ne dopuščajo, da bi bile te propozicije omenjene drugače, temveč kot osnova in temelj, na katerem gradijo svojo religijo ali običaje; tako pridejo do reputacije nedvomnih, samorazvidnih in prirojenih resnic.

§ 23

K čemur lahko dodamo, da ko *ljudje*, tako poučeni, zrastejo in premišljujejo o svojem lastnem duhu, tam ne morejo najti ničesar starejšega od teh mnenj, ki so jih učili, preden je njihov spomin začel vzdrževati register njihovih dejanj ali zapisovati čas, ko se jim je pojavila katerakoli nova stvar, in zato nimajo nobenih predsodkov do *sklepa*, *da so bile tiste propozicije, o katerih vednosti ne morejo najti v sebi nobenega izvira, gotovo vtis Boga in narave na njihovega duha* in da jih ni naučil nihče drug. Jih vzdržujejo in se jim podrejajo, kot se mnogi svojim staršem, z globokim spoštovanjem, ne zato, ker je to naravno, niti otroci tega ne delajo tam, kjer niso tako naučeni, temveč zato, ker so bili vselej tako naučeni in nimajo nobenega spomina o začetku tega spoštovanja, mislijo, da je to naravno.

§ 24

To se bo zdelo zelo verjetno in se bo domala neizogibno zgodilo, če pogledamo naravo človeštva in konstitucijo človeških zadev, kjer *večina ljudi ne more živeti, ne da bi uporabili svoj čas v dnevnem delu svojih poklicev niti imeti mirnega svojega duha brez nekaterih temeljev ali načel, na katera oprejo svoje misli.* Stežka je kdo, ki bi bil tako nestanoviten in površen v svojem razumu, ki ne bi imel nekaterih spoštljivih propozicij, ki so zanj načela, na katerih temeljijo njegova dokazovanja in s katerimi sodi o resnici in neresnici, o pravici in krivici; nekaterim manjka veščina in prosti čas, drugim nagnjenje, nekatere pa so poučili, da jih ne bi smeli preučevati; malo je takih, ki ne bi izpostavljali svoje nevednosti, lenosti, izobrazbe ali prenagljenosti, da *bi jih sprejeli z zaupanjem.*

§ 25

To je očitno primer z vsemi otroki in mladimi; ker pa običaju, večji sili od narave, redko spodleti pri tem, da bi ga častili po božje, kar jih je navadilo, da priklonijo svojega duha in podredijo svoj razum, ni čudno, da odrasli *ljudje*, bodisi zbegani v nujnih zadevah življenja bodisi vneti pri zasledovanju zadovoljstva, *ne* bodo resno sedli, da bi *preučevali svoja lastna načela*, posebej tedaj, ko je eno od njihovih načel, da o načelih ne bi smeli dvomiti. In čeprav bi ljudje imeli čas, zmožnost in voljo, kdo bo domala tisti, ki si bo drznil pretresti temelje vseh svojih preteklih misli in dejanj in se sprijazniti s sramoto, ki bi si jo nakopal, da je bil dolgo časa popolnoma v zmoti in nerazumevanju? Kdo je dovolj trden, da se bo boril z grajo, ki je povsod pripravljena za tiste, ki si drznejo izraziti nesoglasje z sprejetimi mnenji svoje dežele ali stranke? In kje gre najti človeka, ki se lahko potrpežljivo pripravi, da nosi ime muhastega, skeptičnega ali ateista, s katerim se bo gotovo srečal tisti, ki ima najmanjši pomislek o vsakem skupnem/splošnem mnenju? Bolj pa se bo *bal podvomiti o teh načelih*, ko bo zanje mislil, kot misli večina ljudi, da so to standardi, ki jih je Bog postavil v njegovega duha, da bi bili pravilo in temeljni kamen vseh drugih mnenj. Kaj pa ga lahko ovira, da ne misli, da so sveta, ko ugotovi, da so najbolj zgodnje od vseh njegovih lastnih misli in da jih najbolj spoštujejo drugi?

§ 26

Zlahka si je zamisliti, *kako* se tako zgodi, da *ljudje* častijo idole, ki so se oblikovali v njihovem duhu, se spoprijateljijo s pojmi, s katerimi so bili že dolgo seznanjeni in *vtisnejo značaj božanstva nesmislom in napakam*, postanejo goreči častilci bikov in opic in tudi oporekajo, se borijo in umirajo pri obrambi svojih mnenj. *Dum solos credit habendos esse Deos, quos ipse colit.* Ker zmožnosti premišljevanja duše, ki se domala stalno, čeprav ne vselej niti previdno niti modro uporabljajo, ne bodo vedele, kako se gibati zaradi pomanjkanja temelja in opore pri večini ljudi, ki zaradi lenosti ali razvedrila ne prodrejo ali zaradi pomanjkanja časa ali resnične pomoči ali zaradi drugih razlogov ne morejo prodreti v načela vednosti in slediti resnici do njenega vrelca in izvira, je zanje naravno in domala neizogibno, da se oprimejo nekaterih sposojenih načel, za katera se misli, ker so znana in za katera se domneva, da so očitni dokazi drugih stvari, da zase ne potrebujejo nobenega drugega dokaza. Kdorkoli bo sprejel kakšnega od njih v svojega duha in jih tam ohranil z čaščenjem, s katerim običajno nagrajujemo načela, ne da bi se jih kdaj drznili preučiti, temveč se prilagodimo, da vanje verjamemo, ker je vanje treba verjeti, se utegne zaradi svoje izobrazbe in mode svoje države okleniti vsakršnega nesmisla kot prirojenega načela, z dolgim razglabljanjem o istih objektih pa je njegov pogled tako zamegljen, da ima pošasti, nastanjene v svojih lastni možganih, za podobe božanstva in delo njegovih rok.

§ 27

Na tak način lahko zlahka opazimo, koliko je takih, ki pridejo do načel, za katera verjamejo, da so prirojena, v raznolikosti nasprotnih načel, ki se jih držijo in zanje borijo vse vrste in stopnje ljudi. Tisti pa, ki bo zanikal, da je to metoda, s katero večina ljudi pride do jamstva, ki ga imajo o resnici in razvidnosti njihovih načel, bo nemara ugotovil, da je težko na katerikoli drugi način razložiti nasprotna načela, v katera trdno verjamejo, z zaupanjem potrjujejo in katera je veliko število ljudi pripravljenih kadarkoli zapečatiti s svojo krvjo. In res, če je privilegij prirojenih načel, da jih sprejmejo zaradi njihove lastne avtoritete, brez preučevanja, ne vem, v kaj ne bi mogli verjeti ali kako je mogoče dvomiti o kateremkoli *načelu*. Če jih je mogoče in *bi jih morali preučevati* in preizkušati, želim vedeti, kako je mogoče preizkušati prva in prirojena načela ali pa je vsaj razumno zahtevati znake in lastnosti, s katerimi je mogoče prava, prirojena načela razlikovati od drugih; da se tako med veliko raznolikostjo pretendentov lahko obvarujem pred napakami pri tako pomembni točki, kot je tale. Ko je to storjeno, se bom pripravljen okleniti takih dobrodošlih in koristnih propozicij, do takrat pa lahko zmerno dvomim, saj se bojim, da bo univerzalno soglasje, ki je edino ustvarjeno, le stežka pokazalo zadosten znak, da bi vodil mojo izbiro in mi zagotavljal kakršnakoli prirojena načela. Iz tistega, kar je bilo rečeno, mislim, da je onstran dvoma, da ni nobenih praktičnih načel, s katerimi soglašajo vsi ljudje in torej nobenih prirojenih načel.

Četrto poglavje

Drugi premisleki, ki zadevajo prirojena načela, tako spekulativna kot praktična

§ 1

Če bi jih tisti, ki nas hočejo prepričati, da obstajajo prirojena načela, ne vzeli skupaj na debelo, temveč bi ji obravnavali ločeno, po delih, iz katerih so te propozicije sestavljene, morda ne bi bili tako pripravljeni verjeti, da so prirojene. Kajti če *ideje*, ki tvorijo te resnice, ne bi bile prirojene, ne bi bilo mogoče, da bi morale biti propozicije, sestavljene iz njih, prirojene ali da bi se naše njihovo poznavanje rodilo z nami. Kajti če *ideje niso prirojene*, je bil čas, ko je bil duh brez tistih načel in potem ne bodo prirojene, temveč izpeljane iz nekega drugega vira. Kajti tam, kjer same *ideje* niso prirojene, ne more biti nobene vednosti, nobenega soglasja, nobenih mentalnih ali besednih propozicij o njih.

§ 2

Če bi pazljivo obravnavali *novorojence,* bi imeli malo razloga misliti, da s sabo prinesejo na svet mnogo *idej.* Nemara z izjemo nekaj bledih *idej* za lakoto in žejo ter toplino in nekaterih bolečin, ki so jih utegnili *čutiti* v maternici, *ni* v njih sploh najmanjšega pojava kakršnekoli določene *ideje*, posebej *ideje, ki bi ustrezala terminom, ki tvorijo tiste univerzalne propozicije*, ki so cenjena prirojena načela. Kasneje lahko postopno zaznamo, kako *ideje* pridejo v njihovega duha in da ne dobijo niti nič več niti nič drugega, s čimer jih oskrbi izkustvo in opazovanje stvari, ki jim pridejo na pot, kar utegne biti dovolj, da nas zadovolji, da to niso izvirni znaki, vtisnjeni v duha.

§ 3

Ni mogoče, da bi ista stvar bila in ne bila je gotovo (če je sploh kaj takega) prirojeno načelo. Toda ali lahko kdorkoli misli ali bo kdorkoli rekel, da sta *nemožnost* in *identiteta* dve prirojeni *ideji*? Ali sta taki, da ju ima vse človeštvo in ju prinese s sabo na svet? In ali sta tisti, ki sta prvi v otrocih in pred vsemi pridobljenimi idejami? Če sta prirojeni, morata nujno biti taki. Ali ima otrok *idejo nemožnosti* in *identitete*, preden ima idejo *belega* ali *črnega*, *sladkega* ali *grenkega*? In ali iz poznavanja tega načela sklepa, da pelin, če ga drgnemo ob bradavico, nima enakega okusa, ki ga običajno dobiva od njega? Dejansko poznavanje *impossibile est idem esse, et non esse* je tisto, ki stori, da otrok razlikuje med svojo materjo in tujcem ali stori, da je ljubeč do ene in beži pred drugim? Ali pa duh uravnava samega sebe in svoje soglasje z *idejami*, ki jih ni imel še nikoli? Ali razum izpeljuje sklepe iz načel, ki jih doslej še ni poznal ali razumel? Imeni *nemožnosti* in *identitete* stojita za dve *ideji*, tako *daleč od tega, da bi bili prirojeni* ali rojeni z nami, da mislim, da zahtevata veliko skrbi in pozornosti, da bi ju pravilno oblikovali v našem razumu. Sta tako daleč od tega, da bi prišli na svet z nami, tako daleč od misli otroštva in prve mladosti, da bomo, tako sem prepričan, po raziskovanju ugotovili, da manjkata mnogim odraslim.

§ 4

Če naj bo *identiteta* (če ostanemo zgolj pri tem primeru) prirojeni vtis in nam zato tako jasna in očitna, da jo moramo poznati celo iz svoje zibelke, bom vesel, če bom dobil razlago od enega, ki je star sedem ali sedemdeset let, ali je neki človek, ki je bitje, sestavljeno iz duše in telesa, isti človek, ko je njegovo telo spremenjeno? Ali sta bila *Evforbij* in *Pitagora*, ki sta imela isto dušo, isti človek, čeprav sta živela več obdobij narazen? In celo, ali ne bi bil tudi petelin, ki ima isto dušo, ista stvar z obema? S čimer se nemara zdi, da naša *ideja istosti ni tako* trdna in jasna, da bi zaslužila, da bi zanjo mislili, da nam je *prirojena*. Kajti če tiste prirojene *ideje* niso jasne in razločne, tako da so univerzalno znane in da o njih naravno soglašamo, ne morejo biti predmet univerzalnih in nedvomnih resnic, temveč bodo neizogibna priložnost za nenehno negotovost. Torej predpostavimo, da *ideja identitete* kogarkoli, ne bo enaka ideji, ki so jo imeli *Pitagora* in tisoči drugih njegovih privržencev: katera bo potem resnična? Katera prirojena? Ali pa sta dve različni *ideji identitete*, obe prirojeni?

§ 5

Nikar ne komerkoli dopustiti misliti, da so vprašanja, ki sem jih tu zastavil o *identiteti* človeka, čiste, prazne spekulacije, kajti če bi bile, bi zadostovalo pokazati, da v razumu ljudi ni bilo *nobene prirojene* ideje *identitete*. Tisti, ki bo z malo pozornosti premišljeval o vstajenju in bo sodil, da bo božja pravičnost v poslednjem dnevu pripeljala pred sodbo taiste osebe, da bodo srečne ali nesrečne v onstranstvu, ki so delali dobro ali slabo v tem življenju, bo nemara ugotovil, da se v samem sebi ni lahko odločiti, kaj je tisto, kar naredi istega človeka ali iz česa sestoji *identiteta* in ne bo pripravljen misliti, da ima on in vsakdo, celo sami otroci, naravno jasno *idejo* o tem.

§ 6

Preiščimo načelo matematike, *namreč, Da je celota večja od dela.* To se šteje med prirojena načela. Prepričan sem, da je ravno tako upravičeno kot katerokoli drugo, da zanj tako mislimo, ki pa vendarle nihče ne more misliti, da je tako, ko meni, da so *ideje*, ki jih v sebi obsega, *celota* in *del*, popolnoma relativne; toda pozitivne *ideje*, kamor ustrezno in neposredno sodita, sta ekstenzija in število, od katerih sta zgolj *celota* in *del* relaciji. Tako da če sta *celota* in *del* prirojeni *ideji*, mora to biti tudi ekstenzija in število, saj ni mogoče imeti *idejo* relacije, ne da bi sploh imeli kake ideje o stvari, h kateri sodi in na kateri temelji. Torej to, ali ima duh ljudi naravno vtisnjene v sebe *ideji* ekstenzije in števila, prepuščam v premislek tistim, ki so zavetniki prirojenih načel.

§ 7

To *da je Boga treba častiti*, je nedvomno med največjimi resnicami, ki lahko vstopijo v duha človeka in zasluži prvo mesto med vsemi praktičnimi načeli. Vendar zanjo nikakor ni mogoče misliti, da je prirojena, razen če sta *ideji Boga* in *čaščenja* prirojeni. To, da *ideja*, za katero stoji termin *čaščenje*, ni v razumu otrok in znak, vtisnjen v duha v njegovem prvem nastanku, tako mislim, bo zlahka dopustil vsakdo, ki premišlja, kako malo je med odraslimi ljudmi tistih, ki imajo o njej jasen in razločen pojem. Domnevam pa, da ne more biti nobene bolj smešne stvari, kot reči, da imajo otroci to praktično načelo, *Da je Boga treba* častiti, prirojeno in vendarle da ne vedo, kaj je to čaščenje Boga, kar je njihova dolžnost. Toda dovolj o tem.

§ 8

Če si lahko zamislimo, da je lahko kaka *ideja prirojena*, si lahko mislimo, da je taka od vseh drugih idej zaradi mnogih razlogov *ideja Boga*, saj je težko dojeti, kako bi morala obstajati prirojena moralna načela brez prirojene *ideje božanstva*. Brez pojma zakonodajalca ni mogoče imeti pojem zakona in obveznosti, da se ga držimo. Poleg ateistov, ki so bili opaženi v antiki in zapisani v registre zgodovine, ali niso plovbe v poznejših dobah odkrile celotna ljudstva v zalivu (α) *Soldanija*, (β)v Braziliji, (γ) v Borandayju in na karibskih otokih *itn.*, med katerimi ni bilo mogoče najti nobenega pojma boga, nobene religije? *Nicolaus del Techo* v *Literis, ex Paraquaria de Caaiguarum conversione* izreče naslednje besede: (δ) *Reperi eam gentem nullum nomen habere, quod Deum, et Hominis animam significet, nulla sacra habet, nulla Idola*. To so primeri ljudstev, kjer je bila nekultivirana narava prepuščena sama sebi, brez pomoči književnosti in discipline in izboljšav umetnosti in znanosti. Vendar gre najti tudi druga ljudstva, ki so jim bile te stvari v veliki meri všeč, ki pa jim je vendarle zaradi pomanjkanje primerne uporabe svojih misli na ta način, manjkala *ideja* in vednost o bogu. Vsekakor ne dvomim, da bo za druge presenečenje, kot je bilo zame, ugotovitev, da sodijo mednje *Siamci*. Toda za to naj poizvedujejo pri zadnjem odposlancu francoskega kralja tamkaj (ε), ki ne daje nobene boljše razlage za same *Kitajce* (ζ). In če ne bomo verjeli *La Loubere*, misijonarji na Kitajskem, tudi sami jezuiti, veliki poveličevalci *Kitajcev*, vsi do zadnjega soglašajo in nas bodo prepričali, da so sekta *literatov* ali *učenih*, ohranjevalcev stare religije *Kitajcev* in tam vladajoča partija, vsi *ateisti*. Glej *Navarette* v *Collection of Voyages*, 1. zv. in *Historia Cultus Sinensium*. Če pa bi se morali pozorno meniti za življenja in razprave ljudi, ki niso tako daleč, bi morali imeti prevelik razlog za strah, da mnogi v civiliziranih deželah nimajo zelo močnih in jasnih vtisov božanstva na svoj duh in da pritožbe proti ateizmu, izrečene s prižnice, niso brez razloga. In čeprav ga zdaj preveč odkrito priznavajo zgolj nekateri izprijeni nesrečniki, bi vendarle morali o njem nemara slišati več, kot slišimo, od drugih, če se ne bi bali meča magistrata in cenzure sosedov, ki zaveže ljudem jezike, ki bi, če bi bila strah pred kaznijo in sramoto, odstranjena, tako odprto razglašali svoj *ateizem*, kot to počnejo njihova življenja.

§ 9

Če bi vse človeštvo vsepovsod imelo *pojem Boga* (o čemer nam zgodovina doslej govori nasprotno), iz tega *ne* bi sledilo, da je bila *ideja* o njem *prirojena*. Kajti čeprav ne bi zanj našli nobenega naroda brez imena in brez neke peščice temnih pojmov, to vendarle ne bi dokazovalo, da so naravni vtisi na duha, nič bolj kot imena za ogenj ali sonce, vročino ali število ne dokazujejo, da so *ideje*, za katere stojijo, prirojene zato, ker so imena teh stvari in *ideje* o njih tako univerzalno sprejete in znane med človeštvom. Niti ni, nasprotno, izostanek takega imena ali odsotnost takega pojma v duhu ljudi kakršenkoli argument proti bivanju Boga, nič bolj kot ne bi bil dokaz za to, da ni na svetu nobenega magneta, ker velik del človeštva ne bi imel niti pojma o kaki taki stvari niti imena zanjo; kot ne bi bil noben argument za dokaz, da ni nobenih različnih in raznih vrst angelov ali inteligentnih bitij nad nami, dejstvo, da nimamo nobenih *idej* o takih različnih vrstah ali imen zanje. Ker skupen jezik njihovih lastnih dežel ljudem nudi besede, se zato stežka izognejo temu, da bi imeli nekakšno vrsto *idej* teh stvari, katerih imena imajo pogosto priložnost omeniti tistim, s katerimi se pogovarjajo. Če pa *ideja* nosi s seboj pojem odličnosti, sijajnosti ali nekakšne izjemnosti, če jo spremljata bojazen in skrb, če se strah absolutne in nevzdržne moči vtisne v duha, je verjetno, da bo sedla globlje in se razširila dlje, posebej če je taka *ideja*, ki je v soglasju s skupno/splošno lučjo razuma in če je naravno izpeljiva iz vsakega dela naše vednosti, kot je *ideja* Boga. Kajti vidni znaki izjemne modrosti in moči, se pojavljajo tako jasno v vseh delih stvarjenja, da racionalno bitje, ki bo zgolj o njih premišljevalo, ne more zgrešiti odkritja *božanstva*. In vpliv, ki ga odkritje takega bitja mora nujno imeti na duh vseh, ki so zgolj enkrat slišali o njem, je tako velik in nosi s seboj tako težo misli in komunikacije, da se mi zdi nenavadno, da bi se kjerkoli našel tako neomikan celoten narod ljudi, da bi jim manjkal pojem Boga, kot pa da bi morali biti brez vsakršnega pojma števil ali ognja.

§ 10

Ime Boga, ko je enkrat omenjeno v kateremkoli delu sveta, da bi izrazilo superiorno, močno, modro, nevidno bitje, primernost takega pojma za načela skupnega/splošnega razuma in interes ljudi, ki ga bodo vselej imeli, da ga pogosto omenjajo, morajo nujno razširiti daleč in široko in ga prenesti na vse generacije, čeprav *splošen sprejem tega imena in nekaterih nepopolnih in netrdnih pojmov, tako sporočenih* nemislečemu delu človeštva, *ne dokazuje, da je ta ideja prirojena*, temveč zgolj to, da so tisti, ki so prišli do tega odkritja, prav uporabili svoj razum, zrelo razmišljali o vzrokih stvari in jim sledili do njihovega izvira; ko so drugi manj razmišljujoči ljudje od teh enkrat prejeli tako pomemben pojem, ga ni mogoče spet zlahka izgubiti.

§ 11

To je vse, kar bi lahko izpeljali iz pojma *Boga*, če naj bi ga našli univerzalno v vseh plemenih človeštva in bi ga zreli ljudje splošno priznavali v vseh deželah. Kajti splošnost priznanja Boga, kot si domišljam, ne sega nič dlje od tega, kar bo, če zadostuje za dokaz, da je *ideja Boga prirojena*, ravno tako dokazalo, da je prirojena *ideja ognja*, prirojena zato, ker, tako mislim, je mogoče resnično reči, da ni nobene osebe na svetu, ki ima pojem Boga in ne bi imela tudi *ideje* ognja. Ne dvomim, toda če bi morali kolonijo mladih otrok namestili na otok, kjer ne bi bilo ognja, gotovo noben ne bi imel kakršnegakoli pojma take stvari niti imena zanjo, kakorkoli splošno bi jo sprejeli in bi bila znana vsemu svetu poleg in morda bi bila tudi njihovo razumevanje tako daleč od vsakršnega imena ali pojma Boga, dokler ne bi eden od njih uporabil svojih misli za raziskovanje konstitucije in vzrokov stvari, ki bi ga zlahka vodila k pojmu *Boga*, ki ko bi enkrat o njem poučil druge, bi ga potem razum in naravna nagnjenost njihovih lastnih misli širila in nadaljevala med njimi.

§ 12

Trdovratno vztrajajo, da je *primerno za dobroto Boga, da v duh ljudi vtisne znake in pojme samega sebe*, in da jih ne pusti v temi in dvomu v tako pomembnem opravku ter da si s to pomočjo tudi zagotovi sebi poklon in slavljenje, ki mu ga dolguje tako inteligentno bitje kot človek, zato je storil tako.

Ta argument, če naj ima kakšno moč, bo dokazal veliko več kot tisti, ki ga v tem primeru uporabljajo, od njega pričakujejo. Saj če lahko sklepamo, da je *Bog* naredil za ljudi vse tisto, kar bodo ljudje sodili, da je najbolje zanje, ker je za njegovo dobroto primerno, da dela tako, bo to dokazalo ne zgolj, da je Bog vtisnil v duh ljudi *idejo* samega sebe, temveč da je tja jasno vtisnil z lepimi črkami vse, kar bi ljudje morali vedeti ali o njem verjeti, vse, kar bi morali delati v poslušnosti do njegove volje in da jim je dal voljo in ustrezne naklonjenosti. To je, kot bo nedvomno mislil vsakdo, boljše za ljudi, kot da bi morali v temi tipati za vednost, kot nam *sveti Pavel* govori, da so vsi narodi tipali za Bogom, *Apostolska dela*, 17, 27, kot da bi se njihova volja morala spopadati z njihovim razumom, njihova poželenja pa prečiti njihovo dolžnost. *Rimski katoliki* pravijo, da je najbolje za ljudi in tako primerno za dobroto Boga, da bi moral obstajati nezgrešljiv sodnik o kontroverzah na zemlji, zato je eden, jaz pa zaradi istega razloga pravim, da je za ljudi boljše, da bi moral biti vsak človek nezmotljiv. Prepuščam jim premislek, ali bodo z močjo tega argumenta mislili, da je vsak človek tak. Mislim, da je zelo dober argument reči, da je neskončno moder Bog storil tako in je zato najboljše. Vendar *se mi zdi, da je nekoliko preveliko zaupanje v svojo lastno modrost, če rečemo, da mislim, da je to najboljše in zato je Bog storil tako,* v zadevi pa, ki jo imamo pri roki, bi bilo zaman dokazovati iz takega predmeta razprave, da je Bog storil tako, ko nam zanesljivo izkustvo kaže, da tega ni storil. Toda dobrota Boga ni manjkala ljudem, ki niso imeli takih izvirnih vtisov vednosti ali *idej*, vtisnjenih v duha, zato, ker je ljudi oskrbel s tistimi zmožnostmi, ki bodo rabile za zadostno odkritje vseh stvari, potrebnih za smoter takega bitja, in ne dvomim, da bom pokazal, da človek lahko s pravilno rabo svojih naravnih zmožnosti in brez kakršnegakoli prirojenega načela pridobi vednost o Bogu in drugih stvareh, ki ga zadevajo. Ko je Bog podelil človeku tiste zmožnosti vednosti, ki jih ima sam, ni bil dolžan s svojo dobroto vsaditi tiste prirojene pojme v njegovega duha bolj, kot da bi mu moral, potem ko mu je dal razum, roke in materiale, zgraditi mostove ali hiše, ki jih nekatera ljudstva v svetu, sicer v njegovih dobrih delih, bodisi popolnoma pogrešajo bodisi so z njimi slabo preskrbljeni, kot so druga popolnoma brez *idej Boga* in načel moralnosti ali pa so z njimi zgolj zelo slabo oskrbljeni. V obeh primerih je razlog ta, da niso nikoli na tak način marljivo uporabili svojih sposobnosti, zmožnosti in moči, temveč so se zadovoljili z mnenji, običaji in stvarmi svojih dežel, na kakršne so naleteli in niso pogledali nič dlje. Če bi se vi ali jaz rodili v zalivu *Soldanija*, verjetno naše misli in pojmi ne bi presegli tistih neomikanih misli in pojmov *Hotentotov*, ki tam bivajo. In če bi bil kralj *Virdžinije Apochancana* izobražen v *Angliji*, bi bil morda ravno tako učen teolog in dober matematik kot vsakdo v njej. Razlika med njim in bolj izobraženim *Angležem* bi bila zgolj v tem, da je bilo urjenje njegovih zmožnosti omejeno znotraj poti, načinov in pojmov njegove lastne dežele in ni bilo nikoli usmerjeno h kakšnim drugim ali nadaljnjim raziskavam. Če ni imel nobene *ideje* Boga, je bilo to zgolj zato, ker ni sledil tistim mislim, ki bi ga vodile k njej.

§ 13

Priznavam, da *če* bi obstajala *kakršnakoli ideja*, za katero bi bilo ugotovljeno, da je *vtisnjena* v duha človeka, bi imeli razlog pričakovati, *da bi to bil pojem njegovega stvarnika* kot znak, ki ga je BOG vtisnil v svoje lastno delo, da bi človeka spominjal na njegovo odvisnost in dolžnost in da bi se tu morali pojaviti prvi primeri človeške vednosti. Toda koliko časa mora preteči, da kak tak pojem odkrijemo pri otrocih? In ko ga pri njih odkrijemo, koliko bolj je podoben mnenju in pojmu učitelja, kot da bi predstavljal resničnega Boga? Tisti, ki bo pri otrocih opazil napredek, s katerim njihov duh pridobiva vednost, ki jo imajo, bo mislil, da so objekti, s katerimi najprej in najbolj domače občujejo, oni, ki naredijo prve vtise na njihov razum in ne bo našel niti najmanjših sledi kakih drugih vtisov. Zlahka je opaziti, kako se njihove misli razširjajo, zgolj da bi se seznanili z večjo raznolikostjo čutnih objektov, da bi ohranili njihove *ideje* v svojem spominu ter da bi si pridobili veščino, da bi jih uredili/primerjali, razširili in na različne načine sestavili. Kako s temi sredstvi ljudje pridejo do tega, da v svojem duhu izoblikujejo tisto *idejo*, ki jo imajo o božanstvu, bom pokazal v nadaljevanju.

§ 14

Ali lahko menimo, da so *ideje* ljudi, ki jih imajo o Bogu, pisava in znaki njega samega, ki jih je v njihov duh vtisnil s svojim lastnim prstom, ko vidimo, da imajo ljudje v isti deželi pod enim in istim imenom *precej različne, ne, pogosto celo nasprotne in nezdružljive ideje* in pojme *o njem*? Njihove soglasje glede imena ali zvoka, bo le stežka dokazalo, da je pojem Boga prirojen.

§ 15

Kakšen resničen ali še sprejemljiv pojem *božanstva* bi še lahko imeli tisti, ki jih priznavajo in častijo na stotine? Vsako božanstvo, ki presega eno, nezmotljivo kaže njihovo nevednost o njem, in je dokaz, da nimajo nobenega resničnega pojma o Bogu tam, kjer so izključene enotnost, neskončnost in večnost. Če k temu dodamo njihova surova pojmovanja telesnosti, izražena v njihovih podobah in predstavah njihovih božanstev, ljubezni, poroke, kopulacije, opolzkosti, prepire in druge nizke lastnosti, ki jih pripisujejo svojim bogovom, bomo imeli zgolj malo razlogov za razmišljanje, da ima poganski svet, *to je* največji del človeštva, v svojem duhu take *ideje* Boga, katerih avtor je bil on sam Bog iz skrbi, da se ne bi motili o njem. In ta univerzalnost soglasja, tako močno dokazovana, če dokazuje kakršnekoli prirojene vtise, bo zgolj tole: da je Bog vtisnil v duha vseh ljudi, ki govorijo isti jezik, *ime* zase, vendar pa ne nobene *ideje*, kajti tisti ljudje, ki so soglašali v imenu, so imeli hkrati zelo različna razumevanja označene stvari. Če pravijo, da raznolika božanstva, ki jih je častil poganski svet, niso bila nič drugega kot figurativni načini izražanja različnih atributov tega nedojemljivega bitja ali različni deli njegove previdnosti, odgovarjam, da tega, kaj so utegnila biti v svojem izvirniku, tu ne bom raziskoval, vendar da so bili taka v mislih preprostega ljudstva, za to mislim, da tega nihče ne bo zatrjeval. Tisti pa, ki bo iskal svet v *Voyage of the Bishop of Beryte*, 13. pogl. (da ne bi omenili drugih pričevanj), bo ugotovil, da si teologija *Siamcev* izrecno lasti množico Bogov ali pa, kot bolj uvidevno pripominja *Abbé de Choisy* v svojem *Journal du Voyage de Siam* na straneh 107–177, da primerno sestoji iz tega, da ne priznava sploh nobenega Boga.

§ 15 [*še enkrat*] Če bo rečeno, da *modri ljudje* vseh narodov pridejo do tega, *da imajo resnična pojmovanja* enotnosti in neskončnosti *božanstva*, bom to dopustil. Vendar potem to,

Prvič, izključuje univerzalnost soglasja o vsakršni stvari, razen glede imena, saj je ta univerzalnost, ker je tistih modrih ljudi zelo malo, zelo ozka.

Drugič, zdi se mi jasno dokazati, da najbolj resnični in najboljši pojmi, ki jih imajo ljudje o Bogu, niso bili vtisnjeni, temveč so jih pridobili mišljenje in meditacija ter pravilna raba njihovih zmožnosti, ker so modri in preudarni ljudje sveta s pravilno in skrbno rabo svojih misli in razuma pridobili resnične pojme o tej kot tudi vseh drugih stvareh, medtem ko je len in nepremišljen del ljudi, ki jih je bilo veliko več, svoje pojme po naključju pobrali iz skupne tradicije in vulgarnih pojmovanj, ne da bi si z njimi kaj veliko razbijali svoje glave. Če pa naj bi bil razlog za misel, da je *pojem Boga prirojen* zato, ker ga imajo vsi modri ljudje, je treba tudi za vrlino meniti, da je prirojena, saj so jo modri ljudje tudi vselej imeli.

§ 16

To se je očitno primerilo vsemu *poganstvu* in niti med *Judi*, *kristjani in mohamedanci*, ki priznavajo zgolj enega Boga, ta nauk in skrb, ki so jo ti narodi imeli, da bi ljudi naučili, da bi imeli resnične pojme BOGA, nista doslej prevladala, da bi ljudje imeli o njem enake in resnične *ideje*. Za koliko ljudi, celo med nami, se bo z raziskavo ugotovilo, da si ga zamišljajo v obliki moža, ki sedi v nebesih, in da imajo o njem mnogo drugih nesmiselnih in neprimernih pojmovanj? Kristjani kot tudi Turki so imeli cele sekte, ki so priznavale in resno trdile, da je božanstvo telesno in ima človeško obliko. In čeprav med nami najdemo le malo tistih, ki izpovedujejo zase, da so *antropomorfisti* (čeprav so nekateri, s katerimi sem se srečal, ki to priznavajo), vendar verjamem, da bo tisti, ki se bo potrudil, lahko med nevednimi in nepoučenimi kristjani našel mnoge, ki so takega mnenja. Govorite zgolj z podeželani domala vseh starosti ali z mladimi ljudmi domala na vseh položajih in ugotovili boste, da čeprav je ime BOG pogosto na njihovem jeziku, so pojmi, na katere nanašajo to ime, vendarle tako čudni, nizkotni in zaničevanja vredni, da si nihče ne more predstavljati, da jih je poučeval razumen človek, še veliko manj, da so to bili znaki, napisani s prstom samega Boga. Niti ne vidim, kako bi dobroto Boga bolj krnilo to, da nam je dal duh, nepreskrbljen z temi *idejami* njega samega, kot da nas je poslal v svet z nepokritimi telesi in da se z nami ne rodi nobena veščina ali pripravljenost. Ker smo oskrbljeni z zmožnostmi, da jih pridobimo, je zaradi pomanjkanja marljivosti in premisleka v nas, ne pa radodarnosti v njem, da jih nimamo. Gotovo je, da je Bog, kot da sta nasprotna kota, ki jih tvori križanje dveh premih črt, enaka. Nikoli ni bilo nobenega razumnega bitja, ki se je iskreno lotilo raziskovanja resnice teh propozicij, ki bi mu lahko spodletelo privoliti vanju. Čeravno je vendar onstran dvoma, da so mnogi ljudje, ki niso uporabili svojih misli na tak način in so nevedni tako glede enega kot drugega. Če kdo misli, da se spodobi to (kar je njegov skrajen doseg) imenovati univerzalno soglasje, takemu to zlahka dopustim, toda tako univerzalno soglasje, kot je to, nič bolj ne dokazuje, da je *ideja* Boga *prirojena* kot *ideja* takih kotov.

§ 17

Ker je potem vednost o *BOGU* najbolj naravno odkritje človekovega razuma, pa *ideja o njem ni prirojena*, kot je, mislim, razvidno iz tega, kar je bilo rečeno, si zamišljam, da bo stežka najti katerokoli drugo *idejo*, ki bi terjala, da je prirojena, kajti če bi Bog vtisnil kakršenkoli vtis, kakršenkoli znak v razum ljudi, je najbolj razumno pričakovati, ds bi to morala biti neka jasna in uniformna *ideja* njega samega, kolikor bi bile naše slabotne zmožnosti sposobne sprejeti tako nepojmljiv in neskončen objekt. Ker pa je naš duh od začetka brez te *ideje*, ki si jo najbolj prizadevamo imeti, je to *močna domneva proti vsem drugim prirojenim znakom*. Moram priznati, kolikor lahko opazim, da ne morem najti nobenega in bom vesel, če bom poučen o drugih.

§ 18

Priznam, da je še druga *ideja*, ki bi bila v splošno korist človeštva, če bi jo imelo, saj vsi govorijo, kot da bi jo imeli, to pa je *ideja substance*, ki je nimamo in ne moremo imeti s pomočjo *senzacije/občutka* ali *refleksije*. Če bi narava imela skrb, da bi nam ponudila kakršnokoli *idejo*, bi lahko upravičeno pričakovali, da bi morala biti taka, ki je ne moremo pridobiti z našimi lastnimi zmožnostmi. Vendar vidimo nasprotno, da ker ta ideja ne pride v naš duh na tiste načine, kakor so vanj vnesene druge *ideje*, nimamo sploh nobene take *jasne ideje* in zato z besedo *substanca* ne menimo nič, temveč zgolj negotovo domnevo o ne vem čem, (*to je*, o nečem, o čemer nimamo nobene posebne razločne pozitivne) *ideje*, ki jo imamo za *substrat* ali podporo za vse tiste *ideje*, ki jih poznamo.

§ 19

Karkoli že potem govorimo o prirojenih, bodisi *spekulativnih* bodisi *praktičnih*, *načelih*, lahko z ravno tako verjetnostjo rečemo, da ima mož sto funt šterlingov v svojem žepu in vendar zanika, da ima v njem bodisi penije, šilinge, krone ali kakršnekoli druge kovance, ki bi lahko tvorili to vsoto, kot mislimo, da so nekatere propozicije prirojene, ko o *idejah*, o katerih so, nikakor ne moremo domnevati, da so take. Splošno sprejemanje in soglasje, ki je dano, sploh *ne* dokazujeta, da so v njih izražene *ideje prirojene*, kajti v mnogih primerih, kakorkoli že so *ideje* prišle tja, bo privolitev v besede, ki izražajo soglasje ali nesoglasje s takimi *idejami*, nujno sledila. Vsakdo, ki ima resnično *idejo Boga* in *čaščenja*, bo soglasen s propozicijo, da je Boga treba častiti, ko je izražena v jeziku, ki ga razume, in vsak razumen človek, ki nanjo ni mislil danes, utegne biti pripravljen da s to propozicijo soglaša jutri, pa vendarle je upravičeno domnevati, da milijonom ljudi danes manjka ena ali obe od teh *idej*. Kajti če bomo dopustili, da divjaki in večina podeželanov ima *ideji Boga* in *čaščenja* (pogovor z njimi nas ne bo pomaknil bliže temu, da bi o tem verjeli), je vendarle, tako mislim, za malo otrok mogoče predpostavljati, da imajo te *ideje*, ki jih zato morajo začeti imeti v nekem trenutku in potem bodo tudi začeli soglašati s to propozicijo in potem o njej ne bi več dvomili. Toda tako soglasje na prvi mah ne dokazuje tega, da so *ideje* prirojene, nič bolj od tega, da ima nekdo, ki se je rodil slep (z katarakto, ki mu bo odstranjena jutri), prirojene *ideje* sonca ali svetlobe ali žafrana ali rumene zato, ker bo – ko se mu vid povrne – gotovo soglašal s propozicijo, da je sonce svetlo ali žafran rumen. In zato, če tako soglasje na prvi mah ne more dokazati, da so *ideje* prirojene, lahko še manj dokaže propozicije, nastale iz teh *idej*. Če ima kdorkoli kakršnokoli prirojeno *idejo*, bom vesel, da mi bo povedal, kakšna je in koliko jih je.

§ 20

K čemur naj dodam: če naj bi bile kakršnekoli prirojene *ideje*, kakršnekoli *ideje* v duhu, o katerih duh dejansko ne razmišlja, bi morale biti umeščene v spominu in bi jih moralo od tam potegniti na plano spominjanje, *to je*, bi morale biti znane, ko se jih spomnimo, kot prejšnje zaznave v duhu, razen če spominjanje lahko obstaja brez spominjanja. Kajti spominjanje pomeni zaznavanje vsakršne stvari s spominom ali z zavestjo, ki je bila znana ali zaznana že prej; brez tega je katerakoli *ideja*, ki pride v duha, nova in se je ne spomnimo: ta zavest, da je bila v duhu že prej, je tista, ki razlikuje spominjanje od vseh drugih načinov mišljenja. Katerakoli *ideja*, ki je duh nikoli ni zaznal, ni nikoli bila v duhu. Katerakoli ideja, ki je v duhu, je bodisi aktualna zaznava bodisi, ker je aktualna zaznava, je tako v duhu, da spomin lahko stori, da je spet aktualna zaznava. Kadarkoli obstaja aktualna zaznava *ideje* brez spomina, je videti *ideja* popolnoma nova in neznana razumu: kadarkoli spomin potegne katerokoli *idejo* v aktualni pogled, je to z zavestjo, da je bila tam že prej in duhu ni bila popolni tujec. Če to ni tako, se sklicujem na opazovanja vsakogar in si potem želim primer *ideje*, ki se pretvarja, da je prirojena in bi jo vsakdo (pred vsakršnim njenim vtisom s sredstvi, ki jih bomo kmalu omenili) lahko ponovno oživil in se je spomnil kot *ideje*, ki jo je prej poznal; brez zavedanja njenega prejšnjega zaznavanja, ni nobenega spomina in katerakoli *ideja* pride v duha brez tega zavedanja, se je ne spomnimo ali ne pride iz spomina niti ni mogoče reči, da je v duhu pred tem pojavom. Tisto pa, kar ni bodisi aktualno v pogledu bodisi v spominu, sploh nikakor ni v duhu in je natanko tako, kot da nikoli ne bi bilo tam. Predpostavimo, da je otrok uporabljal svoje oči, dokler ni spoznal in razločeval barv, toda potem mu katarakta zapre oči in je štirideset ali petdeset let popolnoma v temi in v tem času popolnoma izgubi ves spomin o *idejah* barv, ki jih nekoč imel. To se je primerilo slepemu možu, s katerim sem enkrat govoril, ki je izgubil svoj vid zaradi koz, ko je bil otrok, in ni nič več imel pojma barv, kot kdo, ki se je rodil slep. Sprašujem se, ali lahko kdo reče, da je ta mož potem imel kakšno *idejo* barv v svojem duhu bolj kot kdo, ki se je rodil slep? In mislim, da ne bo nihče rekel, da je en ali drugi imel v svojem duhu sploh kakšno *idejo* barv. Njegova katarakta je odstranjena in potem ima *ideje* (ki se jih ne spominja) barv *de novo* zaradi svojega znova vzpostavljenega vida in ki so sporočene njegovemu duhu in to brez kateregakoli zavedanja prejšnje seznanjenosti. In te lahko zdaj ponovno oživi in jih prikliče v duha v temi. V tem primeru se za vse te *ideje* barv, ki jih je mogoče, ko so zunaj pogleda, ponovno oživiti z zavedanjem prejšnje seznanjenosti z njimi in so tako v spominu, reče da so v duhu. Iz tega sklepam, da katerakoli *ideja*, ki dejansko ni v pogledu, je v duhu in je tam zgolj tako, da je v spominu, če pa ni v spominu, ni v duhu, in če je v spominu, je spomin ne more prinesti v aktualni pogled brez zaznave, ki izhaja iz spomina, to je, da je bila znana že prej in se je zdaj spomnimo. Če torej naj obstajajo kakšne prirojene *ideje*, morajo biti v spominu ali jih sicer ni nikjer v duhu, če pa so v spominu, jih je mogoče ponovno oživiti brez kakršnegakoli vtisa od zunaj in kadarkoli so prinesene v duh, se jih spomnimo, *to je*, prinesejo s seboj zaznavo, da zanj niso popolnoma nove. To je stalna in značilna razlika med tistim, kar je in kar ni v spominu ali v duhu; tisto, kar ni v spominu se zdi, ko se tam pojavi, popolnoma novo in pred tem neznano, ono pa, kar je v spominu in v duhu, se zdi, kadarkoli ga sugerira spomin, da ni novo, temveč ga duh najde v samem sebi in ve, da je bilo tam že prej. S tem je mogoče preizkušati, ali v duhu obstajajo kakšne prirojene *ideje* pred vtisi iz *senzacij/občutkov* ali *refleksije*. Vesel bi bil srečanja s človekom, ki potem ko je prišel do rabe razuma ali ki se je kadarkoli spomnil kake od njih in ki mu, potem ko se je rodil, nikoli niso bile nove. Če bi kdorkoli rekel, da so *ideje* v duhu, ki niso v spominu, si želim, da si razloži in naredi tisto, kar je rekel, razumljivo.

§ 21

Poleg tega, kar sem že povedal, je še drug razlog, zakaj dvomim, da niti ta niti nobena druga načela niso prirojena. Jaz, ki sem popolnoma prepričan, da je neskončni BOG ustvaril vse stvari v popolni modrosti, se ne morem zadovoljiti s tem, zakaj bi morali predpostavljati, da je v duh ljudi vtisnil neka univerzalna načela, od katerih tista, *ki* se pretvarjajo, da so prirojena in *zadevajo spekulacijo*, *niso zelo koristna in ona, ki zadevajo prakso, niso samorazvidna in nobenih od njih ni mogoče razlikovati od nekaterih drugih resnic, za katere ni dopuščeno, da so prirojene.* Kajti s kakšnim ciljem naj bi bili znaki vrezani v duha z božjo roko, ki tam niso jasnejši od tistih, ki so uvedeni kasneje ali pa jih od njih ni mogoče razlikovati? Če kdo misli, da obstajajo take prirojene *ideje* in propozicije, ki jih je mogoče po njihovi jasnosti in koristnosti razlikovati od vsega, kar je v duhu naključno in pridobljeno, zanj ne bo težko povedati nam, katere so to, in potem bo vsakdo primeren sodnik, ali so take ali ne. Saj če bi obstajale take prirojene *ideje* in vtisi, očitno drugačne od vseh drugih zaznav in vednosti, bi vsakdo odkril, da so res v njem samem. O razvidnosti teh domnevnih prirojenih maksim, sem že govoril, o njihovi koristnosti bom imel priložnost odslej več govoriti.

§ 22

Če sklenem: nekatere *ideje* se voljno ponudijo razumu vseh ljudi in neke vrste resnic izhajajo iz kakršnihkoli *idej*, takoj ko jih duh poveže v propozicije; nekatere resnice zahtevajo sosledje *idej*, postavljenih v red, njihovo primerno primerjavo in pozorno narejene izpeljave, preden jih je mogoče odkriti in z njimi soglašati. Nekatere od tistih iz prve vrste so bile zamenjane za prirojene zaradi svojega splošnega in lahkega sprejemanja, toda resnica je, da se *ideje* in pojmi nič bolj ne rodijo z nami kot umetnosti in znanosti, čeprav se nekatere od njih res ponujajo našim zmožnostim bolj voljno kot druge in so zato splošneje sprejete. Toda tudi to se zgodi v skladu s tem, kako so uporabljeni organi naših teles in zmožnosti našega duha; *Bog je oskrbel ljudi z zmožnostmi in sredstvi, da bi odkrili, sprejeli in ohranili resnice v skladu s tem, kako so uporabljene.* Velika razlika, ki jo gre najti v pojmih ljudi, izhaja iz različne rabe, v katero zaprežejo svoje zmožnosti: medtem ko nekateri (in ti so večina) jemljejo stvari na zaupanje in zmotno uporabljajo svojo moč soglasja z lenim zasužnjevanjem svojega duha narekom in gospostvu drugih v naukih, pri katerih je njihova dolžnost, da jih skrbno preučijo, ne pa da jih slepo in z implicitno vero požrejo, drugi uporabljajo svoje misli zgolj za malo stvari, se z njimi dovolj seznanijo, pridobijo o njih veliko stopnjo vednosti in zanemarjajo vse druge ter nikoli ne dopustijo, da se njihove misli razvežejo v iskanju drugih preiskovanj. Torej to, da so trije koti trikotnika enaki dvema pravima kotoma, je resnica, tako gotova, kot je lahko kakšna stvar in, tako mislim jaz, bolj razvidna od mnogih tistih propozicij, ki veljajo za načela, in vendarle obstajajo milijoni, kakorkoli poznavalci drugih stvari, ki tega sploh ne vedo, ker nikoli niso usmerili svojih misli na to, da bi se ukvarjale s takimi koti. Tisti pa, ki gotovo pozna to propozicijo, utegne biti vendarle neveden o resnici drugih propozicij v sami matematiki, ki so tako jasne in razvidne kot ta, ker je v svojem iskanju tistih matematičnih resnic ustavil svoje misli in ni šel tako daleč. Enako se utegne zgoditi glede pojmov, ki jih imamo o bivanju božanstva, kajti čeravno ni nobene resnice, ki jo človek lahko bolj razvidno naredi samemu sebi, kot je eksistenca Boga, vendarle tisti, ki se bo zadovoljil s stvarmi, kakor jih najde v tem svetu in kot nudijo zadovoljstva in strasti, in ne raziskuje nekoliko dlje o njihovih vzrokih, ciljih in čudovitih darov ter ne sledi tem mislim marljivo in pozorno, lahko dolgo živi brez vsakršnega pojma takega bitja. Če pa je katerikoli osebi ta pojem vbil v njeno glavo z besedami, nemara lahko vanj verjame; če pa ga nikoli ni *preučila*, njena vednost o njem ne bo nič popolnejša od vednosti tistega, ki potem ko mu je bilo povedano, da so trije koti trikotnika enaki dvema pravima kotoma, to sprejme z zaupanjem, ne da bi raziskoval dokaz in utegne dati svoje soglasje kot k verjetnemu mnenju, vendar nima nikakršne vednosti o njegovi resnici, ki bi ji ga bile njene zmožnosti, če bi bile skrbno uporabljene, sposobne razjasniti in narediti razvidnega. Toda to omenjam zgolj mimogrede, da bi pokazal, kako močno je naša *vednost odvisna od prave rabe tistih moči, ki nam jih je podelila narava*, in kako malo od takih prirojenih načel, za katera se zaman domneva, da so v vsem človeštvu za to, da bi ga vodila in za katera si vsi ljudje ne bi mogli kaj, da ne bi zanje vedeli, če bi bila tam, ali pa bi bila tam brez vsakršnega smotra. Ker pa jih vsi ljudje ne poznajo niti jih ne morejo razlikovati od drugih naključnih resnic, lahko upravičeno sklepamo, da takih načel ni.

§ 23

Kakšno grajo tak dvom o prirojenih načelih utegne zaslužiti od ljudi, ki ga bodo pripravljeni imeti za razdejanje starih temeljev vednosti in gotovosti, ne znam povedati: jaz sam sem sicer prepričan, da je pot, ki sem ji sledil in je bila ustrezna resnici, pripravila te temelje tako, da so zanesljivejši. A o eni stvari sem prepričan in to je, da se nisem ukvarjal niti s tem, da bi se odrekel niti sledil kakršnikoli avtoriteti v naslednji razpravi. Resnica je bila moj edini cilj in kamorkoli že se je zdelo, da to vodi, so ji moje misli nepristransko sledile, ne da bi pazil, ali so na tej poti ležale stopinje kakega drugega ali ne. Ni to, da sem želel dolžno spoštovanje mnenj drugih ljudi, toda navsezadnje *gre največje spoštovanje resnici* in upam, da ne bo mišljeno kot aroganca reči, da bi nemara naredili večji napredek pri odkrivanju racionalne in kontemplativne vednosti, če bi jo *iskali* v izviru, *v motrenju stvari samih* in raje uporabljali naše lastne misli kot pa misli drugih ljudi, da bi jo odkrili. Kajti mislim, da lahko racionalno upamo, da bomo tako videli z očmi drugih ljudi kot vedeli z razumom drugih ljudi. Kolikor mi sami motrimo in dojemamo resnico in razum, toliko imamo realno in resnično vednost. Lebdenje mnenj drugih ljudi v naših možganih nas ne naredi niti najmanj bolj vedoče, čeprav se zgodi, da so resnična. Kar je bilo pri njih znanost, je pri nas zgolj trmoglavost, medtem ko dajemo? naše soglasje zgolj častitljivim imenom in ne uporabljamo, tako kot oni, našega razuma, da bi *razumeli tiste resnice,* ki so jim dali sloves. *Aristotel* je bil gotovo učen človek, toda nihče nikoli ni mislil, da je tak zato, ker se je slepo oklenil in zaupljivo dal duška mnenjem drugih. In če ga prevzemanje načel drugih, ne da bi jih preučil, ni naredilo za filozofa, domnevam, da bo le stežka naredilo za kaj takega kogarkoli drugega. V znanostih ima vsakdo toliko, kolikor dejansko ve in dojema: tisto, o čemer zgolj verjame in sprejema na zaupanje, ni nič drugega kot koščki, ki ne, kakorkoli dobro se že prilegajo v celoto, prispevajo kakega znatnega dodatka v zalogo tistega, ki jih zbira. Tako izposojeno bogastvo bo, kot pravljični denar, čeprav je bilo zlato v rokah tistega, od katerega ga je prejel, zgolj prah in pepel, ko ga hoče kdo uporabiti.

§ 24

Ko so ljudje našli nekatere splošne propozicije, o katerih, takoj ko so bile razumljene, ni bilo mogoče dvomiti, je bila to, vem, *kratka in lahka pot do sklepa, da so prirojene*. Ko je bilo to enkrat sprejeto, je lene razrešilo vseh naporov raziskovanja in ustavilo raziskovanje dvomljivcev, kar zadeva vse, kar je bilo enkrat razglašeno za prirojeno. In ni bila mala prednost za tiste, ki so se povzpeli do mojstrov in učiteljev, da so naredili za načelo vseh *načel*, da v načela ne gre dvomiti. Ko so enkrat postavili to načelo, da obstajajo prirojena načela, je to njihove privržence potisnilo v nujnost, da so sprejeli nekatere nauke kot take, kar je pomenilo, da so jih odstranili iz rabe svojega lastnega razuma in sodbe in jih potisnili v verovanje in v to, da so jih jemali na zaupanje, brez nadaljnjega raziskovanja. V tem položaju slepe zaupljivosti jih je mogoče bolj zlahka voditi in narediti koristne za tisto vrsto ljudi, ki imajo zmožnost in zaposlitev, da jim dajejo načela in jih vodijo. Niti ni majhna moč, ki jo en človek da drugemu, da ima avtoriteto biti diktator načel in učitelj nedvomnih resnic in da pripravi človeka, da požre kot prirojeno načelo tisto, kar utegne priti prav cilju tistega, ki jih poučuje. Če bi, nasprotno, raziskovali poti, po katerih ljudje pridejo do vednosti mnogih univerzalnih resnic, bi ugotovili, da so v duhu ljudi posledica bivanja stvari samih, ko so primerno obravnavane, in da jih je odkrila uporaba tistih zmožnosti, ki so po naravi ustrezale, da jih sprejmejo in o njih sodijo, ko so primerno uporabljene.

§ 25

Pokazati, kako v tem razum nadaljuje, je načrt razprave, ki sledi, h kateri se bom napotil, ko bom najprej predpostavil, da je bilo doslej zame nujno za to, da bi odprl mojo pot k tistim temeljem, za katere mislim, da so edino resnični in na katerih gre vzpostaviti tiste pojme, ki jih lahko imamo o naši lastni vednosti, pojasniti razloge, ki sem jih imel za dvom o prirojenih načelih. Ker pa argumenti, vsaj nekateri od njih, ki so proti njim, dejansko izvirajo iz splošno sprejetih mnenj, sem bil prisiljen imeti nekatere stvari za dognane, čemur se vsakdo, čigar naloga je pokazati zmotnost ali neverjetnost kakršnegakoli načela, težko izogne. V protislovnih razpravah se dogaja enako kot v naskoku na mesta, kjer, če so tla dovolj trdna, na katera so postavljene baterije, ni nadaljnjih poizvedovanj, kdo jim je jih dal v posodo ali komu pripadajo, ker imajo zadostno višino za trenutni cilj. Toda v prihodnjem delu te razprave, načrtovane da zgradi uniformno in s seboj konsistentno stavbo, kolikor mi bo pomagalo moje lastno izkustvo in opazovanje, upam, da jo bom zgradil na takem temelju, da mi je ne bo treba podpreti z podporniki in jo okrepiti in da se ne bo opirala na sposojene ali naberačene temelje ali vsaj, če se moja zgradba izkaže za grad v oblakih, si bom prizadeval, da je iz enega kosa in drži skupaj. Pri čemer opozarjam bralca, da ne pričakuje neutajljivo nespodbitnih dokazov, razen če mi bo dovoljen privilegij, ki ni pogosto dopuščen drugim, da štejem moja načela za dognana, in potem ne dvomim, da jih lahko tudi dokažem. Vse, kar lahko rečem o načelih, s katerimi nadaljujem, je, da se lahko *sklicujem* zgolj na lastno *izkustvo* ljudi, ki je brez predsodkov, in opazovanje, ali so resnična ali ne; to pa je dovolj za človeka, ki ne izpoveduje nič več, kot da iskreno in svobodno izpostavlja svoje lastne domneve, ki zadeva predmet, ležeč v nekakšni temi, brez kateregakoli drugega načrta, kot nepristranskem raziskovanju resnice.

Navodila avtorjem

Prispevke oddajte po elektronski pošti (analiza@drustvo-daf.si). Zaradi anonimnega recenzijskega postopka (»double-blind peer-review«) zapišite svoje ime in kontaktne podatke v ločenem dokumentu in poskrbite, da iz samega prispevka ni mogoče razbrati vaše identitete. Prispevku je treba priložiti povzetek (v slovenščini in v angleščini), ki povzema glavne poudarke dela. Povzetku v angleškem jeziku je treba dati tudi angleški naslov. Povzetek ne sme presegati 150 besed. Na koncu povzetka navedite do 5 ključnih besed (deskriptorjev). Na primer, *Ključne besede:* filozofija jezika, metafora, analogija; *Keywords:* philosophy of language, metaphor, analogy.

Prispevki naj praviloma ne presegajo obsega ene in pol avtorske pole (45.000 znakov s presledki). Uporabite urejevalnik besedil *Word*, standardno obliko pisave brez dodatnih slogovnih določil. Napisani naj bodo z dvojnim razmikom med vrsticami; za literaturo, opombe in povzetek pa uporabite enojni razmik. Med odstavki naj bo izpuščena vrstica. Prispevki naj bodo notranje razčlenjeni, torej razdeljeni na razdelke, in opremljeni – če je mogoče – z mednaslovi. Citati v besedilu naj bodo označeni z dvojnimi narekovaji, citati znotraj citatov pa z enojnimi. Citati, daljši od 40 besed, se zapišejo kot samostojni odstavki z levim zamikom 0,8 cm (brez narekovajev). Izpusti so označeni s poševnicami, prilagoditve pa z oglatimi oklepaji. Naslove knjig, periodike in tuje besede je treba pisati *ležeče*. Primeri: *Kritika praktičnega uma, Anthropos, a priori*.

Reference se pišejo v besedilu (»author-date« sistem). Oblikovanje mora biti takole: (avtorjev priimek, letnica: str. ali pogl.); na primer (Davidson, 1967: 312) ali (Blackburn, 1988: III). Na koncu prispevka morate priložiti popoln, po abecednem redu urejen bibliografski opis citiranih virov (**Literatura**). Primer takih opisov:

Davidson, D. (1967). »Truth and Meaning«. *Synthese*, 17, str. 304–323.

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Davidson, D. (1984). »Truth to the Facts«. V Davidson, D., *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press, str. 37–54.

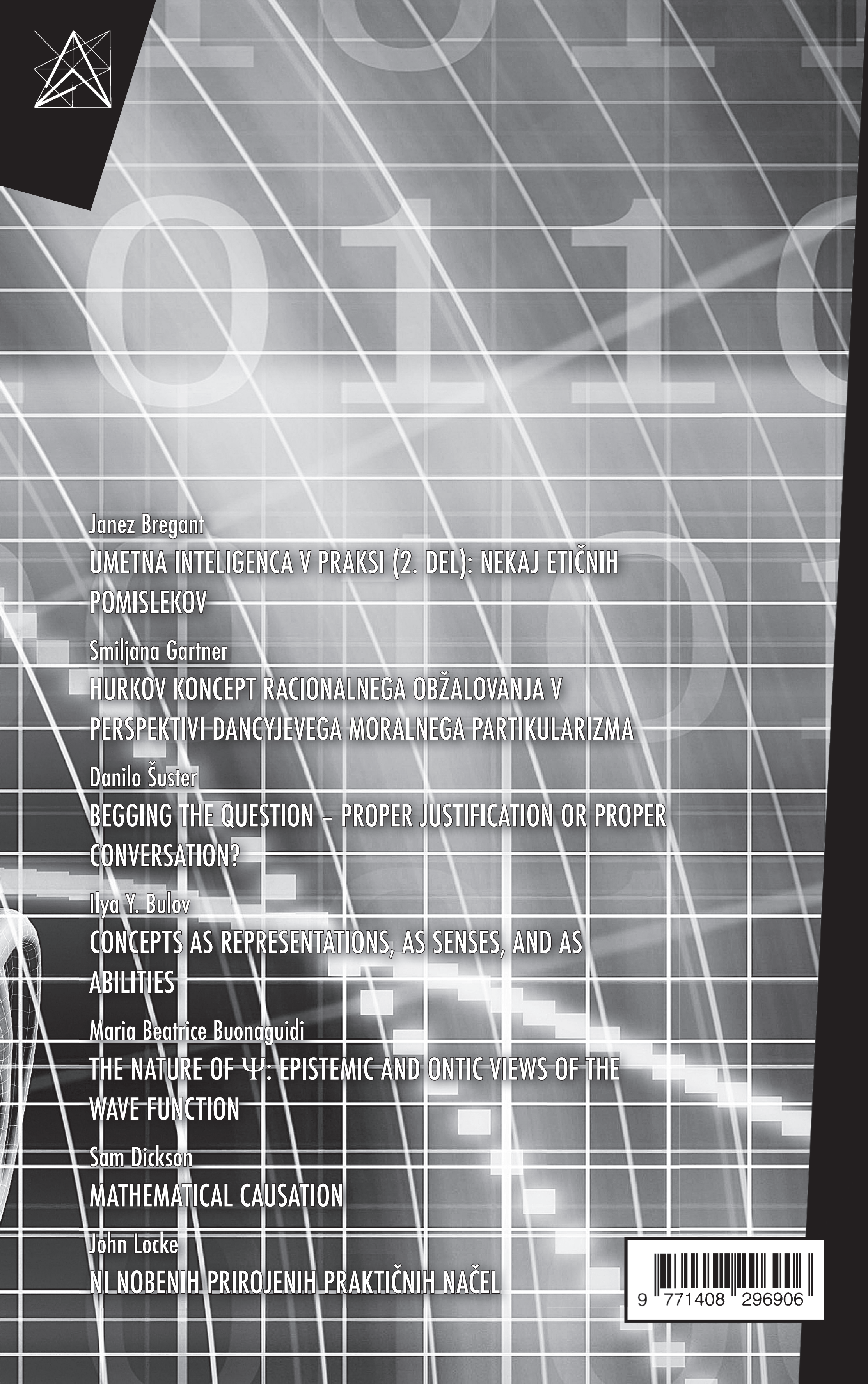
Davidson, D. (1997). »Radical Interpretation«. V Baillie, J. (ur.), *Contemporary Analytic Philosophy*, New Jersey: Prentice Hall, str. 383–393.

Malpas, J. (2015). »Donald Davidson«. V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy*. Dostopno na: https://plato.stanford.edu/archives/fall2015/entries/davidson/ [28. 1. 2017].

Pod črto navajajte *samo* opombe. V besedilu je treba opombe označiti z dvignjenimi indeksi[1]. Ne uporabljajte opomb za navedbo reference.

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

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



9 771408 296906