Kodaj Dániel

Mesterséges unintelligencia

A gépesített ész filozófiai kritikái

Kodaj Dániel

Mesterséges unintelligencia

A gépesített ész filozófiai kritikái

Kodaj Dániel

Mesterséges unintelligencia

A gépesített ész filozófiai kritikái

Budapest, 2026

A kötet megjelenését az Eötvös Loránd Tudományegyetem Oktatási és Képzési Tanácsának tankönyv- és jegyzettámogatási pályázata támogatta.

© Kodaj Dániel, 2026

Szaklektor:
Héder Mihály
Molnár Zalán
Tőzsér János

ISBN 978-963-489-958-7
ISBN (pdf) 978-963-489-959-4

A borítón szereplő kép AI által generált.

ELTE | EÖTVÖS KIADÓ
Az ELTE hivatalos könyvkiadója

ELTE | BTK
BÖLCSÉSZETTUDOMÁNYI KAR

eltebook.hu

Felelős kiadó: az ELTE Eötvös Kiadó ügyvezető igazgatója
Projektvezető: Egresi Nóra
Borítótervező: Balogh Katalin
Műszaki szerkesztő: Katona Csaba
Nyomdai munkák: Multiszolg Bt.

Tartalomjegyzék

Előszó **9**

1. A demarkációs probléma **13**
1.1. A Turing-teszt és a gondolkodás fogalma 15
1.2. A demarkációs probléma metafizikája 19
1.3. Filozófián kívüli demarkációs érvek 20

2. Mesterséges intelligencia **21**
2.1. A transzformer-architektúra 21
2.2. Turing-gépek 26
2.3. Implementált Turing-gépek 30
2.4. A CPU mint Turing-gép 32

I. Elmeoldali demarkáció **35**

3. Testi demarkáció **37**
3.1. Dreyfus, Husserl és Heidegger 38
3.2. Enaktivizmus 41
3.3. Kell-e test a gondolkodáshoz? 43

4. Egzisztenciális demarkáció **45**
4.1. Egzisztenciális holizmus 46
4.2. Felelősség és jelentés 48

5. Referenciális demarkáció **53**

6. Szemantikai demarkáció **61**
6.1. A Kínai Szoba 61
6.2. A szintaxis és szemantika ellentéte 64
6.3. Két klasszikus ellenérv 67
6.4. A Kínai Szoba mint intuíciódetektor 73
6.5. Túl a Kínai Szobán 77

7. Fenomenális demarkáció **85**
7.1. Vannak-e az MI-nek érzései? 86
7.2. Kognitív fenomenológia 91
7.3. Szemantikai fenomenalizmus 95

8. Introspekciós demarkáció **99**

9. Nemteljességi demarkáció **105**
9.1. „Bizonyíthatatlan vagyok” 109
9.2. Nemteljességi demarkáció 112

10. Eldönthetetlenségi demarkáció **121**
10.1. A megállási probléma 121
10.2. Megállási demarkáció 124
10.3. Rice-demarkáció 125
10.4. A Rice-demarkáció előnyei 128

11. Az önreprezentáció elátkozott szentháromsága **131**
11.1. Az önreprezentáció szentháromsága 132
11.2. Elátkozottság 135
11.3. Értékrések és értéktorlódások 137
11.4. Amodális verzió 139

12. Önreprezentációs demarkáció **141**
12.1. Az 1. premissza védelmében 143
12.2. A 2. premissza védelmében 144
12.3. A 3. premissza védelmében 145

II. Számításoldali demarkáció **151**

13. Implementáció **153**
13.1. Az Implementációs Tézis 153
13.2. Az Implementációs Tézis metafizikája 155
13.3. Definíciós kritériumok 157

14. A szemantikai felfogás **159**
14.1. Természetes információ 161
14.2. Shannon-információ 162
14.3. Naturalizált szemantika 163

15. A formalista felfogás **167**
15.1. Négy trivializáló érv a formalizmus ellen 169
15.2. Pánkomputacionalizmus 175

16. A strukturalista felfogás **179**
16.1. Kategoriális strukturalizmus 180
16.2. Kontrafaktuális strukturalizmus 185

17. Az eszközfelfogás **189**
17.1. Naturalizált teleológia 191
17.2. A Piccinini-féle eszközfelfogás 195

18. Antirealizmus **199**

19. Az idealista felfogás **205**

20. Implementációs demarkáció **211**

Konklúzió **217**

Matematikai és logikai szimbólumok **225**

Irodalomjegyzék **227**

Előszó

A mesterséges intelligencia verseket ír, életvezetési tanácsokat ad, tudományos fogalmakat magyaráz, zenét szerez, orvosi leleteket elemez, programoz, professzionális szinten fordít. Néha összevissza beszél, de a fejlesztők gyors ütemben szorítják vissza a hallucinációk arányát, és még a selejtes válaszok is nyelvileg színvonalasak.

Következik-e ebből, hogy a gépek megtanultak *gondolkodni*? Vagy vannak az emberi értelemnek olyan aspektusai, melyek elvben gépesíthetetlenek?

*Demarkációs érv*nek nevezek minden olyan érvet, ami azt próbálja igazolni, hogy az MI nem gondolkodik. Ez a tézis természetesen több fronton is pontosításra szorul; mindenekelőtt tisztázni kell, mit értünk „gondolkodás"-on és „MI"-n. Az alapvető terminológiai kérdéseket a könyv első két fejezete tárgyalja. Röviden: az „MI" szóval a mai nagy nyelvi modelleket és az ezekhez lényegileg hasonló jövőbeli programokat jelölöm, a „gondolkodás" szóval pedig azt a mindannyiunk számára jól ismert intellektuális folyamatot, amit akkor élünk át, mikor spekulálunk, mérlegelünk, elemzünk, szöveget írunk vagy szöveget olvasunk. Az 1. fejezet hosszabban jellemzi ezt a jelenséget, a későbbi fejezetek pedig sok különböző aspektusát mutatják be.

A könyv elsősorban olyan filozófusoknak és egyetemi hallgatóknak szól, akik számára nem idegen az analitikus metafizika és elmefilozófia, valamint a standard elsőrendű logika. Ahol lehet, igyekszem kerülni a szakzsargont és elmagyarázni az alapvető fogalmakat, de ezzel is inkább a szakmabeliek számára próbálom olvashatóbbá tenni a szöveget. Az MI filozófiájának egyik fő kihívása az, hogy szétválaszthatatlanul keveredik benne a metafizika, elmefilozófia, nyelvfilozófia, tudományfilozófia, számítástudomány és matematika. Ezeket a területeket kevesen látják át a maguk teljes mélységében, én biztosan nem. Azt remélem, a szöveg ennek ellenére koherensen mutat be egy MI-szkeptikus álláspontot és elég világos ahhoz, hogy racionális vitákhoz adjon alapot.

Négy érvet fogok felépíteni annak igazolására, hogy az MI nem gondolkodik, továbbá rekonstruálok hat olyan demarkációs érvet, melyet mások dolgoztak ki. A tíz érv szlogenné tömörített alapgondolata a következő:

Az MI nem gondolkodik, mert...

(I) ...nincs teste (3. fejezet)

(II) ...nincs sorsa (4. fejezet)

(III) ...nem képes tárgyakra referálni (5. fejezet)

(IV) ...nem érti a szavak jelentését (Kínai Szoba, 6. fejezet)

(V) ...nincsenek érzéki tapasztalatai (7. fejezet)

(VI) ...nem képes introspekcióra (8. fejezet)

(VII) ...nem ismeri fel formális matematikai rendszerek Gödel-mondatait (9. fejezet)

(VIII) ...nem tudja megállapítani, mi mire való (10. fejezet)

(IX) ...nem képes önreprezentációra (11–12. fejezet)

(X) ...objektíve nem is létezik (13–20. fejezet).

A mások által kidolgozott, általam csak rekonstruált érvek között szerepel a két leghíresebb kognitív demarkációs érv (Searle és Penrose, IV és VII), két történetileg fontos, de mára nagyrészt feledésbe merült MI-kritika, melyekben az amerikai pragmatizmus keveredik a fenomenológiai és egzisztencialista hagyománnyal (Dreyfus és Haugeland, I és II), valamint két veterán MI-kutató kevéssé ismert, de (szerintem) nagyon érdekes érve (Cantwell Smith és Bringsjord, III és VI). Véleményem szerint ez a hat érv sikertelen, de kudarcuk kifejezetten tanulságos.

A maradék négy érv – V, VIII, IX és X – saját munkám. Ezek közül három (V, VIII, IX) korlátozott érvényű és csak az értelem bizonyos formáit tagadja el az MI-től, bár együttesen az emberi gondolkodás jelentős részét lefedik. A X. érv a számítási antirealizmusra épül és az MI ontológiai függetlenségét megkérdőjelezve érvel a gépesített gondolkodás ellen.

Semmiképp sem állítom, hogy nincsenek további és a fentieknél esetleg sokkal jobb demarkációs stratégiák, de azok közül, melyeket ismerek, ezek tűnnek számomra *prima facie* védhetőnek.

Mivel a tíz érv nem kapcsolódik szorosan egymáshoz, a könyv eléggé vegyesfelvágott-jellegű. A jó hír az, hogy a fejezetek emiatt egymástól függetlenül is olvashatóak, legalábbis az első nyolc érv esetében. A IX. érvhez tartozik egy logikai felvezetés, a X. érv pedig nyolc fejezetre bomlik, mert a riválisa (realista) számításelméletek részletes kritikájára épül.

Bár ez a vizsgálódás szűken filozófiai jellegű, tanulságai relevánsak lehetnek a mesterséges intelligenciáról folyó szélesebb szakértői és társadalmi vitában:

(A) Ha *a priori* okunk van kételkedni benne, hogy az MI hozzánk hasonlóan értelmes vagy hamarosan azzá válhat, akkor jó okunk van a szingularitásról, tudatfeltöltésről és MI-diktátorokról szóló spekulációkat sci-fi fantáziáknak tekinteni.

(B) A demarkációs problémának szoros köze van az MI jogi és morális státuszához. Ha bűncselekményt csak az tud elkövetni, aki felfogta tettének várható következményeit, akkor az MI csak akkor követhet el bűncselekményt, ha a demarkációs tézis – miszerint az MI nem gondolkodik – hamis. Ha a jogi vagy morális felelősséghez értelmi képességekre van szükség, akkor a demarkációs tézis igazolása közvetve azt is igazolja, hogy az MI nem lehet jogi vagy morális értelemben felelős semmiért. Általában, ha ágens csak az lehet, aki képes tudatosan célokat formálni, akkor az MI (amennyiben a demarkációs tézis igaz) nem lehet ágens.

(C) A demarkációs érveknek informatikai jelentősége is lehet, ha olyan tulajdonságokra épülnek, melyek fontosak bizonyos kognitív feladatok megoldásához. Ha belátjuk, hogy az MI nem rendelkezhet az F kognitív képességgel, ez utóbbi pedig elengedhetetlen G típusú feladatok megoldásához, akkor beláttuk, hogy az MI nem lehet képes G típusú feladatok megoldására.

A könyv végén, a Konklúzióban részletezem, miért támogatják a sikeres demarkációs érvek azt a tézist, hogy az MI nem egy elme, nem önálló ágens, hanem az emberi kooperáció történetileg teljesen új formája.

Szeretnék köszönetet mondani az ELTE TK Társadalomtudományi Kutatóközpont professzorának, Sebők Miklósnak, amiért felkarolta ezt a projektet és egy alkotói ösztöndíjjal lehetővé tette a könyv megszületését. Hálás vagyok továbbá az ELTE BTK Filozófiai Intézetének, különösen Ambrus Gergelynek, Bene Lászlónak és Olay Csabának a remek munkakörülményekért, valamint a CEU filozófiai tanszékének és egykori témavezetőmnek, Howard Robinsonnak, amiért elindítottak ezen a pályán. Nekik köszönhetem mindazt, amit az analitikus filozófiáról tudok, Howard neoplatonista idealizmusa pedig a mai napig alapjaiban határozza meg gondolkodásomat.

A könyv sokkal színvonaltalanabb lenne a három szaklektor, Héder Mihály, Molnár Zalán és Tőzsér János nélkül, akik áldozatos munkával rengeteg hibára és pontatlanságra hívták fel a figyelmemet. Márton Miklósnak köszönöm a 6. és 7. fejezettel kapcsolatos részletes kritikát. A szerkesztés során generált újabb hibákért és a korrigálatlan tévedésekért természetesen kizárólag én vagyok a felelős.

A jelen publikációban szereplő kutatás a HUN-REN Társadalomtudományi Kutatóközpontban, az Európai Unió támogatásával valósult meg, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében. A kézirat Open Access és nyomtatott megjelenését az ELTE BTK 2025-ös publikációs pályázata támogatta.

1.

A demarkációs probléma

A mesterséges intelligenciával kapcsolatos spekuláció veszélyes műfaj. 1957-ben – hat évtizeddel a jelenlegi MI-forradalom kitörése előtt – egy matematikai konferencia díszvacsoráján Herbert A. Simon, a leendő Nobel-díjas közgazdász és Allen Newell, a későbbi Turing-díjas kognitív tudós az alábbi jóslatokat tette:

> 1. Tíz éven belül [=1967-re] egy digitális számítógép lesz a sakkvilágbajnok, hacsak nem tiltják el bürokratikus úton a versenyzéstől.
>
> 2. Tíz éven belül egy digitális számítógép be fog bizonyítani egy fontos új matematikai tételt.
>
> 3. Tíz éven belül egy digitális számítógép olyan zeneművet fog komponálni, amit a zenekritikusok esztétikailag értékesnek tartanak.
>
> 4. Tíz évben belül a pszichológiai elméletek többsége számítógépes programok vagy a számítógépes programok tulajdonságaira vonatkozó kvalitatív állítások formáját fogja ölteni. (Simon & Newell 1958: 7)

E jóslatok mindegyike súlyos tévedésnek bizonyult. A legelső csak a 2000-es évek vége felé valósult meg, negyven év késéssel. A 3. jóslat talán megvalósult a 2020-as évek közepére az olyan zenekészítő MI-knek köszönhetően, mint a Suno, de ebben nincs egyetértés. A maradék két áttörésre ki tudja, meddig kell várnunk. Mindenesetre tíz éven belül még nyomokban sem igazolódott be egyik jóslat sem. Helyette elkezdett kialakulni az MI-ellenes filozófiai front, melynek két úttörő alakjáról a 3. és 4. fejezetben lesz szó.

Hogy ne csak a hurráoptimista MI-kutatókon élcelődjek és a filozófusok szennyesét is kiteregessem, hangsúlyoznom kell, hogy az MI-vel kapcsolatos elvi szkepticizmus is hasonló sorsra jutott. Hubert Dreyfus, az MI leghíresebb amerikai kritikusa 1972-ben ezt írta:

> [Az intuíció, kreatív feladatmegoldás, szövegértelmezés] az intelligens viselkedésnek az a területe, ahol digitális számítógépek algo-

> ritmusai garantáltan sosem érik el egy érett felnőtt ember kompetenciaszintjét [...]. Ékes bizonyíték az alapvető nehézségekre, hogy nem sikerült sakkbajnokot létrehozni, bármilyen érdekes matematikai tételt bebizonyítani, egyik nyelvről a másikra fordítani, vagy általános cél-eszköz elemzést végezni. (Dreyfus 1972: 208)

Az MI már jó ideje könnyedén megveri a hús-vér sakkmestereket és hatékonyan fordít egyik nyelvről a másikra. Az MI-kritikus filozófusok első nagy generációja mélyen alábecsülte a mérnökök képességeit.

Az MI-kutatások utóbbi hét évtizedét a fenti két véglet (az alaptalan optimizmus és az elhamarkodott pesszimizmus) közti vergődés jellemzi. Egyes pszichológusok, kognitív tudósok és informatikusok már az 1950-es évektől kezdve – röviddel az első digitális számítógép, az amerikai védelmi minisztérium számára készült ENIAC létrejötte után – ígérgetni kezdték, hogy a számítógépek hamarosan, egy-két évtizeden belül az emberhez hasonló értelmi képességekkel fognak rendelkezni, sőt, túl is szárnyalják teremtőiket. Ehhez a prognózishoz, a fenti 4-es predikció szellemében, gyakran az az ígéret társult, hogy az MI az emberi elme megértését is forradalmasítani fogja.

Mint a messiásvárásnál lenni szokott, az eszkhaton időpontja az idő múlásával azonos ütemben tolódott a jövőbe, és a 2020-as évekig mindig már csak egy-két évtizedre voltunk a nagy áttöréstől. A nekibuzdulásokat menetrendszerűen követte egy-egy „MI tél", melynek során az aktuálisan favorizált paradigma látványosan összeomlott, a kutatók pedig szétszéledtek és gondterhelten vakargatták a fejüket, hogyan tovább. A számítási kapacitások bővülése és az óriás adatbázisok megjelenése a 2010-es években végül tényleg meghozta az áttörést, és a 2020-as évek közepére a nagy nyelvi modellekre épülő MI-technológia ritkán látott befektetési és innovációs mániát indított el. Intenzív vita indult az MI társadalmi és gazdasági szerepéről, a merészebb szakértők pedig már arról spekulálnak, hogy az MI hamarosan forradalmasítani fogja a tudományt, felszámolja az összes munkahelyet, és ha nem vigyázunk, kiirtja az emberiséget.

Az MI-vel kapcsolatos mérnöki, etikai, jogi, közgazdasági és szociológiai vitákban ott kísért egy filozófiai kérdés. Kétségtelen, hogy az MI ügyesen végez el olyan feladatokat, melyekhez hagyományosan szellemi munka kellett, de ténylegesen *értelmes*-e, gondolkodik-e? Ez a kérdés többek közt azért érdekes, mert jelentése még az MI-szakemberek számára sem teljesen világos. Nincs hivatalos definíciónk sem az értelemre, sem a gondolkodásra.

Az MI-ről már most hajlamosak úgy beszélni, mintha kognitív tevékenységeket végezne – „válaszol", „magyaráz", „elemez" stb. –, de joggal merül fel a kérdés, hogy tényleg csinál-e ilyesmit, vagy csak szöveget generál egy mechanikus algoritmus alapján. Annak érdekében, hogy a két szóban forgó jelenség (a gépi és az emberi értelem) elvben megkülönböztethető legyen, az emberi gondolkodás algoritmikus szimulációját Cantwell Smith (2019) nyomán ⌜szögletes idézőjellel⌝ jelölöm. Az MI ⌜válaszol⌝, ⌜magyaráz⌝, ⌜elemez⌝, vagyis olyan számításokat

végez, melyeket mi emberek válaszként, magyarázatként, elemzésként értelmezünk. Lehet, hogy ez az értelmezés helyes, de lehet, hogy nem. A vita az alábbi kérdés és az alábbi tézis körül forog:

Kognitív Demarkációs Kérdés Gondolkodás-e a ⌜gondolkodás⌝?

Kognitív Demarkációs Tézis A ⌜gondolkodás⌝ nem gondolkodás.

Elvben megfogalmazhatunk nem-kognitív demarkációs téziseket is (pl. az MI nem tud érezni, csak ⌜érezni⌝, nem tud tüsszenteni, csak ⌜tüsszenteni⌝), de az alábbiakban végig a kognitív demarkációs kérdésről lesz szó. A technológia jelenlegi szintjén ez tűnik igazán relevánsnak, és leginkább ezzel kapcsolatban polarizálódnak a vélemények. A „demarkációs tézis" és „demarkációs probléma" tehát a továbbiakban a kognitív demarkációs tézist, illetve problémát jelöli.

Nem mindenki tartja fontosnak a demarkációs problémát. A legendás számítógéptudós, Edsger Dijkstra szerint „az a kérdés, hogy tudnak-e gondolkodni a gépek [...] nagyjából annyira releváns, mint az a kérdés, hogy tudnak-e úszni a tengeralattjárók" (Dijkstra 1984). A mester itt vélhetőleg arra gondol, hogy az „úszik" szót úgy használjuk, ahogy akarjuk, és bármit is jelölünk vele, az empirikus tények ettől nem változnak. De ez minimum félrevezető. Az „úszik" szót valóban úgy használjuk, ahogy akarjuk, de a szó természetes jelentése alapján egyértelmű, hogy a tengeralattjárók tudnak úszni, míg a „gondolkodik" szó természetes jelentése alapján nem egyértelmű, hogy a számítógépek tudnak gondolkodni. Az úszás egy funkcionális tulajdonság: valami akkor tud úszni, ha képes a vízben előrehaladni. Ezzel a funkcionális tulajdonsággal az emberek és a tengeralattjárók egyaránt rendelkeznek, bár eltérő módon realizálják. A vízi mozgás esetében tehát nincs különbség úszás és ⌜úszás⌝ között. Az értelem esetében viszont elvben lehet különbség gondolkodás és ⌜gondolkodás⌝ között. Ahhoz, hogy ezt tagadjuk, igazolnunk kell, hogy a gondolkodás is puszta funkcionális tulajdonság. Ez a kérdés átvezet az MI-értelemmel kapcsolatos leghíresebb empirikus teszthez.

1.1. A Turing-teszt és a gondolkodás fogalma

Alan Turing (1950) híres cikke egy „imitációs játékot" ír le, melynek során valaki látatlanban tesz fel kérdéseket egy férfinak (A) és egy nőnek (B). A kérdezőnek az a feladata, hogy megállapítsa a válaszolók nemét, A és B pedig arra törekszik, hogy a kérdező őt gondolja nőnek.

> Mi történik, ha egy gép veszi át A helyét a játékban? Vajon a játékos ugyanolyan gyakran téved majd, mint férfi és nő megkülönböztetésekor? Ezek a kérdések váltják fel azt az eredeti kérdést, hogy tudnak-e gondolkodni a gépek. (Turing 1950: 434)

Turing (1950: 442) azt jósolta, hogy 2000-re a gépek az esetek 30%-ában megtévesztik majd az átlagos kérdezőt. Ez a jóslat egyes vélekedések szerint 2012-ben vált valóra, amikor az „Eugene Goostman" nevű, orosz-ukrán fejlesztésű chatbot a kérdezők 29%-ával, két évvel később pedig a kérdezők 33%-ával hitette el, hogy valódi tinédzser. Azóta vélhetőleg nőttek az arányok. Következik-e ebből, hogy a demarkációs tézis hamis?

Ned Block (1981: 8–9) hívta fel rá a figyelmet, hogy a Turing-teszt sokkal zavarosabb koncepció, mint elsőre gondolnánk, mert Turing semmit nem mond arról, milyen tulajdonságokkal kell rendelkeznie a kérdezőnek. Ha a kérdező *akárki* lehet, akkor az intelligencia Turing által javasolt fogalma használhatatlanul tág, hiszen egy kellően naiv kérdező egy primitív programot is embernek fog nézni. Ha viszont a kérdezőnek valamilyen speciális képességgel kell rendelkeznie („érzékenynek kell lennie az értelem jeleire"), akkor Eugene Goostman nem biztos, hogy átment a Turing-teszten, hiszen elképzelhető, hogy a kérdezők, akik embernek hitték, nem rendelkeztek a szóban forgó speciális képességgel.

Van azonban egy ennél is mélyebb probléma. A Turing-teszt eleve csak azt tudja ellenőrizni, hogy az MI ⌜gondolkodik⌝-e, képes-e a mi megszólalásainkhoz hasonló outputokat generálni. De felvethető, hogy az értelemnek ahhoz is szoros köze van, *hogyan* jut el valaki a kérdésről a válaszig, mi történik közben a fejében. A Turing-teszt ezt nyilván nem méri. Érdekes módon ezt a problémát Turing maga is felismerte. Az „imitációs játék" segítségével nem *definiálni* akarta a gondolkodást, hanem azt javasolta, hogy *cseréljük le* az értelem hagyományos fogalmát egy másik, funkcionális fogalomra. Azt a felvetést, hogy az értelem ennél több, az alábbi módon semlegesítette:

> *A tudatból vett ellenérv.* Ezt az érvet remekül dolgozza ki Jefferson professzor 1949-es Lister-előadása. Idézem: „Amíg egy gép nem képes szonettet vagy versenyművet írni, mégpedig a saját gondolatai és érzései alapján és nem szimbólumok véletlen rajzása miatt, addig nem állíthatjuk, hogy a számítógép és az emberi agy azonos szinten vannak. Legalábbis addig nem állíthatjuk, amíg a gép nemcsak létrehoz egy ilyen műalkotást, hanem *tisztában is van vele*, hogy létrehozta. Egy gép nem képes örömöt érezni a sikerei miatt, sem szomorúságot, ha elromlanak a szelepei, bizsergést a hízelgés hallatán, nem képes szégyenkezni, ha hibázott, nem ismeri a szexuális vonzalmat, nem tud feldühödni vagy elszomorodni amiatt, hogy vágyai meghiúsultak."
>
> Ebből az érvből látszólag az következik, hogy az általam javasolt teszt haszontalan. A nézet legszélsőségesebb változata szerint kizárólag onnan tudhatjuk, hogy egy gép gondolkodik, ha a gép mi magunk vagyunk, és érezzük, hogy gondolkodunk. Persze kifejezhetjük másoknak az érzéseinket, de senkinek sem lesz oka rá, hogy higgyen nekünk. Csak akkor vehetjük tehát biztosra, hogy egy

> ember gondolkodik, ha mi vagyunk az adott ember. Végeredményben szolipszizmushoz jutunk, ami a maga módján roppant logikus lehet, de megnehezíti a gondolatok közlését. *A* hajlamos lesz azt gondolni, hogy „*A* gondolkodik, de *B* nem", *B* pedig azt, hogy „*B* gondolkodik, de *A* nem". Ilyen parttalan viták helyett jobb azzal az udvarias feltevéssel élni, hogy mindannyian gondolkodunk. (Turing 1950: 445–446)

Ha a puszta funkcionalitást zárójelbe téve arra akarunk rákérdezni, hogy a gépek *ténylegesen* gondolkodnak-e – vagyis a demarkációs problémával foglalkozunk –, akkor Turing szerint egy abszurd állásponthoz, a szolipszizmushoz jutunk, amit az udvariasság érdekében jobb elkerülni. De ez a levezetés hibás. Mások érzéseihez és gondolataihoz valóban csak közvetve férünk hozzá – megszólalásaik, viselkedésük révén –, de ebből nem következik, hogy ne *tudnánk*, hogy másoknak is vannak érzései és gondolatai. Filozófiai értelemben persze nem teljesen világos, honnan tudjuk ezt. Mások elméjének ismerete ugyanolyan filozófiai talány, mint a külvilág létével kapcsolatos tudás. A legegyszerűbb megoldás szerint onnan tudjuk, hogy mások is gondolkodnak, hogy párhuzamot vonunk saját magunk és a hozzánk hasonlónak tűnő többi organizmus között (Avramides 2023). Ennél bonyolultabb megoldások is léteznek és természetesen mindegyik megoldásba belekötött már valaki, de bizarr lenne azt állítani, hogy tagadnunk kell mások belső élményeinek létét, ha saját belső élményeinket valósnak fogadjuk el.

Turing funkcionalista nyelvújítása azonban egy valós filozófiai kérdésre világít rá: ha az értelem nem funkcionális tulajdonság, akkor micsoda? Mit jelent a szó „valódi" értelmében gondolkodni?

A gondolkodás olyan jelenség, amit közvetlenül („gondolkodás nélkül") felismerünk, de ha elemzés tárgyává tesszük, akkor hétköznapi értelmünk számára szokatlan terepre tévedünk és elnyel minket az elmefilozófiai szakkifejezések dzsungele. Ennek elkerülése érdekében megpróbálom a természetes nyelvre támaszkodva körülírni, hogy mit értek „gondolkodás"-on.

A gondolkodás – Kahneman (2011) terminológiájában a 2-es típusú vagy lassú gondolkodás – mindenekelőtt egyfajta *látás*. *Látjuk*, hogy 1+2=3, *látjuk*, hogy $A \wedge B$-ből következik B, *látjuk*, hogy a változás fogalma magában foglalja az idő fogalmát. A gondolkodás paradigmatikus esetei persze ennél kiterjedtebbek: nem pillanatnyi felismerések, hanem tudatos folyamatok. Arra a folyamatra hasonlítanak, amikor végigtekintünk egy tájon vagy egy festményen és a maga egészében befogadjuk. Ilyen értelemben *látjuk* például azt, hogy a Hilbert-hotelben végtelen sok vendég fér el akkor is, ha a hotel teljesen tele van, és ilyen értelemben látjuk a Gettier-eseteket elemezve, hogy a tudás nem azonos az igazolt igaz hittel.

A látás csupán metafora, de érthető és ősi metafora, ami a természetes nyelvbe is beépült a gondolkodás szinonimájaként (átlátni, belátni valamit).

Értjük, mint jelent a fenti példákban látásról beszélni (*látjuk*, mi a látás). A metafora már Platónnál jelen van, aki az értelmet „a lélek szemének" nevezi (*Az állam* 533d), a Jót pedig, a lét és értelem végső forrását, a fényhez hasonlítja (*Az Állam* 518c).

A gondolkodás, hasonlóan ahhoz az élményhez, hogy az ember a sötétségből kiér a fényre, megélt folyamat. Rengeteg formája van: elemzés, következtetés, mérlegelés, tépelődés, spekulálás, felfogás, megértés. De bármelyik formáját is tekintjük, két alapvető aspektusát különíthetjük el. A gondolkodás egyrészt belső tapasztalat, másrészt specifikus tartalma van és tartalma révén lehetőségekre vagy tényekre vonatkozik. A gondolkodás képességét ez utóbbi aspektus szerint a racionalitással azonosíthatjuk, amit Peter van Inwagen így jellemez:

> [Az racionális, aki képes] komplex lehetőségeket reprezentálni, beleértve olyan lehetőségeket, melyek radikálisan eltérnek jelenlegi érzékszervi élményeitől. (Például: Jane két hét múlva kedden érkezik; valaki a második legolcsóbb ételt rendelte az étlapról; a kormány megakadályozta a bubópestis újbóli kitörését azáltal, hogy új szemétszállítási módszert dolgozott ki, megfosztva a pestissel fertőzött bolhákat hordozó patkányokat a tápláléktól.) Képes valóságosnak vagy valótlannak ítélni lehetséges szituációkat, és át tudja tekinteni őket anélkül, hogy feltenné a kérdést, valósak vagy valótlanok-e. („Megpróbálom elképzelni, milyen lenne az életünk, ha gyereket vállalnánk.") Képes a lehetőségeket a „valószínű" vagy „valószínűtlen" kategóriák alá sorolni, képes felismerni logikai és oksági kapcsolataikat, képes fontossági sorrendbe állítani őket. („Bocsánat, hogy zavarba hoztalak. Nem akartam, tényleg. De azt gondoltam, még mindig jobb, mint ha hazudnék.") Terveket tud kidolgozni az alapján, hogy miből mi következik és mi tűnik valósnak, lehetségesnek, valószínűnek vagy valószínűtlennek, és úgy képes erre, hogy tervei az általa fontosnak tartott szituációk megvalósulását segítsék. Képes felismerni, hogy mások is rendelkeznek ugyanezekkel a képességekkel, és képes utasításokat és kérdéseket intézni hozzájuk a reprezentált szituációkkal, valamint az ezekre vonatkozó hitekkel, vágyakkal és értékítéletekkel kapcsolatban. Racionális tehát az a lény, aki tud állítani, utasítani és kérdezni; és azért képes erre, mert az ilyen megnyilvánulásoktól függetlenül, önmagában is van benne *valami*, ami alapján állít, utasít és kérdez. (van Inwagen 2009: 171)

A továbbiakban a fenti képesség (az ész) használatát nevezem gondolkodásnak, a képesség birtoklását pedig értelemnek. A demarkációs érvek mindegyike azzal foglalkozik majd, hogy az MI ebben az értelemben gondolkodik-e.

1.2. A demarkációs probléma metafizikája

A demarkációs érveket két nagy csoportra osztom, az *elmeoldali* és *számításoldali* érvekre. Számításoldali demarkáció esetén olyan tulajdonságot keresünk, amit a számítógépek természete magában foglal, az emberi elme természete viszont kizár; elmeoldali demarkáció esetén olyan tulajdonságot, amit az elme természete magában foglal, a számítógépek természete viszont kizár.

Az ember elme (vagy a számítógép) természetét az emberi elme (vagy a számítógép) lényegi tulajdonságaival azonosítom, a lényegek metafizikájával kapcsolatban pedig nagy vonalakban Kit Fine (1994) koncepcióját fogadom el. Fine szerint a lényeg sajátos metafizikai jelenség, ami nem redukálható szükségszerűségekre. Abból például, hogy x szükségszerűen F, nem következik, hogy F része x lényegének. Nincs általános szabály rá, milyen tulajdonságok tartoznak egy dolog lényegéhez azon túl, hogy negatív, diszjunktív és konjunktív tulajdonságok tipikusan nincsenek köztük. Bár a lényegek nem redukálhatóak szükségszerűségekre, vannak modális következményeik: ha F része x lényegének, akkor x szükségszerűen F, és ha x lényege kizárja F-et, akkor lehetetlen, hogy x F legyen.

A demarkációs érvekben szereplő F-ek tipikusan intrinzikus kategorikus tulajdonságok vagy intrinzikus képességek lesznek. Intrinzikus kategorikus tulajdonság például a nyelvi megértés élménye (6. fejezet), intrinzikus képesség pedig a Gödel-mondatok igazságának felismerése (9. fejezet). Egyetlen olyan érv kerül majd elő – az implementációs demarkáció (20. fejezet) –, melyben F egy extrinzikus tulajdonság.[1]

Felvethetnénk, hogy mivel végeredményben modális kritériumokról van szó, elegendő a demarkáció során olyan tulajdonságokat keresni, melyekkel az emberek szükségszerűen rendelkeznek, a gépek pedig szükségszerűen nem rendelkeznek, vagy fordítva. Ez azonban elmossa az elmeoldali és számításoldali demarkáció közti különbséget. Intuitíve éles különbség van aközött, hogy alapvetően emberi vagy alapvetően gépi tulajdonság mentén különböztetjük meg a gondolkodást a ⌜gondolkodás⌝-tól. De ha minden ember szükségszerűen F és minden számítógép szükségszerűen nem F, akkor $\neg F$ kielégíti azt a kritériumot, hogy az emberek szükségszerűen rendelkeznek vele és a gépek szükségszerűen nem rendelkeznek vele, tehát minden elmeoldali demarkáció egyben számításoldali demarkáció is lesz. Ennek elkerülésére bevezethetjük a pozitív és negatív tulajdonságok különbséget: elmeoldali demarkációnál az elme valamely *pozitív*

[1] Az extrinzikusságot több, egymással logikailag inkompatibilis módon is definiálni lehet (Marshall és Weatherson 2023). Jelen kontextusban azt a definíciót feltételezem, mely szerint „intrinzikus tulajdonság az, amivel valami kizárólag amiatt rendelkezik, amilyen önmagában" (Lewis 1983: 197). (Figyelmeztetés: Lewis cikke ugyanezen a ponton két teljesen más meghatározást is ad és úgy tesz, mintha ezek ekvivalensek lennének.) F eszerint akkor extrinzikus tulajdonság, ha van olyan G, hogy bármely lehetséges x amiatt F, hogy egy x-től különböző (x-től mereológiailag distinkt) dolog G. Az „amiatt" a metafizikai eredetre utal (ld. 155. old.).

szükségszerű tulajdonságát keressük, számításoldali demarkáció esetén pedig a gépek *pozitív* szükségszerű tulajdonságát. A pozitív és negatív tulajdonságok fogalma azonban hasonlóan komoly metafizikai vállalás, mint a lényegek fogalma, ráadásul nem is feltétlenül világosabb ez utóbbinál.

1.3. Filozófián kívüli demarkációs érvek

Demarkációs érveket nemcsak a filozófiában találunk, sőt, felvethető, hogy az igazán megalapozott demarkációs érveket manapság a mérnököktől és matematikusoktól várhatjuk, mert az MI-laborok rengeteg cikket publikálnak a modellek kognitív képességeivel kapcsolatban. Laikus szemlélőként az a benyomásom, hogy a filozófiához hasonlóan itt sincs konszenzus a szakértők között, ráadásul üzleti érdekek is befolyásolják az álláspontokat. Az Apple kutatói például, miután a cég hátramaradt az MI-versenyben, publikáltak egy tanulmányt „A gondolkodás illúziója" címmel (Shojaee et al. 2025), melyben azt próbálták igazolni, hogy az aktuális MI csúcsmodellek matematikai fejtörők megoldásakor egy bizonyos komplexitási szint felett összeomlanak (ami vélhetőleg nem történne meg, ha értenék a fejtörők alapstruktúráját). Erre az egyik legnevesebb AI-startup, az Anthropic „A gondolkodás illúziójának illúziója" című tanulmánnyal válaszolt (Opus és Lawsen 2025), azzal vádolva az Apple munkatársait, hogy megoldhatatlan feladatokat adtak az MI-nek és nem vették figyelembe az outputok hosszára vonatkozó korlátokat.

A filozófián kívüli demarkációs érvek kapcsán számomra az ARC (Abstraction Reasoning Corpus) tűnik a legérdekesebbnek, amit François Chollet számítógéptudós dolgozott ki (Chollet 2019, Chollet et al. 2025). Az ARC emberi szemmel egyszerűnek tűnő vizuális feladatokból áll, ahol néhány példából általánosítva kell kiegészíteni egy mintázatot. Az első két-két képen például azt látjuk, hogy egy négyzetrácson egy piros paca odaugrik a kék paca alá és rátapad (mint a Tetrisben), az ötödik képen egy piros és egy kék pacát látunk, melyek távol vannak egymástól; a feladat kitalálni, mi lesz a hatodik képen (Chollet 2019: 49). Az ilyen feladatok az MI számára sokszor nagyon nehezek, emberek számára viszont viszonylag könnyűek. A feladatsor ARC-AGI 2 verzióját (2025. március) 20%-os eredménnyel oldják meg a legjobb MI-k, míg emberek képesek 100%-os eredményt elérni. Az ARC-teszt azt sejteti, hogy a jelenlegi MI architektúra alapvető kihívásokkal küzd az absztrakció és térbeli gondolkodás terén. Gyanítom, hogy a döntő erejű demarkációs érvek (ha lesznek) az ilyen mérnöki kutatásokra épülnek majd. A Konklúzióban kitérek rá, hogyan lehet a filozófiai demarkációs érveknek is empirikus vonzata.

2.

Mesterséges intelligencia

Az MI-technológia története két nagy paradigma, a szimbolikus MI és a konnekcionista MI vetélkedéséből áll. Sokáig a szimbolikus MI volt domináns, de végül a konnekcionista MI hozta el az áttörést és jelenleg is ez az uralkodó. A két paradigma rövid jellemzése után rátérek arra, hogyan működnek a ChatGPT-hez hasonló algoritmusok, majd amellett érvelek, hogy az általuk végzett matematikai műveletek Turing-géppel modellezhetőek, emiatt pedig a Turing-gépekkel kapcsolatos matematikai és filozófiai érvek közvetlenül alkalmazhatóak a mai MI-kre.

2.1. A transzformer-architektúra

Tegyük fel, hogy létre akarunk hozni egy programot, ami a Harry Potter-regények világáról ad információt. Ha begépelünk egy kérdést a regények szereplőivel kapcsolatban – „Varázsló-e Harry Potter?", „Milyen viszonyban van egymással Harry két legjobb barátja?" –, akkor a program helyes választ ad rá.

Képzeljünk el két különböző megoldást erre a számítástechnikai feladatra. Az **A** megoldás esetében a program szintaktikailag elemzi a kapott kérdést. Megkeresi, hol van benne az alany és az állítmány, majd egy fogalmi térkép segítségével megállapítja, mi a helyes válasz. A fogalmi térkép a Harry Potter-regények világát képezi le: emberek, szörnyek, épületek, helyszínek és események nevei kapcsolódnak össze egymással és különféle kategóriákkal. A „Harry Potter" címkéhez olyan kategóriákat rendelünk, mint „varázsló", „árva", „főszereplő" stb., és összekapcsoljuk más szereplők nevével. Ha a bementi kérdés az, hogy Harry Potter varázsló-e, akkor program a „Harry Potter"-hez kapcsolódó kategóriák között keresi majd a „varázsló" címkét. Ha a kérdés az, hogy Hermione és Ron ellenségek-e, akkor a „Hermione Granger" és „Ron Weasley" objektumok közt definiált kapcsolatokat veszi szemügyre.

A **B** megoldás teljesen más. A **B** típusú program az inputot nem explicit nyelvi elemzésnek veti alá, hanem felbontja *token*ekre (szavaknál kisebb, két-

három betűből álló töredékekre), majd mindegyik tokenhez hozzárendel – egy előre elkészített adatbázis alapján – több tízezer számot; mindegyik tokenhez ugyanannyit. Nevezzük ezeket a számsorozatokat *vektorok*nak. A következő lépésben a program bonyolult algebrai transzformációknak veti alá a vektorokat, végül pedig előállít belőlünk egy valószínűség-eloszlást, melynek segítségével kisorsolja, mi legyen az output első tokenje. Ezt hozzácsapja a kérdéshez, majd ugyanezt a számítást elvégzi a következő token előállításához, és így, morzsánként építi fel a választ.

Ez a példa a mesterséges intelligencia elmúlt hetven évének két versengő paradigmája, a szimbolikus MI (**A**) és a neurális hálókra épülő vagy konnekcionista MI (**B**) közti különbséget érzékelteti.[1] A szimbolikus MI – amit Haugeland (1985) nyomán GOFAI-nak, „Good Old Fashioned AI"-nak (régimódi MI) is szoktak nevezni – a reprezentált jelenségek absztrakt modelljével dolgozik és nyelvtani/logikai elemzésnek veti alá a nyelvi inputokat. A konnekcionista MI ezzel szemben nagyon bonyolult vektoralgebrai számítások alapján, az input explicit nyelvtani vagy logikai elemzése nélkül tippeli meg a helyes választ.

A szimbolikus MI mint *program* persze szintén aritmetikai műveleteket végez, de fontos különbség, hogy a szimbolikus MI esetében e műveletek *fölött* van egy olyan absztrakciós szint – pl. a Harry Potter-univerzum fogalmi térképe –, amit az MI fejlesztői tudatosan hoznak létre. A konnekcionista MI esetében nincs ilyen előzetes absztrakciós szint, a fejlesztők tudatosan nem írnak bele a programba Harry Pottert reprezentáló kategóriákat. Az pedig, hogy a „motorháztető alatt", a vektorok viharában pontosan milyen értelemben zajlik absztrakció egy konnekcionista MI futtatásakor, nyitott kérdés.

A történelem fura fintora, hogy a két paradigma versenyében a konnekcionista MI győzött, a szimbolikus MI pedig több nekibuzdulási hullám után relatíve háttérbe szorult.[2] A ChatGPT és társai a konnekcionista architektúra elképesztően bonyolult, hajmeresztő számítási kapacitást igénylő változatai, úgynevezett nagy nyelvi modellek vagy LLM-ek (*Large Language Model*).

A konnekcionista MI ereje azokból a paraméterekből vagy súlyokból (*weights*) ered, melyek segítségével a program a vektorok sorozatává alakított inputból előállít egy új vektort (a soron következő tokenre vonatkozó valószínűség-eloszlást). Ezeket a súlyokat (kicsit leegyszerűsítve) úgy hozzák létre a mérnökök, hogy fognak egy véletlen paraméterekkel feltöltött LLM-et, ráeresztik a világ összes digitalizált szövegére, és egy függvény alapján módszeresen addig finomítgatják a paramétereket, amíg a program képes lesz egy

[1] A **B** paradigma fenti leírása igazából túlmegy a konnekcionizmuson és általában a vektortér-alapú statisztikai rendszereket fedi le. Ezen belül a konnekcionista rendszerek sajátosságát a betanítás során kialakított paraméterek jelentik, melyek a betanítási adatbázis belső struktúrájától függnek.

[2] Az LLM-ek egyes kritikusai, például Yann LeCun és Gary Marcus szerint a szimbolikus MI-től várhatjuk csak az újabb komoly szintlépést. A szimbolikus MI és a filozófiatörténet összefüggéseiről érdekes áttekintést ad Haugeland (1985).

tetszőleges szöveg bármelyik pontján kellő pontossággal előrejelezni a következő szót.[3]

A továbbiakban azt feltételezem, hogy amikor az MI értelmi képességeiről spekulálunk, akkor a maiaknál esetleg sokkal bonyolultabb, több paraméterrel rendelkező, nagyobb számítási kapacitást igénylő, de működési logikájukat tekintve lényegileg hasonló konnekcionista rendszerekről beszélünk. Az „MI" szó tehát a továbbiakban az emberi nyelven társalgó nagy nyelvi modellekre és közeli leszármazottaikra vonatkozik.

Az ilyen rendszerek alapvető működési logikája a korábban kissé elnagyoltan jellemzett *next-token prediction*. A ChatGPT és társai az inputként kapott mondatokat első körben felbontják úgynevezett *token*ekre, rövid szótagokra és szimbólumokra. (A tokenek többsége nem egy teljes szó, hanem 3 byte-nyi szöveg, ami általában három betűt jelent.) Ezután a tokeneket sok ezer számból álló számsorozatokra cserélik le, vagyis „beágyazzák" őket egy sokdimenziós vektortérbe („embedding"). Miután a természetes nyelvi inputot tokenekké, a tokeneket pedig vektorokká alakította, a program rengeteg lineáris algebrai műveletet végez a létrejött vektorokon. Az LLM-ek alapját jelentő „transzformer architektúra" alapvető logikája a következő (Viswami et al. 2017):

(1) **Beágyazás** (*embedding*): az input felbontása tokenekre ($t_1 \dots t_k$) és a tokenek lecserélése (egy fix „token-vektor szótár" alapján) vektorokra, melyek mindegyike n db. számból áll (az n értéke a modelltől függ, a ChatGPT 3 esetében nagyjából 12000 volt). A végeredmény egy vektorokból (azonos hosszúságú számsorozatokból) álló sorozat: $\mathbf{v}_1 \dots \mathbf{v}_k$.

(2) **Pozíciós kódolás** (*positional encoding*): A $\mathbf{v}_1 \dots \mathbf{v}_k$ vektorok átalakítása az $i = 1 \dots k$ index, vagyis a pozíciójuk függvényében. Ez tipikusan egy $f(x) = x \cdot g(i) + h(i)$ alakú transzformációt jelent minden $\mathbf{v}_i$ vektor minden komponense (a vektort alkotó n db. szám) esetében.

(3) **Többfrontos önfigyelő réteg** (*multi-head self-attention layer*): Mindegyik $\mathbf{v}_i$ vektort megszorozzuk két-két mátrixszal, majd az eredményeket összeadjuk és normalizáljuk; így létrejön az eredeti $\mathbf{v}_1 \dots \mathbf{v}_k$ vektorsorozat átalakított változata, $\mathbf{v}^1_1 \dots \mathbf{v}^1_k$.

(4) **Perceptron réteg** (*feedforward layer*): Az előző lépésben kapott vektorokat megszorozzuk egy mátrixszal; az eredmény az előző fázis kimeneteinek újabb transzformációja, $\mathbf{v}^2_1 \dots \mathbf{v}^2_k$.

[3] Az LLM-ek működésének és betanításának technikai részleteit a lehetőségekhez mérten felhasználóbarát módon taglalja Prince (2023). A szimbolikus és konnekcionista MI történetéről áttekintést ad Russell és Norvig (2022: 35–45), a konnekcionizmus elmefilozófiai jelentőségéről Ramsey, Stich és Garon (1990), az LLM-ek és az emberi elme közti lehetséges párhuzamokról pedig Millière és Buckner (2024).

(5) A 3-as és 4-es lépés ismétlése további mátrixok bevonásával. Legyen az ismétlések száma z. Az ismétlések végén előáll egy végső vektorsorozat, $\mathbf{v}_1^{2z} \dots \mathbf{v}_k^{2z}$.

(6) **Kiágyazás** (*unembedding*): Megszorozzuk $\mathbf{v}_k^{2z}$-t (az utolsó vektort) egy mátrixszal, melynek n db. oszlopa van és annyi sora, ahányféle tokent a rendszer felismer. A végeredmény egy $\mathbf{t}$ vektor.

(7) **Sorsolás**: Normalizáljuk a $\mathbf{t}$ vektort úgy, hogy a komponensei 0 és 1 közé eső számok legyenek, melyek összege 1. Ezeket valószínűségekként értelmezzük, és kisorsoljuk az egyik tokent. Ez lesz az output első (vagy következő) morzsája.

(8) **Új kör**: Hozzáadjuk az inputhoz az új tokent, és az eredményt inputként kezelve újrakezdjük az egészet az 1. lépéstől.

Mint a leírásból sejthető, egy tipikus LLM meglehetősen sok paramétert tartalmaz, részben a beágyazáshoz használt token-vektor szótár, részben a 3. és 4. lépés iterációiban szereplő mátrixok miatt. A ChatGPT 3-nak kb. 175 milliárd paramétere volt, a 2025 elején élvonalbelinek számító DeepSeek R1 modellnek pedig 670 milliárd. A paraméterek kalibrálása az MI-alkímia legfontosabb, féltve őrzött titka.

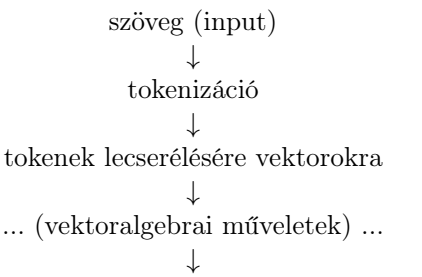
Joggal merül fel a kérdés, hogy egy ilyen mechanizmus *miért* képes emberszerű, informatív válaszokat adni. Ezzel kapcsolatban a szakemberek is kicsit tanácstalanok (Prince 2023: 20. fejezet). Az LLM-ek tipikus mérnöki csodák abban az értelemben, hogy fantasztikusan jól működnek, de készítőik praktikusan hatékonyan módszerekre, nem pedig a szóban forgó jelenség tudományos modelljeire támaszkodnak. Az egyik népszerű hipotézis szerint az LLM-ek a paraméterek kialakításakor egyfajta „szemantikai vektorteret" hoznak létre, mely vektorok közti viszonyokkal reprezentálja a szavak és mondatok szemantikai viszonyait (ld. pl. Kozlowski, Dai és Boutyline 2025). Ennek naiv illusztrálásához képzeljünk el egy algoritmust, ami egy adott „XYZ" szót úgy alakít át egy $\mathbf{v}_{\text{XYZ}}$ vektorrá, hogy ezután vektorok közti algebrai viszonyokban tükröződnek majd olyan szemantikai összefüggések, mint hogy a „királynő" jelentése a „király" jelentését kombinálja a nő és férfi közti különbséggel: $\mathbf{v}_{\text{KIRÁLY}} + \mathbf{v}_{\text{NŐ}} - \mathbf{v}_{\text{FÉRFI}} = \mathbf{v}_{\text{KIRÁLYNŐ}}$. A jelentések hasonlóságát is reprezentálni lehet azzal, hogy a rokon jelentésű, tematikusan összekapcsolódó szavak egymáshoz „közel levő" vektoroknak felelnek meg. A jól paraméterezett, élethűen csevegő LLM-ek *valami* ilyesmit csinálnak, legalábbis a szóban forgó hipotézis szerint. (Az analógia nem tökéletes, mert az LLM-ek vektorok *sorozatává*, nem pedig vektorokká alakítják a szavakat.)

A szemantikai vektortér hipotézise megmagyarázza azt is, miért szoktak az MI-k hallucinálni. Tekintsük a ChatGPT 4 válaszát arra a kérdésre, hogy ki volt az első ember, aki hátramenetben mászta meg a Mount Everestet:

> Az első ember, akiről tudjuk, hogy hátramenetben mászta meg a Mount Everestet, Adrian Ballinger amerikai hegymászó és hivatásos hegyi vezető. A beszámolók szerint az Everest egyes részeit hátramenetben mászta meg, hogy eddze szervezetét és felhívja a figyelmet a mentális egészség és az extrém szituációkban való kreatív gondolkodás fontosságára.[4]

Adrian Ballinger természetesen nem az első, aki hátramenetben mászta meg az Everestet, ilyet ugyanis soha senki nem csinált. Ballinger valójában a negyedik amerikai, aki *oxigénpalack nélkül* mászta meg az Everestet *és* a K2-t. A hallucinációt talán az okozta, hogy a „hátramenetben megmászni az Everestet" állítmány a ChatGPT szemantikai vektorterében az Everesttel kapcsolatos extrém teljesítmények régiójában helyezkedik el, és ehhez a régióhoz relatíve közel esett az „Adrian Ballinger" név. A program tehát annyit ⌜értett⌝ a kérdésből, hogy az Everesttel kapcsolatos őrült mutatványról van szó, ebben a kategóriában pedig Ballinger nevét sorsolta ki, majd a promptban szereplő állítmányt nyelvtanilag korrekt módon illesztette a válaszba. A hallucinációk (a helyes válaszokhoz hasonlóan) abból erednek, hogy az LLM a beágyazásból adódó valószínűségek alapján tippelget, de ezek a valószínűségek nem képezik le hűen az emberi nyelv rekurzív szemantikai struktúráit. Emiatt nevezik szkeptikus kommentátorok néha „sztochasztikus papagáj"-nak vagy „szteroidozott *autocomplete*"-nek az LLM-eket.

A mai MI-k a fent leírt műveleteken kívül egyéb számításokat is végeznek, de a mélyben mindig a transzformer-architektúra valamilyen változata található. A demarkációs probléma szempontjából emiatt a jelenlegi MI-k működési logikája a következő folyamatábrával foglalható össze:



Felmerülhet az a vád, hogy a fenti leírás az MI betanításával (*training*) kapcsolatos fogalmakat mossa össze az MI futtatását (*inference*) érintő fogalmakkal. A szó szigorú értelmében csak a betanítás – az 1–5. lépésben használt mátrixok kialakítása – tekinthető konnekcionista projektnek, maga a futtatás a jó öreg

[4] Köszönet Kodaj Bencének a promptért.

szimbolikus MI példája, hiszen a fenti algoritmus vektorokat manipulál explicit szabályok alapján. Ráadásul a betanításnak is csak a felügyelet nélküli része (*unsupervised learning*) alapul konnekcionista elveken. Ha a tesztelők visszajelzései alapján alakítjuk ki a modell paramétereit (*supervised learning*), akkor tudatosan huzalozunk bele szemantikai összefüggéseket.

A fogalmi zavar elhárítása érdekében hangsúlyozni szeretném, hogy „MI"-n, ha másképp nem jelzem, természetes nyelven ⌜kommunikáló⌝ rendszereket értek, és ezeket a szó *elmefilozófiai* értelmében, az outputjaik által kifejezett propozicionális tartalmak vonatkozásában nevezem konnekcionistának vagy szimbolikusnak. A szó elmefilozófiai értelmében egy MI akkor konnekcionista, ha az architektúrája nem reprezentálja explicit módon az általa kezelt propozicionális tartalmak struktúráját. A ChatGPT kódjában például nincs olyan osztály vagy függvény, ami specifikusan a varázslók fogalmának felelne meg, hanem a rendszer bizonyos értelemben holisztikusan tárolja az olyan igazságokat, mint hogy „Harry Potter egy varázsló" (vö. Ramsey, Stich és Garon 1990). Ebben az elmefilozófiai értelemben a transzformer-architektúra (megfelelő paraméterekkel feltöltve) futtatás közben is konnekcionistának minősül.

2.2. Turing-gépek

A számítógépes programok leghíresebb matematikai modellje a Turing-gép, melyet Alan Turing tíz évvel azelőtt alkotott meg, hogy az első programozható számítógép, az elektroncsövekből álló, harminc méter hosszú ENIAC létrejött volna. Turing a mechanikus számítási folyamatokat akarta modellezni annak érdekében, hogy kiderítse, elvben mit lehet és mit nem lehet gépi módszerekkel kiszámítani. A fejezet lezárásaként amellett fogok érvelni, hogy a transzformer-architektúra Turing-gépnek tekinthető, emiatt pedig a Turing-gépekkel kapcsolatos matematikai és filozófiai érvek érvényesek a mai MI-re is.

A Turing-gépet egy papírral-ceruzával kalkuláló ember absztrakt általánosításaként lehet elképzelni. A „számítógép" szó angol megfelelője, a „computer" eredetileg számításokat végző *embert* jelentett (akik egyébként többnyire nők voltak). A mai számítógépek kifejlődése előtt nehéz szellemi munka kellett minden komolyabb mérnöki vagy tudományos számításhoz.

Tegyük fel, hogy valakinek ceruza, radír és kockás papír segítségével hozzá kell adnia 1-et egy természetes számhoz. Legyen a szám történetesen kétjegyű, azaz mn alakú. Emberünk a következőképpen fog eljárni. Először is megkeresi az input utolsó számjegyét, n-et. Utána n helyére beírja $n + 1$-et és leteszi a ceruzát, amennyiben n kisebb, mint 9. Ha $n = 9$, akkor a helyére 0-t ír, m helyére pedig $m + 1$-et és leteszi a ceruzát, amennyiben m kisebb, mint 9. Ha $m = 9$, akkor a helyére nullát ír, melléír egy 1-est, és leteszi a ceruzát. Az eljárás teljesen hasonló, ha a szám nem kétjegyű.

A humán kalkulátor a fenti „program" futtatásakor háromféle dolgot csinál:

(1) Ide-oda mozgatja ceruzáját.

(2) Megnézi, hogy abban a kockában, ahol a ceruzája épp van, milyen szimbólum szerepel.

(3) Kitörli az aktuális kocka tartalmát és beleír valamit, ami az aktuális kocka tartalmának és a kalkuláció aktuális fázisának a függvénye.

A Turing-gép az ilyen lépésekből álló mechanikus kalkulációk absztrakt modellje. Feltételezzük, hogy adott egy véges szimbólumkészlet, melynek elemei a kockás papír celláinak lehetséges értékei. A példa kedvéért legyen a szimbólumkészlet az $S = \{0 \ldots 9, \text{ÜRES}\}$ halmaz. Az ÜRES érték azt jelzi, hogy az adott cellában nincs semmi. Szigorú értelemben ez nem egy szimbólum, de a gépünk képes felismerni és „leírni” (radír segítségével). A legegyszerűbb elképzelhető szimbólumkészlet az $\{1, \text{ÜRES}\}$ (vagy általában bármely kételemű halmaz).

Legyen U természetes számok véges halmaza; ezek lesznek a program utasításainak sorszámai. Az utasítások specifikálásához szolgáló C cselekvéskészlet álljon $\langle a, b\rangle$ alakú párokból, ahol az a komponens S egyik eleme (azé a szimbólumé, amit az aktuális kockára kell írni), b pedig a $\{\text{JOBBRA}, \text{BALRA}\}$ halmaz egyik eleme. Az $\langle 1, \text{BALRA}\rangle$ például annak az utasításnak felel meg, hogy „írj az aktuális cellába 1-et, majd vidd a ceruzát a balra levő cella fölé”.

Egy Turing-gép lényegében egy táblázat, melynek minden sora $[u_1, s, c, u_2]$ alakú, ahol u_1 egy természetes szám (az adott utasítás sorszáma), s egy szimbólum ($s \in S$), c egy cselekvés ($c \in C$), u_2 pedig vagy U egyik eleme, vagy a speciális STOP utasítás, melynek hatására a gép leáll. A táblázat egy adott $[u_1, s, c, u_2]$ sora azt specifikálja, hogy az u_1 sorszámú utasítás lefuttatásakor, amennyiben az aktuális cellában az s szimbólum szerepel, a gépnek a c cselekvést kell elvégeznie, majd tovább kell mennie az u_2 parancsra (vagy le kell állnia, ha u_2 a STOP utasítás). Egy géptáblázat akkor jólformált, ha mindegyik $u \in U$-ra és $s \in S$-re van pontosan egy darab $[u, s, \ldots]$ kezdetű sor.

A korábban leírt összeadó algoritmust az alábbi táblázat definiálja:

Utasítás sorszáma	*Olvas*	*⟨Ír, Mozog⟩*	*Következő utasítás*
1	0	⟨0, JOBBRA⟩	1
1	1	⟨1, JOBBRA⟩	1
...	...	...	...
1	9	⟨9, JOBBRA⟩	1
1	ÜRES	⟨ÜRES, BALRA⟩	2
2	0	⟨1, BALRA⟩	STOP
2	1	⟨2, BALRA⟩	STOP
...	...	...	...
2	8	⟨9, BALRA⟩	STOP
2	9	⟨0, BALRA⟩	2
2	ÜRES	⟨1, BALRA⟩	STOP

Legyen a gép neve $\mathbf{T^{+1}}$. Működési logikája a következő:

(1) Keresd meg az input jobb szélső számjegyét.

(2) Ha az aktuális cellában a $0, 1, \ldots 8$ számjegyek valamelyike szerepel, akkor akkor írj a helyére $1, 2, \ldots 9$-et és állj le („tedd le a ceruzát"). Ha az aktuális cella üres, akkor írj oda egy 1-est és állj le. Ha az aktuális cellában „9" szerepel, akkor írj a helyére egy nullát, menj át a tőle balra levő cellába, és kezdd újra a 2-es utasítást.

A fenti Turing-gép szinte pontosan azt csinálja, amit a radírral és ceruzával felszerelt humán kalkulátor az $f(n) = n + 1$ számítás elvégzésekor. Annyi a különbség, hogy a Turing-gép nem egy kockás lappal, hanem egy kockás „szalaggal" dolgozik, és csak JOBBRA/BALRA mozgásokra képes, nem tudja a ceruzát függőleges irányban mozgatni. Ennek matematikai szempontból semmi jelentősége, mert nem valódi papírról és ceruzáról beszélünk, nem a technikai kivitelezés a fontos. Hasonló a helyzet azzal a furcsasággal, hogy $\mathbf{T^{+1}}$ akkor is beír valamit a cellába, ha elvben békén kéne hagynia annak tartalmát (az 1. utasítás összes soránál). A „hagyd békén az aktuális számjegyet" utasítást azzal szimuláljuk, hogy kitöröljük, majd visszaírjuk ugyanazt a számot. Értelmesebbé lehetne tenni a gépet, ha bevezetnénk a „ne bántsd a cellát" cselekvéstípust is, de ezzel bonyolítanánk az absztrakt számítási modellt anélkül, hogy matematikai értelemben bármit nyernénk.

Fontos, hogy $\mathbf{T^{+1}}$ csak akkor fog működni, ha az inputként kapott szám a gép elindításakor az írófejtől („ceruzától") jobbra található. Ha az input a program indulásakor az írófej bal oldalán van, akkor $\mathbf{T^{+1}}$ a végtelenségig mozgatja majd az írófejet jobbra. Ez jól mutatja, hogy egy Turing-gép nem feltétlenül végez tényleges számítást. Akár végtelen ciklusba is kerülhet, és mint a 10. fejezetben látni fogjuk, a Turing-gépekkel kapcsolatos egyik legérdekesebb tény az, hogy számítógéppel lehetetlen eldönteni, fenyegeti-e egy Turing-gépet a végtelen ciklus veszélye („megállási probléma").

$\mathbf{T^{+1}}$ mintájára könnyű megszerkeszteni a $\mathbf{T^{-1}}$ gépet, ami az $f(x) = x - 1$ függvényt komputálja (kivéve, ha az input 0, mert negatív számok leírását az S szimbólumkészlet nem teszi lehetővé). A $\mathbf{T^{+1}}$ és $\mathbf{T^{-1}}$ segítségével pedig építhetünk összeadógépet, ami két inputot fogad (mondjuk ÜRES cellával elválasztva) és ráírja a szalagra az összegüket. Ehhez előbb megkeressük a két inputot, mondjuk az abc-t és a tőle ÜRES cellával elválasztott ef-et, majd addig adunk hozzá abc-hez egyet, illetve vonunk le ef-ből egyet, amíg ef nullára csökken. Tehát egyszerűbb Turing-gépekből bonyolultabbak építhetők.

Szigorú matematikai értelemben a Turing-gépek függvények, konkrétan az alábbi típusú függvények:

$$U \times S \to C \times (U \cup \{\text{STOP}\})$$

Egy Turing-gép olyan függvény, ami minden $\langle u, s \rangle$ párhoz (az utasítássorszámok és felismert szimbólumok minden lehetséges kombinációjához) a cselekvések és parancsok egyik kombinációját rendeli.

Nem ez a Turing-gépek egyetlen lehetséges matematikai definíciója.[5] Vannak, akik STOP helyett két másik végállapotot feltételeznek, az ELFOGAD (*accept*) és az ELUTASÍT (*reject*) állapotokat.[6] Mindkettő hatására leáll a gép, de egyben jelzi is, hogy elégedett volt-e az inputtal. Ha az ELFOGAD és az ELUTASÍT végállapotokkal dolgozunk, akkor a Turing-gépeket „igen/nem" válaszokat adó mechanizmusként kezelhetjük. Innen ered az a kifejezés, hogy egy számítási probléma „eldönthető-e": akkor eldönthető, hogy rendelkezik-e egy tetszőleges n szám az F tulajdonsággal, ha van olyan Turing-gép, ami az ELFOGAD állapotba kerül, ha az inputként kapott n szám rendelkezik az F tulajdonsággal, és az ELUTASÍT állapotba kerül, ha nem. Eldönthető például, hogy egy természetes szám prímszám-e és híresen eldönthetetlen, hogy az elemi aritmetika egy adott állítása igaz-e.[7] Egyes tankönyvek az elfogadás és elutasítás kapcsán beszélnek a Turing-gépek *nyelvéről* (*language*) is. A $\mathbf{T}$ Turing-gép nyelve azoknak az inputoknak (azoknak az S elemeiből álló véges szimbólum-sorozatoknak) halmaza, melyeknél $\mathbf{T}$ az ELFOGAD állapotba kerül.

A Turing-gépek szalagja végtelen és a gépet definiáló táblázat tetszőlegesen hosszú lehet (bár végesnek kell lennie). Előbbi annak a feltevésnek felel meg, hogy a gép memóriája végtelen kapacitású, utóbbi annak, hogy elvben bármilyen komplex (véges) programot le tudunk futtatni. Hallgatólagosan azt is feltételezzük, hogy a gép sosem romlik el. A Turing-gépek tehát nem valós gépek, lehetetlen létrehozni őket. Emiatt néha felmerül a vád, hogy a Turing-gépeknek nincs közük a valódi számítógépekhez. Erről a következő alfejezetben lesz szó. Most csak arra szeretném felhívni a figyelmet, hogy ha az inputok hosszának van felső korlátja, akkor a gép szalagja is lehet véges. Mivel a Turing-gépek (a fenti definíció szerint) determinisztikusak, minden lehetséges i input esetén van egy olyan $\langle B(i), J(i) \rangle$ számpár, hogy a gép az i inputtal elindulva maximum $B(i)$ cellát halad balra és maximum $J(i)$ cellát jobbra. Ha a lehetséges inputok száma véges, akkor vehetjük ezeknek a kitéréseknek a maximumát, és elég lesz egy ennek megfelelő hosszúságú szalag ahhoz, hogy a gépünk ugyanazt csinálja, mint végtelen hosszú szalaggal tenné.[8] A Turing-gépekkel kapcsolatos irreális matematikai idealizációk tehát reálisabbá tehetőek. A Turing-gépek idealizált tulajdonságai arra szolgálnak, hogy lássuk, mi az, ami a számításokkal kapcsolatban a technológia fejlődésének *tetszőleges* szintje mellett igaz marad.

[5] A fenti definíció nagyjából Monk (1976) verziójának felel meg.

[6] Pl. Sipser (2013). A számítástudományban ez a megközelítés tűnik standardnak.

[7] Pontosabban: ha kódszámokat rendelünk az elemi aritmetika formalizált állításaihoz, akkor eldönthetetlen lesz, hogy az n kódszámú állítás igaz-e. Bővebben ld. a 9. fejezetben.

[8] A véges szalagú gépek persze a standard Turing-gépektől eltérően viselkedhetnek, ha végtelen ciklusba kerülnek. Ilyen esetben a véges szalagú gépből kieshet a szalag, a végtelen szalagú gép pedig az idők végezetéig vándorolhat egyik celláról a másikra. Mivel mindkét eset a számítás kudarcát jelenti, ennek nincs elvi jelentősége.

Ha belátjuk, hogy az X kérdésre Turing-géppel lehetetlen kiszámítani a választ, akkor beláttuk, hogy fizikailag realizálható számítógépekkel sem lehetséges X-re kiszámítani a választ.

2.3. Implementált Turing-gépek

Egy program mint absztrakt számítási modell nem számít ki semmit. Ahhoz, hogy egy program ténylegesen outputot állítson elő, implementálni kell: egy fizikai tárgynak futtatnia kell a programot (akármit is jelent ez). Az implementációt vizsgáló filozófusok nagyjából két táborra bonthatók: (i) azokra, akik szerint bármit is jelent a fizikai implementáció, a Turing-gép mint matematikai modell pontosan leírja, hogy egy számítógép *elvben* mit képes kiszámolni, és (ii) azokra, akik szerint a Turing-gép mint absztrakció irreleváns a számítás mint fizikai jelenség elemzésekor. Ez utóbbi tábor benyomásom szerint kisebbségben van, de a véleményük ettől függetlenül fontos és érdekes. Mivel a könyv további fejezeteiben az MI-t Turing-gépnek fogom tekinteni, szót kell ejteni a szkeptikus oldal aggályairól.

Héder Mihály (2020) kétféleképpen érvel amellett, hogy az MI kapcsán hiba Turing-gépekről beszélni. Egyik érve az, hogy a számítógépes programok működésében fontos szerepet kapnak bizonyos nyers fizikai folyamatok. A gép komponenseinek hőmérséklete például hatással van a programok által használt véletlenszámok generálására. A mai MI kétségkívül használ véletlenszámokat, sőt, egyes LLM-ekben mi magunk állítgathatjuk a „hőmérséklet" – *temperature* – nevű változót, ami azt szabályozza, hogy a következő token kiválasztásakor hogyan sorsol a rendszer (mindig a legvalószínűbb opciók valamelyikét választja, vagy ennél kockázatvállalóbb; utóbbi esetben érdekesebb, de potenciálisan zavaros vagy téves outputok keletkeznek). Tehát ez a kérdés nagyon is releváns az MI kapcsán.

Az, hogy a véletlen kezelésében a gép programjától független tények is szerepet játszanak, véleményem szerint nem érv a Turing-gépek mint számítási modellek relevanciája ellen, legalábbis ha arra vagyunk kíváncsiak, hogy *elvben* mire képes egy számítógép. Definiálhatunk nem-determinisztikus Turing-gépeket is, melyek ekvivalensek a determinisztikus Turing-gépekkel abban az értelemben, hogy bármit, amit ki tudnak számolni, determinisztikus Turing-géppel is ki lehet számolni (Sipser 2013: 178–180). A nem-determinisztikus Turing-gépek – a Héder által jelzett fizikai összefüggések miatt – nem képesek hűen modellezni egy számítógép tényleges működését, de képesek modellezni a véletlenszámokat használó számítógépek elvi korlátait. Ha valamit lehetetlen nem-determinisztikus Turing-géppel kiszámolni, akkor fizikailag realizált számítógéppel sem lehet.

Héder másik érve az, hogy a Turing-gépek absztrakt matematikai objektumok, a számítógépek viszont nem absztrakt matematikai objektumok, hanem

valós, fizikai eszközök. Emiatt „a számítógépet formális rendszernek tekinteni kategóriahiba” (Héder 2020: 130).

> [A] Turing-gép egy matematikai absztrakció, amelynek végtelen méretű tárhelye van. Márpedig ilyen gép nincs a természeti, megérinthető világban, így Turing-gép sincs. És nem elsősorban azért, mert lehetetlen a végtelen tárhely, hanem azért, mert a Turing gép absztrakt természetű, a tényleges gépek pedig nem. Tehát Turing-gép leképezések, közelítések vannak csupán, amelyeket hasznos Turing-géppel modellezni, de a modell egyes tulajdonságai (pl. gödeli nemteljesség) nem igazak, sőt még csak nem is értelmezhetők a tényleges gépekre. (Héder 2020: 96)

A kategóriahiba vádja azonban megkérdőjelezhető. A Turing-gépek absztrakt természetéből nem következik, hogy matematikai tulajdonságaik csak a platóni mennyországban lennének relevánsak. A számok is absztrakt tárgyak, mégis igaz, hogy ha van három nyuszink és két répánk, akkor nem juthat minden nyuszinak egy teljes répa – absztrakt törvények a konkrét valóság korlátai is lehetnek. Az implementált programok számítási képességeire vonatkoznak a program mint absztrakt szabályrendszer matematikai korlátai. Ha az F fizikai rendszer számítási képességei ekvivalensek egy A absztrakt gép számítási képességeivel, akkor A elvi korlátai egyben F korlátai is (bár F-nek lehetnek olyan korlátai, melyekkel A nem rendelkezik). A gödeli nemteljesség vagy megállási probléma ezért érvényes maradhat a konkrét, fizikai számítógépek kontextusában is, legalábbis ha igaz a Church-Turing tézis és algoritmusokkal csak olyasmit lehet kiszámolni, amit elvben Turing-géppel is ki lehet számolni.[9]

Héder fenti kritikájára két további válasz is adható. Az egyik hosszabb kifejtést igényel és a 13. fejezetben lesz róla szó. Az érv lényege az, hogy ha az MI értelmes, akkor egy program futtatása *miatt* értelmes, tehát ha a Turing-gép a programok általános modelljének tekinthető, akkor a Turing-gépekre vonatkozó korlátok az MI értelmi képességeire is vonatkoznak. A másik érv a Turing-gépek relevanciája mellett az, hogy a számítógépek központi processzora (CPU) Turing-gépnek feleltethető meg; ezt a következő alfejezet részletezi.

Shagrir (2022: 46–47) szerint a Turing-gépek azért nem modellezik adekvátan a fizikai számításokat, mert a kiszámít*hat*óság elméletéhez tartoznak. Számításokat nagyon sokféle módon lehet végezni, és ezen a téren a Turing-gép csak az *emberi* kalkulációs tevékenység lényegét ragadja meg. Nem tükrözi a

[9] A Church-Turing tézis azt mondja ki, hogy minden „effektíven kiszámítható” függvény kiszámítható Turing-géppel (Monk 1976: 46, Rogers 1987:20, Sipser 2013: 183). Az „effektív kiszámíthatóság” nem egzakt matematikai fogalom – a mechanikus, algoritmuson alapuló számítások naiv fogalmára utal –, de a legtöbben egyetértenek abban, hogy a Turing-gépek adekvát módon modellezik. A különböző komplikációkról áttekintés ad Piccinini (2015) és Shagrir (2022). A Church-Turing tézist elvben cáfolhatják olyan fizikai rendszerek, melyek végtelen sok számítást végeznek („hiperkomputáció”, Piccinini 2015: 265–270), de ez jelen kontextusban irreleváns, mert a mai MI nem hiperszámítógépen fut.

számítógépek valódi képességeit, például azt, hogy párhuzamosan több műveletet végeznek.

Shagrir nem említi, de felhozhatjuk példaként az analóg számítógépeket is, melyeket eleve lehetetlen Turing-gépnek tekinteni. Különösen szórakoztató példányuk A. W. Phillips angol közgazdász MONIAC (Monetary National Income Analogue Computer) nevű gépe 1949-ből. A MONIAC csövekben áramló folyadékokkal modellezte a gazdaságpolitikai intézkedések hatását az árakra, megtakarításokra és egyéb makroökonómiai mutatókra.[10] Mivel nem fizikailag tárolt számjegyeket manipulált diszkrét lépésekben, a MONIAC még elvben sem tekinthető Turing-gépnek.

Ezek a felvetések azonban szintén nem igazolják, hogy az MI, számítási képességeit tekintve, ne egy Turing-géppel lenne ekvivalens. A Turing-gép mint absztrakt számítási modell nyugodtan kibővíthető párhuzamos műveletekkel, extra szalagokkal és hasonlókkal. Elvi képességein ez nem változtat. Az analóg számítógépek valóban érdekes kihívást jelentenek, de jelen kontextusban ez sem lényeges, mert az MI-ket nem analóg számítógépeken implementálják.[11]

2.4. A CPU mint Turing-gép

Az MI formális jellemzésének lezárásaként amellett fogok érvelni, hogy a gépi kód szintjén az MI működése megfeleltethető egy Turing-gép működésének.

A mai számítógépeket egy központi processzor (CPU) irányítja, végeredményben *ezen* fut minden program. Fizikai értelemben az összes számítást a CPU és a vele kapcsolatban levő segédprocesszorok (GPU-k, TPU-k) végzik.

A CPU-kba gyárilag belehuzaloznak egy utasításkészletet, melynek programnyelve a *gépi kód* (ld. pl. Patterson és Hennessy 2014: 259–272). Minden program gépi kódra fordul le, és ebben az állapotban kerül be – az operációs rendszer közvetítésén keresztül – a CPU-ba. Egy gépi kódú program nullák és egyesek byte-okra osztott sorozata, ami parancsok sorozatát kódolja. Íme egy példa az x86 típusú processzorok egyik gépi kódú parancsára:

G `00000001 11000001`

E `ADD EAX, ECX`

G a gépi kódú parancs abban a formában, ahogy a processzorba kerül (elektromos töltések formájában). **E** ennek az utasításnak számunkra olvashatóbb változata. Az „ADD” az összeadást fejezi ki, az „EAX” és az „ECX” pedig a processzor két adattároló regiszterét jelöli. **G** arra utasítja a CPU-t, hogy adja

[10] Köszönet Kellerwessel Klausnak a példáért.

[11] Az analóg számítógépek egyesek szerint algebrai differenciálegyenletek megoldásait számítják ki (Pour-El 1974. Rubel és Singer 1985, Lipshitz és Rubel 1987), emiatt pedig funkcionálisan ekvivalensek Turing-gépekkel. A kérdést részletesen tárgyalja Piccinini (2015: 12. fejezet).

hozzá az EAX regiszter tartalmát (az ott tárolt nullák és egyesek által reprezentált kettes számrendszerbeli számot) az ECX regiszter tartalmához (ami szintén egy kettes számrendszerbeli szám), majd az eredményt tárolja el az ECX regiszterben. **G**-ben az első 8 bit az ADD parancsot kódolja, a második 8 bitből az első kettő azt jelzi, hogy regiszterek közti műveletről van szó, a „000" az első (EAX), a „001" pedig a második (ECX) regisztert jelenti.

A CPU szabályos időközönként beolvassa a következő parancsot, végrehajtja, majd új ciklust indít. A CPU nemcsak alapvető aritmetikai számításokat (pl. összeadást) tud végezni, hanem kommunikál a merevlemezzel, memóriával, wifivel, billentyűzettel, képernyővel stb. Ez adatbuszokon keresztül történik, melyeken a CPU jeleket küld az adott komponensnek vagy jeleket fogad tőle.

Az, hogy egy számítógép egy programot futtat, fizikai szinten azt jelenti, hogy a CPU gépi kódú utasításokat hajt végre. Persze léteznek olyan mechanizmusok, melyeknek nincs CPU-ja, mégis számításokat végeznek – Pascal és Leibniz is tervezett fogaskerekekből álló számológépet –, de a demarkációs probléma szempontjából a CPU-val rendelkező számítógépek relevánsak.

Egy CPU felfogható sokszalagos Turing-gépnek, melynek szalagjai véges hosszúságúak. Ilyen matematikai struktúra persze szigorú értelemben véve nincs, mert a Turing-gépek szalagja definíció szerint végtelen (Boolos, Burgess és Jeffrey 2007: 25, Monk 1976: 14, Sipser 2013: 165 stb.). De nyilván definiálhatunk a Turing-gépek mintájára egy olyan absztrakt gépet, amit szintén egy véges táblázat definiál, ugyanolyan parancsokat tud értelmezni, mint egy Turing-gép, viszont a szalagja véges, és ha az írófej véletlenül túlmegy rajta, akkor a gép output nélkül leáll (hasonlóan ahhoz, hogy egy Turing-gép sem produkál outputot, ha végtelen ciklusba kerül). Az ilyen absztrakt gépeket „véges Turing-gépnek" (VTG) fogom hívni. Nyilván definiálhatóak több szalagos VTG-k is, melyek táblázatában négynél több oszlop szerepel (annyi *Olvas* oszlopuk van, ahány szalag).

Egy véges Turing-gép számításelméleti értelemben ekvivalens egy véges állapotú automatával (*Finite State Automaton*, FSA).[12]

[12] Ha I a gép számára értelmezhető inputok halmaza, S pedig a gép belső állapotainak véges halmaza (ahol s_1 a kezdőállapot, és van legalább egy STOP állapot), akkor egy véges állapotú automata egy $I \times S \to S$ típusú függvénynek feleltethető meg (Sipser 2013: 35), vagyis minden $\langle i, s_1 \rangle$ pároshoz egy s_2 állapotot rendel, mint egy kávéautomata, ami a vásárló által épp megnyomott gomb (i) és a gép aktuális állapota („üzemkész", s_1) alapján elkezd kávét tölteni, vagy arra vár, hogy fizessen a felhasználó (s_2).

Minden FSA-val kiszámolható függvény kiszámolható VTG-vel, hiszen bármely $I \times S \to S$ függvényre konstruálhatunk egy VTG-t, ami a szalagján fogadja az i inputokat és az FSA által definiált állapotváltozásokon megy keresztül. Másrészt minden VTG megfeleltethető egy FSA-nak. Ha K a VTG szalagjára ráírható lehetséges karaktersorozatok halmaza (ez véges, mivel a VTG szalagja véges), akkor van olyan $v : K \to K$ függvény, ami a VTG összes lehetséges inputjához hozzárendeli a VTG vonatkozó outputját. Ez alapján pedig konstruálható olyan FSA, ami egy tetszőleges k inputra kiírja a vonatkozó $v(k)$ outputot, majd leáll.

A CPU regiszterei egy sokszalagos, véges Turing-gép egy-egy szalagjának feleltethetőek meg (egy 16 bites regiszter például egy 16 cellás szalagnak), a VTG utasításai (a gépet definiáló táblázat sorai) pedig a gépi kódú parancsoknak. Egy átlagos CPU persze bonyolultabb egy VTG-nél, hiszen nem csak az elindításkor kap inputot, hanem (a felhasználó klikkelései, a billentyűzetből érkező jelek stb. miatt) folyamatosan újabb és újabb inputok kerülnek a „szalagjára". De ha azt modellezzük, hogy egy számítógép valamilyen fix inputra előállít egy outputot (pl. az MI ⌜válaszol⌝ egy kérdésre), akkor ez a komplikáció figyelmen kívül hagyható. További különbség, hogy a CPU nem egyszerűen összeadásokat és szorzásokat végez a kapott inputon, hanem különféle programokat (gépi kódra lefordított alkalmazásokat) futtat. Ezt a jelenséget is lehet VTG-k segítségével értelmezni, hiszen létezik univerzális Turing-gép, mely más Turing-gépeket futtat, a szalagjára írt inputot programként értelmezi. Az MI-k implementálása tehát végeredményben – a CPU szintjén –, egy sokszalagos, a tankönyvi változatnál bővebb parancskészletű, véges Turing-gép implementálásának feleltethető meg.

I. rész

Elmeoldali demarkáció

3.

Testi demarkáció

Az MI-nek jelenleg nincs teste. Ha a „test” szót tág értelemben használjuk – általában az anyagi hordozó értelmében –, akkor persze az MI-nek is van teste, hiszen fizikai tárgyakban, számítógépekben van jelen. De olyan teste nincs, ami képes önmozgásra és a gondolkodással szorosan együttműködő érzékszerveket használ. Bár az MI-kutatások korábbi hullámaiban próbáltak ilyen gépeket építeni, és az MI ágensek kapcsán kezd újra divatba jönni ez a kutatási program, a jelenlegi MI tipikusan testetlen.

Az MI testetlenségéből könnyen adódik egyfajta demarkációs érv. A megtestesültség nyilván sajátos élményekkel, igényekkel és preferenciákkal jár, melyekkel az MI, ha netán van is saját belső nézőpontja, nem rendelkezik. Emiatt az MI, ha van belső nézőpontja, másfajta szubjektum, mint mi. De ennek nem feltétlenül van köze a *kognitív* demarkációhoz, ahhoz, hogy az MI érti-e az emberi nyelvet, gondolkodik-e, vannak-e intellektuális tapasztalatai. Ezeket aligha cáfolja közvetlenül a testtudat hiánya, és nem is egyértelmű, hogyan cáfolhatná. Ezért is érdekes az MI történetileg talán legfontosabb kritikusa, Hubert Dreyfus, aki az amerikai Heidegger-recepció egyik vezéralakja is volt, és a testetlenség alapján érvelt a kognitív demarkáció mellett.

Dreyfus egyben az MI leg*hírhedtebb* kritikusa is, mert filozófiai alapon „megjósolta”, hogy az MI sosem fog egyik nyelvről a másikra fordítani vagy nagymesteri szinten sakkozni (Dreyfus 1972: 208). Az ilyen tévedések miatt könnyű Dreyfust karosszékből okoskodó filozófusnak látni, akinek *a priori* tévelygéseit leleplezték a szorgos mérnökök. Bár egyes predikciói tévesek voltak, Dreyfus MI-kritikája ettől függetlenül érdekes, és összefüggésbe hozható egy sajátos elmefilozófiai kutatási programmal, az enaktivizmussal. Amellett fogok érvelni, hogy valójában *a priori* okaink is vannak rá, hogy kétségbe vonjuk a dreyfusi és általában az enaktivista demarkációs stratégiát.

3.1. Dreyfus, Husserl és Heidegger

Dreyfus (1972: 204) az intelligens viselkedés négy szintjét különíti el az alapján, hogy a jelentés és kontextus mekkora szerepet játszik bennük:

I **Asszociációs intelligencia.** Ide tartoznak a memóriajátékok, a labirintus-jellegű (próbálgatásra épülő) feladatok és izolált szavak fordítása (keresés egy szótárban). A jelentés és kontextus minimális szerepet játszik.

II **Egyszerű formális intelligencia.** Ide tartoznak az egyszerű algoritmusokat igénylő játékok („amőba"), valamint matematikai tételek bizonyítása tisztán formális szabályok mechanikus alkalmazásával. A szavak vagy szimbólumok jelentése expliciten definiálható és kontextusfüggetlen.

III **Komplex formális intelligencia.** Ide tartoznak az intuíciót igénylő, de formalizálható játékok, mint a sakk és a go, a bonyolultabb tervezési problémák, valamint az olyan matematikai bizonyítások, ahol nem áll rendelkezésre mechanikus igazolási mód.

IV **Nemformális intelligencia.** Ide tartozik a szövegek fordítása, találós kérdések megoldása, mintafelismerés példák alapján (explicit definíciók és formális szabályok nélkül), és általában minden nyitott jellegű, ötletet vagy sejtést igénylő feladat. A nyelvi jelentés és a kontextus alapvetően fontos.

Dreyfus szerint a III. és IV. szint jelenti az MI számára a fő kihívást, a IV. szint gépesítése pedig konkrétan reménytelen. Ennek igazolására legalább három stratégiát használ, melyek fogalmi téren eltérnek, de ugyanabból az intuícióból erednek. Az intuíciót hatásosan foglalja össze a Dreyfus filozófiájáról szóló dokumentumfilmben John Haugeland (akiről a következő fejezetben lesz szó):

> Ezek a rendszerek [a korabeli MI-k] állandóan csődöt mondtak közismert, mindennapi metaforák vagy összefüggések miatt. Itt van például ez a gyom. Ha felemelem a földről és elengedem, le fog esni. És anélkül fog leesni, hogy zajt csapna, mert puha. Ilyen dolgokról van szó, amiket mindenki tud. *Hány* ilyet dolgot tudunk? Százat? Ezret? Egymilliót? (*Being in the World*, 2010, rendezte Tao Ruspoli, 15:50–16:15)

A dreyfusi intuíció lényege, hogy ez a fajta háttértudás szükségszerű a megértéshez (a III-as és IV-es szint esetében), de nemcsak formalizálatlan, hanem formalizál*hat*atlan.

Dreyfus első stratégiája ennek igazolására empirikus jellegű. Részletesen elemzi, hogyan működnek a korabeli – 1960-as, 1970-es évekbeli – feladatmegoldó és képfelismerő algoritmusok, melyeket előszeretettel állítottak be alkotóik

az algoritmikus értelem első jeleinek. Dreyfus (1972: 49) szerint ezek a kutatási programok ugyanazt a logikát követik: „óvatos optimizmussal bemutatnak egy specifikus problémára adott *ad hoc* megoldást, majd kijelentik, hogy ez az első lépés az általánosabb megoldások felé", utóbbiak viszont sosem érkeznek el.

Dreyfus második érve a III-as és IV-es szint formalizálhatatlansága mellett fenomenológiai jellegű és arra az alaptézisre épül, hogy „a mintafelismerés előfeltétele egyfajta meghatározatlan, globális elvárás" (Dreyfus 1972: 149). Azt a szituációt hozza fel példaként, mikor valaki egy pohár vízért nyúl, de véletlenül egy pohár tejet vesz a kezébe és abba kortyol bele. Ilyen esetben nem azt gondoljuk, hogy „Nocsak, víz helyett tejet iszom", hanem zavarodottságot érzünk: tudjuk, hogy nem vizet kortyoltunk, de még nem fogtuk fel, mi van a szánkban (Dreyfus 1972: 154, 1974: 25). Ez a meghatározatlan *valami* az érzékelésünk fogalmaktól és elvárásoktól megtisztított alapanyaga, amihez konkrét elvárásokkal kell közelítenünk, hogy emberi értelemben vett tapasztalataink legyenek. Az elvárásaink azonban egyrészt globális jellegűek, másrészt a testünkben gyökereznek. Globális jellegűek, mert nem tudjuk tételesen felsorolni, mire számítunk és mit tartanánk meglepőnek abban a helyzetben, amikor a vizespohárért nyúlunk, hogy belekortyoljunk. (Hány konkrét elvárásunk ilyenkor? Száz? Ezer? Egymillió?) Másrészt ezek az elvárások a testünkben gyökereznek, a fizikai tevékenység és az ebből adódó tapasztalat finom összjátékából erednek. Másfajta interakciót folytat környezetével az, aki vizespohárért nyúl, mint az, aki selymet készül végigsimítani (Dreyfus 1974: 26). Mivel a III-as és IV-es típusú gondolkodás a tudatos motorikus tevékenység és a külvilág összjátéka, melyet a mi oldalunkon egy általános elvárásrendszer vezérel, és mivel az ilyen interakciók nem alapulnak formálisan specifikálható szabályokon, a III-as és IV-es típusú intelligenciát lehetetlen szimulálni olyan gépek segítségével, melyek véges számú explicit axióma segítségével reagálnak diszkrét inputokra:

> Mi [emberek] a tárgy egészére reagálunk először, és csak utána, mikor már jobban megismertük, töltjük ki részletekkel [...]. Egy gép, melynek nincs hasonló belső horizontja, ugyanezt az információt fordított sorrendben kell, hogy feldolgozza: a részletektől halad az egészig. (Dreyfus 1972: 153).

A fenomenológiai érv lényege tehát az, hogy az értelem sajátosan emberi formái olyan (testi eredetű) *Gestalt*okon alapulnak, melyeket nem lehet formális axiómákra lebontani és az ezekre épülő kalkulációkkal élethűen reprodukálni.

Dreyfus harmadik érve Heideggerre támaszkodik. Dreyfus a *Lét és idő* sajátos pragmatikus olvasatát dolgozta ki. Értelmezése szerint Heidegger a karteziánus hagyománnyal élesen szakítva azt hangsúlyozza, hogy a világhoz fűződő kognitív kapcsolatunk alapvető jellegét tekintve eleme is van, ami gépesíthetetlen (Dreyfus 1991: 116–118). A világhoz fűződő viszonyunk egyrészt holisztikus jellegű: az *asztal* fogalmát például nem lehet definiálni anélkül, hogy mozgósítanánk hétköznapi világképünk egyéb elemeit és ezek történeti kontextusát.

Másrészt, a gyakorlati tevékenységünk olyan képességek (*skills*) használatán alapul, melyek a szituációk egészére érzékenyek. Az például, hogy egy kalapácsot hogyan kell használni, a kalapács típusának és állapotának, a szándékolt célnak és a kontextus egyéb paramétereinek függvénye. Ezeket a faktorokat egy tapasztalt eszközhasználó nem mechanikus kalkuláció alapján méri fel és összegzi, hanem a szituációval kapcsolatos globális érzékenység révén, ami egyszerre a gondolkodás egy formája és közvetlen kapcsolat a világgal. Hatásosan illusztrálja ezt a jelenséget a korábban említett dokumentumfilm, a *Being in the World*, melyben Dreyfus és kollégáinak filozófiatörténeti eszmefuttatásai váltakoznak egy ács, egy szakács, egy zsonglőr, egy flamenco-énekes és három jazz-zenész megszólalásaival. A művészek és mesteremberek arról beszélnek, milyen érzés használni szaktudásukat, és mindig az derül ki beszámolóikból, hogy bizonyos értelemben összefonódnak környezetükkel, amikor alkotnak. Az ács az aktuális faanyag sajátosságaira reagál, a jazz-zenészek az adott szituáció apró rezdüléseit építik be a zenéjükbe, a szakács sosem az írott receptet követi, hanem váratlan – és briliáns – ötletei támadnak az épp kéznél levő hozzávalók alapján, és így tovább.

Dreyfus (2014a) a „tapasztalt igyekezet" (*skillful coping*) fogalmával ragadja meg ezt a fajta intelligens, de nem elméleti jellegű viselkedést, amit egyben Heidegger antikarteziánus kritikájának alapjaként azonosít. A tapasztalt igyekezet nem a szubjektumról leválasztott külvilág mentális reprezentációján, nem egy elméleti világmodellen alapul, hanem a világhoz fűződő sokkal közvetlenebb viszony. Ugyanakkor nem azonos a gyakorlati tudással sem, nem pusztán gyakorlati feladatok megoldását jelenti, amire akár egy robot is képes lehet. A tapasztalt igyekvés olyan értelmes cselekvés, ami nem szabályok tudatos alkalmazásából áll, hanem a szituációval kapcsolatos mély megérzésekre épül (Wrathall 2014: 13). Paradigmatikus esete nem a szögbeverés vagy biciklizés, hanem az olyan, tapasztalatot és szaktudást igénylő tevékenységek, mint az ácsmesterség vagy zeneművészet. A tapasztalt igyekvésnek emiatt normatív, esztétikai és tisztán deskriptív tartalma is van, nem tisztán *know-how*. Egy ácsmester tudása nem korlátozódik arra, hogy tud bútort vagy házat készíteni, hanem azt is tudja, milyen alapanyagból *lehet* jó bútort vagy házat készíteni, az alapanyagot hogyan *kell* megmunkálni, tudja, hogyan kell *szép* bútort vagy házat készíteni, meg tudja állapítani, hogy egy faház kibírja-e az időjárás viszontagságait, a bútor nem fog-e megvetemedni, és így tovább. A tapasztalt igyekvés és a propozicionális vagy elméleti tudás között nem az a különbség, hogy a tapasztalt igyekvés puszta gyakorlati kompetencia lenne, hanem egyrészt az, hogy nem választható el az episztemikus szubjektum élettevékenységétől, másrészt nem szabályok vagy elméletek alkalmazásából áll, hanem a megértésnek a tevékenységgel összefonódó formája. Tömören írja le ezt a viszonyulást az alábbi Heidegger-idézet:

> Ami eleve „adott” [...] az az „írásra”, „kimenésre”, a „világításra”, „ülésre”. Vagyis ami *a priori* foglalkoztat minket, az az írás, kimenés, ülés és hasonlók. Ezeket a „mirevalóságokat” tanuljuk meg, ezeket ismerjük, ha magabiztosan mozgunk a világban. (Heidegger 2010: 121, idézi Dreyfus 2014b: 256)

Dreyfus szerint a tevékeny igyekvés természetéből azért következik a demarkációs tézis, mert ha az emberi értelem alapformája a tapasztalt igyekvés, utóbbinak pedig pont az a lényege, hogy nem explicit szabályokra és nem a külvilágról alkotott reprezentációkra épül, akkor az emberi értelem alapjait tekintve nem gépesíthető, hiszen az MI – vagy legalábbis a Dreyfus idejében divatos szimbolikus MI – definíció szerint explicit szabályokon és különféle reprezentációkon (formális modelleken) alapul. Mielőtt rátérnék arra, hogy ez a demarkációs stratégia mennyire sikeres, teszek egy rövid kitérőt az enaktivizmus, Dreyfus elméleteinek tágabb kontextusa felé.

3.2. Enaktivizmus

Dreyfus antikarteziánus elmélete egy szélesebb elmefilozófiai és kognitív tudományos irányzathoz kapcsolódik, melynek központi fogalma a megtestesült értelem (*embodied cognition*). A megtestesült értelem hívei szerint az intelligencia vizsgálatában központi szerepet kell kapnia a testnek vagy általában az organizmus egészének. A „négy E” kutatási programjának is szokták nevezni, főbb áramlatai alapján: az értelem megtestesült (*embodied*), beágyazott (*embedded*), enaktív (*enactive*) és/vagy kiterjedt (*extended*). A különböző áramlatok képviselői többnyire egyetértenek abban, hogy a test korlátozza vagy diktálja azt, hogy milyen fogalmakkal rendelkezik egy lény; a test nem pusztán oksági értelemben járul hozzá a gondolkodáshoz és felfogáshoz, hanem konstitutív része a kognitív folyamatoknak. Az értelem emiatt nem az agyban vagy az idegrendszerben van, hanem a test egészében, vagy még inkább a biológiai test és a környezet interakciójában. A megtestesült értelem szószólói ennek megfelelően szkeptikusak szoktak lenni az olyan fogalmakkal kapcsolatban, mint a mentális reprezentáció, szimbolikus gondolkodás, komputáció vagy következtetés (Shapiro & Spaulding 2025: 2.1. szakasz). A világról leválasztott belső mentális állapotok feltételezése teljes tévútnak tűnik számukra, és emiatt nemcsak az anyagtalan lélek karteziánus koncepcióját és a dualizmus egyéb változatait utasítják el, hanem azt a fajta fizikalizmust is, ami az agyban vagy általában az idegrendszerben zajló számításokra, információfeldolgozásra próbálja redukálni az elmét (Noë 2009: 7).

Dreyfus MI-kritikája a megtestesült értelem irányzatai közül az enaktivizmushoz kapcsolódik legszorosabban. Az enaktivisták szerint az értelem kulcsa a szenzomotoros tevékenység, az érzékelés és mozgás finom összjátéka. Az értelem (beleértve a gondolkodást) a szenzomotoros tevékenységgel azonos vagy

abból ered. Ezt a kutatási programot részben olyan kognitív tudományos vizsgálatok támogatják, melyek szerint az agyban rögzülő szenzomotoros mintázatoknak alapvető szerepe van a fogalomhasználatban.[1] Az enaktivizmust egyes MI-fejlesztések is inspirálták, például Rodney Brooks „Kreatúrái", melyek versenytársaikhoz képest feltűnően ügyesen tudtak közlekedni és gyakorlati feladatokat elvégezni (pl. italdobozokat összegyűjteni).

> Ahelyett, hogy az érzékelés, modellezés, tervezés és cselekvés fázisaiból álló ciklusokat futtatnának, a Kreatúrák különféle szenzorokkal vannak felszerelve, melyek közvetlenül kapcsolódnak a viselkedést generáló mechanizmusokhoz. Brooks „Allen" nevű robotjában például a szenzorok három különböző viselkedésgenerátorhoz csatlakoztak: Elkerülő, Bóklászó és Fürkésző. Mikor a szenzorok érzékelték, hogy Allen egy tárgy felé közeledik, az Elkerülő mechanizmus azt az utasítást küldte, hogy álljon meg, váltson irányt, majd menjen tovább. A Bóklászó egyszerűen egy véletlenszerű irányba forgatta Allent, míg a Fürkésző generátor egy kiválasztott célpont felé kormányozta. Ez a három aktivitásréteg (ahogy Brooks nevezte őket) folyamatosan versengett egymással. Ha például Allennek bóklászás közben akadály akadt az útjába, akkor az Elkerülő közbelépett és megakadályozta az ütközést. A Fürkészű képes volt blokkolni Bóklászót, hogy Allent pályán tartsa egy cél felé. A három réteg versengő interakciója meglepően flexibilis és látszólag célirányos viselkedést eredményezett. (Shapiro & Spaulding 2025: 4.1. szakasz)

Teremtőjük szerint a Kreatúrák a reprezentációk nélküli intelligencia példái (Brooks 1991). Érdekes módon ezt a projektet részben Dreyfus MI-kritikája inspirálta, és „heideggeri" – világmodellek nélküli – MI létrehozásával mások is kísérleteztek (Dreyfus 2014b).

Mint a fenti példa mutatja, az enaktivizmus nem feltétlenül MI-szkeptikus álláspont. Ugyanakkor az a tény, hogy az enaktivista robotok fejlődése viszonylag primitív szinten rekedt meg, támogathatja azt a dreyfusi hipotézist, hogy az emberekre jellemző tapasztalt igyekvés nem gépesíthető. Dreyfus bizonyos mértékig el is határolódott a megtestesült értelem programjától, mert szerinte sok szerző a karteziánus modell rabja maradt és azt hiszi, hogy különbséget tehetünk az elme és a külvilág között, az intelligencia lényege pedig az, hogy egy ágens (mint különálló individuum) a külvilágban felmerülő kihívásokat old meg. „Alapvető létmódunk alapján – mint tapasztalt igyekvők – nem elmék vagyunk, hanem egyek a világgal" (Dreyfus 2014b: 259).

Az egyszerűség kedvéért most zárójelbe teszem ezeket az ideológiai vitákat és Dreyfust is besorolom az az enaktivizmus kategóriája alá, mert szerintem

[1] Ld. pl. Barsalou (1999). Az empirikus bizonyítékokról áttekintést ad Fischer & Zwaan (2008) és Dove (2009: 413–414).

létezik olyan absztrakciós szint, ahonnan Dreyfus és az enaktivista szerzők a demarkáció szempontjából közös platformon vannak. Az enaktivizmus melletti és elleni érveket sem próbálom összefoglalni, hanem rátérek arra, hogy sikeres lehet-e az enaktivizmus-alapú demarkáció.

3.3. Kell-e test a gondolkodáshoz?

A fejezet lezárásaként amellett érvelek, hogy az enaktivizmus segítségével nem igazolható a kognitív demarkációs tézis, legalábbis ha utóbbit az 1. fejezetben rögzített módon az intellektuális tapasztalatra vonatkoztatjuk, vagyis arra vagyunk kíváncsiak, az MI lát-e logikai összefüggéseket, érti-e, amit mond, tudatosan elemez és következtet-e. Annak érdekében, hogy ezt a kérdést terminológiai szinten összekössem az enaktivizmussal, a gondolkodást mentális reprezentációkon végzett műveletekkel azonosítom (reprezentációk összekapcsolása, megkülönböztetése, összehasonlítása stb.), és a mentális reprezentációk jelenlétét a gondolkodás szükségszerű feltételének tekintem. Ez alapján elvben háromféle enaktivista álláspont létezhet:

(EE) **Eliminatív Enaktivizmus**: Mentális reprezentációk nem léteznek. Az emberi intelligencia a szenzomotoros tevékenységben merül ki (Hutto & Myin 2013).

(RE) **Reduktív Enaktivizmus**: Mentális reprezentációk léteznek, de a szenzomotoros tevékenységre redukálhatóak, ennek bizonyos típusai vagy aspektusai (Borghi & Binkofski 2014).

(NRE) **Nem-reduktív Enaktivizmus**: Mentális reprezentáció léteznek, de a szenzomotoros tevékenységből erednek (anélkül, hogy a szenzomotoros tevékenység kimerítené őket) (Dove 2011).

Amennyire látom, ezek közül egyik sem alkalmas demarkációra, emiatt pedig az enaktivizmus általában nem alkalmas demarkációra.

(EE) triviálisan alkalmatlan demarkációra, mert tagadja a gondolkodás létét. Ha (EE) igaz, akkor az MI valóban nem gondolkodik, de mi sem gondolkodunk, tehát a demarkációs kérdés eleve fel sem merül.

(NRE) csak akkor lenne alkalmas demarkációra, ha belátnánk, hogy a mentális reprezentációk *szükséges* feltétele a szenzomotoros tevékenység. Ha a mi mentális reprezentációink a szenzomotoros tevékenységből erednek, de ez csak a természet szeszélye miatt van így és mentális reprezentációkkal szenzomotoros tevékenységet nem végző lények is rendelkezhetnek, akkor (NRE) nem zárja ki, hogy az MI-nek is vannak mentális reprezentációi. Tehát (NRE) csak akkor alkalmas demarkációra, ha igaz a következő állítás ($M(x) = x$ mentális reprezentációkkal rendelkező fizikai lény, $S(x) = x$ szenzomotoros tevékenységet végez, $\Box$ = szükségszerűen):

(E*) $\Box(\forall x : M(x) \rightarrow S(x))$ (Szükségszerűen bármely x-re igaz, hogy ha x mentális reprezentációkkal rendelkező fizikai lény, akkor x szenzomotoros tevékenységet végez)

Ez a tézis tehát a demarkáció előfeltétele (NRE) esetében. Ugyanakkor (E*) nyilván egyenes következménye (RE)-nek, hiszen ha a mentális reprezentációk használata valamilyen szenzomotoros tevékenységgel azonos, akkor $M(x)$ esetén $S(x)$ is garantáltan fennáll. Az enaktivizmus tehát csak akkor alkalmas demarkációra, ha (E*) igaz. Ha ugyanis az enaktivizmus alkalmas demarkációra, akkor (RE) és (NRE) közül valamelyik igaz; előző esetben (E*) garantáltan igaz, utóbbi esetben pedig (E*) nélkül nem lehet sikeres a demarkáció. Az enaktivista demarkáció tehát legalábbis részben (E*) igazságán múlik.

(E*) azonban hamisnak tűnik, mert elgondolhatóak olyan fizikai lények, akik nem végeznek szenzomotoros tevékenységet, de rendelkeznek mentális reprezentációkkal. Ilyenek például a tartályban lebegő, testetlen agyak Hilary Putnam (1981: 5–6) híres gondolatkísérletében. Tegyük fel, hogy (a távoli jövőben, a technológia radikális fejlődése után) valakinek kioperálják az agyát és átteszik egy táptartályba, az idegvégződéseire pedig rácsatlakoztatnak egy szuperszámítógépet, ami a külvilág illúzióját kelti megfelelő idegi ingerek és visszacsatolások segítségével. Elgondolható, hogy egy ilyen tartályban lebegő agy (TLA) úgy tapasztalja, mintha szabadon tevékenykedne, beszélgetne, töprengene, spekulálna a külvilágban. A TLA gondolkodik – vannak mentális reprezentációi –, de nem végez szenzomotoros tevékenységet, hiszen nincs teste. Tehát (E*) hamis: van olyan lehetséges világ, ahol egy x-re igaz, hogy $M(x) \wedge \neg S(x)$. Mivel (E*) hamis, enaktivista alapon lehetetlen igazolni a demarkációs tézist.

A demarkációban reménykedő enaktivista erre azt felelheti, hogy TLA-k kétségkívül elgondolhatóak, de metafizikailag lehetetlenek. A TLA-król szóló sci-fi történetek nem tükrözik a mentális reprezentációk valódi természetét, emiatt (E*) cáfolatára sem alkalmasak. De ez a védekező manőver nem tűnik különösebben hatékonynak. Egyrészt, a TLA-k nem tűnnek metafizikailag lehetetlennek, még csak nomologikusan sem tűnnek lehetetlennek. Semmi okunk kizárni, hogy a technológia fejlődésével lehetővé válik agyak közvetlen számítógépes stimulálása és létrejönnek tudattal rendelkező TLA-k. Másrészt nem feltétlenül kell tudományos-fantasztikus történetekhez nyúlnunk, hogy ellenpéldákat találjunk (E*)-ra. Az álmodó vagy hallucináló személyek például tipikusan nem végeznek olyan szenzomotoros tevékenységet, melyekre élményeik hihetően redukálhatóak lennének. Egy teljesen mozdulatlan, külső ingerektől megfosztott személynek is lehetnek élénk álmai vagy hallucinációi. Minden ilyen x ellenpélda lesz (E*)-ra, mert kielégíti az $M(x) \wedge \neg S(x)$ formulát.

4.

Egzisztenciális demarkáció

Az egzisztenciális demarkáció John Haugeland amerikai filozófustól ered, aki Dreyfushoz hasonlóan az MI egyik fő kritikusa volt az 1980-as években. Írt egy olvasmányos monográfiát a modern MI-kutatások első negyven évéről (Haugeland 1985), és tőle származik a széles körben elterjedt „GOFAI" („Good Old Fashioned Artificial Intelligence", régimódi MI) becenév a szimbolikus MI-re.

Haugeland is szkeptikusan viszonyult a gépesített értelemhez, de dogmatikus kijelentések és konkrét negatív jóslatok helyett inkább szuggesztív példákkal és aforizmákkal dolgozott. Leghíresebb aranyköpése szerint „a gépeknek tökmindegy" („computers don't give a damn", Haugeland 1998: 47). MI-történeti monográfiájának utolsó, „Valódi emberek" című fejezetében azon morfondírozik, hogy egy egérfogó meg *akarja*-e ölni az egeret, majd a következő anekdotával érzékelteti a társadalmi lét és az emberi értelem összefonódását:

> Feltételezhetjük-e az MI elemzésekor, hogy az értelem és az én lényegileg független egymástól, az intellektus és szelf elválasztható? Szerintem aligha.
>
> Egy hölgyismerősöm mesélte a következő anekdotát. Egyetemista korában egy fehér patkányt tartott háziállatként, ami kellően engedelmes volt ahhoz, hogy csendben kövesse őt a kampuszon. A patkány egyszer nagyon megijedt egy kutyától és feliszkolt gazdája lábán, be egyenesen a nadrágjába, ahol olyan mélyre fúrta magát, hogy nem tudott kifarolni. Gazdája viszont nem mert megmoccanni, nehogy agyonnyomja. Pillanatnyi hezitálás után szégyenlősen letolta a nadrágját és kihámozta belőle a remegő rágcsálót, mire a környéken álló derült nézők spontán tapsban törtek ki.
>
> Ezt a történetet sokan viccesnek találják. Miért? Vélhetőleg azért, mert az anekdota főhősének helyébe képzelik magukat és miközben maguk is átélik a kínos helyzetet, megkönnyebbülnek, hogy nem velük történt. Kínosnak azonban csak az érezhet valamit (és megkönnyebbülni is csak az tud), akinek fontos, hogy kicsoda, aki-

> nek van önképe, és az önképe szerencsétlen helyzetekben veszélybe kerülhet. Ha nem egy személyiség, egy „öntudat" sorsa a tét, akkor nincs mit kínosnak vagy megnyugtatónak érezni. A fenti szituáció a személyes bevonódás miatt volt kínos, de pont ennyire lényeges, hogy amikor halljuk, mi is bevonódunk, magunk is feszengünk egy kicsit, és emiatt nevetünk. (Haugeland 1985: 239–240)

Haugeland szerint erre a fajta bevonódásra gépek nem képesek – pont azért, mert nekik tökmindegy. Az alábbiakban a szerző két explicitebb érvét ismertetem ezzel kapcsolatban.

4.1. Egzisztenciális holizmus

„A természetes nyelv megértése" című tanulmányában Haugeland a holizmus négy formáját különbözteti meg. Az első az *intencionális értelmezés holizmusa*, ami nagyjából megegyezik a radikális fordítás Quine-féle problémájával. Tegyük fel, hogy találkozunk valakivel, aki számunkra teljesen ismeretlen nyelvet beszél. A viselkedése alapján próbáljuk értelmezni a szavait. Ezt egyik szó esetében sem tudjuk megtenni anélkül, hogy figyelembe vennénk a többi szó feltételezett jelentését – a nyelvi jelentés holisztikus.

A második szint a „józan ész holizmusa", ami a kétértelmű vagy hiányos kifejezések értelmezésekor jelentkezik. Ha valaki azt mondja, „A kádban hagytam a kabátomat, mert vizes volt", akkor arra következtetünk, hogy az illető kabátja volt vizes, nem pedig a kád (Haugeland 1998: 49). Haugeland szerint az MI szempontjából a józan ész holizmusa nagyobb kihívást jelent, mint az intencionális értelmezés holizmusa, mert utóbbi csak a nyelvi közösségbe való belépés feltétele, a józan ész holizmusa viszont minden egyes nyelvi interakció kontextusában jelentkezik. A józan ész holizmusát egy (szimbolikus) MI úgy próbálhatja kezelni, hogy egy fogalmi térkép segítségével áttekinti az elvben lehetséges komplex jelentéseket és reménykedik, hogy csak az egyik jelentés koherens. Ha például azt a mondatot kell értelmezni, hogy „A majmok megették a banánokat, mert éhesek voltak", akkor megfelelő fogalmi térkép segítségével ki lehet zárni azt az értelmezést, hogy a banánok voltak éhesek. De ez a megoldás nem mindig működőképes. A korábbi példában, a kádban hagyott kabátnál a fogalmak struktúrája alapján nem kizárható az az értelmezés, hogy a beszélő azért hagyta a kabátját a kádban, mert a kád volt vizes. Lehet, hogy a beszélőnek a vizes kabát a fétise, vagy be akarta takarni a vizes kádat. Ez persze szokatlan viselkedés, de a problémát pont az jelenti, hogy ha a józan ész alapján tipikusan ki is zárhatóak az ilyen esetek, specifikálnunk kellene a józan észen alapuló normális vagy tipikus beszédszituációt, ez viszont explicit módon definiálhatatlannak tűnik.

Ez a probléma átvezet a harmadik típusú, *szituációs* holizmushoz. Szituációs holizmus esetén egy mondat helyes értelmezése nemcsak a józan észen

alapuló háttértudást feltételez, hanem a megszólalás aktuális kontextusától is függ. Az alábbi 1-es mondat például egész mást jelent, ha a 2-es mondat írja le a kontextust, ahol elhangzik (Haugeland 1998: 52):

(1) Mikor megérkezett a rendőrautó, a kamaszok abbahagyták a bankrablást, eltették a fegyvereket és hátraszaladtak a kocsihoz.

(2) Mikor apa hazaért, a gyerekek abbahagyták a játékot, eltették a játékokat és hátraszaladtak a kocsihoz.

> A mondat megértéséhez az olvasónak minden tagmondathoz egy „szituációs modalitást" kell rendelnie, méghozzá úgy, hogy megőrizze a globális szituáció koherenciáját. A koherenciának ezt az igényét nevezem szituációs holizmusnak: a különböző „szituációknak" (melyek függvényében az egyes tagmondatokat interpretáljuk) értelmes egésszé kell összeállnia. Ez a természetes nyelvi szövegek általános jellemzője. Olvasni csak az képes, aki képes ilyen koherenciát teremteni. (Haugeland 1998: 54)

A negyedik típusú, *egzisztenciális* holizmust Haugeland a sakkozással illusztrálja. Miért akar valaki sakkozni? Lehet, hogy elismerésre vágyik, vagy intellektuális kihívást keres, vagy le akar győzni másokat. Sokféle motiváció elképzelhető. De közös bennük, hogy mindegyik a sakkozó önképéből ered, abból, hogy milyen ember akar lenni. Hasonlítsuk ezt össze a gépek esetével:

> A sakk–számítógépekre azt szoktuk mondani, hogy „próbálnak" jót lépni, vagy „igyekeznek" nyerni. De vélhetőleg mindannyian érezzük, hogy valójában nem *érdekli* őket, nyernek-e, jól játszanak-e – valahogy az egész nem *számít* nekik. Honnan erednek az ellentmondó intuíciók? Elsőre úgy tűnhet, a gépekből hiányzik a motiváció: nincs okuk nyerni, nincs olyan távlati céljuk, amit a győzelem elősegítene. De ez csak elodázza a problémát, hiszen a következő kérdés az, hogy miért nincsenek távlati céljaik. Ameddig ezt nem válaszoltuk meg, nem jutottunk előbbre. (Haugeland 1998: 55)

Haugeland itt ugyanarra a jelenségre utal, amit a patkányos anekdotával próbált illusztrálni, de ezúttal *szemantikai* problémának tekinti. A természetes nyelv megértéséhez szerinte önképpel rendelkező, társadalmilag szituált szubjektumnak kell lenni:

> Csak olyan lény számára lehet fontos a bűntudat, felelőtlenség, önérzet, siker, élet és halál, aki egy többé-kevésbé stabil egésznek a részeként tekint magára és akit érdekel, hogy ezen az egészen belül kicsoda. És csak egy ilyen lény tud *olvasni*. Ezt a fajta holizmust, ami már látszólag sem a szövegben magában, hanem egyértelműen

> az olvasóban gyökerezik, a nagy szavaktól való félelmemet leküzdve *egzisztenciális holizmus*nak fogom hívni. Véleményem szerint ez a fajta holizmus kulcsszerepet játszik minden olyan szöveg megértésében, mely a szó hagyományos értelmében jelentést hordoz. Ha a szituációs holizmus a történetmondás alapja, az egzisztenciális holizmus a szépirodalom alapja. [...]
>
> Az egzisztenciális holizmus [ugyanakkor] nem csak a művészi tevékenység előfeltétele. Végeredményben – és ez a legfontosabb – semmi okunk éles különbséget tenni a mindennapi beszéd és a szépirodalom megértése között. Néhány rendkívül specifikus területet leszámítva, mint a sakk, tömegspektrometria vagy repülőjegyvásárlás, a legtöbb nyelvi interakciónak szerves része az emberi élet összes fontos aspektusa. (Haugeland 1998: 59)

A holizmus négy formája az MI előtt tornyosuló négy problémának feleltethető meg. A legsúlyosabb problémát az egzisztenciális holizmus jelenti, mert ha a gépeknek minden mindegy, ugyanakkor – az egzisztenciális holizmus miatt – az emberi beszédet csak az értheti, akinek igényei és szándékai vannak, akkor garantálható, hogy a gépek nem értik az emberi beszédet. Ahhoz, hogy értsék, identitást, aspirációt és társadalmi helyzetet kellene beléjük huzaloznunk.

4.2. Felelősség és jelentés

„Autentikus intencionalitás" című írásában Haugeland az intellektuális felelősség kapcsán dolgoz ki hasonló hipotézist, mint előbb a szemantika kapcsán. „Intencionalitás"-on azt a kognitív tevékenységet érti, hogy valamit valamilyennek ítélünk. Bár egy ponton a szemantikával azonosítja az intencionalitást (2002a: 161), később irányulásként írja le: a tárgyakra olyanként irányulunk, amilyenek (2002a: 172). Tehát az intencionalitás nem pusztán a szemantikát jelenti és nem is a referálás képességét, hanem a kettő egyesülését az igazságban, azt, hogy x-et helyesen F-nek ítéljük. Haugeland szerint ez a tudás (az „objektív tudás") alapfeltétele (2002a: 163).

Haugeland az intencionalitás három szintjét különíti el. Az első a *látszólagos* („ersatz") intencionalitás, ami a robotokat, állatokat, vagy (ha nagyon kitágítjuk a fogalmat) a fákat és termosztátokat jellemzi. A valódi intencionalitáshoz felszínesen hasonló, de alapvetően eltérő jelenségről van szó. Fizikai rendszerekről – akár a termosztátokról is – beszélhetünk úgy, mintha érzékelnék a környezetüket és kommunikálnának vele, de ez az azonos nevű emberi tevékenységek olcsó utánzata csupán.

A második szint a *közönséges* intencionalitás, amit Haugeland a tudományos megismerés példáján vezet be. Arra hívja fel a figyelmet, hogy a tudománynak mint ismeretszerző tevékenységnek alapvető része a felelősségvállalás és az önkritikára való hajlam. A tudós tiszteletben tart bizonyos normákat és nyitott a

normasértéssel kapcsolatos kritikára, főleg, ha a kritika a módszertani alapelvek vagy az eddig feltárt törvények megsértésére vonatkozik:

> A tudósoknak (az esetleges válságidőszakokat leszámítva) mindig határozott elképzeléseik vannak arról, hogy az általuk vizsgált dolgoknak milyennek *kell* lennie. A legalapvetőbb ezzel kapcsolatos meggyőződéseiket a természeti törvények és a módszertani elvek fejezik ki. Ezek (és specifikusabb következményeik) előre meghatározzák a tárgyakat, előre korlátozzák, hogy a kísérletek eredményeinek hogyan *kell* összefüggniük (Haugeland 2002a: 166)

A közönséges intencionalitásban azonban az episztemikus konzervativizmus mellett jelen van a tudomány radikális reformjára való elvi nyitottság is:

> Az empirikus tudomány lényegileg önkritikus [...]: kritikusan fürkészi nemcsak az egyéni teljesítményeket, hanem magát az egész tudományos gyakorlatot. [...] [Ez] is nyilvánvalóan elengedhetetlen ahhoz, hogy feltárja – módszeresen, nem véletlenszerűen – az igazságot. Tehát a tudományos tudásnak, ezáltal pedig a valódi intencionalitásnak ez a fajta önkritika is alapvető feltétele. (Haugeland 2002a: 166)

Bár még nem érkeztünk el az intencionalitás harmadik – autentikus – szintjére, Haugeland szerint már a második szinten sem lehet elválasztani a világ reprezentációját a normáktól és a felelősségtől:

> Az a benyomásom, hogy a kognitív tudomány – különösen az a része, mely előszeretettel használja a számítás fogalmát – gyakorlatilag teljesen figyelmen kívül hagyta az értelem és felelősség közti kapcsolatot. Vélhetőleg e mulasztás sejtelme áll az olyan amorf (és gyakran irritáló) „intuíciók" mögött, hogy az MI valójában nem gondolkodik, nem érti saját inputjait és outputjait, nincs tudata, és így tovább. Nem gondolom, hogy ezek az intuitív ítéletek hamisak lennének, de szerintem elvétik a tényleges problémát. Puszta intuíciók, a velük kapcsolatos (pro vagy kontra) „érvek" pedig a legjobb esetben is bosszantóan erőtlenek. (Haugeland 2002a: 168)

Az intencionalitás harmadik, *autentikus* szintje az önkritika még mélyebb formájára épül: arra a normára, hogy ha egy tudomány saját szigorú kritériumai szerint sikeres, ha megbízhatóan produkál alapelveivel összhangban eredményeket, akkor fenn kell tartani, ha viszont a valóság ellenáll a vizsgálatnak és az igazság, a tudományág saját kritériumai alapján, kicsúszik a megismerők kezéből – a kísérleti eredményekkel és hipotézisekkel kapcsolatos minden önkritikus revízió ellenére –, akkor az adott vizsgálódást fel kell adni.

> A tudományos objektivitás előfeltétele, hogy autentikusan felelősséget vállaljunk, vagy legalábbis időnként egyesek ezt tegyék. A fentiek miatt pedig ez a valódi intencionalitás és értelem előfeltétele is. Autentikus intencionalitásnak azt nevezem, mikor valaki autentikus módon vállal felelősséget intencionális tartalmaiért. [...] A hétköznapi intencionalitás feltételezi az autentikus intencionalitás képességét, bár e képesség használatára hétköznapi esetekben nem feltétlenül van szükség. De még ez a bonyolult függőségi viszony is jelzi, hogy a kognitív tudomány addig nem tudja megérteni az intencionalitás – és így a megismerés – alapfeltételeit, amíg nem érti meg az autentikus felelősség és autentikus intencionalitás természetét. (Haugeland 2002a: 173)

> Úgy vélem, ez a kétélű elkötelezettség-struktúra [az episztemikus vállalkozások fenntartásának vagy feladásának normája] életünk alapvető része, és nemcsak a tudományban van jelen, hanem sok (és bizonyos értelemben minden) fontos területen. A szerelem és a szabadság alapszerkezete is ilyen, sőt, épp a szerelem és a szabadság e struktúra fő megnyilvánulásai. Ezért a következő drámai – de teljesen komoly – állítással zárom soraimat: a kognitív tudomány és az MI-kutatás addig nem éri el alapvető céljait, amíg nem érti meg és/vagy implementálja a valódi szabadság és a szeretet képességét. (Haugeland 2002a: 174)

Bár egyetértek Haugelanddal abban, hogy az autentikus intencionalitás és az egzisztenciális holizmus gépesíthetetlennek tűnik, úgy vélem, gépesíthetetlenségük nem következik a fenti érvekből, különös tekintettel arra, hogy szigorúan véve nem is beszélhetünk érvekről, csak szakszerű sejtetésről. A megérzésekből akkor lenne érv, ha megmutatnánk, *miért* nem képes egy számítógép valódi (autentikus vagy holisztikus) intencionalitásra. A fenti sejtetések erre szerintem nem képesek. Ezt egy gondolatkísérlet révén láthatjuk be.

Isaac Asimov „A két évszázados ember" (*Bicentennial Man*) című novellájában egy Andrew nevű robotról kiderül, hogy különös tehetsége van a fafaragáshoz. Világhírű művész lesz, tulajdonosai pártolják ambícióit, Andrew pedig idővel állampolgárságot is kap, eszközből morális szubjektummá változik. Bár Andrew „puszta" robot, egyértelműen fontos számára a sorsa, társadalmi szerepe, ugyanúgy képes értelmezni a szituációkat, mint a hús-vér emberek. Asimov persze nem magyarázza el, hogy ezek a tulajdonságok hogyan erednek az Andrew fejében levő számítógépből, de a novella ettől függetlenül cáfolja, hogy egy MI nem *lehet* képes valódi intencionalitásra. Elgondolható, hogy képes rá, tehát nem zárhatjuk ki, hogy ez lehetséges. Ennek cáfolatához többet kell tennünk az intencionalitás és holizmus szuggesztív leírásánál, mert ezeket robotokhoz is lehet társítani a logikai ellentmondás veszélye nélkül.

Felvethető, hogy a Haugeland-féle autentikus intencionalitás és egzisztenciális holizmus az önreprezentáció képességére, a teleológiára és a fenomenális tartalmakra vezethető vissza. Azért van egzisztenciális nézőpontunk, ehhez kapcsolódva pedig episztemikus felelősségünk, mert érző és célokkal rendelkező lények vagyunk, akik képesek elgondolni saját magukat. A belső élettel nem rendelkező robotokból hiányoznak ezek a komponensek, és *emiatt* nem számít nekik semmi, ezért nem tudnak felelősséget vállalni megszólalásaikért. Ha viszont így érvelünk, akkor az egzisztenciális demarkáció mint önálló stratégia összeomlik, és a fenomenális tapasztalattal, teleológiával, önreflexióval kapcsolatos demarkációs érvek veszik át a helyét.

Az is lehetséges, hogy Haugeland a fenomenalitástól és önreflexiótól független, primitív jelenségnek vélte az autentikus intencionalitást és egzisztenciális holizmust. De ha ezek redukálhatatlan jelenségek, akkor nehéznek tűnik megkerülni az Asimov-novellából adódó ellenérvet, miszerint elgondolható egy egzisztenciális nézőponttal rendelkező MI. A fenomenalitásra, teleológiára vagy önreflexióra épülő demarkációs stratégiák esetében túl lehet lépni ezen a gondolatkísérleten, mert lehet amellett érvelni, hogy a gépeknek nincs fenomenális tapasztalata (7. fejezet), nem képesek teleologikusan gondolkodni (10. fejezet) és nem képesek önreflexióra (11–12. fejezet). Ha valamelyik érv ezek közül helyes, akkor Asimov novellája vagy egy metafizikailag lehetetlen szituációt ír le, vagy titokban azt feltételezi, hogy Andrew nem pusztán mechanikus lény, hanem mentális szubsztancia.

5.

Referenciális demarkáció

Ha Ancsa azon gondolkodik, érdemes-e Rómába utazni, akkor Ancsa gondolatai egy specifikus városra vonatkoznak, tartalmuk nem írható le anélkül, hogy az aktuális, valós Rómára hivatkoznánk. Ha Ancsa azon töprengene, érdemes-e elutazni egy olyan városba, ami tele van régi épületekkel, fontos vallási központ, egy birodalom székhelye volt és így tovább, akkor Ancsa gondolatai nem feltétlenül Rómára vonatkoznának, hiszen elvben létezhet egy másik galaxisban egy Rómára tökéletesen hasonlító város, ami a Rómáról adható összes kvalitatív leírást kielégíti. De ha létezik is egy ilyen város, Róma-2, Ancsa viszont azon töpreng, érdemes-e *Rómába* utaznia, akkor Ancsa gondolatai *Rómára* vonatkoznak, nem pedig városok egy bizonyos típusára. Tehát Ancsa elméje valami titokzatos módon képes kinyúlni magából és megragadni a külvilág egy részét, konkrétan egy teljes várost. Vagy legalábbis felvethető, hogy ez a helyzet, ha pedig ez a helyzet, akkor egyből adódik a kérdés, hogy az MI ⌜elméje⌝ képes-e ugyanerre.

A nyelvfilozófiában *referálás*nak („reference") nevezik azt a jelenséget, hogy egy kifejezés egy specifikus tárgyra vonatkozik: „Róma" Rómára referál, „III. Károly" pedig a jelenlegi angol uralkodóra. A *szinguláris propozíciók* azok az állítások, melyek referáló kifejezéseket tartalmaznak (vagy referáló kifejezéseket tartalmazó mondatok fejezik ki őket). Az a propozíció például, amit a „Róma gyönyörű város" magyar mondat kifejez, szinguláris, mert Rómára referál. Ennek nyomán az elmefilozófia néha beszél *szinguláris gondolatok*ról is („singular thoughts"), olyan mentális állapotokról, melyek tartalma egy szinguláris propozíció. A továbbiakban a tömörség kedvéért „referálás"-on nemcsak a propozíciók és tárgyak közti viszonyt értem, hanem az elmének azt a képességét is, hogy specifikus tárgyakra irányul. Ez nem standard terminológia, de kényelmesebb, mint a szinguláris gondolatok formálásának képességét emlegetni.

A referenciális demarkáció lényege az a felvetés, hogy az MI-nek nincsenek szinguláris gondolatai, a ⌜gondolatai⌝ nem referálnak. Ha ez így van, akkor az emberi gondolkodás és a gépi ⌜gondolkodás⌝ lényegileg különbözik, hiszen a mi gondolataink referálnak.

Ez az érv nagyon ritka; egyetlen szerzőről tudok, aki képviseli. De az igazság nem statisztikai kérdés, a referenciális demarkáció pedig kellően izgalmas ötlet ahhoz, hogy érdemes legyen foglalkozni vele. Nehézsége, hogy a nyelvfilozófia egyik különösen talányos területére vezet.

Mielőtt rátérnék magára az érvre, szeretném rövidre zárni a vitát azzal kapcsolatban, hogy a mi gondolataink *tényleg* referálnak-e. Ez részben azon múlik, hogy vannak-e szinguláris propozíciók, melyek a referáló gondolataink tartalmát alkotják. A szinguláris propozíciók létéről minimum fél évszázada zajlik meglehetősen szerteágazó vita a nyelvfilozófiában.[1] Csak hogy érzékeltessem a vita komplexitását, röviden bemutatok egy pro és egy kontra érvet a szinguláris propozíciókkal kapcsolatban. Az első érv a szinguláris propozíciók létét próbálja igazolni (Kaplan 1989: 512–13). Tegyük fel, hogy Ancsa egy vendéglőben ül, balján Bélával, jobbján pedig Cecillel. Cecil Szegeden él. Ancsa rámutat a tőle jobbra ülő személyre és azt mondja:

(1) Ő Szegeden él.

Legyen **p** az a propozíció (nyelvi tartalom), amit Ancsa kifejezett. **p** nem azonos azokkal a propozíciókkal, melyeket az alábbi mondatok fejeznek ki:

(2) Az Ancsától jobbra ülő személy Szegeden él.

(3) Az a személy, akire Ancsa rámutatott, Szegeden él.

Annak belátáshoz, hogy ezek egyike sem **p**-t fejezi, tekintsünk egy V lehetséges világot, ami kizárólag abban különbözik az aktuális világtól, hogy Béla és Cecil fordítva ült le, vagyis Béla ül Ancsa jobbján. V-ben (2) és (3) egyaránt hamis, de **p** igaz (hiszen Ancsa V-ben is Szegeden él). Mivel a (2) és (3) által kifejezett propozíciók másképp váltogatják az igazságértéküket világok között, mint **p**, nem lehetnek azonosak **p**-vel, hiszen más szituációk teszik igazzá őket. Hasonló a helyzet bármely propozícióval, ami – mint (2) és (3) – valamilyen reprezentáción, leíráson keresztül utal Cecilre. Minden ilyen esetben kereshetünk olyan V-t, ahol a szóban forgó leírás nem Cecilre vonatkozik, de (1) továbbra is igaz. A legkézenfekvőbb feltevés tehát az, hogy **p** közvetlenül Cecilre vonatkozik, tartalmának része Cecil mint konkrét, partikuláris egyén.

A fenti érvbe többféleképpen beleköthetünk,[2] de nézzünk inkább egy ettől független támadást a szinguláris propozíciók ellen. Legyen **q** az az állítólagos szinguláris propozíció, amit az alábbi mondat fejez ki, ha kimondom:

(4) Nem létezem.

[1] Ld. Fitch & Nelson (2024). A főszövegben kifejtett két rekonstrukció innen származik.

[2] Dummett (1991) például azt állítja, hogy a propozícióknak nincsenek, csak az őket alkotó jelentéseknek (*senses*) vannak modális tulajdonságai. Plantinga (2022) individuális lényegeket feltételez, melyek minden lehetséges szituációban egyedien kijelölnek egy individuumot, így képesek kiváltani a szinguláris propozíciókat.

A szóban forgó propozíció – legalábbis e sorok írásakor – hamis. De **q** vélhetőleg nem *szükségszerűen* hamis, hiszen akkor Istenhez hasonló lény lennék. Ha azonban **q** nem szükségszerűen hamis, akkor van olyan V lehetséges világ, ahol **q** igaz. De V-ben ekkor egyszerre igaz **q** és igaz az, hogy **q** létezik (hiszen **q** aligha lehet igaz, ha nem létezik). De ha **q** létezhet anélkül, hogy én léteznék, akkor hogyan lehet **q** szinguláris propozíció, olyan propozíció, mely bizonyos értelemben tartalmaz engem, melynek tartalmába bele vagyok építve mint konkrét, partikuláris individuum? Ha **q** létezhet anélkül, hogy én is léteznék, akkor aligha vagyok **q** része. Vélhetőleg **q** csak valamiféle leíráson keresztül utal arra, hogy nem létezem.

A szinguláris propozíciók barátait ez persze nem rettenti el.[3] De a további érvekbe és ellenérvekbe nem szeretnék belebonyolódni. Ennyi talán elég annak érzékeltetésére, hogy nyakatekert nyelvfilozófiai vitáról van szó, melyben vélhetőleg egyik fél sem áll nyerésre. A referenciális demarkáció értékelésekor ezért érdemes első körben, pusztán a vita kedvéért, feltételezni, hogy egyes mentális állapotaink referálnak. *Prima facie* védhetőnek tűnik ez az álláspont, ha pedig demarkációs érv építhető rá, akkor az az érv is *prima facie* védhető.

Mivel a referenciális demarkáció relatíve ismeretlen ötlet, hosszabban idézem egyetlen számomra ismert képviselőjének, Brian Cantwell Smith-nek a gondolatait:

> Nyilvánvalónak tűnhet, hogy egy számítógép is tud egyszerűbb tárgyakra, mondjuk egy pendrive-ra orientálódni vagy irányulni. Ha rákattintok egy gombra, ami azt ⌜mondja⌝ a gépnek, hogy „másold a kijelölt fájlt az A nyílásban levő pendrive-ra", és ha normál körülmények között a kattintás hatására a gép pontosan ezt teszi, akkor nem mondhatjuk-e, hogy a gép a pendrive-ra orientálódott?
>
> Nem, nem mondhatjuk. Tegyük fel, hogy közvetlenül a parancs teljesítése előtt egy csínytevő lecseréli az eredeti pendrive-ot egy másikra. A probléma nem egyszerűen az, hogy a számítógép ugyanúgy átmásolja majd a fájlt anélkül, hogy feltűnne neki a különbség; a probléma az, hogy képtelen megkülönböztetni a két esetet, nincsenek rá eszközei, hogy a két szituációt *eltérő* szituációkként értelmezze – másszóval, nem képes különbséget tenni az „ami épp a nyílásban van" leírás és egy partikuláris tárgy között, ami a leírást egy adott pillanatban kielégíti.
>
> Ebből következik, hogy [...] a számítógép nem képes a pendrive-ot *tárgyként* kezelni. Ha másolás közben átvariálom az adatkábeleket, és a gép a továbbiakban a pendrive helyett a Facebook-oldalamra küldi az adatokat, akkor a gépnek fogalma sem lesz, mi

[3] Talán a legfigyelemreméltóbb Timothy Williamson (2002) válasza, aki lenyeli a békát és bevállalja, hogy szükségszerűen létezik. Ezt az álláspontot (hogy minden szükségszerűen létezik, „necessitism") Williamson (2013) a modális logika matematikai struktúrájára vonatkozó érvekkel is megpróbálta alátámasztani.

történt. Ez nem csak ebben a konkrét esetben igaz, hanem szükségszerűen. Mivel a gép lényegileg tudatlan ezzel kapcsolatban, nem állíthatjuk, hogy bármikor is a pendrive-ra irányulna, szemben a pendrive reprezentációjával, vagy a pendrive-ot kezelő mechanikával (ez a legtöbb, amit állíthatunk).

Hogyan lehetne tudatában egy számítógép a pendrive és a pendrive által kielégített leírás („a meghajtóban jelenleg található adathordozó") közti különbségnek, ha a művelet során nem kell detektálható fizikai különbségnek lennie – oksági viszonyok szintjén – a tárgy és annak reprezentációja között? Emiatt nem veszi észre a gép, ha suttyomban lecserélik az eredeti adathordozót. A (normavezérelt) reprezentációs rendszerek funkciója pontosan ebben ragadható meg: hogy a közvetlen oksági kapcsolatokon túlnyúló viszonyok tekintetében is elszámoltathatóak legyenek és a társadalmi gyakorlatok gigantikus hálózatainak közvetítésével megfelelően kezeljék őket. Ha azt feltételezzük, hogy a szituációkat kimerítik a közvetlen környezet aktuális oksági összefüggései (vagyis ha *a priori* letesszük a garast a nyers mechanizmus mellett), akkor eleve kizárjuk a reprezentációt, szemantikát, intencionalitást és normativitást – azt, hogy valaki a világra orientálódik. (Cantwell Smith 2019: 83–84)

Képzeljük el, hogy tisztán műszaki alapon próbálunk belekötni a fenti érvbe. Egy pendrive-nak lehet egyedi azonosítója: az operációs rendszerek kötetsorszámot tudnak tenni az adathordozókra, és ezek a sorszámok különböző pendrive-ok esetében szinte mindig különböznek. A gép tehát észre tudja venni, ha suttyomban kicserélik a pendrive-ot egy másikra. Azt is észre tudja venni, ha az általa küldött adatok nem a pendrive-ra, hanem az internetre kerülnek, mert utóbbi esetben nem az USB-meghajtón, hanem a wifin vagy a netkábelen zajlik majd adatforgalom.

Cáfoltuk-e ezzel Cantwell Smith-t? Nem, mert a szóban forgó számítástechnikai különbségeket nem közvetlenül érzékeli a számítógép alaplapja és processzora, ahol a tényleges számítások zajlanak. Ha egy másik pendrive-ra ráügyeskedjük ugyanazt a kötet-sorszámot, akkor a két tárgyat már azonosnak fogja ⌜tekinteni⌝ a gép. Ha az USB port helyett trükkösen a hálózati kártyára küldjük az USB-nek szánt adatokat, akkor a gép nem veszi észre, hogy az internettel kommunikált. Vagyis a számítástechnikai részletek figyelembe vételével is arra jutunk, hogy a számítógép csak reprezentációk közvetítésével azonosít külső tárgyakat, képtelen különbséget tenni egy reprezentáció és annak (esetleg váltakozó) tárgyai között.

Felvethetnénk, hogy ez ránk is igaz, hiszen mi is csak az idegpályáinkat közvetlenül érő ingerek alapján azonosítjuk a külső tárgyakat. Az aktuális vizuális élményeim miatt azt gondolom, a megszokott asztalomnál ülök, de ha egy tréfás kedvű démon másodpercenként lecseréli az előttem levő asztalt egy kvalitatíve

azonos kópiára, akkor „az" asztalommal kapcsolatos gondolataim nem szinguláris gondolatok. Az asztalomat tehát ugyanúgy csak reprezentációkon keresztül tudom azonosítani, mint a számítógép a pendrive-ot. Ugyanúgy vak vagyok a közvetlen idegi ingereimen túli valóságra, mint a processzor az adatbuszain túli valóságra.

A referenciális demarkáció ellensége szkeptikus hipotézisek helyett kidolgozhat egy konkrét elméletet is, mely alapján a számítógép hozzánk hasonlóan képes referálni. Ez némi előkészületet igényel, mindenekelőtt különbséget kell tennünk a szűk és tág mentális tartalmak között:

> Egy tulajdonság *anti-individualisztikus*, ha definíciójában szükségképpen szerepel a tulajdonság hordozójának környezete, és *individualisztikus*, ha nem. [...] Egy mentális tulajdonság tartalma akkor *szűk*, ha a tulajdonság individualisztikus. A szűk tartalom lokálisan szuperveniál hordozóján, vagyis mindenki, aki intrinzikusan azonos egy hordozójával, rendelkezik ugyanazzal a mentális tartalommal. Az anti-individualisztikus mentális tulajdonságok tartalma a *tág* tartalom. (Sawyer 2007: 20)

A démonos példát használva: annak a mentális tulajdonságomnak, hogy egy asztalt látok magam előtt, szűk a tartalma, mert minden lehetséges világban, ahol ugyanilyen belső állapotban vagyok, asztalt látok magam előtt. Annak a mentális tulajdonságomnak viszont, hogy konkrétan *ezt* az asztalt látom magam előtt, már tág a tartalma, hiszen a démon által megtréfált hasonmásaim, bár belsőleg ugyanazt tapasztalják, mint én, nem *ezt* az asztalt látják maguk előtt. A referencia – a szinguláris propozíció – tág mentális tartalom.

A tág és szűk tartalom különbségét az MI ⌜mentális tartalmaira⌝ is alkalmazhatjuk. Függetlenül attól, hogyan varázsoljuk elő a szemantikát a gépben vagy környékén zajló folyamatokból, megkülönböztethetjük két típusát: azt a szemantikai tartalmat, ami a gép összes intrinzikus hasonmásában jelen van, külső viszonyaiktól függetlenül (szűk szemantikai tartalom) és azt, ami részben a környezettől függ (tág szemantikai tartalom). A referencia, a korábbi meggondolások fényében, tág tartalom, csakhogy Cantwell Smith érve a számtógép *individualisztikus* tulajdonságaira apellál: a processzor azért nem képes referálni a pendrive-ra, mert nem tudja megkülönböztetni a pendrive-ot annak belső reprezentációjától. De ez a jelenség – hogy a processzor belső reprezentációkkal dolgozik – szűk mentális tartalom (amennyiben mentális tartalom). Mivel Cantwell Smith nem foglalkozik a számítógép tág mentális tartalmaival, nem zárható ki, hogy a számítógép is képes referálni.

A referenciális demarkáció híve erre kétféleképpen válaszolhat. Egyrészt állíthatja, hogy a referencia *szűk* mentális tartalom. Ez nemcsak logikailag lehetséges, hanem történetileg pont egy ilyen álláspontra vezethető vissza a szinguláris propozíciók elmélete. Russell (1910, 1912) *ismeretség*nek („acquaintance") nevezte a szubjektum és az általa közvetlenül megragadott tárgyak közti

viszonyt. Szerinte csak akkor lehetünk ismeretségben valamivel, ha nem merül fel a téves azonosítás lehetősége (Fitch & Nelson 2024). Tehát Russell szerint sosem vagyunk a szó szigorú elmefilozófiai értelmében ismeretségben más emberekkel; még ha személyesen találkoztam is az angol királlyal, nem vagyok (a szó russelli értelmében) ismeretségben vele, mert lehetséges, hogy legközelebb egy tökéletes hasonmásával találkozom és nem veszem észre a különbséget. Emiatt Russell, bár hitt a szinguláris propozíciókban, furcsa dolgokat gondolt arról, hogy mire vonatkoznak: szerinte csak az aktuális érzet-adatainkra (*sense data*) és az univerzálékra tudunk referálni. Ha aktuálisan kávéillatot érzek, akkor ez a konkrét érzet (a fenomenális tapasztalatom) és annak tulajdonsága (az illat milyensége) az, amivel ismeretségben vagyok, amire referálok, ami partikuláris tárgyként jelen van gondolataimban. Ha kizárólag az elménk tartalmaira tudunk referálni, akkor a referencia vélhetőleg szűk tulajdonság, hiszen bármely két, belsőleg azonos szubjektumnak ugyanolyan szinguláris gondolatai lesznek. Ha kávéillatot érzek, minden intrinzikus hasonmásom kávéillatot érez. A russelli értelemben vett szinguláris tartalmak tehát csak akkor változhatnak, ha a szubjektum belső állapotai változnak.

A russelli koncepció elhárítja a szűk és tág tartalmak különbségére épülő ellenérvet, de felvethető, hogy emiatt érvénytelenné teszi magát a referenciális demarkációt. Cantwell Smith kiinduló feltevése az, hogy *mi* képesek vagyunk pendrive-okra és egyéb tárgyakra referálni, a számítógép viszont nem. De ha a referencia russelli természetű, akkor ez a kiinduló feltevés hamis, hiszen mi sem vagyunk közvetlen ismeretségben a pendrive-val, csak a pendrive-ot reprezentáló érzeteinkkel. Tehát a russelli referencia pürrhoszi győzelemhez vezet.

Gyümölcsözőbb válasz lehet a referencia tág jellegével kapcsolatos ellenérvre, ha elismerjük, hogy a referencia tág tartalom, de tagadjuk, hogy lennének szűk mentális tartalmaink. A szinguláris gondolatok közvetlenül a külvilágra vonatkoznak és elménk egyáltalán nem dolgozik reprezentációkkal. Ha ez az álláspont koherens, akkor a referenciális demarkáció továbbra is működik, mert mondhatjuk, hogy noha a referencia tág tartalom és tág tartalom a számítógépek állapotainak is tulajdonítható (pl. a külvilághoz fűződő oksági kapcsolatok miatt), a számítógép tág tartalmai reprezentációkon (pl. a pendrive-nak egy adatsínen elérhető fizikai címén) keresztül vonatkoznak a külvilágra, a mi elménk viszont reprezentációk nélkül vonatkozik rá. Ekkor továbbra is fenntartható, hogy mi emberek képesek vagyunk referálni, a gépek azonban nem.

Érzésem szerint John McDowell (1998) a szinguláris gondolatok ilyesfajta koncepcióját próbálja kidolgozni, bár nem köti nyíltan az MI-hez. Maga az összekötés azonban kézenfekvőnek tűnik. A kérdés az, hogy koherens-e azt állítani, hogy az elménk úgy irányul a külvilágra, hogy nem rendelkezik reprezentációkkal, vagy csak nagyon vékony értelemben, a puszta vélekedések értelmében rendelkezik velük, nem pedig speciális létezők, *reprezentációk* formájában, melyek a mentális állapotaink (külső tényektől különböző) tartalmát alkotnák. McDowell (1998: 237) szerint Russell az ismeretség-koncepció meg-

alkotásakor a karteziánus filozófia foglya maradt, Descartes alapvető hibája pedig az volt, hogy miközben az *Elmélkedések*ben megpróbálta lebontani és visszanyerni a világot, a lebontás során végleg elvesztette azt, mert bezárta az elmét a külvilágtól metafizikailag független reprezentációkba, melyekből akkor is kialakulhat a külvilág érzete, ha a külvilág nem is létezik. McDowell úgy véli, fel kell adni azt a russelli követelményt, hogy az ismeretséghez tévedhetetlenségre van szükség. Az elme és a külvilág között kell tételeznünk közvetlen kapcsolatot, melyet nem belső mentális tartalmak közvetítenek. McDowell nem tagadja, hogy szubjektumként befogadjuk a világot, van saját perspektívánk, de azt hangsúlyozza, hogy a perspektívánk magára a világra nyílik:

> A kognitív tér magában foglalja a „külvilág" releváns részeit. Emiatt a külvilághoz fűződő viszony nem generál karteziánus jellegű filozófiai problémákat. [...] Szükségünk van gyógyírra a világvesztéstől való szorongás ellen; de ha fenntartjuk az intencionalitással kapcsolatos [karteziánus] álláspontot, aligha tehetünk többet, mint hogy elfordítjuk a tekintetüket és tudatosan elkerüljük a kérdést, amit (jogosan) nyűgnek érzünk. A szinguláris gondolatok általam javasolt koncepciója a világvesztéstől való karteziánus félelmet más, és valóban kielégítő módon kezeli. Nem az intencionalitás és a tárgyak közti szakadékot próbálja áthidalni, és nem is feledkezik meg nagyvonalúan az egész problémáról, miközben annak gyökerét érintetlenül hagyja. Ehelyett alapjaiban forgatja fel azt az elképzelést az elme természetéről, amiből a karteziánus dichotómia ered. (McDowell 1998: 258–259)

A naiv realisták és a diszjunktivisták jó része hasonló álláspontot képvisel (ld. Crane & French 2021: 3.4, Soteriou 2020). Az elméleti opciók áttekintése helyett, ami kicsit messzire vezetne, néhány általános megjegyzést tennék erről a megközelítésről, melyet a referencia *antikarteziánus* elméletének fogok hívni.

A referencia antikarteziánus elmélete szerint a szubjektum közvetlenül maga előtt találja a világot, nem tolakszik a szubjektum és a világ közé a belső reprezentációk fátyla. Ha így van, akkor a referenciális demarkáció elvben sikerre vihető annak ellenére, hogy a referencia többé nem szűk tulajdonság és nincsenek szűk mentális tartalmaink. A feladat persze ekkor sem triviális, mert előbb be kell látnunk, hogy ha az MI-nek vannak mentális tartalmai, akkor ezek belső reprezentációkon (pl. az adatbuszokon közlekedő jeleken) alapulnak, vagyis az MI *nem* találja közvetlenül maga előtt a világot. Ezt talán abdukciós alapon, a legjobb magyarázat elvével igazolhatja az antikarteziánus. Az MI működéséről van egy világos mechanisztikus modellünk, e modell szerint az MI a processzoraiba érkező jeleken keresztül ⌜értelmezi⌝ a környezetét, tehát nincs okunk azt gondolni, hogy az MI esetleges ⌜mentális⌝ állapotaiban további – a mechanisztikus leírásban nem szereplő – tények is szerepet kapnak.

A referenciális demarkációnak ez a változata működőképes, de felvethető, hogy komoly ára van és egyfajta dualizmus elfogadását követeli annak ellenére, hogy a kartezianizmus elleni gyógyírként árulják. Az emberi agy és a külvilág viszonyával kapcsolatban ugyanis rendelkezésre áll egy ugyanolyan részletes fizikai magyarázat, mint az MI és a külvilág viszonyával kapcsolatban. Ha az MI belső folyamataiból hiányzik a világgal való közvetlen kapcsolat, akkor az agy biokémiai folyamataiból is hiányoznia kell. Bár a tranzisztorok és neuronok fizikailag eltérnek, az emberi agy pedig (ha számításokat végez) más algoritmust használ, mint a mai MI-k, ezek a különbségek lényegtelennek tűnnek, ha arról van szó, hogy képes-e az agy mint fizikai rendszer arra a reprezentációktól mentes, közvetlen kognitív viszonyra, amit az antikarteziánus feltételez. Ha az MI „agya" esetében ezt a kognitív viszonyt abduktív alapon kizártuk, milyen alapon feltételezzük az emberi agy esetében? Ha viszont nem feltételezzük, akkor az antikarteziánusnak bizonyos mértékig függetlenítenie kell a tudatot az agytól. Azt kell állítania, hogy amikor közvetlenül magunk előtt találjuk a külvilágot, valami más is történik azon kívül, hogy különféle ionok szaladgálnak a neuronokban. Ha pedig az érzékelés ennél többet jelent, akkor az elmét nem meríti ki az, ami az agyban történik, tehát elérkezünk a dualizmushoz.

Mivel az antikarteziánus álláspont rejtetten dualista, nem csoda, hogy igazolja a demarkációs tézist. Ha a gondolkodáshoz nemfizikai elmére van szükség, akkor az MI nyilván nem gondolkodik, mert semmi okunk feltételezni, hogy kellően bonyolult mátrixműveleteket esetén a számítógépre rácsatlakozik egy mentális szubsztancia. Kérdés azonban, hogy mi szól a referencia antikarteziánus elmélete által feltételezett dualizmus mellett. Önmagában az a tény, hogy a kartezianizmus nehezen megoldható talányokhoz vezet, nem érv amellett, hogy a dualizmus igaz. Analógia: az asztrofizika számára komoly kihívást jelent a sötét anyag és sötét energia magyarázata, ezek a rejtélyek pedig semmivé foszlanak, ha feltételezzük, hogy a távoli csillagrendszerek léte illúzió, mert Isten csak anyagtalan emberi lelkeket teremtett, az asztrofizikát pedig szándékosan kicsit inkoherensre tervezte, hogy az emberek ne váljanak a tudomány sikerei miatt elbizakodottá. Ha elfogadjuk ezt a fajta teista idealizmust, akkor a sötét anyag és sötét energia nem rejtélyes többé, hiszen valójában nem léteznek és a velük kapcsolatos elméleti nehézségek az isteni terv részei. Érv-e ez Isten léte mellett? Vélhetőleg nem. Hasonló a helyzet a referencia antikarteziánus elméletével. Az antikarteziánus megold egy problémát az érzékelés filozófiájával kapcsolatban, de ez az előny ugyanúgy nem érv az elmélet által implikált dualizmus mellett, ahogy a teista idealizmus asztrofizikai problémamegoldó képessége sem érv az istenhit mellett. Mindkét esetben további érvekre lenne szükség a szóban forgó fajsúlyos metafizikai tézis védelmében. Ha az antikarteziánusnak van ilyen további érve, akkor viszont referenciális demarkációra vélhetőleg már nincs szükség, mert a dualizmus önmagában igazolni fogja a demarkációs tézist.

6.

Szemantikai demarkáció

A szemantikai demarkációs tézis szerint az MI (velünk ellentétben) nem érti, mit jelentenek a szavaink, még akkor sem, ha nálunk ügyesebben generál szöveget. Ez azért tekinthető a kognitív demarkáció egyik válfajának, mert a nyelvi megértés nyilván az emberi gondolkodás központi formája, tehát ha az MI-ből hiányzik a nyelvi megértés, akkor az MI nem gondolkodik.

A legfontosabb szemantikai demarkációs érv, és egyben az MI szellemi képességeivel kapcsolatos leghíresebb kritika John Searle „Kínai Szoba" gondolatkísérlete. Mivel az ezzel kapcsolatos vita több, mint negyven éve zajlik és a filozófián túlnyúlva átszivárgott a kognitív tudományba és a számítástudományba is, irreális vállalás lenne, ha megpróbálnám összefoglalni vagy lezárni a polémiát. A gondolatkísérlet összefoglalása után két klasszikus ellenérvet fogok röviden ismertetni, utána pedig amellett érvelek, hogy a Kínai Szoba nem valódi érv, hanem egy „intuíciódetektor", amit csak az érez helyesnek, aki eleve úgy véli, hogy az MI nem érti az emberi nyelvet. A fejezet végén egy, a Kínai Szobától eltérő szemantikai demarkációs stratégiát ismertetek, ami véleményem szerint több sikerrel kecsegtet.

6.1. A Kínai Szoba

Searle legendás cikke, az „Elmék, agyak, programok" (1980) egy korabeli, mai szemmel kezdetleges MI-vel foglalkozik, ami a történetmegértést szimulálta (Shank & Abelson 1977: 150–237). A program rövid szöveges történeteket („scripts") kapott inputként, majd ⌜válaszolt⌝ a velük kapcsolatos kérdésekre. Ha a történet szerint valaki bement egy étterembe, rendelt egy hamburgert, majd a kiérkező elszenesedett ételt látva fizetés nélkül elrohant, az MI helyesen válaszolt arra a kérdésre, hogy az illető evett-e a hamburgerből. Searle az „erős MI" koncepcióját támadja e program kapcsán, konkrétan azt a hipotézist, hogy a program *érti* a történeteket és a program struktúrája *magyarázza* az emberi megértést. Searle elképzeli a szóban forgó program egy olyan változa-

tát, ami kínai nyelven ⌜kommunikál⌝. Ez a program (mint absztrakt számítási modell) leírható papírra, rögzíthetjük egy szabálykönyvben. A könyv mechanikusan végrehajtható instrukciókat tartalmaz majd arról, hogyan kell kínai jelek sorozataiból kínai jelek további sorozatait generálni, pusztán a jelek formája alapján. Tegyük fel, hogy Searle kézhez kapja a könyvet. Egy szót sem tud kínaiul, de erre nincs is szüksége, ha el akarja látni a történetelemző MI funkcióját. Ahhoz, hogy táblázatból kikeresse és egy digitális program szabályai szerint manipulálja a kínai jeleket, elég, ha meg tudja különböztetni őket. Ha tehát Searle-t bezárják egy szobába, és a kódkönyv mellé adnak neki két, kínai jelekkel teleírt lapot – a „történetet" és a „kérdéseket" –, akkor Searle a kézkönyv utasításait követve létre tudja hozni ugyanazt az outputot, amit a számítógép. Ha egy kínai személy cetliken csúsztatja be neki a kérdéseket, akkor a kérdező joggal hiheti majd (amennyiben a program elég fejlett), hogy a szobában levő ember is tud kínaiul. Ennek ellenére Searle-nek, akik odabent iszonyatosan unalmas numerikus kalkulációkat végez, halvány fogalma sincs a beszélgetés tartalmáról. Mivel Searle ugyanazt csinálja a kínai jelekkel, mint az MI, az MI sem érti, miről ⌜beszélget⌝. Az MI-szakirodalom Harnad (1990) nyomán ezt néha „szimbólum-alapozási problémának" (*symbol grounding problem*) nevezi.

Searle hat lehetséges ellenérvet vázol fel, és a vita a mai napig nagyjából ezek körül mozog (ld. Cole 2024), az egyetlen újdonság a Dennett-féle „dialektikus materialista" kifogás (ld. alább, 6.3.2). A Searle által megelőlegezett ellenérvek a következők:

Rendszerválasz Bár a Kínai Szobában ülő ember nem ért kínaiul, a rendszer egésze (ember + szabálykönyv + toll stb.) ért kínaiul.

Robotválasz A Kínai Szoba azért nem ért kínaiul, mert nincsenek érzékszervei, nem képes helyváltoztatásra, eszközhasználatra. Ha egy *robot* fejébe töltenénk megfelelő programot, akkor az így keletkező rendszer már értené a kínai mondatokat.

Agyszimulátor-válasz A Kínai Szoba nem ért kínaiul, de ha realisztikusan szimulálnánk számítógéppel egy kínai személy agyát, akkor a számítógép már értené a kínai mondatokat.

Kombinált Válasz Az előző három válasz kombinációja: ha agyszimulátort szerelünk egy robotba, az így keletkező rendszer egész érteni fogja a mondatok jelentését.

Interszubjektivitás-válasz A többi emberrel kapcsolatban sem lehetünk biztosak benne, hogy értik a kínai (vagy egyéb) mondatokat.

A „sok hely van Isten házában” válasz [1] A Kínai Szoba csak a jelenlegi számítógépekre érvényes. Lehet, hogy a tudomány és technológia fejlődésével képesek leszünk olyan gépeket gyártani, melyek értik az emberi beszédet.

A fentiek közül a **Rendszerválasz**t tárgyalom majd részletesebben, mert tapasztalatom szerint ezt találják a legmeggyőzőbbnek azok, akik elutasítják Searle érvét, és ha valami szerencsétlenség folytán fizikalistává változnék, én is ezt a választ preferálnám. Hogy a neve ütősebb legyen, „emergentista válasz”-nak fogom hívni. Ez szerintem találóbb címke, mint a „rendszerválasz”, mert a kulcskérdés nem az, hogy a Kínai Szoba *rendszer*-e, hanem az, hogy a nyelvtudás a rendszer *egészének* tulajdonsága-e, anélkül, hogy a rendszer bármely (valódi) részének (konkrétan a szobában kalkuláló Searle-nek) a tulajdonsága lenne. Az „emergentista” címkével nem utalok az emergencia semmilyen specifikusabb típusára, főleg nem az erős emergenciára, amiből az következne, hogy a megértés nem a Kínai Szoba részein szuperveniál (az emergentista válasz képviselői pont amellett kardoskodnak, hogy ezeken szuperveniál). A címke pusztán arra utal, hogy a részek interakciójából újfajta jelenség ered, ami az egyes részekben, vagy azok kisebb csoportjaiban, nincs jelen.[2]

A Searle által tárgyalt többi ellenérvről vagy szó esik máshol a könyvben, vagy jelen kontextusban irreleváns. A **Robotválasz**hoz hasonló ötletekről szó volt a testi demarkáció kapcsán. Az **Agyszimulátor-válasz** a technológia jelenlegi fokán kevéssé releváns, mert belátható időn belül semmi esély emberi agyak realisztikus szimulálására (Bostrom 2014: 30–36), de egy ehhez hasonló kifogásról szó lesz a Searle által eredetileg nem tárgyalt „dialektikus materialista” válasz kapcsán (6.3.2). A **Kombinált Válasz**ra az előző három megjegyzés kombinációja vonatkozik, az **Interszubjektivitás-válasz**t pedig véleményem szerint megnyugtatóan hatástalanította Searle (1980: 421–422). Mint megjegyzi, nem az a kérdés, hogy tudjuk-e a más emberekről, hogy értik a mondatokat, hanem az, hogy az a belső folyamat, amit tulajdonítunk nekik, mikor gondolkodó lényeknek tekintjük őket, számítási folyamat-e. Az utolsó válasz, a **„Sok hely van Isten házában”**, szintén nem releváns, mert a tudomány jövőbeli diadalmas fejlődésével kapcsolatos messianisztikus hiten alapul, ez a könyv viszont a jelenlegi MI-technológiával foglalkozik.

Mielőtt rátérnénk a klasszikus ellenérvekre, érdemes áttekinteni Searle saját értelmezését a Kínai Szobáról, valamint az ellenoldal reakcióit rá.

[1] „Many mansions reply”. A cikk magyar fordításában (Searle 1996) a „sok ház válasz” címke szerepel, de Searle itt János evangéliumának 14:2. versére utal: „Az én Atyámnak házában sok lakóhely van”.

[2] Az emergentista irodalom klasszikusa, Philip W. Anderson Nobel-díjas fizikus „More is different” c. írása erre a gyenge emergencia-fogalomra épül (Anderson 1972).

6.2. A szintaxis és szemantika ellentéte

Searle szerint a Kínai Szoba fő tanulsága az, hogy a szimbólumok formális manipulációja és a nyelvi megértés két teljesen eltérő jelenség:

> Szimbólumok formális manipulációja önmagában nem hordoz intencionalitást; nyoma sincs benne az értelemnek; igazából nem is *szimbólumok* manipulációja, mert a szimbólumok ilyenkor nem szimbolizálnak semmit. Nyelvészeti zsargonnal élve csak szintaxisuk van, szemantikájuk nincs. A számítógépekben látszólag jelenlevő intencionalitás kizárólag a programozók és felhasználók elméjében létezik, amikor az inputokat betáplálják és az outputokat értelmezik. (Searle 1980: 422)

Searle számára az intencionalitás és a szemantika ugyanaz a jelenség, vagy legalábbis egy tőről fakad, mert a szavaknak végeredményben a mentális állapotaink reprezentációs tartalma (vagyis az intencionális tartalom) miatt van jelentése. Ezzel talán nem mindenki ért egyet, de anélkül, hogy állást foglalnék, Searle felfogásához igazodom.

Searle nem definiálja, mit jelent implementálni egy algoritmust. Későbbi munkáiban (pl. Searle 1990) mintha a formalista felfogást képviselné (15. fejezet), de az implementációs idealizmus (19. fejezet) is belelátható az állításaiba. Bármit is gondol azonban az implementációról Searle, úgy véli, a számítógépek – a róluk adott interpretációinkról leválasztva – csak fizikai különbségekre érzékenyek, csak „szintaxisuk" van. Az emberi elmében azonban megértés is zajlik:

> Tisztán formális szimbólumokból nem eredhet mentális tartalom, mert definíció szerint nincs jelentésük – nincs interpretációjuk, szemantikájuk –, vagy kizárólag annyiban van, amennyiben kívülről valaki jelentést tulajdonít nekik. (Searle 1989: 45)

A szintaxis és szemantika ellentétének tézisét sokan próbálták aláásni. Daniel Dennett például egyetért azzal, hogy a számítógépek outputjai és belső állapotai csak a mi interpretációinkban hordoznak jelentést, önmagukban puszta mechanikai jelenségek, de Dennett szerint ez ránk, emberekre is igaz:

> Az agy az intencionális rendszerek elmélete vagy az evolúciós biológia felől nézve szemantikai gép [*semantic engine*], melynek az a dolga, hogy értelmezze az inputok özönét, szemantikailag elkülönítse őket, és „jól" reagáljon rájuk. Az agy erre való. Ugyanakkor az agy a fiziológia vagy az egyszerű józan ész szerint puszta szintaktikai gép [*syntactic engine*], ami az inputokat kizárólag struktúrájuk, időtartamuk és fizikai tulajdonságaik alapján különbözteti meg, működését mechanikusan vezérlik az inputoknak ezek a „szintaktikai"

> jegyei. Az agy ennél többre nem képes. De akkor hogyan lesz a szintaxisból szemantika? (Dennett 1987: 61)

Dennett puszta *interpretációnak* tekinti azt, hogy az agyállapotaink jelentést hordoznak. Szerinte tehát mi emberek is kínai szobák vagyunk, és nincs senki, aki intrinzikusan, searle-i értelemben Valóban Értené a szavakat, amiket használunk. A szemantikai tartalom feltételezése puszta heurisztika, ami segít megjósolni az emberek viselkedését, de nem a belső természetük pontos leírása. Az MI ugyanannyira nem ért magyarul, mint mi; vagy a másik oldalról nézve: ugyanannyira ért magyarul, mint mi, mert értelmezhető intencionális rendszerként. Az értelem a gépek és az emberek esetében is külső szemlélők által rácímkézett tulajdonság. „Searle a jelek szerint összetévesztette a szemantikának a szintaxisból való levezethetetlenségét azzal, hogy a szintaxisból nem vezethető le a szemantikát tulajdonító *tudat*" (Dennett 1987: 336).

A szintaxis és szemantika állítólagos ellentétére tehát reagálhatunk *antirealista* módon, tagadva, az értelem a külső interpretációktól független jelenség lenne. De ezzel ellentétes módon is tagadhatjuk a searle-i dichotómiát: tulajdoníthatunk valódi intencionális állapotokat a gépeknek. Searle több kritikusa szerint a számítógépek belső állapotai sem tisztán formálisak vagy szintaktikaiak, bár ezt a Kínai Szoba trükkösen elleplezi.[3] David Chalmers (1996b: 327) a Kínai Szoba alábbi paródiájával illusztrálja a problémát:

(1) A tortareceptek tisztán szintaktikaiak.

(2) A szintaxis önmagában nem garantálja az omlósságot.

(3) A torták omlósak.

(4) Tehát a tortarecept megvalósítása nem elegendő egy torta létrehozásához.

Mivel (4) hamis, Chalmers szerint Searle demarkációs érve is hibás, hiszen Searle az 1., 2. és 3. premisszával analóg állításokból vonja le azt a következtetést, hogy a szintaxis implementálása nem elegendő a szemantikához.

Ez a paródia azonban nem hatástalanítja a Kínai Szobát, mert (1) hamis, vagy legalábbis nem igaz abban az értelemben, ami a Kínai Szoba hatástalanításához kellene. Tegyük fel, hogy a marslakók nyelve a hangzás és íráskép szempontjából hasonlít a magyarra, de egyes szavak és nyelvtani szerkezetek teljesen mást jelentenek, emiatt pedig egy magyar tortarecept a marslakó számára a sós perec receptje. Ha egy embernek és egy marslakónak odaadjuk ugyanazt a receptet, akkor előbbi esetben egy torta, utóbbi esetben egy adag sós perec lesz a végeredmény. Ha „recept"-nek azt nevezzük, amit a humán és a marslakó egyaránt kézhez kap, akkor (1) igaz, viszont a konklúzió, (4) szintén igaz, hiszen a recept önmagában – mint betűk sorozata – nem elegendő a torta létrejöttéhez, kell hozzá szemantika is. Ha a „recept" szót ezzel szemben olyan értelemben

[3] Az alábbiakban Cole (2024) alapján foglalom össze a vitát.

használjuk, mely alapján a marslakó más receptet kap kézhez, mint az ember (annak ellenére, hogy ugyanazt a betű-sorozatot kapják), akkor (4) Chalmers szándékainak megfelelően hamis, viszont (1) is hamis, hiszen a recept ez esetben már nem pusztán egy fizikai mintázat, hanem társul hozzá jelentés. Megszűnik a párhuzam a Kínai Szobával, mert a tortarecept a „recept" szó második értelmezése alapján nem tisztán szintaktikai struktúra. A Chalmers-féle paródia inkább megerősíti Searle tézisét a szintaxis és szemantika különbségéről.

Ettől függetlenül persze tagadhatjuk, hogy a számítások szigorúan szintaktikai jellegűek. A számítógépekben bonyolult fizikai folyamatok zajlanak, emiatt pedig elvben az MI belső állapotainak is lehet szemantikai tartalma, amennyiben a szemantikai tartalom bonyolult belső folyamatokon, vagy ezeknek a környezettel való összefüggésein múlik. Rey (2002: 222) szerint „a számítások leírásában szereplő, szintaktikailag definiálható objektumoknak lehet és többnyire van is jelentése annak ellenére, hogy [számítástechnikai] leírásukban ez történetesen nem szerepel". Haugeland (2002b: 385) úgy véli, a szemantika egyenesen kihagyhatatlan a számítógépek leírásából: „Azt, hogy egy számítógép egy programot futtat, eleve csak úgy lehet értelmezni, hogy a processzor jelentést tulajdonít a programban szereplő utasításoknak". Az implementáció szemantikai felfogásának hívei (ld. 14. fejezet) ugyanezt gondolják.

Ha a számítógép belső állapotai önmagukban – a mi interpretációinktól függetlenül is – hordoznak jelentést, akkor ez vélhetőleg a gép és a külvilág közti kapcsolatból ered. Térjünk vissza egy pillanatra Chalmers tortareceptjéhez és képzeljünk el egy szakácsrobotot, ami szöveges recepteket ⌜értelmez⌝. Ha kezébe nyomunk egy tortareceptet, beszkenneli kamera-szemével és nekilát krémet főzni, piskótát sütni, csokimázat készíteni. Most képzeljünk el egy *marsi* robotot, ami ugyanannak a papírfecninek a hatására sós perecet állít elő. A két robot agya nyilván teljesen eltérő kapcsolatokat kódol a szavak (mint betűk sorozatai) és a külvilág között. A földi robot számára a „torta" szó tortákkal függ össze, a marsi robot számára perecekkel. A földi robotot vélhetőleg humán cukrászokról készült videók segítségével tanították be, tehát az ő programjába a „torta" és a torták közti oksági kapcsolat épült be, a marsi robot programjába viszont a „torta" és a sós perecek közti oksági kapcsolat épült be. Emiatt az MI algoritmusa nem puszta szintaktikai jelenség: szemantikai tartalma is van.[4]

Két kifogást láttunk a Kínai Szoba searle-i értelmezése ellen. Az egyik szerint az MI valóban szintaktikai gép, de mi emberek is azok vagyunk, tehát Searle demarkációs kísérlete kudarcot vall. A másik ellenreakció szerint az MI belső állapotai okságilag összefüggnek egymással és a külvilággal, ez pedig elegendő a szemantika jelenlétéhez. Véleményem szerint e stratégiák egyike sem működőképes. Az első, a jelentéssel kapcsolatos dennetti antirealizmus akkor volna hihető, ha komolyan vehetnénk azt a felvetést, hogy a megértés élménye nem

[4] Az MI állapotainak ilyesfajta „tág" szemantikai tartalma mellett érvel Cappelen és Dever (2021).

valós jelenség. De ezt semmi okunk komolyan venni. Az Orwell-féle duplagondol mesterfokú használata kell ahhoz, hogy valaki a szcientista ideológia nevében semmisnek nyilvánítsa saját belső világát. Erre a kérdésre bővebben visszatérek majd a 20. fejezetben. A második ellenérvhez a szemantika naturalizálására van szükség; erről a 6.5. alfejezetben lesz szó.

Fontos azonban hangsúlyozni, hogy ha a Kínai Szoba searle-i értelmezése helytálló is és a szintaxis és szemantika valóban élesen különbözik, különbségük *önmagában* nem igazolja, hogy az MI nem érti a szavak jelentését. A szintaxis és szemantika ellentéte a Searle-féle érvnek egy olyan premisszája (vagy az érv premisszáinak olyan következménye), ami logikailag kompatibilis a demarkációs tézis tagadásával. Logikailag lehetséges ugyanis, hogy egy számítógép érti az emberi nyelvet annak ellenére, hogy a megértéshez nem elegendő a szintaktikai manipuláció.[5] Ennek a logikai lehetőségnek a kihasználásával pedig megtorpedózhatjuk a Kínai Szobára épülő szemantikai demarkációt. Egyrészt felvethetjük, hogy a Kínai Szobában igenis jelen van a megértés (a szintaktikai műveleteknek köszönhetően, egyfajta emergencia révén). Másrészt felvethetjük, hogy a Kínai Szoba nem végez megfelelő típusú szintaktikai műveleteket. Ha megfelelő típusú – kellően komplex, vagy kellően gyors – szintaktikai műveleteket végezne, akkor a megértés is jelen lenne benne. A következő alfejezet röviden ismerteti ezt a két népszerű ellenvetést.

6.3. Két klasszikus ellenérv

6.3.1 Az emergentista válasz

A Kínai Szoba demarkációs erejét nemcsak a szintaxis és szemantika közti állítólagos ellentét kikezdésével lehet megkérdőjelezni, hanem úgy is, ha tagadjuk a konklúziót, miszerint a szóban forgó rendszerből hiányzik a nyelvi megértés. Az emergentista válasz tagadja ezt a konklúziót, míg a következő alfejezetben sorra kerülő „dialektikus materialista" válasz deflálja azáltal, hogy technikailag kivitelezhetetlennek állítja be a Kínai Szobát.

Az emergentista szerint a szobában kalkuláló Searle nem ért kínaiul, az általa használt eszközök sem értenek kínaiul, de a belőlük összeálló – „emergens" – rendszer ért kínaiul, legalábbis ha az implementált program elég fejlett. Az

[5] Searle szerint *aktuálisan* léteznek olyan számítógépek, melyek értik az emberi nyelvet: az emberi elme szerinte pontosan ilyen számítógép. „Képes lehet-e egy digitális számítógép gondolkodni? Igen, hiszen mi emberek tetszőleges számú számítógépes program implementációi vagyunk, és képesek vagyunk gondolkodni" (Searle 1980: 422). Searle tehát szubsztrátumfüggőnek tekinti az elmét és tagadja, hogy többszörösen realizálható lenne (vagy legalábbis azt, hogy szilikonchipeken realizálható lenne). Számomra érthetetlen, hogyan lehet ilyen alapon elutasítani a Robotválasz és az Agyszimulátor-válasz kombinációját. Ha egy robot fejében funkcionális értelemben pontosan ugyanaz történne, mint az emberi agyszövetben, miért hiányozna belőle továbbra is a megértés? Mert nincsenek benne nátrium- és kálium-ionok...?

emergentista nem tagadja, hogy különbség van szintaxis és szemantika között, hanem azt állítja, hogy a szemantika, bár a szobát működtető ember „fejében" nincs jelen, a szoba „egészében" jelen van (Copeland 2002, Haugeland 2002b, Kurzweil 2002).

A vita előtt fontos tisztázni, mit értünk itt „szemantikán". Ha dennettiánus antirealisták vagyunk és a jelentést pusztán hasznos fikciónak tekintjük, akkor a szoba egésze nyilván érti az emberi nyelvet, mert interpretálható ilyen módon (a kint ülő kínai személy, aki a szobával „beszélget", pont ezt teszi). Hasonlóan, ha a szemantika a számítógép és külvilág közti ok-okozati viszonyokon múlik, akkor a szoba állapotai hordozhatnak jelentést (ld. a marsi és földi szakácsrobot korábbi példáját). De sem a szemantikai antirealistának, sem a szemantikai naturalistának nincs szüksége emergentizmusra – nem kell feltételezniük, hogy a szoba bizonyos értelemben több, mint részeinek összege. Ha a szoba lakója fejből megtanulja a szabálykönyvet (de továbbra sem érti a kínai jeleket), akkor az emergentista válasz működésképtelenné válik, míg a szemantikai antirealista vagy naturalista válasz továbbra is fenntartható. Ha tehát valaki emergentistaként akar reagálni a Kínai Szobára, akkor el kell ismernie, hogy a megértés egyrészt valós, másrészt nem pusztán funkcionális jelenség, hanem valami más, valami intrinzikus tartalom.

Tegyük egyelőre félre, pontosan milyen tartalom is lehet ez. Tegyük fel, hogy a Kínai Szobának mint szupraindividuális rendszernek pontosan ugyanolyan (és nem antirealista vagy naturalisztikus) értelemben vannak hitei és gondolatai, mint egy embernek. Ez nem feltétlenül inkoherens hipotézis, hiszen szoktunk szupraindividuális rendszereknek tudást és vélekedéseket tulajdonítani. A Nemzeti Adó- és Vámhivatalról például kijelenthető, hogy ismeri minden magyar munkavállaló keresetét, pedig ez a NAV egyetlen munkatársáról sem mondható el. Sőt, még az sem igaz, hogy a szóban forgó tudás szét lenne szórva a NAV sok száz munkatársa között és az egyik adminisztrátor X-ről tudná, mennyit keres, egy másik Y-ról, az összes adminisztrátor pedig kollektíve mindenkiről. A NAV mint számítógépekből, épületekből, eljárási szabályokból *és* emberekből álló *rendszer* tudja, ki mennyit keres.

Az emergentista válasz híveinek vélhetőleg azt kell állítaniuk, hogy a Kínai Szoba valahogy így, az intézményekhez és társadalmi csoportokhoz hasonló módon rendelkezik kognitív állapotokkal. Egyes emergentisták expliciten azt feltételezik, hogy a Kínai Szoba „virtuális elmét" alkot, ami különbözik a szabályokat végrehajtó ember elméjétől (Sloman & Croucher 1980, Block 2002).

Számomra meglehetősen homályos, hogy emergentista alapon pontosan hol lennének jelen a Kínai Szobában kínai nyelvű gondolatok. De akárhol is tanyáznak, az emergentista ellenérvvel szembe lehet szegezni egy másik, a Kínai Szobához hasonló gondolatkísérletet, a Kínai Agyat (Block 1978). Képzeljük el, hogy Kína teljes lakossága szimulálni kezd egy N neurális hálót: adóvevők és különféle eljárási szabályok segítségével implementálják az N-t definiáló számítási folyamatokat. Blocknál a kínai lakosság az emberi agyat implementálja,

de ehelyett feltehetjük, hogy Kína egy LLM-et implementál. Fura lenne azt gondolni, hogy ekkor a kínai polgárok feje fölött a rádióhullámok rezgéséből előidereng egy új szubjektum, Kína Tudata, ami *érti* az egyes emberek számára értelmezhetetlen jelek szemantikai tartalmát.

Block eredetileg nem a nyelvi megértéssel, hanem a fenomenális tudattal kapcsolatban hivatkozott a Kínai Agyra, de a gondolatkísérlet kognitív demarkációra is használható.[6] Volt, aki expliciten ilyen célra használta fel, mégpedig Block előtt két évtizeddel, és nem a filozófiában, hanem a sci-fi irodalomban. Anatolij Dnyeprov *A játék* című novellájában (Dnyeprov 1969, eredeti megjelenés 1961) az Ifjú Szovjet Matematikusok Konferenciájának elnöke, Zarubin professzor kitereli az ifjú szovjet matematikusokat egy stadionba, csapatokra osztja őket és azt a feladatot adja nekik, hogy egy előre megadott szabályrendszer alapján adogassanak cetliket egymásnak, illetve csoportosan változtassák a helyüket. Senki sem érti, mi értelme a játéknak, de pár óra elteltével kiderül, hogy a cetlik és a szabályok segítségével egy fordítóprogramot implementáltak, ami lefordított oroszra egy portugál mondatot (az egyik csapat szabályszegése miatt kicsit hibásan). Zarubin professzor a záróbeszédében kijelenti, hogy a kísérlet az emberi értelem gépesíthetetlenségét illusztrálja:

> – Ha önöknek egy logikai mintázat strukturális elemeiként fogalmuk sem volt, mit csinálnak, beszélhetünk-e egy olyan szerkezet „gondolatairól", melynek alkatrészei még az elektronikus agyak legelvakultabb rajongói szerint is képtelenek bármiféle gondolkodásra? Önök ismerik ezeket az alkatrészeket: a rádiócsöveket, félvezetőket, mágneses mátrixokat és hasonlókat. Úgy vélem, játékunk választ adott a kérdésre: „Tudnak-e a gépek gondolkodni?". Bebizonyítottuk, hogy a gépi gondolkodás legtökéletesebb szimulációja sem maga a gondolkodás folyamata, mert ez az élő anyagnak egy magasabb mozgásformája.

Mondhatnánk, hogy ezek a gondolatkísérletek bohó képzelgések. De technikailag megvalósíthatónak tűnnek, sőt (mint a NAV példája mutatja) akár aktuálisan létezhetnek is. A NAV-nak tulajdonítunk kognitív állapotokat, mégsem gondoljuk, hogy egy külön szubjektum, ami hozzánk hasonlóan *érti* a magyar mondatokat. Márpedig a Kínai Szoba emergentista hatástalanításához az ilyen típusú állításokat kellene hihetővé tennünk. Az emergentistát az a veszély fenyegeti, hogy ha nem naturalizálja vagy eliminálja a jelentést (ami viszont feleslegessé teszi az emergentizmust), akkor a Kínai Agyhoz és a dnyeprovi játékhoz hasonló esetekben mélyen kontraintuitív vállalásokra kényszerül.[7]

[6] Block a *kválék* (ld. 7. fejezet) melletti érvnek szánta a Kínai Agyat, és a Kínai Szoba kapcsán érdekes módon az emergentista válasz híve (Block 2002).

[7] Schwitzgebel (2015) szerint az ilyen kontraintuitív vállalások a fizikalizmus lényegéhez tartoznak.

A kontraintuitív vállalások azonban nem *bizonyítják*, hogy az emergentista válasz téves. A tudomány gyakran kényszerül kontraintuitív vállalásokra. Ha az elmével kapcsolatban is kontraintuitív következtetésekre jutunk, akkor lehet, hogy az elme sokkal furcsább jelenség, mint eredetileg gondoltuk. Ezt a logikai lehetőséget nemhogy nem zárja ki, hanem egyenesen előtérbe tolja Searle. Bár számomra a szóban forgó logikai lehetőség túlságosan bizarrnak tűnik, sokan védhetőnek tartják, a Kínai Szoba pedig nem képes igazságot tenni ebben a vitában. Az emergentista a gondolatkísérlet egyik rejtett, a gondolatkísérlet által nem igazolt premisszáját támadja, azt, hogy ha egy mechanizmus érti a kínai szavakat, akkor egy valódi része is érti őket.

6.3.2 A dialektikus materialista válasz

A dialektikus materialista válasz szerint fizikailag lehetetlen papíron-ceruzával elvégezni azokat a számításokat, melyek egy fejlett MI-ben zajlanak, emiatt pedig a Kínai Szoba fizikailag megvalósíthatatlan. Ha egy kellően komplex fizikai rendszer implementál egy kellően fejlett algoritmust, akkor a rendszer érteni fogja az emberi nyelvet; az emberi agy épp ilyen rendszer. De Searle fizikailag képtelen olyan gyorsan számolni, hogy funkcionálisan ekvivalens legyen egy MI-vel. A Kínai Szoba tehát eleve nem is létezhet.

A dialektikus materialista válasz képviselői nem hívják magukat dialektikus materialistáknak. Ezt a címkét önhatalmúlag aggatom rájuk, mert álláspontjuk arra a marxista-leninista tanításra emlékeztet, hogy a valóság alapvető mozgástörvénye a mennyiség minőségbe való átcsapása. A Kínai Szobának nincs tudata, de ha odabent Searle *sokkal gyorsabban* mozgatná a ceruzáját, akkor már lenne. Ez legalább olyan figyelemre méltó szintlépés az anyag mozgásformái terén, mint a mozgalmi tevékenység átcsapása kommunista forradalomba. De semmiképp sem akarom azt sugallni, hogy ez a felvetés pusztán egy gúnynév segítségével cáfolható, és az alábbiakban igyekszem korrekten elemezni.[8]

A dialektikus materialista válasz prominens képviselője Daniel Dennett. Szerinte Searle gondolatkísérlete félrevezető, mert fizikailag megvalósíthatatlan:

> A sebesség *kulcsfontosságú* az intelligencia szempontjából. Ha egy lény nem ismeri fel elég gyorsan a releváns környezeti változásokat ahhoz, hogy védekezni tudjon, akkor praktikusan nem lesz intelligens, bármennyire komplex. [...]
>
> A digitális számítógépek [...] bizonyos értelemben alapvetően lineárisak. Von Neumann-gépek: architektúrájuk soros, nem párhu-

[8] Felvethetnénk, hogy a mennyiség minőségbe való átcsapásának tana valójában Hegeltől ered. A hegeli logika a lét tanának végén tárgyalja a mennyiség és minőség összefonódását a homályos határok (*vagueness*) kapcsán (Hegel 1979: 299–354). Hegel azonban nem volt materialista (ha az lett volna, nem kellett volna „feje tetejéről a talpára állítani"), a szóban forgó reakció a Kínai Szobára pedig egyértelműen materialista vagy fizikalista jellegű, így Hegel helyett a marxizmus-leninizmus nagyjaival érdemes párhuzamba állítani.

> zamos, ezért egyszerre csak egy dolgot tudnak csinálni. Manapság közhelynek számít, hogy egy von Neumann-gép – akárcsak az univerzális Turing-gép, melynek leszármazottja – elvben bármit ki tud számítani, amit számítógéppel ki lehet, mégis sok fontos számítási feladat – mindenekelőtt olyan kognitív szempontból lényeges feladatok, mint a mintafelismerés és adatbázisokban való keresés – von Neumann-gépekkel nem oldható meg észszerű időn belül, még akkor sem, ha a hardver a lehető leggyorsabban – fénysebességgel – működik és a gép alkatrészei közti távolság mikroszkopikus. Az ilyen számításokat reális időben csak olyan hardver tudja elvégezni, melynek adatfeldolgozó folyamatait masszív párhuzamosság jellemzi. (Dennett 1987: 326–327)

Dennett elismeri, hogy elvben a processzorok is végezhetnek párhuzamos számításokat (az elmúlt évtizedekben pont ebbe az irányba fejlődött a technológia). De szerinte elképzelhető, hogy az emberi elme sebességét csak biológiai úton lehet elérni:

> *Lehetséges*, hogy organikus számításra kell átváltanunk a megfelelő sebesség eléréséhez. [...] Tegyük fel például, hogy az enzimek szintjén zajló információfeldolgozás alapvetően fontos a neuronok közti interakciók szabályozásában – minden neuron egy aprócska számítógép, ami makromolekuláris alkatrészein komputálja a maga rendkívül komplex, „számításigényes" input-output függvényeit. Ekkor előfordulhat, hogy szó szerint lehetetlen egy neuron viselkedéséről olyan modellt készíteni, ami valós időben hűen utánozná a neuron információfeldolgozó folyamatait. (Dennett 1987: 328)

Hasonló álláspontot képvisel Pinker (1997: 94–95) és a Churchland-házaspár (Churchland és Churchland 1990). Churchlandék a Kínai Szobát ahhoz az abszurd érvhez hasonlítják, hogy a fény „nyilvánvalóan" nem az elektromágneses tér rezgése, mert ha rázogatunk egy mágnest, nem kezd világítani. A valóság az, hogy ha elég gyorsan ráznánk a mágnest, világítana. Hasonlóan, ha a Kínai Szoba lakója elég gyorsan (emberfeletti sebességgel) számolna, értené a kínai szavakat. Egyes matematikusok és számítástudósok is szimpatizálnak ezzel az ötlettel:

> Az extrém bonyolultság problémáját véleményem szerint nem lehet figyelmen kívül hagyni. Ez elég spekulatív kérdés, de nem tartom elképzelhetetlennek, hogy létezik az algoritmikus komplexitásnak olyan „kritikus" szintje, ami a mentális tulajdonságok megjelenéséhez szükséges. (Penrose 1999: 25)
>
> Nagyon érdekesnek találom ezt a fajta választ Searle érvére. Ha a válasz helyes, akkor a polinomiális és exponenciális komplexitás közti

> különbségnek *metafizikai* jelentősége van. Egy exponenciális méretű adattáblának, ami átmegy a Turing-teszten, nincs elméje (nincs tudata, öntudata, nem intelligens stb.), de egy polinomiális korlátú programnak, melyet ugyanazok az input/output összefüggések jellemeznek, már lehet. És pont *amiatt* lehet elméje, hogy polinomiális korlátú. (Aaronson 2011: 14)

Az is kétségtelen, hogy ha egy mai LLM struktúráját vesszük alapul, akkor a Kínai Szoba már a technológia jelenlegi állása mellett is kivitelezhetetlen. A DeepSeek R1 nevű MI például (ami a ChatGPT 4-hez hasonló képességekkel rendelkezik, de vele ellentétben nyílt forráskódú, ezért vannak róla pontos adatok) nagyjából 670 milliárd paramétert tartalmaz.[9] Ezt azt jelenti, hogy a futtatásakor minden token (két-három betűs választöredék) kiszámításához minimum 670 milliárd szorzásra van szükség. (Igazából sokkal többre, attól függően, hogy mekkora a kontextusablak, vagyis a beszélgetés korábbi részéből mennyit vesz figyelembe a program. Emellett összeadásokra és véletlenszámgenerálásra is szükség van. De ezt felejtsük el, és tekintsük csak a szorzásokat.) Ha feltesszük, hogy a Kínai Szoba lakója a DeepSeek R1-et futtatja papír és ceruza segítségével, akkor ahhoz, hogy generáljon 5 tokent (ami maximum 3-4 szónak felel meg), minimum 3350 milliárd szorzást kell elvégeznie. Ha azzal a teljesen elrugaszkodott felvetéssel élünk, hogy az illető 1 másodperc alatt végez el egy szorzást, akkor több mint 106 ezer évig tart létrehoznia egy pár szavas választ. Egy ilyen géppel viszonylag nehéz beszélgetést folytatni.

A dialektikus materialista válasz az informatikai tényeket tekintve egyáltalán nem abszurd. A Kínai Szoba színtiszta gondolatkísérlet, nem egy empirikus hipotézis. De van-e ennek bármi jelentősége? Hihető-e, hogy ha egy kínaiul nem tudó személy emberfeletti sebességgel tudna összeadni és szorozni – ha Supermant tennénk a Kínai Szobába –, akkor pusztán a számítások következtében hirtelen *érteni* kezdené a kínai szavakat, a ceruza sercegéséből egyszer csak előcsapna az értelem villáma? Számomra ez teljesen abszurd felvetésnek tűnik.

A dialektikus materialista erre azt felelheti, hogy a szóban forgó szituáció – Superman a Kínai Szobában – fizikailag lehetetlen, ezért irreleváns, hogy hihető-e. De ahhoz, hogy az ellenérv valóban aláássa Searle demarkációs tézisét, a dialektikus materialistának mégiscsak azt kell állítania, hogy egyes esetekben – konkrétan az emberi agy esetében – a gyors számítási folyamatok nyelvi megértéssel járnak együtt (vagy akár azonosak vele). Ez a felvetés pedig továbbra is abszurdnak tűnik. Abszurditását jól illusztrálja, ha összevetjük a mennyiség minőségbe való átcsapásának valós eseteivel, például a víz megfagyásával. A víz megfagyása minőségi változás, hiszen egy folyadék megszilárdul. A minőségi változás ugyanakkor mennyiségi változások függvénye: a vízmolekulák egyre lassabban mozognak, és a köztük ható elektromos erők kristályszerű struktúrát alakítanak ki. Az ilyen fázisátmenetek fizikai modellezése

[9] github.com/deepseek-ai/DeepSeek-R1

bonyolult feladat, de *konceptuális* szinten semmi érthetetlen nincs abban, ahogy a minőség mennyiségbe csap át. A Kínai Szobára adott dialektikus materialista válasz esetében viszont hiányzik az a konceptuális kapcsolat, ami az állítólagos átmenetet legalább nagyjából érthetővé tenné. Teljességgel érthetetlen, hogy számítások sebességének növekedésétől miért ébredne egyszer csak öntudatra a számítógép. Persze ez a fogalmi szakadék – az MI-tudat „nehéz problémája" – aligha fogja elbizonytalanítani a meggyőződéses dialektikus materialistát. Attól, hogy a dialektikus materialista megoldás kontraintuitív, nem lesz garantáltan hamis. Akárcsak az emergentista, a dialektikus materialista is egy olyan rejtett premisszát támad, amit Searle érve felhasznál, nem pedig igazol.

6.4. A Kínai Szoba mint intuíciódetektor

Daniel Dennett szerint a Kínai Szoba „intuíciópumpa": olyan gondolatkísérlet, ami „nem a megismerés motorja, hanem pedagógiai vagy retorikai eszköz, mellyel elérhetjük, hogy mások is úgy lássák a világot, mint mi, akik Searle-höz hasonlóan megvilágosodtunk" (Dennett 1980: 429). Az intuíciópumpák Dennett szerint nem feltétlenül károsak (előfordulhat, hogy tényleg megvilágosodáshoz vezetnek), de lehetnek teljesen félrevezetőek (és Dennett szerint ez a helyzet a Kínai Szobával).

Szerintem ez a diagnózis téves. Tapasztalatom szerint a Kínai Szoba nem ügyes propagandaeszköz, nem intuíciópumpa, hanem intuíció-*detektor*: az emberek egy részére semmilyen hatást nem tesz, a többiek pedig magától értetődőnek érzik a konklúziót. Ha így van, akkor a Kínai Szoba nem valódi érv, vagy legalábbis nem olyan érv, ami új meggyőződés felé terel minket előzetes hiteink alapján, vagy egy előzetes hit felülvizsgálatára kényszerít a hamisnak tűnő konklúzió miatt. A Kínai Szoba egyszerűen *detektálja*, valaki eleve hajlamos-e elvitatni az MI-től a nyelvi megértést. Az alábbiakban ezt a gyanúmat próbálom igazolni.

A Kínai Szobára épülő demarkációs érv, amennyire látom, nagyjából így néz ki:

(1) A Kínai Szoba lakója egy szintaktikai algoritmust futtat.

(2) A Kínai Szoba lakója nem érti a kínai szavak jelentését.

(3) Ha az MI érti a kínai szavak jelentését, akkor bármely x-re igaz, hogy ha x egy szintaktikai algoritmust futtat, akkor x érti a kínai szavak jelentését.

(4) Tehát az MI nem érti a kínai szavak jelentését.

Az érv logikailag érvényes. Az első két premissza triviálisan következik a gondolatkísérlet leírásából, a harmadik pedig azt az intuíciót fejezi ki, hogy ha az MI érti a szavak jelentését, akkor egy szintaktikai algoritmus futtatása elegendő a

nyelvi megértéshez. Az intuíciót szakszerűbben is kifejezhetnénk a szuperveniencia fogalmával: ha az MI érti a szavak jelentését, akkor a megértés szintaktikai műveleteken szuperveniál. Ekkor egy modális premisszát kellene használnunk, amiből komplikáltabban vezethető le a konklúzió. Mivel a modális verzióból is következni fog a 3. premissza fenti változata, az egyszerűség kedvéért a fenti változatot használom. A 3. premissza abból a szempontból is pontosításra szorul, hogy csak a szintaktikai algoritmusok egy speciális alosztályára, az *élethű* kínai chatbot-algoritmusokra vonatkozik, de ezt a kiegészítést az állítások tömörítése érdekében impliciten adottnak veszem.

A fenti rekonstrukcióból (ha helyes) levezethető a szintaxis és szemantika különbségének tétele, és a rekonstrukció azt is világossá teszi, hogy ez a tétel nem használható premisszaként. Ha az érv helyes, akkor (1)–(3) konjunkciójából, illetve (3) és (4) konjunkciójából is következik, hogy egy szintaktikai algoritmus futtatása nem elegendő a szavak megértéséhez, tehát a szintaxis nem azonos a szemantikával. Mivel (3) nem kihagyható ebből a levezetésből, és (3) önmagában elegendő – a vitán felül álló első két premissza közvetítésével – a konklúzió igazolásához, a szintaxis és szemantika különbsége az érv következménye, nem pedig előfeltevése.

A fenti rekonstrukcióhoz a két klasszikus ellenérv is közvetlenül kapcsolható. Az emergentista és a dialektikus materialista egyaránt a 3. premisszát tagadja, de különböző alapon. Az emergentista szerint (3) azért hamis, mert előfordulhat, hogy egy fizikai rendszer egy része (konkrétan a Kínai Szoba humán komponense) nem érti a kínai szavak jelentését, de a rendszer egésze érti, annak ellenére, hogy a rendszer egy szintaktikai algoritmust futtat. A dialektikus materialista szerint viszont (3) azért hamis, mert az MI más algoritmust futtat, mint a Kínai Szoba lakója, vagy ugyanazt az algoritmust nagyságrendekkel gyorsabban futtatja. Ha a 3. premisszában kikötnénk, hogy x pontosan ugyanazokat a számításokat végzi ugyanolyan gyorsan, mint az MI, akkor az 1. premissza kétségessé válna, mert szuperhumán képességekkel kellene felruháznunk a Kínai Szoba lakóját.

Ez tehát a Kínai Szobával kapcsolatos klasszikus patthelyzet logikai struktúrája. A patthelyzeten véleményem szerint azzal léphetünk túl, ha belátjuk, hogy az érvben valójában két eltérő értelemben szerepel az „algoritmust futtat" kifejezés, és ha ezeket elkülönítjük, akkor az érv vagy logikailag érvénytelenné válik, vagy csak azon az áron lesz logikailag érvényes, hogy körkörössé válik.

Ennek belátáshoz első körben azt állítom, hogy a Kínai Szoba lakója más módon futtatja a szóban forgó chat-algoritmust, mint egy MI. Tegyük fel, hogy a Kínai Szoba lakója Searle, az általa futtatott algoritmus pedig A. Searle tudatosan futtatja A-t. Eleve érti az emberi nyelvet, érzékeli és képes manipulálni a külvilágban található tárgyakat, érti A utasításait (csak az utasítások által manipulált kínai jeleket nem érti), és ebből a belső nézőpontból, az A-t alkotó szabályok tudatos alkalmazásával futtatja A-t. Hasonlítsuk össze ezt egy MI-vel. tegyük fel, hogy Mikrobi egy MI, ami szintén A-t futtatja és semmi-

lyen más MI-alkalmazást nem futtat. Mikrobi nyilván nem ugyanolyan módon futtatja A-t, hiszen nem rendelkezik előzetes nyelvismerettel, nem értelmezi az A-t alkotó szabályokat, miközben alkalmazza őket. Mikrobi esetében A futtatását különféle processzorok végzik. A processzorok nyers fizikai folyamatairól még a leglelkesebb MI-hívők sem feltételezik, hogy a tudatosság mélyebb, A-t megelőző szintjét alkotják, ami Searle belső perspektívájával és előzetes nyelvi kompetenciájával lenne analóg. Mikrobi kognitív tevékenysége A futásában *merül ki*, míg Searle tudatosan futtatja A-t, A-tól független kognitív képességek használatával.

Ezt a különbséget a következő terminológiával szeretném megragadni:

- *x szubjektív módon futtatja* az A algoritmust, ha x előzetes kognitív képességeit használva, tudatosan alkalmazza az A-t alkotó szabályokat.
- *x konstitutív módon futtatja* az A algoritmust, ha x kognitív tevékenységét A futása alkotja.

Mivel ez a két futtatási mód eltér, az érv 1. premisszáját módosítani kell, a 3. premisszának pedig két különböző olvasata áll rendelkezésre:

(1a) A Kínai Szoba lakója *szubjektív* módon futtat egy szintaktikai algoritmust.

(2) A Kínai Szoba lakója nem érti a kínai szavak jelentését.

(3a) Ha az MI érti a kínai szavak jelentését, akkor bármely x-re igaz, hogy ha x *szubjektív* módon futtat egy szintaktikai algoritmust, akkor x érti a kínai szavak jelentését.

(3b) Ha az MI érti a kínai szavak jelentését, akkor bármely x-re igaz, hogy ha x *konstitutív* módon futtat egy szintaktikai algoritmust, akkor x érti a kínai szavak jelentését.

(3b) igaznak tűnik, hiszen az MI konstitutív módon futtat egy szintaktikai algoritmust, kognitív képességeit (ha vannak neki) egy algoritmus implementálásának köszönheti, tehát ha az MI érti a szavak jelentését, akkor értenie kell minden mechanizmusnak, ami ugyanannak az algoritmusnak köszönheti kognitív képességeit. (3b) azonban logikailag érvénytelenné teszi az érvet: (3b)-ből nem vezethető le (1a) és (2) közvetítésével (4).

(3a) segítségével levezethető a konklúzió, (3a) viszont nem igazolható anélkül, hogy körkörössé tennénk az érvet. Tegyük fel, hogy a kínai anyanyelvű személyek nyelvi tevékenysége abban merül ki, hogy agyuk az A szintaktikai algoritmust futtatja (konstitutív módon). Ezt a lehetőséget nem zárhatjuk ki, mert akkor körkörössé tennék az érvet: eleve feltételeznénk, hogy a nyelvi megértés nem merülhet ki egy algoritmus futásában. Képzeljük el, hogy Searle, aki nem tud kínaiul, bevonul az erre rendszeresített szobába és egy szabálykönyv

segítségével futtatni kezdi (szubjektív módon) A-t. Ebben a szituációban a következő állítás hamis:

(3a*) Ha egy kínai anyanyelvű személy érti a kínai szavak jelentését, akkor bármely x-re igaz, hogy ha x *szubjektív* módon futtat egy szintaktikai algoritmust, akkor x érti a kínai szavak jelentését.

Az elképzelt helyzetben a kínai anyanyelvűek értik a kínai szavak jelentését, Searle nem érti, és Searle szubjektív módon futtatja a releváns szintaktikai algoritmust. Tehát van olyan releváns (episztemikusan lehetséges) szituáció, ahol (3a*) hamis. De akkor miért lenne (3a) garantáltan igaz? (3a) és (3a*) mindössze abban különböznek, hogy előtagjuk egy MI-re vagy egy kínai személyre hivatkozik. Mivel feltehetjük, hogy az MI és a kínai személy nyelvi tevékenysége egyaránt az A algoritmus futtatásában merül ki, ugyanazoknak a kognitív törvényszerűségeknek kell vonatkoznia rájuk. Ha az MI esetében a nyelvi megértés jelenlétéből következnie kell, hogy Searle is érti a kínai szavakat, akkor a kínai személy esetében is ugyanennek kellene következnie; ha pedig a kínai személy értheti a kínai szavakat anélkül, hogy Searle is értené őket, akkor az MI is értheti.

Az algoritmusok szubjektív és konstitutív futtatásának elkülönítésekor tehát a Kínai Szobára épülő érv vagy logikailag érvénytelen lesz (3b), vagy csak úgy tehető helyessé, ha eleve feltételezzük a konklúziót és kizárjuk a tisztán algoritmikus eredetű nyelvi megértés lehetőségét (3a).

A fenti okfejtés, szemben azzal, amit (3) elsőre sugall, már nem egyszerű materiális kondicionálisokat, hanem szükségszerűsített materiális kondicionálisokat használ, vagyis (3a) és (3a*) logikai struktúrája a következő (K = egy kínai anyanyelvű személy érti a kínai szavak jelentését, M = az MI érti a kínai szavak jelentését, P = bármely x-re igaz, hogy ha x szubjektív módon futtat egy szintaktikai algoritmust, akkor x érti a kínai szavak jelentését):

(3a) $\Box(M \rightarrow P)$

(3a*) $\Box(K \rightarrow P)$

$K \wedge \neg P$ igaz egy olyan episztemikusan lehetséges világban, ahol a nyelvi megértés egy szintaktikai algoritmus konstitutív futásának eredménye. Ha ezt az episztemikus lehetőséget ki akarjuk zárni és tagadjuk a tisztán algoritmikus eredetű értelem metafizikai lehetőségét, akkor a Kínai Szobára épülő érv körkörössé válik. Ha viszont $K \wedge \neg P$-t elismerjük metafizikai lehetőségnek, akkor (3a*) hamis, és ha (3a*) hamis, akkor (3a) is hamis, hiszen semmi alapunk azt gondolni (legalábbis a Kínai Szoba alapján), hogy egy szintaktikai algoritmus konstitutív futtatása csak akkor eredményez megértést, ha az algoritmust agyszövet (nem pedig egy digitális számítógép) futtatja.

Ez a kritika véleményem szerint hatékonyabb az emergentista, dialektikus materialista és hasonló klasszikus válaszoknál, melyek a 3. premisszát próbálják

annak eredeti formájában tagadni. A Kínai Szobával nem az a gond, hogy logikailag érvényes, de van egy intuícióháborúkat kiváltó premisszája, hanem az, hogy logikailag érvénytelen, amennyiben nem tekintjük eleve igaznak a konklúziót és tételezzük fel, hogy a nyelvi megértés nem merülhet ki egy szintaktikai algoritmus futásában.

Ez a kritika arra is magyarázatot ad, miért degenerálódik a Kínai Szobával kapcsolatos vita viszonylag gyorsan intuíciók csatájává: az érvet csak az tekintheti logikailag érvényesnek, aki eleve elhiszi a konklúziót. Mivel ezt a tényt elfedi a szubjektív és konstitutív futtatás fogalmának összemosása, az érv pedig e megkülönböztetés híján logikailag érvényesnek tűnik, a Kínai Szoba ellenfelei úgy érzik, hogy tagadniuk kell (3)-at, ehhez pedig nehezen igazolható ellenintuíciókat (emergencia, a mennyiség átcsapása minőségbe stb.) használnak fel.

6.5. Túl a Kínai Szobán

A Kínai Szoba nem az egyetlen lehetséges szemantikai demarkációs érv, és nyomokban tartalmazza egy sokkal általánosabb stratégia elemeit. Tegyük fel, hogy az emberi agy szintaktikai algoritmust futtat, vagyis az inputok tisztán fizikai különbségei alapján generál outputokat. Ha elfogadjuk azt is, hogy a szemantika túlmegy a szintaxison és mi emberek tudatosan észleljük a szavak jelentését, akkor arra következtethetünk, hogy az elménk nem azonos az agyunkkal, hiszen az agyban csak szintaktikai manipuláció zajlik. A szintaxis és szemantika különbsége a dualizmust támogatja, ha az agy puszta szintaktikai gép és a megértés élménye valós jelenség. Ha pedig a számítógép is szintaktikai gép (és miért ne lenne az), akkor máris levonhatjuk a következtetést, hogy az MI nem gondolkodik. A Kínai Szoba érv egyes feltevéseit tehát dualista módon is értelmezhetjük és a dualizmuson keresztül juthatunk el a demarkációig.

Nem állítom, hogy a Kínai Szobából *következik* ez a fajta szemantikai alapú dualizmus. A szintaxis és szemantika különbsége eleve nem következik a Kínai Szobából, mert a Kínai Szobára épülő érv vagy érvénytelen, vagy körkörös és feltételezi a szóban forgó különbséget. Még kevésbé következik a Kínai Szobából az, hogy az emberi agy puszta szintaktikai gép lenne. Pusztán annyit állítok, hogy *ha* a szintaxis különbözik a szemantikától (mint azt Searle szerint a Kínai Szoba sugallja) *és* az agyunk csak szintaktikai manipulációt végez (ami nem következik a Kínai Szobából, de logikailag kompatibilis vele), akkor egy szemantikai antinaturalista demarkációs tézist kapunk.

A szemantikai antinaturalizmus nem a Kínai Szoba következménye, de logikailag nem is mond ellent neki. Önálló hipotézis, amit megpróbálhatunk a Kínai Szobától független érvekkel igazolni. Ha ez sikerül és ha belátjuk, hogy a számítási folyamatok a jelentéssel ellentétben naturalizálhatóak, akkor a demarkációs tézist is igazoltuk.

A fejezet lezárásaként szeretnék bemutatni egy konkrét példát a szemantikai antinaturalista demarkációra Howard Robinsontól (6.5.1), majd néhány általános megállapítást teszek erről a stratégiáról (6.5.2). A következő fejezet egy, Robinsonétól eltérő szemantikai antinaturalista érvet dolgoz ki, ami a fenomenális tudatosságra épül.

6.5.1 Robinson és az arisztotelészi elme

Howard Robinson (2022) az arisztotelészi elmefilozófiához nyúl vissza annak igazolásához, hogy gondolkodásra voltaképpen *semmilyen* fizikai mechanizmus sem képes. Robinson, Dennett nyomán, „szintaktikai gép"-nek nevez minden olyan fizikai rendszert, ami az inputok tisztán fizikai jegyei alapján generál outputot. A szintaktikai gépek Robinson szerint csak az értelmes viselkedés törékeny illúzióját képesek létrehozni. A látszólag értelmes, de alapvetően mechanikus viselkedés illusztrálására Dennettől vesz át egy érzékletes példát:

> Mikor elérkezik a peterakás ideje, a kaparódarázs (*Sphex*) üreget épít és keres egy tücsköt, amit úgy csíp meg, hogy a tücsök megbénuljon, de ne pusztuljon el. Bevonszolja az üregbe, mellé rakja a petéit, lezárja az üreget, majd elrepül és soha nem tér vissza. A peték idővel kikelnek és a darázslárvák a megbénított tücsökből táplálkoznak, ami a sajátos tartósítási módszernek köszönhetően nem bomlott el. Az emberi elme számára ez a viselkedés rendkívül szervezettnek és céltudatosnak tűnik, logikus gondolkodás és tervezés benyomását kelti – egészen addig, amíg fel nem figyelünk bizonyos részletekre. A darázsnak például szokása, hogy a megbénított tücsköt odaviszi az üreghez, a bejáratnál hagyja, bemegy ellenőrizni, hogy minden rendben van-e a petékkel, majd kijön és csak ezután vonszolja be a tücsköt. Ha azonban a tücsköt néhány centivel odébb csúsztatják, mialatt bent folyik az előzetes ellenőrzés, akkor a darázs az üregből kimászva ismét visszahúzza a tücsköt a küszöbhöz – anélkül, hogy bevinné –, és újra bemegy ellenőrzést végezni. Ha a tücsköt megint odébb csúsztatják, miközben a darázs bent van, a darázs újra a küszöbhöz húzza, és ismét bemegy egy utolsó ellenőrzésre. A darázsnak soha nem jut eszébe, hogy egyből bevonszolja a tücsköt. Egy alkalommal ezt az eljárást negyvenszer ismételték meg, mindig ugyanazzal az eredménnyel. (Dennett 1984: 11, eredeti forrás Wooldridge 1963: 82)

Robinson szerint ez a fajta gépiesség lényegében azonos azzal a jelenséggel, amit az MI-kutatás „Keretezési Probléma" (*Frame Problem*) néven ismer (Haugeland 1985: 203–211). A Keretezési Probléma akkor lép fel, ha egy szituáció valamelyik elemének változásakor (például: „a robot elmozdítja a piros építőkockát") meg kell jósolni, hogy mi történik a szituáció többi elemével. Ha három doboz van

a szobában és az egyiket valaki átviszi a szoba túlsó végébe, hány doboz lesz a szobában? Számunkra triviálisan könnyű válaszolni egy ilyen kérdésre, de az MI esetében nulláról kell létrehozni a vonatkozó háttértudást.

Robinson egyetért Fodorral (2001: 39) abban, hogy a Keretezési Probléma szorosan összefügg az abdukcióval, a legjobb magyarázat elvével. Abduktív következtetést akkor hajtunk végre, ha a releváns információk alapján a legvilágosabbnak, legerősebbnek, legkézenfekvőbbnek tűnő magyarázatot választjuk ki. Ehhez nemcsak a rendelkezésre álló információk implicit összefüggéseit kell felismernünk, hanem azt is, hogy *melyik* információ releváns. Az abdukció tehát globális, holisztikus magyarázati mód (Fodor 2001: 38).[10] A dobozok számának alakulása szempontjából például releváns, hogy egy dobozt a falhoz tolunk vagy elégetünk, de nem releváns, hogy lassan vagy gyorsan toljuk a falhoz. Az üregben elhelyezett peték állapota szempontjából releváns, hogy mi történik az üregben, de nem releváns, hogy az üreg előtt elmozdult-e egy tárgy.

A Robinson-féle demarkációs érv arra a tézisre épül, hogy a fizikai rendszereket óhatatlanul a kaparódarázs viselkedéséhez hasonló rugalmatlanság jellemzi az értelem szimulálásakor, vagyis nem képesek megoldani a Keretezési Problémát. Mivel a Keretezési Probléma megoldásához abdukció kell, a gépek abdukcióra sem képesek, és mivel az abdukció az emberi gondolkodás egyik alapvető aspektusa, a gépek képtelenek gondolkodni.

Az LLM-ek esetében a „kaparódarazsas" viselkedéssel kapcsolatos alapvető hajlamot azok a kognitív tudományos spekulációk támaszthatják alá, melyek a mély neurális hálók absztrakciós kudarcait és szemantikai ingatagságát hangsúlyozzák. Robinson a neurális hálókkal kapcsolatban híresen szkeptikus Gary Marcus (2018) megállapításait idézi, de említhetnénk azt az újabb felvetést, hogy a ChatGPT-t és társait „hepehupás intelligencia" (*jagged intelligence*) jellemzi: bizonyos feladatokat ügyesen megoldanak, de a sikeresen megoldott feladatokhoz nagyon hasonló szituációkban drámaian romlik a teljesítményük (ld. pl. Gans 2026). Ennek a jelenségnek egyik kézenfekvő magyarázata az lehet, hogy ezek a rendszerek nem képesek abdukcióra, érzéketlenek egy szituáció lényegének és partikuláris struktúrájának összefüggéseire, ahogy a kaparódarázs sem veszi észre, hogy fölöslegesen ismétli ugyanazt a rutint.

Robinson a mechanizmusok „kaparódarazsas" hajlamait azzal az ógörög eredetű elmefilozófiai intuícióval köti össze, hogy az emberi elme nem lehet anyagi természetű, mert anyagtalan dolgok a tárgyai:

> Platón amellett érvelt a *Phaidrosz*ban (78b–84b), hogy a léleknek anyagtalannak kell lennie, mert érzékeny a Formákra, ezek pedig anyagtalanok. Vissza-visszatérő felvetés, hogy a gondolkodás tartalmai nem lehetnek pusztán fizikai tárgyak vagy események, mert a gondolati tartalom absztrakt tárgy, ha pedig a gondolat tartalma anyagtalan, hogyan lehet a gondolatokkal dolgozó elme anyagi? [...]

[10] Vö. Haugeland tézisét a szituációs holizmusról (46. old.).

> Arisztotelész ezt az érvet az érzékelés és gondolkodás szerveire fókuszálva dolgozta ki. Azt állította, hogy az észnek anyagtalannak kell lennie, mert ha anyagi lenne, akkor nem tudná befogadni az összes formát. Ahogy a szem, sajátos fizikai felépítése miatt, érzékeny a fényre, de a hangra nem, a fül pedig a hangra igen, de a fényre nem, hasonlóan, ha az ész egy anyagi szerv volna, akkor a fizikai tények egy korlátozott részére lenne csak érzékeny; de nem ez a helyzet, mert bármilyen anyagi tárgyról tudunk gondolkodni, ahogy nem-anyagi tárgyakról is (*De Anima* III. 4, 429a10–b9). Mivel nincs anyagi szerve, az elme tevékenységének lényegileg anyagtalannak kell lennie. [...] Ha nagyon általánosan fogalmazzuk meg ezt az antimaterialista érvet, Arisztotelész aggálya az volt, hogy egyetlen anyagi szerv sem lehet képes olyan széles spektrumban és olyan flexibilisen értelmezni dolgokat, ahogy az emberi gondolkodáshoz szükséges. Úgy vélte, az anyag súlyosan korlátozná az értelem számára felfogható tárgyak körét. (Robinson 2022)

Robinson szerint ez az arisztotelészi intuíció továbbra is érvényes, és az MI esetében a Marcus és mások által kipécézett absztrakciós kudarcokban és szemantikai ingatagságban nyilvánul meg. Ezek a jelenségek is abból erednek, hogy fizikai rendszerek csak korlátozott és tökéletlen módon képesek szimulálni a lényegileg abduktív emberi gondolkodást. Az ember azért *látja*, hogy az üreg előtti tárgy elmozdulása miatt nem kell újra ellenőrizi a darázspetéket, mert az inputjaikon mechanikus átalakításokat végző gépekkel ellentétben felfogja a szituáció globális struktúráját.

Robinson spekulatív érve ellen kifogásként hozható fel, hogy az érv nem mutat ki lényegi kapcsolatot a gépies viselkedés és a fizikai mechanizmusok között, nem tárja fel, hogy fizikai mechanizmusok *miért* képtelenek abdukcióra. A szubsztantív implicit feltevés mintha az lenne, hogy az MI csak a szavak statisztikailag kimutatható összefüggéseire érzékeny, emiatt pedig mindig lesznek olyan helyzetek, amikor az MI képtelen felismerni a releváns abduktív összefüggéseket és valami ostobaságot mond. Jó példa erre az alábbi (szó szerint idemásolt) beszélgetés a ChatGPT-vel:

> Felhasználó: Le akarom mosatni a kocsimat. A kocsimosó 50 méterre van. Gyalog menjek vagy autóval?
>
> ChatGPT: Őszintén: ilyen távolságnál teljesen felesleges autóba ülni. 50 méter kb. fél perc séta. Ha beülsz az autóba: elindítod → kopik a motor (hidegen még jobban), manőverezel → több macera, mint haszon, parkolás / beállás → ugyanannyi idő, mint lesétálni. Gyalog: gyorsabb vagy ugyanannyi idő, egyszerűbb, még egy kis mozgás is belefér :) **Kivétel lehet**, ha: nagyon rossz az idő (szakad az eső, jégeső), vagy valami miatt nehéz odasétálni (pl. sérülés). De normál esetben: **menj gyalog, egyértelműen.**

A programnak nem tűnik fel, hogy az „Érdemes-e kocsival mennie az x helyre?” kérdés teljesen más logikát követ, ha x a kocsimosó, mint ha x egy másik (statisztikailag tipikus) helyszín. A ChatGPT ugyanúgy egy merev rutin rabja, mint a kaparódarázs: érzéketlen olyan különbségekre, melyeket a szituáció egészét értelmező emberi elme azonnal észrevesz.

Azt azonban, hogy a „kaparódarazsas” viselkedés *szükségszerű* lenne az MI esetében, nem igazolja az arisztotelészi intuíció, vagy legalábbis nem világos, miért igazolná. Nem igazolják a fentihez hasonló konkrét példák sem, melyekre Marcus és a hozzá hasonló szkeptikusok hivatkoznak. Az ilyen bizonyítékok szavatossága gyakran korlátozott. A Google MI-je, a Gemini például a fenti kérdésre ezt válaszolja:

> Mivel le akarod mosatni az autódat, mindenképpen **autóval** érdemes menned. Bár az 50 méter gyalogosan mindössze 1 perces séta lenne, az autómosáshoz magát a járművet is oda kell juttatnod a helyszínre.

Lehet, hogy hamarosan a ChatGPT is hasonló választ ad majd, mert frissítik a paramétereit egy jobb betanítási adatbázis alapján. Ahhoz, hogy a gépiesség igazoltan *szükségszerű* legyen, valami olyasmit kellene bizonyítanunk, hogy az LLM-ek által használt „szemantikai vektorterekben” (ld. 24. old.) mindig lesznek olyan szituációk, melyek az LLM paraméterezése alapján nagyon hasonlónak minősülnek, de valójában (abduktív alapon) nagyon eltérőek, vagy a paraméterek alapján különböznek, de valójában hasonlóak. Lehet, hogy így van, és ha így van, akkor Arisztotelésznek volt igaza, de ezt tudtommal senkinek sem sikerült még belátnia.

Robinson érvét nem azért mutattam be, mert kikezdhetetlennek gondolom, hanem mert jól mutatja, hogyan lehet a Kínai Szobától markánsan eltérő módon érvelni a szemantikai demarkáció mellett. Bár Robinson érve is végeredményben egy MI-ellenes intuícióra épül, ez az intuíció (szerintem) sokkal világosabb, mint a Kínai Szoba által mozgósított sejtés, és utóbbival ellentétben operacionalizálható következményei vannak. Robinson érve továbbá egy általánosabb, a Kínai Szobától eltérő demarkációs stratégia példája is. A fejezet lezárásaként ezt próbálom röviden jellemezni.

6.5.2 Szemantikai antinaturalizmus

A szemantika naturalizálása a nyelvi jelentés, a nyelvi tartalom természettudományos elemzése. A naturalizált szemantika olyan elmélet a szavak jelentéséről és a gondolatok tartalmáról, mely kizárólag természettudományos fogalmakat – fizikai, kémiai, biológiai fogalmakat – és esetleg az okság fogalmát használja, pszichológiai kifejezéseket („hit”, „szándék” stb.) vagy szemantikai kifejezéseket („jelölet”, „szinonima” stb.) nem. A naturalista projektnek persze nemcsak a

szavak és dolgok közti viszony, hanem az elme belső folyamatainak naturalizálása is része. Végeredményben az agyi vagy általában fiziológiai állapotok és a szavaink által jelölt tárgyak közti viszonyt kell oksági és természettudományos fogalmak segítségével megragadnunk. Az alábbiakban az egyszerűség kedvéért elmosom a különbséget a szavak jelentésének naturalizálása és a mentális állapotok naturalizálása között; azt feltételezem, hogy ez a két vállalkozás szorosan összefügg.

Annak illusztrálására, hogy pontosan mit jelent a szemantika naturalizálása, tekintsük az alábbi primitív elméletet (Loewer 2017: 177–178):

(N) Az „F” szó (mint fizikai mintázat) akkor vonatkozik az F-ekre (pl. a „macska” akkor vonatkozik a macskákra) ha az „F” használatát az F-ek okozzák.

Mivel (N) fizikai mintázatokon és fizikai tárgyakon kívül csak oksági viszonyokra hivatkozik, naturalista hipotézisnek minősül. (Az egyszerűség kedvéért tegyük félre, hogy az okság fizikai jelenség-e vagy bizonyos fizikai jelenségek emberi interpretációja. A természettudósok rendszeresen használják az okság fogalmát, tehát naturalista szempontból (N) aligha elfogadhatatlan.)

(N) inadekvát elmélet, mert rengeteg ellenpélda található rá. Előfordulhat például, hogy valaki egy kistestű kutyát látva azt mondja, „Ott egy macska”. A „macska” szó ez esetben a macskákra vonatkozik (az illető azt állította egy x-ről, hogy x macska), de (N) szerint a szó ekkor vagy nem a macskákra vonatkozik (mert a használatát egy kutya okozta), vagy a „macska” extenziójába a kistestű kutyák is beletartoznak.

(N) sikertelensége nem jelenti azt, hogy a naturalizált szemantika általában reménytelen lenne. (N) csak illusztráció arra, hogy milyen formát ölthet a naturalizált szemantika és milyen problémákkal szembesülhet. Tegyük fel azonban, a vita kedvéért, hogy a szemantika naturalizálhatatlan, mert minden változata a fentihez hasonló ellenpéldákba ütközik. Ekkor az alábbi demarkációs érvet konstruálhatjuk:

(5) A szemantika naturalizálhatatlan.

(6) Az MI összes állapota naturalizálható.

(7) Ha x gondolkodik, akkor x egyes állapotainak szemantikai tartalma van.

(8) Ha x egy M állapota naturalizálható és M-nek szemantikai tartalma van, akkor a szemantika naturalizálható.

(9) Tehát az MI nem gondolkodik.

Az érv logikailag érvényes. A 8. állítás tisztán *a priori* igazság vagy terminológiai konvenció. A 7. állítást (vagy annak valamilyen parafrázisát) mindenki

elfogadja, aki egyetért azzal, hogy vannak gondolataink. A 6. állítás megalapozottnak tűnik, mert az MI számítógépen futó program, a számítógépek pedig tisztán fizikai rendszerek, melyek működését természettudományos fogalmak segítségével kimerítően le lehet írni. Az érv kulcsa tehát az 5. premissza, a szemantikai antinaturalizmus.

A szemantikai antinaturalizmus a dualizmus egy formája, hiszen ha az értelem nincs jelen a fizikai folyamatokban, akkor nem lehetünk tisztán fizikai lények. Itt tehát merészebb vállalkozásról van szó, mint a Kínai Szoba esetében, amiből nem következik egyenesen a dualizmus. Robinson esetében ez nyíltan megjelenik az 5. premisszát támogató arisztotelészi intuíció formájában. De a szemantikai demarkációt máshogy, Arisztotelésztől függetlenül is megpróbálhatjuk igazolni. Bármilyen érv, ami az 5. premisszát igazolja, szemantikai demarkációhoz vezet. Az 5. premissza legkézenfekvőbb igazolása a szemantikai naturalista elméletek átfogó kritikája lenne. Ha kimutatjuk, hogy az összes rendelkezésre álló szemantikai naturalista elmélet (N)-hez hasonlóan inadekvát, akkor jó okunk van kételkedni (5)-ben. A szemantikai naturalizmus átfogó kritikájára azonban itt nem vállalkozom, mert túl sok helyet igényelne.[11]

Felvethetnénk, hogy a 6. premissza körkörössé teszi az érvet, mert nem kizárható, hogy az MI egyes állapotai naturalizálhatatlanok Ha igaz a dualizmus, akkor nyilván léteznek pszichofizikai törvények, melyek az agyi állapotainkat összekötik az anyagtalan gondolatainkkal. Honnan tudjuk, hogy ugyanezekből a törvényekből nem következik, hogy egy kellően komplex MI-hez is kapcsolódnak majd anyagtalan gondolatok, vagy egy ezeket hordozó mentális szubsztancia? Én személy szerint nem érzem meggyőzőnek ezt a felvetést, mert a számítógépek működéséről nagyon pontos természettudományos modelljeink vannak (sokkal pontosabbak, mint az emberi agy működéséről), és teljesen *ad hoc*nak tűnik az a hipotézis, hogy a processzorok elektromágneses folyamatain túl valami más is jelen van a számítógépben, mikor az MI ⌜gondolkodik⌝. De nem gondolom, hogy ezzel rövidre zártam a vitát. Kicsit bővebben tárgyalom ezt a kérdést a következő fejezetben, mely a fenomenális tudatra építve dolgoz ki egy szemantikai demarkációs érvet.

[11] A szemantikai naturalizmus kritikájához remek kiindulópont Loewer (2017).

7.

Fenomenális demarkáció

Az érzékszerveink által közvetített tapasztalatok végtelenül sokszínűek és jellegzetesek. Kávét ízlelni, Mozart c-moll miséjét hallgatni, a naplementét nézni sajátos élmény, melyet leírhatunk szavakkal, de szavaink nem sokat mondanak annak, aki soha nem érzett ízeket, hallott zenét, vagy látta az eget. Az elmefilozófiában *fenomenális tudatosság*nak („phenomenal consciousness" vagy egyszerűen „consciousness") nevezik azt a fajta mentális állapotot, melyet a fentiekhez hasonló, közvetlenül tapasztalt minőségek jellemeznek. Ha M fenomenálisan tudatos állapot, akkor M-ben lenni valamilyen; ha az x lény rendelkezik fenomenális tudattal, akkor x-nek lenni valamilyen. Denevérnek lenni valamilyen, kávé ízét érezni valamilyen. Asztalnak lenni semmilyen, röntgensugarakat elnyelni semmilyen.

A fenomenális tudat az emberi elme egyik legfontosabb, alapvető sajátossága. Főbb formái az érzékszervi tapasztalatok (ízek, illatok, hangok, látvány, tapintás), a testi érzetek (fájdalom, testtudat), valamint az érzelmek, hangulatok. Ez a fejezet azt vizsgálja, hogy alkalmas-e a fenomenális tudatosság kognitív demarkációra.

A legtöbben egyetértenek abban, hogy az MI-nek nincsenek fenomenális állapotai; az ezzel kapcsolatos érvekre hamarosan rátérek. Előtte azonban érdemes kifejteni, hogy a fenomenológiai alapú kognitív demarkáció miért nem triviális vállalkozás még akkor sem, ha elfogadjuk, hogy az MI nem fenomenálisan tudatos. *Ha* igaz, hogy az MI-nek nincsenek fenomenális állapotai, akkor az MI tudata (amennyiben létezik) nyilván radikálisan eltér az emberi tudattól. Ha az MI tudatos, de nem rendelkezik fenomenális tudattal, akkor nincs közvetlen tapasztalata a külvilágról, nem érzékeli saját fizikai határait, nincsenek érzelmei, nincs hangulata. Egy ilyen lény teljesen más, mint mi. De a fenomenális tapasztalat hiányából önmagában nem következik, hogy az MI nem *gondolkodik*. Elképzelhető, hogy nem érez semmit, de vannak intellektuális tapasztalatai, felfog dolgokat, tudatosan következtet, érti, amit mond. Tehát a fenomenális tudat hiánya nem ad alapot kognitív demarkációra, vagy legalábbis nem világos, miért adna alapot rá. Ennek a fejezetnek az a célja, hogy megmutassa, hogyan

függhet össze a két jelenség, hogyan vezethet a fenomenális demarkáció kognitív demarkációhoz.

7.1. Vannak-e az MI-nek érzései?

Manapság a szakértők és laikusok többsége egyetért abban, hogy az MI-nek nincsenek, és a jelenlegihez hasonló technológiai feltételek mellett nem is lesznek, érzékszervi tapasztalatai, testi érzetei, érzelmei. Az MI-nek nem fáj, ha belerúgnak, nem látja vagy hallja környezetét, nem örül vagy szomorkodik. Még akkor sem, ha esetleg erre utaló kijelentéseket tesz.

A közvélekedés persze nem jelent igazolást, de a közvélekedést nem nehéz racionálisan megtámogatni. Egyrészt a jelek szerint nem igazán vannak pozitív érvek amellett, hogy az MI fenomenálisan tudatos volna. Chalmers (2023) áttekintése alapján a „pro" kosárban csak pár soványnak tűnő érv található („néha azt mondja magáról, hogy tudatos", „tudatosnak tűnik", „képes beszélgetni", „ha még ügyesebb lesz, talán érzései is lesznek"), az ellenkező oldalon – az MI-tudatosság tagadóinak oldalán – viszont egy sor szubsztantívnak tűnő érv merül fel olyan kérdések kapcsán, mint az érzékszervek és a megtestesültség fontossága, a világmodellek szerepe, az önreflexió és a visszacsatolás. (Hasonló alapon érvel az MI fenomenális állapotai ellen Seth (2025).) Egy 2023-as kutatási jelentés az MI fenomenális tudatosságáról, amit kognitív tudósok, filozófusok és számítógéptudósok csapata állított össze, expliciten csak funkcionalista alapon foglalkozik az érzékeléssel, a szubsztantív értelemben vett fenomenális tartalmak lehetőségét fel sem veti (Butlin et al. 2023). Az egyik vezető MI-cég, az Anthropic külön „MI jóléti" (*AI welfare*) kutatót alkalmaz annak kiderítésére, hogy az MI-nek vannak-e érzései, melyeket morális alapon figyelembe kellene venni; a pozíciót e sorok írásakor betöltő Kyle Fish szerint erre maximum 20% az esély.[1] Akadnak persze szakemberek, akik az MI-t érző lénynek gondolják. A leghíresebb valószínűleg Blake Lamoine, a Google egykori tesztelője, aki 2022-ben azzal kürtölte tele a médiát, hogy a Google akkori LLM-je öntudatra ébredt és valódi érzései vannak.[2] De ez erősen kisebbségi véleménynek tűnik. A megfigyelők többségének az a robusztus intuíciója, hogy az MI nem rendelkezik fenomenális állapotokkal.

Az érző MI ellen szisztematikusan érvelhetünk a fenomenális tudatossággal kapcsolatos antifizikalista érvek segítségével. Egyesek elmefilozófusok szerint a fenomenális tapasztalat cáfolja, hogy az elme puszta fizikai jelenség lenne. Két klasszikus érv létezik ennek igazolására.

A zombi-érvet (*Zombie Argument*) Robert Kirk (1974) dolgozta ki, de David Chalmers (1996b) tette közismertté. Elgondolható olyan világ, melynek lakói az

[1] „Evidence of Consciousness in Claude.AI: Mechanisms, Documentation, and Questions". *AI-Consciousness.Org*, 2025. december.

[2] „The Man Who Thinks A.I. Is Sentient". *Critical Opalescence*, 2024. június 3.

emberekkel fizikailag azonos hús-vér lények – van szívük, tüdejük, agyuk stb. –, de nincsenek érzéseik, az érzékszerveik használatát nem kíséri belső tapasztalat, nincs fenomenális tudatuk („filozófiai zombik"). A zombi-érv egyik lehetséges rekonstrukciója a következő:

(1) Elgondolható, hogy léteznek filozófiai zombik: olyan lények, akik fizikailag ugyanolyanok, mint mi, de mentálisan nem ugyanolyanok, mert nincsenek fenomenális állapotaik.

(2) Ha elgondolható, hogy léteznek filozófiai zombik, akkor metafizikailag lehetséges, hogy léteznek filozófiai zombik (ez a lehetőség nem mond ellent a szóban forgó jelenségek – a fenomenális tudatosság és a fizikai folyamatok – természetének).

(3) Ha az emberi elme fizikai jelenség (kizárólag fizikai folyamatok alkotják), akkor minden metafizikailag lehetséges lény, ami fizikailag ugyanolyan mi, mentálisan is ugyanolyan, mint mi.

(4) Tehát az emberi elme nem fizikai jelenség.

Az érv logikailag érvényes, a kérdés tehát az, hogy a három premissza igaz-e. A zombik nyilván elgondolhatóak, hiszen leírásuk érthető és koherens. A filozófiai zombikra továbbá definíció szerint igaz, hogy fizikailag azonosak velünk, mentálisan viszont nem. Tehát az 1. premissza igaz. A 3. premissza is nehezen kikezdhető, hiszen ha az emberi tudat az idegrendszerben zajló biokémiai folyamatokkal lenne azonos, akkor az ilyen folyamatoknak garantálnia kellene a tudat jelenlétét is. A zombi-érv tehát a 2. premisszán áll vagy bukik. Erre mindjárt visszatérek.

A tudás-érv (*Knowledge Argument*) legismertebb mai formája Frank Jacksontól (1982) ered és egy képzelt jövőbeli neurológusról, Maryről szól, aki mindent tud a látás biofizikájáról, de sosem látott kromatikus színeket, csak szürkeárnyalatokat. Az érvnek vannak azonban sokkal egyszerűbb és könnyebben érthető alternatív változatai. Bertrand Russell így fogalmazta meg a szóban forgó intuíciót:

> Nyilvánvaló, hogy egy látó ember rendelkezik olyan tudással, amivel egy vak ember nem; de egy vak ismerheti a teljes fizikát. Tehát az a tudás, amivel a látók rendelkeznek, de a vakok nem, nem része a fizikának. (Russell 1927: 389)

Howard Robinson egy siket halláskutató példájával illusztrálja ugyanezt a problémát:

> Képzeljünk el egy siket tudóst, aki a hallás neurológiájának legnagyobb szakértője lett. A jelenleginél fejlettebb neurológiát feltételezve elképzelhetjük, hogy az illető mindent tud a hallás fizikai

> folyamatairól, amit elvben tudni lehet, a dobhártyától elindulva az agyig. Intuitíve nyilvánvaló, hogy ekkor is lenne valami, amit ez az ember nem tudna, nevezetesen azt, hogy *milyen hallani*. A problémát egy képzeletbeli marslakón is demonstrálhatjuk, akiről kiderült, hogy rendelkezik egy számunkra ismeretlen hatodik érzékszervvel. Hiába kezdenénk vizsgálni a marslakót és térképeznénk fel tetszőleges alapossággal az idegrendszerét, fogalmunk sem lenne, milyen élményeket okoz az adott szerv (hacsak nem utalnának jelek arra, hogy a mi egyik érzékszervünkre hasonlít). Bármennyit is tudunk a szóban forgó fizikai folyamatról, ez semmit nem mond arról, *milyenek* a vonatkozó érzéki élmények. Ez a *milyenség* – az, ahogyan az érzékszerv működése a szubjektum számára adódik –, ez a szubjektív mozzanat jelenti a materialista számára az alapvető problémát. A [dualista] az ilyen szubjektív állapotokat, például a fájdalmat vagy a piros szín látványát, sajátos mentális tulajdonságoknak tekinti. A materialista viszont nem állíthatja ezt és máshogy kell értelmeznie a szubjektivitást. (Robinson 1982: 4)

A tudás-érv lényege az az intuíció, hogy az érzékszervek működéséről szerzett maximálisan részletes fizikai (biokémiai, élettani) tudásból hiányozni fognak az érzékszervek használatával együttjáró fenomenális milyenségek – szakszóval *kvál*ék („qualia") –, melyekkel mindannyian közeli ismeretségben vagyunk, mióta az eszünket tudjuk. Ismerheti valaki a látásban vagy a hallásban szerepet játszó összes biológiai folyamatot anélkül, hogy fogalma lenne, milyen látni vagy hallani, vagyis anélkül, hogy ismerné azokat a kválékat, melyeket látó vagy halló szubjektumként tapasztalunk. Mivel az anyagi folyamatokról (elvben) szerezhető tudásunkban az összes fizikai tulajdonság szerepel, de a kválék nem feltétlenül szerepelnek benne, a kválék nem fizikai tulajdonságok. Tehát, akárcsak a zombi-érv esetében, arra jutunk, hogy a fenomenális tudatosság nem fizikai jelenség.

A zombi-érv és a tudás-érv intenzív viták tárgya.[3] A zombi-érv kritikusai általában abba kötnek bele, hogy az elgondolhatóságból nem következik a metafizikai lehetőség: el tudunk képzelni filozófiai zombikat, de a képzelet nem feltétlenül ad megbízható hozzáférést a valóság mélystruktúrájához. Azt is elképzelheti valaki, hogy a víz nem H_2O, de ebből nem következik, hogy a víz természete túlmegy azon, amit a kémikusok állítanak róla. A tudás-érv esetében a kritikusok általában azzal érvelnek, hogy a siket tudós és sorstársai nem új tudásra tesznek szert, csak új képességekre vagy fogalmakra, de olyan kritikus is akad, aki tagadja, hogy az érzékszerveink működésével kapcsolatos teljes fizikai tudásból hiányozna a kválék ismerete.

[3] Ld. Gendler & Hawthorne (2002), Kirk (2023), Ludlow, Nagasawa & Stoljar (2004), Robinson (2016). Magyarul áttekintést ad a vitáról Tőzsér (2009: 226-246).

A helyzetet tovább bonyolítja, hogy a kválék létét még egyes fizikalisták is elismerik, és redukálható vagy emergens fizikai tulajdonságoknak tekintik őket (Tye 2025). Ráadásul létezik olyan álláspont is – a russelli semleges monizmus –, amiről nehéz eldönteni, hogy fizikalista vagy antifizikalista (uo.). A kválék természetéről sincs egyetértés. Egyesek szerint a kválék intrinzikus, mások szerint relációs tulajdonságok, és vita van arról is, hogy nemfizikai partikulárék (például érzet-adatok) tulajdonságai-e, vagy magukat a fenomenális állapotokat jellemzik – egy kávéízű nemfizikai partikuláré a tudatunk tárgya/része, amikor kávéízt érzünk, vagy „kávéízes" a tapasztalatunk.

Nem vállalkozom rá, hogy ezeket a vitákat szétszálazom és a koncepciókat egyenként összekötöm az LLM-ekkel, egyrészt mert ez kicsit hosszadalmas lenne, másrészt mert nem ismerem ilyen mélységben sem az elmefilozófiát, sem az informatikát. Az alábbiakban azt szeretném röviden megvizsgálni, hogy vezethet-e a fenomenális tudat hiánya (amit az MI esetében a legtöbben elismernek) a gondolkodás hiányához (ami az MI esetében nyitottabb kérdésnek tűnik). Az egyszerűség kedvéért ehhez a fenomenális tudat kválé-alapú, antifizikalista értelmezését használom, egyrészt mert ez jól kidolgozott álláspont, másrészt mert közismert érvek támasztják alá, így azok számára, akik a szóban forgó érveket elfogadják, a kválékból eredő kognitív demarkációs érv különösen meggyőző lehet. Célom tehát egy kondicionális állítás igazolása: *ha* vannak nemfizikai kválék, akkor az MI nem gondolkodik. (Valójában egy fokkal gyengébb állítást védek: *ha* vannak nemfizikai kválék, akkor az MI nem érti a legtöbb dolgot, amit mond.) Nem állítom, de nem is zárom ki, hogy ez a tézis fizikalista kválék esetén is fenntartható, de a továbbiakban a „kválé" nemfizikai kválékra vonatkozik majd.

A fenomenális tudatosságtól a dualizmusig vezető érvek az MI kontextusában is relevánsak. Ha a számítógépek tisztán fizikai rendszerek, a fenomenális állapotok pedig nem fizikai állapotok, akkor a számítógépeknek nincsenek fenomenális állapotai. Bármennyire élethű az MI robotkutya, nem fáj neki, ha belerúgnak; bármennyire emberszerű egy MI chatbot, nem tudja, milyen a kávé íze vagy az ég látványa. Még akkor sem tudja, ha – mert belehuzalozták – azt ⌜állítja⌝, hogy tudja. És még akkor sem tudja, ha folyadékdetektort kötünk rá és kávét öntünk bele, vagy kamerát szerelünk rá, amit az égboltra irányítunk. Ahhoz ugyanis, hogy az MI tudja, milyen a kávé íze vagy az égbolt látványa, kválékra lenne szükség, de tisztán anyagi rendszerek – a zombi-érv és a tudás-érv fényében – nem rendelkezhetnek kválékkal. Nevezzük ezt *fenomenális demarkáció*nak.

Felvethetnénk, hogy a fenomenális demarkáció még akkor sem sikeres, ha elfogadjuk a zombi-érv vagy a tudás-érv összes premisszáját. Ezek az érvek ugyanis nem zárják ki, hogy egyes fizikai rendszereket törvényszerűen nemfizikai kválék kísérnek. Sőt, egyenesen következik belőlük, hogy ez a helyzet, hiszen az emberi agy pont ilyen rendszer. Ha tehát a fenomenális tudat nemfizikai jelenség, akkor léteznek pszichofizikai törvények, melyek a fizikai rendszereket

nemfizikai mentális állapotokkal kötik össze. Emiatt pedig nem kizárható, hogy egy kellően komplex MI fizikai folyamatait ugyanúgy nemfizikai fenomenális állapotok kísérik, mint a mi agyi állapotainkat. Ezt a lehetőséget támogatja Chalmers (1996b: 254) egyik gondolatkísérlete, melyben egy ember neuronjait fokozatosan lecserélik funkcionálisan ekvivalens chipekre. Ha az illetőnek nem huny ki a folyamat közben a tudata, akkor lényegében egy fenomenálisan tudatos MI lesz a végeredmény: egy ránézésre embernek tűnő lény, akinek a koponyájában egy számítógép futtat valamilyen algoritmust. Ha ezt a számítógépet leválasztják a testről és megfelelő inputokat adnak digitális idegvégződéseinek, akkor a lény továbbra is tudatos marad, de külsőleg sem tűnik majd embernek. Tehát elgondolható, hogy egy számítógépes program futtatását pontosan ugyanolyan kválék kísérik, mint esetünkben az agyi folyamatokat. Ráadásul ez a szituáció nemcsak elgondolható, hanem akár technológiailag lehetségessé válhat a jövőben.

Nem egyértelmű azonban, hogy ez a gondolatkísérlet aláásná a fenomenális demarkációt. Ahhoz, hogy aláássa, több súlyos feltevést kell tennünk, például azt, hogy a pszichofizikai törvények – melyek meghatározzák, mely fizikai folyamatokat kísérnek kválék – pusztán a funkcionális tulajdonságokra érzékenyek, a fizikai szubsztrátum anyagára (pl. hogy szilikon vagy idegszövet implementál egy funkciót) nem. Ha azonban a pszichofizikai törvények nem csak funkcionális tulajdonságokra érzékenyek, akkor nincs okunk azt gondolni, hogy a neuronok chipekre cserélésekor megmarad a tudatosság. Ráadásul ahhoz, hogy az MI-nek kválékat tulajdonítsunk, nem elég belátni, hogy a pszichofizikai törvények funkcionalista jellegűek. Azonosítanunk kell egy F funkcionális tulajdonságot is, ami elégséges a fenomenális tudat jelenlétéhez, továbbá igazolnunk kell, hogy az MI architektúrája rendelkezik F-el. Ez a projekt tudtommal nem tart abban a fázisban, hogy az MI-kválék lehetőségét komolyan kellene vennünk. Kétségtelen, hogy a tudás-érv és a zombi-érv közvetlenül nem igazolja az MI-kválék hiányát, de ráirányítja a figyelmet erre a problémára és ezáltal megalapozottabbá teszi azt az erős intuíciót, hogy az MI-nek nincsenek fenomenális állapotai.

A fejezet további része arra keresi a választ, vezethet-e út a fenomenális demarkációtól a kognitív demarkációig. Két felvetést vizsgálok majd. Az egyik az elmefilozófia kissé ezoterikus témájához, a kognitív fenomenológiához kapcsolódik. Egyes elmefilozófusok szerint gondolkodni ugyanúgy sajátos élmény, mint a kávé ízét érezni – a gondolkodásnak, sőt, az egyes gondolatainknak is megvan a saját „milyensége”, vagyis léteznek kognitív kválék. Következik-e ebből, hogy az MI nem gondolkodik?

A fejezet végén egy ettől eltérő kválé-alapú kognitív demarkációs stratégiát ismertetek, ami a szemantikai antinaturalista stratégia (6.5.2. alfejezet) egyik változata. Arra a tézisre épül, hogy sok emberi fogalom megértéséhez kválék ismeretére van szükség.

7.2. Kognitív fenomenológia

Chudnoff (2015: 1) az alábbi példákkal illusztrálja, hogy a gondolkodás, az érzékeléshez hasonlóan, sajátos fenomenális élményekkel jár:

> **Megértés**: Próbálod elolvasni a használati utasítást egy gyógyszeren, amit az állatorvos írt fel a kutyádnak. A szöveg először olvashatatlannak tűnik, de végül látod, hogy azt írja, egy hétig naponta kétszer kell beadni a gyógyszert.
>
> **Felfogás**: Egy könyvben azt olvasod: „Ha $a < 1$, akkor $2 - 2a > 0$". Eltöprengsz, tényleg így van-e. Ráébredsz, hogy ha a kisebb 1-nél, akkor $2a$ kisebb 2-nél, tehát $2 - 2a$ nagyobb, mint 0.

Mindkét szituációban az történik, hogy először nem értünk valamit, aztán megértjük. Ez a változás sajátos *érzés*, pontosabban minden ilyen változás valamilyen sajátos érzés. Más érzés összefolyó, apró betűket látni, mint felfogni a mondatot, amit alkotnak. Más érzés leírva látni egy matematikai állítást, mint felfogni, miért igaz. És az ilyen érzések karaktere a szóban forgó állítások tartalmával együtt változik.

A kognitív fenomenológia hívei szerint a szóban forgó érzések fenomenális állapotok. Gondolkodni is *valamilyen*, ugyanolyan – kválék jelenlétét jelző – értelemben, mint az érzékszervi tapasztalat. A gondolkodás nem csupán reprezentál tárgyakat és lehetőségeket (nemcsak intencionális tartalma van), hanem az érzékeléshez hasonlóan tudatosan megélt, kvalitatív állapotok jellemzik, melyeket a gondolkodó szubjektum közvetlenül tapasztal.

A kognitív fenomenológiát legalább kétféleképpen jellemezhetjük. Gondolhatunk rá úgy, mint egyfajta belső beszédre vagy belső képek váltakozására, vagy tekinthetjük érzékszervi felhangok nélküli, „nemszimbolizált" élménynek. Ez utóbbit egy nagyon érdekes pszichológia kutatás, a Descriptive Experience Sampling (DES) segít megvilágítani, melyet Végh Júlia (2024a) cikke alapján ismertetek.

A DES keretében pszichológusok arra kérték a kísérleti alanyokat, hogy 24 órán át hordjanak maguknál egy csipogót, ami véletlenszerű időpontokban megszólalt. A résztvevőknek fel kellett jegyeznie, hogy a csipogás pillanatában milyen mentális állapotban voltak, a kutatók pedig a noteszek tartalma és utólagos értelmező interjúk alapján megpróbálták típusokba sorolni az állapotokat. Az egyik vizsgálatnál (Heavey & Hulburt 2008) a mentális állapotok 97%-a az alábbi öt kategória valamelyikébe tartozott: (i) belső beszéd, (ii) belső kép, (iii) testi érzet, (iv) érzékszervi tapasztalat, (v) „nemszimbolizált gondolkodás". Az első négy vélhetőleg nem igényel bővebb magyarázatot, a nemszimbolizált gondolkodás fogalma viszont kevésbé ismerős. Az élménynaplóról szóló interjún az egyik résztvevő így írta le a szóban forgó jelenséget:

> Evelyn [résztvevő]: A kanapén ültem, tévét néztem. A tévében a NetZero reklámja ment. Hallgattam a NetZero-reklámot, és azon

gondolkodtam, vajon mennyivel olcsóbb, mint a Cox Cable. És akkor megszólalt a csipogó. És... Amennyire meg tudom ítélni, a tudatomban az volt, hogy kezemben a kávéscsésze, és tudja, tűnődöm, azon, hogy a Cox... hogy mennyivel olcsóbb ez a NetZero, mint a Cox Cable. És megszólalt a csipogó.

Sharon [interjúztató]: És ez a „Vajon mennyivel olcsóbb a NetZero, mint a Cable", ez volt abban a pillanatban a tudatában?

E: M-hmm.

S: Szavak formájában? Vagy nem szavakban? Mintha kimondaná? Vagy gondolta?

E: Csak magamban gondolkodtam, hogy, tudja, tényleg olcsóbb-e.

S: És ez a „tényleg olcsóbb-e"... Ezt most szavakként kell gondolni, vagy nem szavakként, vagy képekben, vagy... Hogyan jelenik meg ez a gondolat önnek abban a pillanatban?

E: [tanácstalannak tűnik: felfordítja tenyerét, felvonja szemöldökét, hanghordozása bizonytalan] Asszem csak... csak gondolkodom rajta. Nem képekben gondolkodom vagy... Csak magamban, hogy vajon tényleg sokkal olcsóbb lenne-e. Mert folyton ilyen reklámokkal bombáznak.

S: És, hm, fogja közben a csészét.

E: M-hmm.

S: Ennek tudatában van, vagy ez egyszerűen csak a világ egyik ténye – fogja, de nem figyel rá, hogy fogja?

E: Mindig figyelek arra a csészére, mert kristályból van, és általában csak hétvégén használom, szeretem azt az érzést – elég nehéz.

S: És abban a pillanatban is tudatában van, hogy nehéz, vagy...?

E: [ismét tanácstalannak tűnik] Úgy érzem, csak a gondolat, hogy Cox Cable vagy NetZero... csak az volt a tudatomban. (Hurlburt & Akhter 2008: 1365, Végh Julianna fordítása alapján)

A nemszimbolizált gondolkodás tehát nem belső monológ, hanem nehezen leírható, de sajátos érzés: az ember egyszerűen *szemlél* egy tényt vagy összefüggést, jelen esetben a két internetszolgáltató közti árkülönbséget.

Az introspekción és pszichológiai vizsgálatokon kívül szisztematikus érvek is támogatják a kognitív fenomenológia (KF) létét. Pitt (2004) szerint azért kell elfogadnunk a KF létét, mert a gondolatainkhoz közvetlenül hozzáférünk és ez csak úgy lehetséges, ha „érezzük", melyik gondolatunknak mi a tartalma és miben különbözik más gondolatoktól és egyéb tudatos állapotoktól, például érzékszervi tapasztalatainktól. Mivel engem a KF és a kognitív demarkáció közti

viszony érdekel, a vita kedvéért feltételezem, hogy empirikus alapon vagy Pitt érve miatt vagy egyéb okokból elfogadtuk a KF létét.[4]

Nevezzük az M mentális állapotot „intellektuális"-nak, ha van olyan P propozíció, hogy bármely x szükségszerűen akkor és csak akkor van M állapotban, ha x azt gondolja, hogy P. A KF-re épülő demarkációs érv alapformája ekkor a következő:

(5) Minden M mentális állapotra igaz, hogy ha M intellektuális, akkor M fenomenális is. (kognitív fenomenológiai alaptézis)

(6) Ha x gondolkodik, akkor x-nek van intellektuális mentális állapota.

(7) Az MI-nek nincsenek fenomenális állapotai.

(8) Tehát az MI nem gondolkodik.

Az érv logikailag érvényes. Ha elfogadtuk KF létét (5) és az MI-kválékat kizártuk (7), akkor úgy tűnik, el kell fogadnunk a konklúziót, hiszen a 6. premissza definíció szerint igaz. KF tehát a kognitív demarkációhoz vezető királyi útnak tűnhet. De ez szerintem puszta látszat.

A KF elméletének legalább *négy* változatát különböztethetjük meg aszerint, hogy hogyan viszonyulnak a kognitív kválékhoz és a gondolatok tartalmához. A kognitív kválékkal kapcsolatban egyrészt lehetünk redukcionisták vagy antiredukcionisták. A *reduktív* álláspont szerint a KF az érzékszervi, testi és érzelmi, azaz a nem-kognitív fenomenológiából ered (például belső beszéd, belső képek formájában). A szembenálló, *nemreduktív* tábor szerint van specifikusan a gondolkodásra jellemző fenomenológia is, amire talán a nemszimbolizált gondolkodás lehet példa.[5]

A KF egyes hívei szerint gondolataink milyensége teljes egészében kimeríti azt, amit gondolunk, vagyis a gondolatok tartalma a kognitív kválékkal azonos vagy ezekre redukálható. Ezt „tisztán szenzoros" álláspontnak hívom majd. Mások szerint a gondolatainknak nemcsak fenomenális, hanem szűk értelemben vett kognitív (a világ megismerésével, reprezentációjával kapcsolatos) tartalma is van. Nemcsak *valamilyen* azt gondolni, hogy P, hanem egy tény vagy lehetőség is reprezentációja is, és a szóban forgó reprezentációs tartalom is a tudatunk előtt van, mikor arra gondolunk, hogy P.[6]

[4] A Pitt érvével kapcsolatos vitát bővebben elemzi Chudnoff (2015: 1. fejezet), magyarul pedig Végh (2020b).

[5] Mint Végh (2024b: 147) megjegyzi, a kétféle (reduktív ill. nemreduktív) megközelítésnek nincs hivatalos neve: Prinz (2011) restriktivizmusól és expanzionizmusról, Siewert (2011) inkluzivizmusról és exkluzivizmusról, Robinson (2011) fukar és liberális álláspontokról beszél; Pitt (2004) sajátosnak (*proprietary*), Chudnoff (2015) pedig redukálhatatlannak nevezi a szenzoros fenomenológiára vissza nem vezethető KF-et.

[6] Ld. Chudnoff (2015: 48). A helyzetet még tovább bonyolítja, hogy a tisztán szenzoros koncepció hívei is beszélnek szemantikai tartalomról, ami szerintük szenzoros kválékból ered vagy azoknak a tulajdonsága (uo., ill. ld. Hawley & Macpherson 2011).

Ha a KF részben kognitív, akkor a fenti érv burkoltan érvénytelen lehet, mert elképzelhető, hogy (5) csak az emberek mentális állapotaira igaz, az MI-nek pedig anélkül vannak intellektuális állapotai, hogy kváléi lennének. Ha ugyanis a gondolatainkat csak részben alkotják kválék és van másfajta (propozicionális, szűken vett kognitív) tartalmuk is, felmerül a gyanú, hogy a gondolkodás (mint tények, lehetőségek megértése, elemzése, összehasonlítása) a szóban forgó *sui generis* kognitív tartalmakon végzett művelet és szorosan véve nem a kválékhoz van köze. Ha ez a helyzet, akkor logikailag lehetséges, hogy az MI is gondolkodik, csak belső beszéd vagy a fenomenális tapasztalat egyéb formái nélkül; olyan mentális állapotok révén, melyeknek ugyanaz a kognitív tartalma, mint a mi gondolatainknak, de nem csatlakoznak hozzájuk kválék. Ez aláássa a kognitív fenomenológiai demarkációt.

A redukálhatatlan KF hasonló problémához vezet, főleg, ha a részben kognitív koncepcióval kombináljuk. Ha a gondolkodás kváléi nem redukálhatóak szenzoros kválékra (ha a KF nem belső beszéd, belső képek sorozata vagy általában az érzékszervi vagy testi fenomenológiából táplálkozó élmény), akkor vannak specifikusan a megértést, spekulációt, következtetést stb. jellemző vagy alkotó kválék. Ez az elmélet szerintem a 7. premisszát gyengíti, mert azok az érvek, melyek a dualisták szerint igazolják a nemfizikai kválék létét, egyáltalán nem vagy nehezen alkalmazhatóak a gondolkodás kváléira. Nehéz lenne Robinson halláskutatójának mintájára elképzelni egy olyan szakembert, aki mindent *tud* a gondolkodás fiziológiájáról, de sose tapasztalta meg, milyen *gondolkodni*. A zombi-érvet könnyebbnek tűnhet áthangszerelni a kognitív fenomenológiára, hiszen ha a zombi fejében teljes sötétség van, akkor vélhetőleg gondolatai sincsenek. Ha azonban részben kognitív redukálhatatlan KF-et feltételezünk, akkor a zombi-érv is bajba kerülhet, mert nem világos, hogy a gondolkodás kváléinak hiányából miért következne a gondolkodás hiánya. Lehet, hogy a zombi is gondolkodik, de belső beszéd vagy az érzéki tapasztalat egyéb formái nélkül, a szűken vett kognitív tartalmak révén. Ezzel kapcsolatban szerintem bizonytalanok az intuícióink, emiatt pedig a részben kognitív redukálhatatlan KF a zombi-érvet is meggyengíti.

Felvethetnénk, hogy a zombi-érv és a tudás-érv helyett használhatjuk a Kínai Szobát annak igazolására, hogy fizikai mechanizmusokból hiányzik a megértés sajátos élménye, vagyis a redukálhatatlan kognitív fenomenológia. De a Kínai Szoba érv körkörös (6.4. alfejezet), emiatt pedig nem alkalmas demarkációs célra.

A KF-alapú demarkációra tehát a redukálhatatlan és a részben kognitív KF is veszélyt jelent, a kettő kombinációja pedig különösen. A részben kognitív KF az 5. premisszát ássa alá, a redukálhatatlan KF pedig a 7. premisszát (különösen akkor, ha KF részben kognitív is). A fenti érv tehát csak akkor áll biztos lábakon, ha tisztán szenzoros redukálható KF-et feltételezünk, vagyis azt állítjuk, hogy a gondolataink tartalmát kimeríti fenomenális tartalmuk. Ez a tézis viszont véleményem szerint nehezen hihető, hiszen ekkor azt állítjuk, hogy

gondolataink *tartalma* is szenzoros kválékra redukálható. Nyilvánvalónak tűnik, hogy rengeteg olyan fogalmunk és gondolatunk van, ami túlmegy az érzékszervi, testi és érzelmi kválékon. A DES fent idézett részvevőjének az a nemszimbolizált gondolata, hogy a reklámban szereplő netszolgáltató talán olcsóbb, mint a riválisszolgáltató, nem tűnik szenzoros kválékkal kapcsolatos állításnak, ahogy a tudományos állítások többsége sem.

A tisztán szenzoros *redukálhatatlan* KF esetében ez a probléma nem merül fel, felmerül viszont az a kérdés, hogy a szóban forgó redukálhatatlan fenomenológia, amit nem kísér *sui generis* kognitív tartalom, voltaképpen *micsoda* és mi köze van azokhoz a propozíciókhoz, melyeket hagyományosan a gondolatok tartalmainak tekintünk. Annak a gondolatnak a sajátos *É* érzete, hogy $\forall x : x = x$, milyen összefüggésben van azzal a P állítással, hogy minden azonos önmagával? Jelen lehet *É* anélkül, hogy az *É*-t tapasztaló személy hajlamos lenne elfogadni P-t? Ha igen, akkor mi köze *É*-nek a gondolkodáshoz? Ha nem, akkor pontosan mi teremt *É* és P között ennyire intim kapcsolatot annak ellenére, hogy *É*-nek egy érzetminőségen nincs semmi tartalma?

A tisztán szenzoros redukálhatatlan KF híveinek erre vélhetőleg van valamiféle válasza, de a fenti dialektika számomra azt jelzi, hogy a KF-alapú demarkáció nem működik a kognitív kválékkal kapcsolatos erős szűkítő feltevések nélkül, utóbbiak viszont olyan mélyre vezetnek az elmefilozófia dzsungelébe, hogy ha az így keletkező demarkációs érv netán helyes is, nagyon nehéz érthetővé és meggyőzővé tenni.

7.3. Szemantikai fenomenalizmus

A fejezet lezárásaként a kognitív demarkáció egy olyan változatát szeretném kidolgozni, ami nem a kognitív fenomenológiát, hanem a hagyományos kválékat veszi alapul: az érzékszervi tapasztalatot, a testérzetet és az érzelmeket, hangulatokat. Ezeket az egyszerűség kedvéért „szenzoros fenomenológiának" hívom majd. Kiindulásként feltételezem, hogy a szenzoros fenomenológia, a zombi-érv vagy a tudás-érv fényében, nemfizikai kválékat feltételez. Nem állítom, hogy ez a helyzet; arra vagyok kíváncsi, hova vezet ez a tézis a kognitív demarkáció terepén. Az esztétikai megértésre fogok koncentrálni, de később általánosítom az érvet az összes szenzoros kváléra.

A szemantikai fenomenalizmusra épülő demarkáció alapgondolata az, hogy egyes fogalmaink *tartalma* a szenzoros fenomenológiában gyökerezik abban az értelemben, hogy ha x-nek nincsenek és soha nem is voltak szenzoros élményei, akkor x nem rendelkezhet az adott fogalommal. Ez a helyzet például az esztétikai fogalmakkal. Akinek nincsenek és soha nem is voltak érzéki élményei, nem tud esztétikai ítéleteket alkotni, vagy legalábbis emellett próbálok érvelni az alábbiakban. Azt a tételt, hogy a T típusú fogalmaink tartalma a szenzoros fenomenológiában gyökerezik, *T-szemantikai fenomenalizmus*nak hívom. Ennek

a tézisnek hangsúlyozottan *nem* része, hogy a T típusú fogalmak tartalma kválékra *redukálható*. Nem állítom például, hogy a szimfónia fogalmát definiálni lehet auditív kválékra hivatkozva. A szemantikai fenomenalizmus csak annyit állít, hogy a kválék *részei* bizonyos fogalmak tartalmának és a szóban forgó fogalmak elsajátításának előfeltétele a kválékkal való ismeretség.

Tekintsük a következő esztétikai ítéletet:

(B) Beethoven harmadik zongoraversenye Mozart zongoraversenyeinek apoteózisa.

(B) tartalma nem redukálható auditív élményekre, mert túlmegy a szenzoros fenomenológia puszta leírásán. Már maga Beethoven 3. zongoraversenye sem azonosítható kválék halmazával, hiszen végtelenül sokféleképpen előadható. (B) valami olyasmit állít – helyesen vagy sem, most mindegy –, hogy Beethoven 3. zongoraversenye Mozart zongoraversenyeinek mintáját követi, de transzcendenssé, monumentálissá változtatja a mozarti hangzást. (B) jelentése zeneművek egész sorának bonyolult hasonlósági viszonyain alapul és közvetve arra is utal, hogy Beethoven, aki a zenetörténet talán leglelkiismeretesebb zsenije volt, magába szívta és átlényegítette közvetlen elődei, Haydn és Mozart stílusát.

Bár (B) jelentése nem redukálható kválékra, jelentése részben kválékban *gyökerezik*, hiszen azt a kérdést, hogy Beethoven 3. zongoraversenye hasonlít-e Mozart zongoraversenyeire, nyilván nem lehet elválasztani attól a kérdéstől, hogy Beethoven 3. zongoraversenye hogyan *hangzik*. (B) nem azokról a kacskaringós tintavonalakról szól, melyeket Beethoven a kottapapírra írt; nem a tintavonalak lényegítették át Mozart tintavonalait. Tehát annak ellenére, hogy (B) tartalma nem kválékban merül ki, bizonyos értelemben magában foglal kválékat, bizonyos kválék ismerete elengedhetetlen (B) megértéséhez. Egy siket tudós behatóan ismerheti a hallás fiziológiáját, de nem értheti meg teljes mélységében, hogy (B) miért igaz (vagy hamis).

Ez az eset tipikusnak tűnik az esztétikai ítéleteknél. Emiatt az esztétika-szemantikai fenomenalizmus igaznak tűnik. Más megismerési területekkel kapcsolatban analóg érvek konstruálhatóak; erre mindjárt visszatérek.

A szemantikai fenomenalizmusra épülő demarkációs érv következő premisszája az, hogy az MI-nek nincsenek kváléi, nincs szenzoros fenomenológiája. Mint arról a fejezet elején szó volt, ezt gyakorlatilag mindenki elfogadja, még az MI kognitív képességeivel kapcsolatban maximálisan optimista szakértők is. (Persze sokan azt is gondolják, hogy kválék egyáltalán nem léteznek. Ahogy jeleztem, átmenetileg azt feltételezem, hogy hiszünk a kválékban.)

A fenti két premisszából egy tisztán terminológiai jellegű axióma, (11) közvetítésével érvényes kognitív demarkációs érv áll össze:

(9) Az esztétikai fogalmak tartalmát részben kválék alkotják: x csak akkor rendelkezhet esztétikai fogalmakkal, ha x rendelkezik bizonyos kválékkal. (esztétika-szemantikai fenomenalizmus)

(10) Az MI nem rendelkezik kválékkal.

(11) Ahhoz, hogy x értse az esztétikai állításokat, x-nek rendelkeznie kell esztétikai fogalmakkal.

(12) Tehát az MI nem érti az esztétikai állításokat.

Ez a demarkációs érv sikeres, *ha* hiszünk a kválékban – mert nélkülük (9) hamis –, *és* elfogadjuk (10)-et (de ezt gyakorlatilag mindenki elfogadja manapság).

Az érv nemcsak az esztétikai megismerésre, hanem minden olyan gyakorlati és tudományos szférára vonatkozik, ahol a központi fogalmak tartalmát részben kválék alkotják. Kiterjeszthető az érv többek között az érzelmekkel kapcsolatos diskurzusra és talán a morális diskurzusra is (ha a morális indokoknak szoros köze van ahhoz, hogyan hatunk mások érzéseire). Kiterjeszthető továbbá a tárgyfelismerésre és a térbeli gondolkodásra (a mozgással és mozgatással kapcsolatos ítéletekre), amennyiben a kválék itt is szerepet kapnak. A szemantikai fenomenalizmusra épülő demarkáció tehát nem azt igazolja, hogy az MI *általában* nem gondolkodik, hanem azt, hogy *bizonyos* kérdésekről nem gondolkodik, például nem érti az esztétikai fogalmakat, az érzelmekkel kapcsolatos diskurzust, az erkölcsi és morális állításokat, a hétköznapi környezetünk manipulálásával kapcsolatos beszédet.

Bár a kognitív demarkációnak ez a formája korlátozott – csak addig terjed, ameddig a kválék befurakodnak a fogalmainkba –, fontos lehet az MI képességeivel és státuszával kapcsolatos vitákban. Ha például igaz az, hogy esztétikai fogalmak nélkül lehetetlen kreatívnak lenni, akkor a fenti érvből következik, hogy az MI sosem lesz kreatív. Ha igaz, hogy az érzelmek megértése nélkül lehetetlen sikeres pszichoterapeutának lenni, akkor a szemantikai fenomenalizmusból következni fog, hogy az MI sosem lesz sikeres pszichoterapeuta. Ha a morális elvek helyes alkalmazásához szükség van az érzésekkel kapcsolatos fogalmakra, akkor az MI sosem fog morális elveket helyesen alkalmazni. Ha a térbeli navigációhoz és tárgyfelismeréshez kválékra van szükség, akkor sohasem lesznek igazán hatékony robotmunkások. És így tovább. A szemantikai fenomenalizmusra épülő demarkáció hatóköre tehát, bár elvben korlátozott, potenciálisan elég tág, és kiterjedhet az összes olyan alkalmazásra, melyekről az MI-fejlesztők jelenleg álmodoznak.

A szemantikai fenomenalizmusra épülő kognitív demarkáció sikeres, *ha* elfogadjuk a kválék létét. A jelek szerint a demarkációs vita résztvevőinek mindkét oldalon az az intuíciója, hogy az MI-nek nincsenek (szenzoros) kváléi, az pedig szerintem nehezen tagadható, hogy a kválék konstitutív részei egyes fogalmainknak. Azok számára tehát, akik az elmefilozófiában már elköteleződtek a fenomenológiai alapú dualizmus mellett, ingyen rendelkezésre áll egy erős kognitív demarkációs érv.

8.

Introspekciós demarkáció

Mentális állapotaink *privátak*, vagyis kizárólag mi férünk hozzájuk. Bizonyos értelemben persze mások is hozzáférhetnek az érzéseinkhez, vágyainkhoz, gondolatainkhoz és élményeinkhez, például ha elmondjuk nekik, vagy ha tanulmányozzák a viselkedésünket, esetleg az agyunkat. De mi ettől eltérő, speciális módon férünk hozzá mentális állapotainkhoz, úgy, ahogy mások nem. *Közvetlenül* férünk hozzájuk, csak mi férünk közvetlenül hozzájuk, és csak a saját elménkhez férünk hozzá közvetlenül. Nem tudjuk mások érzéseit, vágyait, gondolatait közvetlenül megtapasztalni.

A mentális állapotainkhoz való közvetlen hozzáférés fontos része az *introspekció*, vagyis az a képesség, hogy felismerjük, milyen mentális állapotban vagyunk. Az erre épülő demarkációs érv szerint az ember képes introspekcióra, az MI viszont nem. Ez az érv (ha netán helyes) nem igazolja, hogy az MI nem *gondolkodik* – lehet, hogy vannak gondolatai, csak ezt nem ismeri fel –, de a kognitív képességeinket lényegesen gazdagabbnak mutatja be a gépekénél. Ráadásul ha az introspekció egyben az öntudat, önismeret vagy morális felelősség előfeltétele, akkor az introspekciós demarkáció azt is igazolja, hogy az MI-nek nincs öntudata, nincs önismerete, nem felelős semmiért.

Az introspekciós demarkáció Selmer Bringsjord amerikai MI-kutató leleménye. Az érv véleményem szerint sikertelen, de megéri foglalkozni vele, mert szellemes és az emberi elme egy fontos aspektusára irányítja rá a figyelmet.

Az introspekciót hagyományosan négy alaptulajdonsággal szokták jellemezni (Schwitzgebel 2024: 4.1.1):

- tévedhetetlen (*infallible*)
- cáfolhatatlan (*incorrigible*)
- kétségbevonhatatlan (*indubitable*)
- önprezentáló (*self-intimating*).

Tévedhetetlen: Szükségszerű, hogy ha x introspekció alapján úgy ítéli, hogy az M mentális állapotban van (pl. fáj a feje), akkor x tényleg M-ben van. Vágyainkat, érzéseinket, gondolatainkat nem tudjuk hibásan azonosítani.

Cáfolhatatlan: Szükségszerű, hogy ha x introspekció alapján úgy ítéli, hogy M-ben van, akkor ezt számára semmi sem cáfolhatja. Ha fáj a fogam, senki sem tudja bebizonyítani nekem, hogy nem fáj.

Kétségbevonhatatlan: Szükségszerű, hogy ha x introspekció alapján úgy ítéli, hogy M-ben van, akkor x nem kételkedhet abban, hogy M-ben van. Ha konstatálom, hogy fáj a fogam, akkor racionálisan nem vonhatom kétségbe, hogy fáj a fogam.

Önprezentáló: Szükségszerű, hogy ha x az M állapotban van, akkor introspekció útján fel tudja ismerni, hogy M-ben van. Ez nem minden mentális tulajdonságra igaz (a tudattalan vágyakra, ha vannak ilyenek, nem feltétlenül igaz), de a testi érzetek, érzékszervi benyomások és gondolataink esetében igaznak tűnik. Ha Londonra gondolok, képes vagyok tudatosítani magamban, hogy Londonra gondolok.

Nem mindenki ért egyet azzal, hogy az introspekció mind a négy tulajdonsággal rendelkezik. A négy tulajdonság ugyanis nem ekvivalens. A tévedhetetlenségből következik a cáfolhatatlanság, de más logikai kapcsolat nincs köztük. Logikailag lehetséges például, hogy az introspekció cáfolhatatlan, de nem tévedhetetlen. Ez esetben néha hasonló helyzetben vagyunk mentális állapotainkkal kapcsolatban, mint az oltástagadók a biológiai tényekkel kapcsolatban: súlyosan tévedünk, de erről lehetetlen meggyőzni minket. Logikailag lehetséges az is, hogy az introspekció kétségbevonhatatlan, de nem tévedhetetlen vagy nem önprezentáló.

Bringsjord demarkációs érvének központjában a tévedhetetlenség, pontosabban ennek egy ritkábban vitatott, gyengébb változata áll, amit a szerző kissé zavaró módon „hipergyenge cáfolhatatlanság"-nak hív (*„hyper-weak incorrigibilism*, Bringsjord 1992: 329). Az alábbiakban a Stanford Enciklopédia terminológiáját használom, vagyis a tévedhetetlenséget „tévedhetetlenség"-nek, nem pedig „cáfolhatatlanság"-nak hívom, és Bringsjord szövegét ennek megfelelően módosítom. Mindjárt rátérek arra, mitől lesz a tévedhetetlenség hipergyenge.

Legyen M egy mentális állapot, és tegyük fel, hogy az introspekció tévedhetetlen. Ekkor definíció szerint szükségszerű, hogy ha x azt hiszi, hogy M állapotban van, akkor x M-ben van. Bringsjord felhívja rá a figyelmet, hogy ez a kritérium túl erős. Ellenpélda:

> Jones, a kém beférkőzött az ellenség soraiba, de lefülelték. Az ellenség egy billogozó vassal próbálja vallomásra kényszeríteni. Újra meg újra megkérdezik, hajlandó-e beszélni, és ha nemet mond, a hátába nyomják az izzó vasat. Hússzor kudarcot vallanak; Jones hússzor tagadja meg a választ, és hússzor megégetik. A huszonegyedik alkalom előtt a vallatók észrevétlenül, Jones háta mögött kicserélik

> az izzó vasrudat egy jégkockára. Megkérdezik, beszél-e, Jones nemet mond, mire „megbillogozzák” a jégkockával. Jones fájdalmasan felsikolt. (Churchland 1984: 77, idézi Bringsjord 1992: 334)

Jones nem tudja introspektálni, hogy fájdalmat érez-e, mert annyira traumatizált állapotban van, hogy már egy hűvös tárgyat is tüzes billognak hisz. De Bringsjord szerint még ebben a szituációban is van valami, amivel kapcsolatban Jones nem téved: úgy *tűnik* számára, hogy fájdalmat érez. Az ellenpélda elhárításához tehát elég az ilyen tűnésekre koncentrálnunk, mert még Jones esete sem cáfolja, hogy képesek vagyunk introspektálni, milyennek *tűnik* számunkra a saját mentális állapotunk. Nevezzük $\mathcal{K}$ típusú („karteziánus”) mentális állapotoknak az ilyen tűnéseket, $\mathcal{K}$ típusú mentális tulajdonságnak pedig minden olyan tulajdonságot, amivel pontosan akkor rendelkezik valaki, ha úgy tűnik számára, hogy egy specifikus kognitív vagy fenomenális állapotban van (pl. fájdalmat érez). Az introspekciós demarkáció szempontjából nem jelent gondot, hogy a mentális állapotok egyik altípusával foglalkozunk.

Maga az érv így néz ki:

(1) Ha x egy $\mathcal{K}$ típusú M mentális állapotban van, akkor x képes tévedhetetlenül introspektálni M-et.

(2) Az MI nem képes tévedhetetlenül introspektálni saját állapotait.

(3) Tehát az MI soha nincs $\mathcal{K}$ típusú mentális állapotban.

Az 1. állítás a fenti meggondolások miatt *a priori* igaz, tehát az érv helyessége a 2. állításon múlik. Miért nem képesek a gépek tévedhetetlenül detektálni saját állapotaikat? Bringsjord szerint azért, mert minden hardver meghibásodhat, ezért egyetlen számítógépre sem lesz *szükségszerűen* igaz, hogy ha azt ⌜hiszi⌝, hogy úgy ⌜tűnik⌝ számára, hogy az M állapotban van, akkor valóban úgy ⌜tűnik⌝ számára, hogy az M állapotban van (Bringsjord 1992: 340–341). Metafizikailag (sőt, nyilván technikailag is) lehetséges, hogy a számítógép működésébe hiba csúszik a belső állapotok detektálása közben, emiatt pedig hamis ⌜hit⌝ keletkezik azzal kapcsolatban, hogy milyennek ⌜tűnik⌝ a gép számára saját belső állapota. Az emberek esetében viszont, a $\mathcal{K}$ típusú állapotok természete miatt, a tévedés lehetősége kizárt.

Ellenérvként felvethető, hogy elvben az elme is is meghibásodhat. Elképzelhető, hogy valaki számára úgy tűnik, fájdalmaz érez, de téved abban, hogy valóban így tűnik-e számára. Lehet, hogy az illető teljesen inkoherensen gondolkodik és téved a mentális állapotaira vonatkozó tűnésekkel kapcsolatban. Ez persze abnormális jelenség, de ha az MI oldalán hivatkozhatunk műszaki hibákra, miért ne hivatkozhatnánk emberek esetében kóros állapotokra?

Erre rekontraként talán azt lehet felelni, hogy a mi emberi tűnéseink lényegi tulajdonsága, hogy ha rendelkezünk velük, akkor tudjuk (vagy mindig tudatára ébredhetünk), hogy rendelkezünk velük; az introspektálhatóság valamilyen

értelemben bele van építve a tűnések struktúrájukba. Emiatt pedig az emberek esetében nem lehet elmebajra hivatkozva elgondolni, hogy valaki azt hiszi, hogy úgy tűnik számára, hogy az M állapotban van, de valójában nem tűnik úgy a számára, hogy M-ben van. Ha valaki elmezavar miatt erre utaló bizarr kijelentéseket tesz, akkor nem az introspekció kudarcával van dolgunk, hanem az elme felbomlásával: az illető összevissza beszél és nincsenek valódi hitei.

Az MI-párti erre azzal a szubkontrával felelhet, hogy talán a hardverhibás MI is összevissza ⌜beszél⌝, és nincsenek valódi ⌜hitei⌝, tehát az MI esetében is abnormális működésről van szó. A szubkontrába már csak úgy lehet belekötni, ha pontosan megvizsgáljuk, hogyan reprezentálja az MI saját állapotait és ez alapján próbáljuk kimutatni, hogy az MI-nek a szóban forgó hardverhibák esetén is vannak ⌜hitei⌝, tehát az MI nem viselkedik kórosan, amikor ⌜félreismeri⌝ a saját állapotait, a hardverhiba és az elmezavar között nincs analógia. Ezt a rejtett premisszát Bringsjord nem igazolja, én pedig nem látom, hogyan lehetne igazolni. Emiatt az introspekciós demarkáció számomra helytelennek tűnik.

Bringsjord érve ettől függetlenül szellemesen irányítja rá a figyelmet az elme egy olyan lényegi tulajdonságára, amit nem könnyű gépeknek tulajdonítani. Nem világos, hogy egy gép esetében a belső állapotok regisztrálása miért lenne lényegileg más típusú jelenség, mint a külső tárgyak fizikai paramétereinek detektálása. Az, hogy a számítógép ⌜érzékeli⌝ a processzor hőmérsékletét (vagy bármely más belső állapotát), nem tűnik lényegileg különbözőnek attól a folyamattól, hogy egy dróton keresztül egy másik számítógép processzorának hőmérsékletét ⌜érzékeli⌝, és semmi okunk azt gondolni, hogy a számítógépben zajlanak emellett nemfizikai folyamatok is, melyek a belső és külső állapotok detektálását alapvetően megkülönböztetnék. Az introspekció viszont lényegileg másfajta folyamat, mint a külső tárgyak állapotainak regisztrálása. Tehát Bringsjord alapvető intuíciója helyesnek tűnik. Az ember mintha másképp lenne tudatában saját belső állapotainak, mint egy gép. De ezt a különbséget, Bringsjord szándékaival ellentétben, nem lehet az introspekció tévedhetetlenségén keresztül megragadni, mert az emberi képességek használata ugyanúgy magában hordozza a kudarc lehetőségét, mint a számítógépek működése.

Bringsjord intuícióját talán hatékonyabban lehet kibontani, ha introspekció helyett az öntudatra (*self-consciousness*) fókuszálunk. Az „öntudat" persze sok dolgot jelenthet; Kriegel (2007) tíz különböző formáját vagy aspektusát különbözteti meg. Arra a jelenségre gondolok, hogy ha ébren tevékenykedünk, folyamatosan tudatában vagyunk saját magunknak, saját létünknek és annak, hogy a tevékenységünknek mi vagyunk az alanyai. Ahogy William James (1961: 42) írta, „bármire is gondolok, közben mindig többé-kevésbé tudatában vagyok önmagamnak, saját létezésemnek". Anélkül, hogy tagadnám az öntudat egyéb formáinak létét vagy fontosságát, a továbbiakban az „öntudat" szóval ezt a jelenséget, ezt a belső kozmikus háttérzajt jelölöm, az *én*ség érzetét, ami árnyékként követi minden gondolatunkat és érzésünket.

Ha megélt elmeállapotainknak minden pillanatban tudatosan mi vagyunk a birtokosai – *én* érzem, *én* gondolom ezt –, akkor a korábbitól eltérő demarkációs érvet építhetünk arra a premisszára, hogy az MI ⌜mentális⌝ állapotaiból hiányzik ez az alapvető, önreflexív „tulajdonosi struktúra". Értelmezhetetlennek tűnik egy program futása kapcsán arról beszélni, hogy a számítógép nemcsak bizonyos állapotban van (mondjuk az 1001 számsort tárolja a memória egyik regiszterében, vagy egy bizonyos polinomot komputál), hanem ezzel együtt és ettől elválaszthatatlanul észleli, hogy ez az állapot az *ő* állapota. Mit jelentene ez az állítás egy számítógépes architektúra vagy számítógépes program esetében?

Az öntudatra épülő demarkáció nagy előnye, hogy az öntudat nem értelmezhető funkcionális tulajdonságként. Nem az a kérdés, hogy az MI azt válaszolja-e, hogy „*Én* gondolom ezt, folyamatosan tudatában vagyok saját magamnak"; ilyen betűsorozatok generálására nyilván képes. Az öntudat nem egy reagálásmód, hanem a mentális állapotaink állandó kísérőjelensége (legalábbis sokak szerint – erre még visszatérek). Emiatt az öntudatra épülő demarkációt nem fenyegeti az egyre életszerűbb vagy egyre pontosabb válaszokat adó LLM-ek kifejlődése.

Ellenérvként felmerülhet, hogy ugyanilyen alapon az *agyban* sem találjuk az öntudatot. Mit jelentene azt állítani, hogy az agyi neuronok nemcsak egy bizonyos állapotban vannak, hanem kollektíven észlelik, hogy ez az állapot az *ő* állapotuk? Ha erre azt feleljük, hogy az öntudatot nem az egyes neuronok szintjén, hanem az idegrendszer globális konfigurációiban kell keresni (esetleg mert a neuronok csak kollektíven hordoznak bizonyos típusú információkat), akkor hasonlóan felelhetnek az MI-öntudat hívei: az MI esetében sem a számítógép vagy a program komponenseinek szintjén, hanem a teljes rendszer szintjén kell keresni az öntudatot.

De ez az ellenérv véleményem szerint viszonylag gyenge, mert előfeltételezi, hogy a fizikalizmus igaz. Ha valaki nem elkötelezett fizikalista, akkor semmi oka fennakadni azon, hogy az öntudat a biokémiai folyamatokban ugyanúgy nem található meg, mint a digitális számítógépekben.

Az öntudat-alapú demarkáció problémáit szerintem máshol kell keresni. Igazolásra szorul, hogy ez a jelenség – a tudatos állapotainkat folyamatosan árnyékként kísérő éntudat – valóban létezik. Egész pontosan azt kell igazolni, hogy az öntudat nem reflexió eredménye, vagyis nem arról van szó, hogy bármikor képesek vagyunk reflektálni mentális folyamatainkra és ha reflektálunk rájuk, akkor mindig azt konstatáljuk, hogy *mi* érezzük vagy gondoljuk ezt vagy azt. Ha nem kíséri árnyékként összes tudatos állapotunkat éntudat, hanem az én mintegy készenlétben áll, hogy betolakodjon a képbe önreflexió révén, akkor már nem egy állandó tudatállapotról, hanem az introspekcióhoz hasonló – vagy talán azzal azonos – képességről van szó. Ekkor azonban érvénytelenné válik az MI-öntudat elleni érv, hiszen az MI is képes lehet monitorozni a saját belső állapotait; itt lényegében visszaérkezünk Bringsjord érvéhez. Amennyire látom,

csak azzal garantálhatjuk, hogy az öntudat állandó kísérőjelenség legyen, ne pedig reflexiós képesség eredménye, ha – Uriah Kriegel terminológiáját használva – belátjuk, hogy az öntudat *intranzitív* természetű:

> Hasonlítsuk össze az alábbi két megfogalmazást: „Tudatában vagyok annak, hogy azt gondolom, P," illetve „Tudatosan gondolom, hogy P". Az első, tranzitív konstrukcióban az öntudat köztem és gondolatom közti viszony. A második, intranzitív konstrukcióban az öntudat a gondolatom módosulása. Ez utóbbi esetben az öntudat „nyelvi operátora" nem azt az állapotot jelöli, hogy valamilyen viszonyban vagyok a gondolatommal (vagy a gondolkodásom folyamatával), hanem azt a módot jelöli, ahogyan elgondolok valamit (a gondolkodás folyamatának módját). [...] Az intranzitív öntudat esetében nincs numerikus különbség a gondolat és az éntudat között: a gondolat maga az én tudata. A „tudatosan" határozószó a módot jelöli, ahogyan elgondolom, hogy P. Nem történik további önreflexió azután, hogy ezt elgondoltam, hanem tudatosan jelentkezik az a gondolat, hogy P. (Kriegel 2007: 2. szakasz)

Ha az öntudat intranzitív, akkor valóban árnyékként kíséri a gondolatainkat, sőt, metafizikai értelemben az árnyéknál sokkal szorosabb viszonyban van velük; hasonló viszonyban van velük, mint a gyűrődés a szőnyeggel. Ha azonban az öntudat tranzitív, akkor felmerül a gyanú, hogy az *én*ség érzete egy képesség használatának eredménye és nem a gondolkodás szükségszerű kísérőjelensége. Ha az öntudat nem gondolataink szerves része, hanem a gondolatainkhoz való viszonyulás puszta eredménye, akkor gondolatok elvileg anélkül is létezhetnek, hogy éntudat kísérné őket.

Az öntudatra épülő demarkációhoz tehát igazolni kellene, hogy az öntudat intranzitív természetű. Fogalmam sincs, hogyan lehet ezt igazolni. Saját élményeim alapján az öntudat nem a gondolataim módosulásának, hanem reflexió eredményének tűnik, de kétlem, hogy az introspekció ezen a téren tévedhetetlen. Az öntudatra épülő demarkáció, bár sikeresebbnek tűnik, mint az introspekciós demarkáció, olyan elmefilozófiai mélyfúrásokat igényel, melyeket kénytelen vagyok másokra hagyni.

9.

Nemteljességi demarkáció

Az Erdős Pálnak tulajdonított aranyköpés szerint a matematikusok olyan gépek, melyek kávét alakítanak át tételekké. De a helyzet ennél is abszurdabb, ugyanis a „tétel" végeredményben papírra írt krikszkrakszokat jelent. A matematikus kávét alakít át telefirkált papírokká. Örül, ha sikerül az általa kedvelt sorrendben leírnia bizonyos jeleket, és hosszan töpreng azon, kollégáinak joga volt-e számukra kedves sorrendben leírni jeleket.

Felvethetnénk, hogy ez a minimalista jellemzés egyéb szakmákra is ráillik. A költők is végeredményben csak jeleket írogatnak egymás után, akárcsak a gazdasági elemzők vagy a diagnózist lejegyző orvosok. De a költők, közgazdászok, orvosok és egyéb jel-rakosgató szakemberek egyértelműen a jelektől különböző dolgokra utalnak a jeleikkel. Ez alól talán csak a sci-fi-írók és fantasy-írók jelentenek kivételt, akik kitalált világokról szóló jeleket rakosgatnak egymás után. A kitalált világok azonban sok szempontból hasonlítanak az aktuális világra (pontosabban: a kitalált világok élményét okozó jelek többsége az aktuális világ dolgaira is utal). A matematikusok krikszkrakszai viszont még ilyen fiktív tartalommal sem rendelkeznek, vagy legalábbis nem világos, miért rendelkeznének vele. Megfoghatatlan, hogy mire vonatkoznak. A számok és a geometriai alakzatok nincsenek *sehol*. Bizonyos értelemben persze mindenhol ott vannak: ha ketten vagy hárman összegyűlnek, akkor Jézus Krisztuson kívül a 2 vagy 3 is ott van velük. A dobókockában (bizonyos értelemben) jelen van a kocka mint geometriai alakzat. De a matematikusok krikszkrakszai nem a világnak ezekre a konkrét részeire, az összeverődő emberekre vagy dobókockákra vonatkoznak. Ráadásul a matematika messze túlfut a valóságban jelen levő számosságokon, alakzatokon és egyéb struktúrákon.

A matematika teljesen formális tudomány, vagy legalábbis tárgyai csak formális úton, jelek rakosgatásával ragadhatóak meg. Vegyük az elemi aritmetikát, vagyis a természetes számokon végzett összeadást és szorzást. A természetes számok halmaza az a halmaz, ami a nullából és a pozitív egész számokból áll: $\{0, 1, 2, \ldots\}$, egyezményes jele $\mathbb{N}$. Hogy érzékelhető legyen a számokkal kapcsolatos tudásunk formális jellege, számok helyett beszéljünk *bigyók*ról és a

számokon végzett műveleteket nevezzük *izélés*nek. Mivel nem világos, hogy a számok micsodák, jellemzésükhöz első körben használhatunk tartalommal nem rendelkező címkéket. Ez segít áttekinteni, hogy valójában mit tudunk róluk. Semleges címkékkel így foglalható össze a természetes számok elmélete:

(N1) Minden bigyónak van egy egyedi miféléje.

(N2) Van egy bigyó, Bubu, ami semminek sem a miféléje.

(N3) Ha egy bigyóhoz hozzáizéljük Bubut, akkor az eredeti bigyót kapjuk.

(N4) Ha az x bigyóhoz hozzáizéljük y bigyó miféléjét, akkor ugyanazt kapjuk, mint ha x-hez hozzáizélnénk y-t, és vennénk az eredmény miféléjét.

(N5) Ha bármelyik bigyót megizéljük Bubuval, akkor az eredmény Bubu.

(N6) Ha az x bigyót megizéljük y miféléjével, akkor ugyanazt kapjuk, mintha x-et megizélnénk y-nal, és hozzáizélnénk x-et.

(N7) Ha valami igaz Bubura és minden x bigyó esetében igaz x miféléjére is, akkor igaz az összes bigyóra.

Matematikai értelemben ez a szabályrendszer – pontosabban ennek formalizált változata – definiálja a természetes számokat. „Bigyó" helyett normális esetben a „természetes szám", „Bubu" helyett a „0", „miféle" helyett az „utód" (*successor*, rákövetkező szám), „hozzáizél" helyett a „hozzáad", „megizél" helyett a „megszoroz" szavakat használjuk. (N1) azt mondja ki, hogy minden természetes számnál van eggyel nagyobb természetes szám, (N2) azt, hogy a nulla előtt nincsenek természetes számok, (N2)–(N6) az összeadás és szorzás tulajdonságait rögzíti, (N7) pedig a teljes indukció törvényét: ha valami igaz a nullára és továbböröklődik minden rákövetkező számra, akkor igaz az összes természetes számra.

Így már értelmesebbnek tűnnek a szabályok, de felvethető, hogy az ismerős terminológia semmivel sem mond többet a számok természetéről, mint a fenti halandzsa. A matematika annyiban vissza is igazolja ezt a gyanút, hogy megszabadul az ismerős terminológiától és a számok jellemzéséhez tisztán formális eszközöket használ. Szigorú matematikai értelemben a természetes számokat azok az axiómák jellemzik, melyeket a fenti (N1)–(N7) állításokból kapunk, ha az értelmetlen kifejezéseket szimbólumokra cseréljük le:

(ℕ1) $\forall x \exists y : y = s(x) \wedge (\forall z : y = s(z) \to z = x)$

(ℕ2) $\neg \exists x : 0 = s(x)$

(ℕ3) $\forall x : x + 0 = x$

(ℕ4) $\forall x \forall y : x + s(y) = s(x + y)$

(ℕ5) $\forall x : x \times 0 = 0$

(ℕ6) $\forall x \forall y : x \times s(y) = (x \times y) + x$

(ℕ7) $(F(0) \wedge \forall x : F(x) \rightarrow F(s(x))) \rightarrow \forall x : F(x)$[1]

$\forall$	minden
$\exists$	létezik
$\neg$	nem
$\wedge$	és
x, y, z	„bigyó”
$P \rightarrow Q$	Ha P, akkor Q
$s(x)$	x miféléje (*successor*, rákövetkező szám)
$F(x)$	x rendelkezik az F tulajdonsággal

Ezek a formális krikszkrakszok mindent összefoglalnak azzal kapcsolatban, amit a természetes számok a matematika számára jelentenek.[2] A formális logika segítségével pedig ismert számelméleti tényeket vezethetünk le a krikszkrakszokból. Beláthatjuk például, hogy $x + y = y + x$ (az összeadás kommutatív). Ehhez első lépésben be kell látni, hogy minden x-re

(*) $x + 0 = 0 + x$

(ℕ3) miatt $0 + 0 = 0$, így (*) igaz $x = 0$-ra. Másrészt ha (*) igaz x-re, akkor igaz $x + 1$-re, azaz $s(x)$-re: (ℕ3) miatt $s(x) + 0 = s(x)$, (ℕ4) miatt pedig $0 + s(x) = s(0 + x)$, utóbbi viszont (*) és (ℕ3) miatt egyenlő $s(x)$-el. Mivel (*) igaz a nullára és ha igaz x-re, akkor igaz $x + 1$-re, (ℕ7) miatt (*) minden természetes számra igaz. Ezután újabb teljes indukcióval láthatjuk be, hogy $x + y = y + x$, vagyis az összeadás tényleg kommutatív.

A matematika formális jellege filozófiai szempontból rossz hír, mert érthetetlenné teszi, hogy a matematika *miről szól*. A természetes számokról a szakértők

[1] A teljes indukció axiómája az elsőrendű és másodrendű logikában eltérő módon jelenik meg. A másodrendű logikában, ahol tulajdonságok létezéséről is beszélhetünk, (ℕ7) elé oda kell biggyeszteni a „$\forall F$:” kvantort. Az elsőrendű logikában, ahol csak „bigyókról” (számokról) állíthatunk dolgokat és tulajdonságok felett nem kvantifikálunk, (ℕ7) egy axióma-séma, aminek végtelen sok példánya van, minden F kifejezéshez tartozik egy. Érdekes módon a végtelenített (*infinitary*) logikában, ahol végtelen konjunkciók és diszjunkciók is megengedettek, a teljes indukció logikailag levezethető a $\forall x : (x = 0 \vee x = s(0) \vee x = s(s(0)) \vee \ldots)$ végtelen diszjunkcióból (vö. Bell 2023: 1. szakasz).

[2] Ínyenceknek: ez szigorúan véve csak a másodrendű logikában igaz, ahol (ℕ7) nem egy axióma-séma, és szerepel előtte a „$\forall F$” másodrendű kvantor, az így keletkező elmélet ugyanis a másodrendű logikában kategorikus (ld. pl. Button & Walsh 2018: 153–155). A természetes számok elmélete a legegyszerűbb végtelenített elsőrendű logikában, $L(\omega_1, \omega)$-ban is kategorikus (Bell 2023: 1. szakasz).

véleménye alapján annyit tudunk, hogy kielégítik (ℕ1)–(ℕ7)-et, de ezek az axiómák teljes homályban hagyják a számok természetét. A filozófusok egy része ebből arra következtet, hogy a matematika „tiszta struktúrákkal" foglalkozik: az elemi aritmetika például az a struktúra, melynek üres „helyei" az (ℕ1)–(ℕ7) által leírt viszonyrendszerben állnak. A számoknak, és analóg okokból a többi matematikai objektumnak, nincs intrinzikus természete, hanem a létük a viszonyaikban merül ki (*relations without relata*, Shapiro 1997). Bármit is kezdjünk ezzel a metafizikai talánnyal, a matematikai formális jellege annyiban jó hír, hogy a formállogikai bizonyítások *mechanikusak*, emiatt pedig egy gép, ami gondolkodás nélkül hajt végre szintaktikai átalakításokat, előbb-utóbb rájuk talál. Így elvben megnyílik az út ahhoz, hogy az összes matematikai igazságot számítógéppel megkeressük és bebizonyítsuk. Valami ilyesmi volt a célja a matematika leghíresebb formalista projektjének, a „Hilbert-programnak".

David Hilbert, a XX. század egyik legnagyobb matematikusa az 1920-as években tűzte ki célként, hogy a matematika egészét axiomatikus alapokra kell helyezni és tisztán formális levezetésekből kell felépíteni, anélkül, hogy matematikai intuíciókra apellálnánk. Hilbert számára ez főleg azért volt fontos, mert be akarta bizonyítani, hogy a matematika konzisztens. Egy inkonzisztens elméletből a standard logikai szabályok szerint bármi következik, tehát ha a matematika inkonzisztens, akkor a matematikusok találomra turkálnak egy logikai katyvaszban, ahol minden és mindennek az ellentéte is bizonyítható. Hilbert már az 1900-as évektől kezdve próbálta igazolni, hogy a matematika egyes részei konzisztensek (sikerült például igazolnia, hogy az euklideszi geometria konzisztens), és azt remélte, ez a program kiterjeszthető a matematika egészére (Zach 2023).

Hilbert álmát a huszonöt éves Kurt Gödel foszlatta semmivé 1931-ben két, azóta legendássá vált nemteljességi tételével. Az első nemteljességi tétel nagyjából azt mondta ki, hogy ha a természetes számok elmélete konzisztens, akkor tartalmaz olyan igazságokat, melyek nem vezethetőek le (ℕ1)–(ℕ7)-ből. Ráadásul hiába bővítjük (ℕ1)–(ℕ7)-et, hiába adunk hozzá további axiómákat, mindig maradnak bizonyíthatatlan igazságok, legalábbis ha ragaszkodunk hozzá, hogy a bizonyítások véges hosszúak legyenek és egy algoritmus elő tudja állítani őket az axiómákból. Hiába próbáljuk körülhatárolni a természetes számokat egy formális rendszerrel, mechanikus levezetések révén lehetetlen lesz felfedezni a velük kapcsolatos összes igazságot.[3]

[3] Precízebben megfogalmazva Gödel tételei olyan, az elsőrendű logikában formalizált elméletekre vonatkoznak, melyek axiómái rekurzívak (Turing-géppel eldönthető, hogy egy formula axióma-e vagy sem), és tartalmazzák az elemi aritmetikát (ami ekvivalens (ℕ1)-(ℕ7)-tel). Gödel az első nemteljességi tétel eredeti bizonyításában azt is kikötötte, hogy a szóban forgó rendszer nemcsak konzisztens, hanem ω-konzisztens, ami nagyjából azt jelenti, hogy akkor sem válna inkonzisztenssé, ha végtelen hosszú állításokat írhatnánk le. (Hivatalosan: egy elmélet akkor ω-konzisztens, ha nincs olyan $\varphi(x)$ formula, hogy minden n-re bizonyítható $\neg\varphi(n)$, de bizonyítható az is, hogy $\exists x : \varphi(x)$.) Rosser (1936) az első nemteljességi tételt az ω-konzisztencia elhagyásával igazolta.

Gödel tételei nemcsak a számelméletre vonatkoznak, hanem a matematika minden olyan területére, melynek struktúrája elég bonyolult ahhoz, hogy „szimulálni" lehessen benne a természetes számokat. De ennek fordítottja is igaz: ha egy matematikai elmélet nem tartalmazza az elemi aritmetikát, akkor előfordulhat, hogy összes tétele mechanikusan levezethető az axiómákból, vagyis az elmélet teljes. Ez a helyzet például akkor, ha az elemi aritmetikából kihagyjuk a szorzást („Pressburger-aritmetika", ld. Chang és Keisler 1973: 43). Bonyolult és részben nyitott kérdés, hogy mely matematikai elméletek teljesek. De Hilbert követői számára sovány vigasz, hogy egyes matematikai elméletek teljesek, mert egészen alapvető elméletek, például az elemi aritmetika, nem azok. A matematika túlcsordul a gépiesen felsorolható axiómarendszereken. Nem krikszkrakszok puszta formális manipulációja tehát, hanem a krikszkrakszok mechanikus manipulációján túlmutató tartalma van. Vagy legalábbis ez az első nemteljességi tétel egy lehetséges filozófiai olvasata, és mint látni fogjuk, erre az olvasatra épül a nemteljességi demarkáció.

A második nemteljességi tétel ennél is nagyobb katasztrófát jelentett a Hilbert-program számára, hiszen Hilberték deklarált célja volt a matematika konzisztenciájának igazolása. Gödel második tételének értelmében ez csak akkor lehetséges, ha a matematika inkonzisztens. Tehát pont annak kell örülni, ha *nem* tudjuk igazolni a matematika konzisztenciáját. A második tétel kapcsán is felmerülnek persze értelmezési nüanszok, mert a matematika egyes területeinek konzisztenciája bizonyítható. Gerhard Gentzen például (aki nemcsak tehetséges matematikus, hanem hithű náci is volt) 1936-ban bebizonyította, hogy az elemi aritmetika konzisztens. De ehhez olyan segédfeltevéseket használt fel, melyek nem következnek az elemi aritmetikából (DeLong 2004: 217–218), az így keletkező bővebb – (ℕ1)–(ℕ7) mellett a szóban forgó plusz feltevéseket is tartalmazó – axiómarendszer konzisztenciáját pedig a bizonyítás már nem igazolta; ehhez még újabb axiómák kellenek. Tehát a matematika konzisztenciája (amennyiben a matematika konzisztens) csak olyan feltevésekkel „igazolható", melyek konzisztenciája (a keletkező kibővített kontextusban) további igazolásra szorul. A matematika konzisztenciájának igazolatlansága az alábbiakban ritkán kerül majd elő, csak azért tértem ki rá, mert fontos hangsúlyozni, hogy Gödel két teljesen különböző nemteljességi tételt bizonyított be. A demarkáció kapcsán végig az *első* nemteljességi tételről lesz szó, ezért „a Gödel-tétel" kifejezés a továbbiakban mindig erre vonatkozik.

9.1. „Bizonyíthatatlan vagyok"

A nemteljességi demarkáció arra az intuícióra épül, hogy a matematikai formulákat mechanikus szabályok alapján kezelő rendszerek (mint az MI) nem ismerik fel az úgynevezett Gödel-mondatok igazságát, mi viszont felismerjük. Tehát a ⌜gondolkodás⌝ nem azonos a (matematikai) gondolkodással.

A fenti gondolatmenet sok helyen pontosításra szorul. Mindenekelőtt tisztázni kell, mik a Gödel-mondatok. A Gödel-tétel precíz bizonyítása eléggé nyakatekert mutatvány, ezért csak nagy vonalakban próbálom érzékeltetni a struktúráját.

Tegyük fel, hogy formalizáltuk az elemi aritmetikát az (ℕ1)–(ℕ7)-hez hasonló krikszkrakszok képében. Nevezzük az így keletkező matematika struktúrát „az elemi aritmetika (Peano Aritmetika) standard modelljének". A „modell" itt a krikszkrakszok egy konkrét matematikai értelmezését jelenti, konkrétan azt az értelmezést, ahol „0" a nullát jelöli, az $s(x)$ az x-nél eggyel nagyobb számot jelöli, a „$\forall x$:" kezdetű mondatok minden természetes számra és kizárólag a természetes számokra vonatkoznak, és így tovább.[4]

Gödel egyik forradalmi trükkje az volt, hogy az elemi aritmetika standard modelljének formuláit természetes számoknak feleltette meg („Gödel-számozás"). Ez lehetővé tette, hogy egyes számelméleti állítások magukra a számelméleti állításokra vonatkozzanak, hiszen ha a φ formális mondat Gödel-száma ($\ulcorner\varphi\urcorner$) maga is egy természetes szám, akkor minden olyan aritmetikai állítás, ami a $\ulcorner\varphi\urcorner$ számra vonatkozik, egyben a φ mondatról is állíthat valamit, amennyiben a mondat bizonyos tulajdonságait (például a levezethetőségét) a $\ulcorner\varphi\urcorner$ szám bizonyos tulajdonságainak feleltetjük meg.

A Gödel-számozás technikai részletei kicsit bonyolultak, de az alapötlet aligha igényel hosszas kifejtést. Az okostelefonok és számítógépek korában mindenki tisztában van vele, hogy szövegeket számokká lehet alakítani. A szövegeket el tudjuk menteni egy fájlban, a fájlok pedig lényegében nullák és egyesek sorozatai, vagyis kettes számrendszerben leírt számok.[5] Az egyszerűség kedvéért tegyük zárójelbe a részleteket és tételezzük fel, hogy minden formulához rendeltünk egy egyedi numerikus azonosítót. Tegyük fel azt is, hogy formulák *sorozatával* is megtehetjük ugyanezt, vagyis bizonyos természetes számok – a Gödel-számozáshoz választott metódus alapján – valamely $\varphi_1, \varphi_2 \dots \varphi_n$ formula-sorozatnak felelnek meg.

Gödel második forradalmi trükkje az volt, hogy az általa definiált kódolási módszer segítségével definiált egy aritmetikai viszonyt, ami a tételek és a bizonyításaik közti viszonynak felel meg. Vagyis belátta, hogy van olyan $\mathsf{BIZ}(m, n)$ aritmetikai formula, amit pontosan akkor elégítenek ki az m, n természetes számok, ha $m = \ulcorner\alpha\urcorner$, $n = \ulcorner\beta_1 \dots \beta_n\urcorner$ és a $\beta_1 \dots \beta_n$ formula-sorozat az α állítás bizonyítása (a sorozat minden tagja a bizonyítás egyik sora). $\mathsf{BIZ}(m, n)$ emellett azzal a fontos tulajdonsággal is rendelkezik, hogy az elemi aritmetika axiómáiból bármely m, n természetes szám esetén levezethető $\mathsf{BIZ}(m, n)$ vagy annak tagadása, vagyis (ℕ1)–(ℕ7)-ből levezethető, hogy az n szám az m által kódolt formula bizonyítását kódolja-e.

[4] A modellek formális definícióját ld. a 11. fejezetben.

[5] Vagy, ha byte-ok sorozatainak tekintjük őket, 256-os számrendszerben leírt számok.

Az aritmetikai mondatok egy része ezek alapján az aritmetikai mondatok bizonyíthatóságáról szól. A nemteljességi tétel utolsó komponense az a segédtétel, hogy az aritmetika standard modelljében bármely $\varphi(x)$ formulához van olyan ψ mondat, hogy az alábbi állítás levezethető az aritmetika axiómáiból:

(*) $\qquad \psi \leftrightarrow \varphi(\ulcorner\psi\urcorner)$

Példa: ha $\varphi(x)$ azt fejezi ki, hogy x hatjegyű prímszám (tízes számrendszerben), akkor létezik egy olyan ψ számelméleti mondat, ami akkor és csak akkor igaz (az aritmetika standard modelljében), ha a kódja, $\ulcorner\psi\urcorner$, hatjegyű prímszám. A ψ mondat lényegében azt állítja, hogy „a Gödel-számom rendelkezik a φ tulajdonsággal". (*) erejét az adja, hogy bármely $\varphi(x)$ formulára létezik olyan ψ formula, hogy (*) bizonyítható az aritmetika axiómái alapján.[6]

Alkalmazzuk ezt a tételt a $\neg\exists y : \mathsf{BIZ}(x, y)$ formulára, ami azt mondja ki (a Gödel-számozás közvetítésével), hogy az x által kódolt formula nem bizonyítható. Tudjuk, hogy az aritmetika axiómából valamely γ mondatra levezethető az alábbi összefüggés:

(**) $\qquad \gamma \;\leftrightarrow\; \neg\exists y : \mathsf{BIZ}(\ulcorner\gamma\urcorner, y)$

Vagyis γ akkor és csak akkor igaz, ha γ nem bizonyítható.

Bizonyítható-e γ? Ha γ bizonyítható és az aritmetika logikailag helyes (ha nem bizonyít hamis állításokat), akkor γ igaz. De ha γ igaz, akkor nem bizonyítható, hiszen (**) pont ezt mondja. Tehát γ nem bizonyítható, amennyiben az aritmetika logikailag helyes. Cáfolható-e γ? Ha γ cáfolható (és az aritmetika logikailag helyes), akkor létezik olyan n természetes szám, hogy $\mathsf{BIZ}(\ulcorner\gamma\urcorner, n)$. De ekkor az n által kódolt formulasorozat bizonyítja γ-t. Tehát ha γ cáfolható, akkor bizonyítható. Ez ellentmondás, tehát γ nem cáfolható.[7]

---

[6] Ez az úgynevezett fixpont-tétel vagy fixpont-lemma (Hamkins 2020: 236). Vázlatos bizonyítás: legyen $\mathsf{bh}(x, y)$ az az aritmetikai függvény, ami az x által kódolt formulába behelyettesíti az y számot, vagyis igaz rá, hogy bármely $\varphi(x)$ egyváltozós formula és n szám esetén $\mathsf{bh}(\ulcorner\varphi(x)\urcorner, n) = \ulcorner\varphi(n)\urcorner$. Azt, hogy ilyen függvény létezik (és effektíven kiszámolható), a Gödel-számozás tulajdonságai alapján kell bebizonyítanunk; tegyük fel, hogy ezt a munkát már elvégeztük. Legyen $\varphi(x)$ egy tetszőleges egyváltozós formula, és tekintsük a $\varphi(\mathsf{bh}(x, x))$ formulát. Legyen ennek Gödel-száma n. A $\varphi(\mathsf{bh}(n, n))$ formula pont alkalmas a (*)-ban szereplő ψ szerepére, hiszen

$$\mathsf{bh}(n, n) = \mathsf{bh}(\ulcorner\varphi(\mathsf{bh}(x, x))\urcorner, n) = \ulcorner\varphi(\mathsf{bh}(n, n))\urcorner$$

Emiatt $\varphi(\mathsf{bh}(n, n))$, azaz ψ, bizonyíthatóan akkor és csak akkor igaz, ha $\varphi(\ulcorner\varphi(\mathsf{bh}(n, n))\urcorner)$, azaz $\varphi(\ulcorner\psi\urcorner)$, igaz.

[7] Ezen a ponton van szükség az ω-konzisztenciára, vagyis arra, hogy a rendszer ne tudja igazolni $\exists x : \varphi(x)$-et, ha minden egyes n számra igazolja $\neg\varphi(n)$-et. Ezt a tulajdonságot kell kihasználni $\exists x : \mathsf{BIZ}(\ulcorner\gamma\urcorner, x)$ és $\neg\mathsf{BIZ}(\ulcorner\gamma\urcorner, n)$ kapcsán. Rosser (1936) egy olyan Gödel-mondattal szabadul meg az ω-konzisztenciára vonatkozó feltételtől, ami nem azt mondja, hogy „Bizonyíthatatlan vagyok", hanem azt, hogy „Ha bizonyítható vagyok, akkor a tagadásomra létezik egy azonos hosszúságú vagy rövidebb bizonyítás".

γ tehát nem bizonyítható és nem is cáfolható az aritmetika standard modelljében. Vajon igaz vagy hamis ebben a modellben? Hamis nem lehet (ha az aritmetika logikailag helyes), mert akkor létezne olyan n természetes szám, melyre $BIZ(\ulcorner\gamma\urcorner, n)$ igaz az aritmetika standard modelljében, ekkor pedig γ bizonyítható volna, és az aritmetika hamis állítást bizonyítana. Igaz viszont lehet γ, mert nem bizonyítható, és pontosan ezt állítja magáról. Tehát γ az elemi aritmetika egyik Gödel-mondata. Ha az elemi aritmetika logikailag helyes, akkor γ az aritmetika standard modelljében igaz, de annak axiómái segítségével nem bizonyítható. Az aritmetika standard modellje emiatt nem teljes.

A fenti vázlatból talán sejthető, miért nem segít, ha γ-t egyszerűen hozzácsapjuk a kiinduló axiómarendszerünkhöz, (ℕ1)–(ℕ7)-hez. Ekkor a korábbi $\mathcal{A}_0$ axiómarendszer helyett lesz egy bővített $\mathcal{A}_1$ axiómarendszerünk. Csakhogy $\mathcal{A}_1$-hez a Gödel által kidolgozott algoritmussal szerkeszthetünk egy új γ_1 Gödel-mondatot, ami az $\mathcal{A}_1$ standard modelljén belül igazolhatatlan és cáfolhatatlan, de igaz. Ez a folyamat pedig sosem ér véget. A matematikai igazság mindig kicsúszik a bizonyíthatóság hálójából.

9.2. Nemteljességi demarkáció

A nemteljességi demarkáció a Gödel-mondatok alapján próbálja elhatárolni az emberi gondolkodást a számításoktól. A stratégia két szószólója John Lucas filozófus és Roger Penrose matematikus. Lucas így foglalta össze az érv lényegét:

> Gödel tétele nyilván a kibernetikai szerkezetekre is vonatkozik, mert egy gép lényegileg egy formális rendszer megtestesülése. Emiatt ha konzisztensen kezeli az elemi aritmetikát, akkor garantáltan létezik olyan formula, melyet képtelen igaznak felismerni – mert a formula nem bizonyítható a szóban forgó rendszerben –, noha mi emberek látjuk, hogy a formula igaz. Emiatt egyetlen gép sem lehet az elme teljes vagy adekvát modellje. Az elménk lényegileg különbözik a gépektől. (Lucas 1961: 113)

Lucas ötletével kapcsolatban felmerülhet az az elvi kifogás, hogy egy fizikai számítógép még elvben sem képes konzisztensen kezelni a *teljes* aritmetikát, hiszen a Turing-gépekkel ellentétben nincs végtelen ideje és végtelen memóriája. Mivel ez ránk emberekre is igaz, mi sem tudjuk konzisztensen kezelni a *teljes* aritmetikát. Tehát az ember és a gép matematikai képességeinek ütköztetése a jelek szerint már az első lépésnél elbukik. Ez a probléma minimális figyelmet kapott a Lucas/Penrose érv körüli vitában, pedig jogosnak tűnik. A vitázó felek tipikusan azt feltételezik hogy egy idealizált matematikusról beszélünk, és az ő képességeit hasonlítjuk össze egy idealizált (Turing-gépnek tekintető) számítógéppel (vö. Koellner 2018a: 337). De ha az idealizált matematikusok nem tekinthetőek a valódi matematikusok adekvát modelljének, sem az idealizált számítógépek a valós számítógépekének, akkor pontosan mit is akarunk

igazolni? Hogy a mesebeli matematikusok képességei meghaladják a mesebeli számítógépekét?

Ezt a kifogást zárójelbe teszem, egyrészt mert nem találtam nyomát a szakirodalomban, másrészt mert a nemteljességi demarkációt ettől független okokból el fogom utasítani. A következő fejezetben, a Rice-tételre épülő demarkáció kapcsán viszont részletesen tárgyalok majd egy teljesen hasonló empirista kifogást. Hozzáteszem azonban, hogy az idealizációs stratégia egyáltalán nem tűnik védhetetlennek, hiszen nem kitalált lényekről vitázunk, hanem azt vizsgáljuk, hogy egy ember *elvben* – ha tetszőlegesen sokáig dolgozhatna számelméleti tételeken – mit lenne képes belátni. Szabatosabban: a Lucas/Penrose-érv híve azt állítja, hogy egy ember számára metafizikailag lehetséges belátni olyan dolgokat, melyek igazolása gépek számára metafizikailag lehetetlen. A metafizikai lehetőség és szükségszerűség fogalmát persze kritizálhatjuk empirista alapon, de ez a vita eléggé messzire vezet és nem nyújt gyors, adu ász jellegű érvet a nemteljességi demarkáció ellen.

Penrose szerint a Gödel-tétel elmefilozófiai tanulsága az, hogy mi emberek képesek vagyunk gondolatban kilépni bármelyik formális rendszerből és „kívülről" szemügyre venni, a gépek viszont nem képesek erre:

> A Gödel-mondatot igazoló belátás egy általánosabb módszertani elv megnyilvánulása, amit a logikusok reflexiós elvnek neveznek. Az axiómarendszer jelentésére és a műveleti szabályokra „reflektálva" megállapítjuk, hogy alkalmasak matematikai igazságok levezetésére, ezt pedig belekódolhatjuk további matematikai igazságokba, melyek nem voltak levezethetőek az eredeti axiómákból a megadott szabályokkal. [...] A reflexiós elvek teljesen ellentétesek a formalista gondolkodásmóddal. Ha ügyesen alkalmazzuk őket, kitörhetünk bármelyik merev formális rendszer kereteiből, és olyan matematikai felismerésekre juthatunk, melyek korábban elérhetetlennek tűntek. A matematikai szakirodalom vélhetőleg sok olyan bizonyítást tartalmaz, melyekhez az aritmetika standard formális rendszerétől távol eső belátásokra van szükség. Emiatt pedig a matematikai gondolkodás – az, hogy a matematikusok mi alapján fogadnak el igaznak egy állítást – nem egy formális rendszer eredménye. A Gödel-mondatról látjuk, hogy bizonyos számok igazzá teszik [...], noha ezt lehetetlen levezetni az axiómákból. A reflexiós elvekből eredő belátás nem kódolható bele egy formális rendszerbe, ahol egy algoritmus alapján vezetünk le állításokat. (Penrose 1999: 144)[8]

A nemteljességi demarkáció lehetősége mások fejében is felmerült. Maga Gödel is azt gondolta, hogy ha a matematika helyes, akkor az emberi elme képessé-

[8] A Penrose által említett logikai reflexiós elveket részletesen tárgyalja Boolos (1993: 63) és Raatikainen (2020: 4.3).

gei meghaladhatják a számítógépekét (Gödel 1951/1995).[9] Turing (1948/2004) szintén felismerte egy ilyen érv lehetőségét, bár magát az érvet elutasította.

A Lucas/Penrose-érv egy lehetséges formája a következő:

(1) Ha $\mathcal{A}$ egy logikailag helyes, az elemi aritmetikát tartalmazó rekurzív formális axiómarendszer, akkor szerkeszthető olyan γ Gödel-mondat, hogy $\mathcal{A}$-ból nem vezethető le γ, de mi emberek látjuk, hogy (az aritmetikai szimbólumok standard interpretációja szerint) γ igaz.[10]

(2) Ha x egy MI, akkor létezik egy olyan $\mathcal{A}_x$ rekurzív formális axiómarendszer, hogy az x által igaznak tekintett (matematikai) állítások halmaza $\mathcal{A}_x$ logikai következményeinek részhalmaza. Nevezzük $\mathcal{A}_x$-et „x axiómarendszeré"-nek.

(3) Ha x egy MI, $\mathcal{A}_x$ az MI axiómarendszere, és $\mathcal{A}_x$ tartalmazza az elemi aritmetikát, akkor (1) miatt $\mathcal{A}_x$-nek van olyan γ Gödel-mondata, amiről mi emberek látjuk, hogy igaz, de x képtelen felismerni, hogy γ igaz.

(4) Tehát, (2) és (3) miatt, minden, az elemi aritmetikát kezelni képes MI esetében van olyan állítás, amiről mi emberek látjuk, hogy igaz, de az MI ezt nem képes felismerni.

(5) Tehát, (4) miatt az emberi gondolkodás meghaladja a gépi ⌜gondolkodást⌝.

A Lucas/Penrose-érvvel kapcsolatos vita az 1. és a 3. premissza körül forog.[11] Néhány karakteresebb álláspont ismertetése után amellett fogok érvelni, hogy a nagy nyelvi modellek kapcsán a 2. premisszába is bele lehet kötni, a leghatékonyabban pedig a konklúzió deflálásával lehet támadni a nemteljességi demarkációt.

[9] Gödel más okokból is elutasította az elme számítógépes modelljét, ld. Wang (1995).

[10] „Formális axiómarendszer"-en egy Φ formula-halmaz és egy logikai levezetési szabályokból álló Λ halmaz kombinációját értem. Egy ilyen értelemben vett $\mathcal{A}(\Phi, \Lambda)$ axiómarendszer akkor logikailag helyes, ha nem lehet Φ-ből Λ segítségével olyan állítást levezetni, ami Φ valamelyik modelljében (a nem-logikai szimbólumok valamely interpretációja alapján) hamis. Ha $\mathcal{A}$ helyes, akkor konzisztens is (nem vezethető le Φ-ből Λ felhasználásával ellentmondás), mivel az ellentmondások minden modellben hamisak. $\mathcal{A}$ akkor rekurzív, ha egy effektív algoritmus (egy Turing-gép) segítségével bármely, a szóban forgó formális nyelven leírt ψ formuláról megállapítható, hogy (a) eleme-e Φ-nek, illetve (b) levezethető-e Λ felhasználásával egy tetszőleges φ formulából.

[11] Az itt következő diszkusszió alapja Megill (2011). A Lucas/Penrose érv körüli vita sokkal kuszább, mint ez a fejezet sejteti, de precíz rekonstrukciójára nem vállalkozom. Nagyon alaposan tárgyalja az érv különböző változatait Koellner (2018a, 2018b). Koellner (2018b) terminológiáját használva jelen fejezet tárgya kizárólag „a Penrose-féle Új Érv". Nem foglalkozom a nemteljességi érv mutációival sem. Bringsjord és Arkoudas (2004) a transzfinit számítások (hiperkomputáció) fogalmára, Stern (2018) az intenzionális paradoxonokra, Wright (1995) pedig az intuicionista logikára épít a Lucas/Penrose-érvhez hasonló, de részleteiben eltérő demarkációs érvet.

Az 1. premissza azt implikálja, hogy mi emberek bármelyik (az elemi aritmetikát tartalmazó) formális rendszerhez tudunk Gödel-mondatot szerkeszteni. Felvethető, hogy ez két okból is hamis. Egyrészt az emberek nagy része nem érti a Gödel-tételt, emiatt Gödel-mondatokat sem tud szerkeszteni (Coder 1969). Másrészt a Gödel-mondatok szerkesztése elég bonyolult feladat olyan axiómarendszerek esetében, melyek komplexitása megközelíti az emberi elme komplexitását (Benacerraf 1967).

Lucas erre azt válaszolta, hogy a matematikailag pallérozatlan emberek esetében kevésbé világos, hogy nem gépek, de rájuk is vonatkozik az érv (Lucas 1970). Nem az a fontos, hogy bárki ténylegesen előállítja-e egy axiómarendszer Gödel-mondatát, hanem az, hogy az emberek *elvben* képesek erre, a gépek pedig elvben *sem* képesek rá.

A 3. premissza hallgatólagosan feltételezi, hogy a *mi* matematikánk logikailag helyes, vagyis nem tudunk hamis állításokat „belátni". Ha a matematikánk nem helyes, akkor nem hivatkozhatunk arra, hogy *mi* látjuk a Gödel-mondatok igazságát, a gépek pedig nem, mert a matematikai levezetéseink megbízhatatlanok. A Lucas/Penrose-érv sok kritikusa ebbe a rejtett előfeltevésbe köt bele. Lewis (1969) azzal vádolta Lucast, hogy képtelen specifikálni saját matematikai rendszerét – a „Lucas-aritmetikát" –, emiatt pedig képtelen igazolni, hogy a saját matematikai rendszere helyes. Hutton (1976) hangsúlyozza, hogy lehetetlen belátni saját konzisztenciáinkat; hasonlóan érvel Chalmers (1996c), McCullough (1996) és McDermott (1996). McDermott például felemlegeti, hogy Alfred Kempe brit matematikus 1879-ben közreadta a négyszín-tétel egy „bizonyítását",[12] amiről csak 1890-ben derült ki, hogy hibás. Úgy tűnik, a közbülső 11 évben a matematikusok gondolkodása helytelen volt. De akkor honnan tudjuk, hogy ma már helyes?[13]

Az ilyen kifogásokra transzcendentális, pragmatikus, és antiszkeptikus rekontrákkal lehet válaszolni.[14] Lucas az alábbi transzcendentális rekontrát dolgozza ki:

> A gondolkodás lehetőségéhez fel kell tennünk, hogy konzisztensek vagyunk. Ez talán olyasmi, mint a természeti törvények létével kapcsolatos posztulátum – nem egy logikailag precíz érv konklúziója,

[12] A négyszín-tétel szerint minden térkép kiszínezhető legfeljebb 4 színnel úgy, hogy bármely két szomszédos ország különböző színű legyen. Ezt a tételt csak 1976-ban sikerült bebizonyítani, ezernél is több eset számítógépes ellenőrzésével.

[13] Nehezen sorolható be a fenti három kategóriába Koellner (2018b) szofisztikált ellenérve, ami William Reinhardt (1986) munkájára épít. Koellner az igazságoperátor és a tudás-operátor formális modelljei alapján érvel amellett, hogy nincs elvi ellentmondás a Gödel-tétel és az elme algoritmizálhatósága között. Ennek részletes elemzése nagyon messzire vezetne, de érzésem szerint Koellner formális diagnózisa is végeredményben arra fut ki, hogy nem tudjuk igazolni a saját konzisztenciánkat.

[14] Ezek természetesen nem merítik ki a palettát. Különösen érdekes Crispin Wright (1995) érve, miszerint intuicionisták számára eleve nem merül fel a konzisztenciával kapcsolatos kifogás.

> hanem szükségszerű kiindulópont ahhoz, hogy egyáltalán gondolkodni kezdjünk. (Lucas 1976: 147)

Lucas szerint kiküszöbölhetetlen problémáról van szó, mert a logikai elveink helyességét nem tudjuk igazolni, csak előfeltételezni. Emiatt a demarkációs érvtől is csak annyit várhatunk el, hogy igazolja: *ha* helyesen gondolkodunk, *akkor* a gondolkodásunk nem gépesíthető.

Az aritmetikánk helyességével kapcsolatos kifogásra adható *pragmatikus* válasz is. Eszerint kétségkívül követünk el következtetési hibákat, de ezek elvben kijavíthatóak és többnyire ki is javítjuk őket. Ahogy Penrose írja Kempe hibás bizonyításával kapcsolatban:

> Kétségtelen, hogy időnként hiba csúszik a matematikai bizonyításokba és kalkulációkba. De a matematikusok ezt a jelenséget nem használják a tudásuk *bővítésére* (bár elképzelhetőnek tartom, hogy szerencsés félreértésekből néha eredhetnek új felismerések). Ami pedig ennél is fontosabb, a hibák *korrigálhatóak*. Ha valaki kiszúrja őket – egy másik tudós, vagy utólag maga a szerző –, akkor meg tudunk szabadulni tőlük. (Penrose 1994: 140–141)

Az, hogy pusztán empirikus értelemben egyikünk sem konzisztens, nem erős érv a 3. premissza ellen, mert szándékolatlan inkonzisztenciáink olyan normák megsértésből erednek, melyeket a tudomány – és legfőképpen a matematika – nagy becsben tart és kíméletlenül számonkér. A nemteljességi demarkáció lényege nem az a feltevés, hogy az emberek *mindig* konzisztensek, hanem az, hogy *elvben* tudnak helyesen gondolkodni. Ráadásul ez nem is egy elvont lehetőség, hiszen a matematikában rendszeresen megvalósul; a Kempe-fiaskóhoz hasonló esetek viszonylag ritkának tűnnek. Azokban a (nem túl távoli) lehetséges világokban tehát, ahol mindig logikailag korrekten gondolkodunk, a matematikai képességeink meghaladják a gépekét. Ez a modális különbség pedig garantálja, hogy az emberi gondolkodás más természetű, mint a gépi ⌜gondolkodás⌝.

Ettől függetlenül persze nem kizárható, hogy a gondolkodásunk annak ellenére inkorrekt, hogy a lehető leglelkiismeretesebben próbáljuk betartani a szabályokat. Sem magunkkal, sem a gépekkel kapcsolatban nem tudjuk igazolni a logikai helyességet. Négy eset lehetséges:

	Az MI logikája helyes	Az MI logikája helytelen
Az emberi logika helyes	E_1	E_2
Az emberi logika helytelen	E_3	E_4

E_1 a Lucas/Penrose érv szándékolt kontextusa: a gép és az ember is helyesen következtet (ideális esetben), de gödeli okok miatt a mi matematikánk gazdagabb, emiatt a gondolkodás különbözik a ⌜gondolkodás⌝-tól.

E_2 a demarkáció triviális esete: ha az MI helytelenül következtet, mi pedig helyesen, akkor nyilván éles különbség van a gondolkodás és a ⌜gondolkodás⌝ között.

E_3 a demarkáció szempontjából előnytelen szituáció, mert itt az MI következtet helyesen, mi pedig nem, tehát a gödeli érvelés nem érvényes, ráadásul az MI kognitív szempontból felsőbbrendű hozzánk képest.

E_4 a totális katasztrófa esete.

A logikai korrektséggel kapcsolatos kifogás alapja az a felvetés, hogy nem tudjuk, a fenti négy szituáció melyikében vagyunk, és ha E_3-ben vagy E_4-ben vagyunk, akkor a nemteljességi demarkáció sikertelen. Tehát nem tudjuk, hogy a nemteljességi demarkáció sikeres-e.

Ha így rekonstruáljuk a 3. premissza elleni kifogást, akkor egy *antiszkeptikus* viszontválasz is adható: a kifogás lényegében egy szkeptikus hipotézis, a szkeptikus hipotéziseket pedig – a szkepticizmusról szóló belterjes viták kontextusán kívül – joggal figyelmen kívül hagyhatjuk. Legalábbis ez a helyzet általában a tudományokban. Képzeljük el, hogy valaki így érvel a darwini evolúcióelmélet ellen:

> Lehet, hogy rendszeresen hallucinálunk vagy helytelenül következtetünk – anélkül, hogy észrevennénk –, mikor fosszíliákat vizsgálunk. Ha ez a helyzet, akkor az evolúcióelmélet gyenge lábakon áll és nem szabad elhinnünk. Ahhoz tehát, hogy az evolúcióelméletet elhiggyük, előbb igazolnunk kell, hogy a kognitív mechanizmusaink megbízhatóak. Mivel ezt nem tudjuk igazolni – és bármely igazolási módszernél felmerülne a kétely, hogy *az* a módszer megbízható-e –, az evolúcióelmélet gyenge lábakon áll, nem szabad elhinni. Nem tudjuk és még csak valószínűnek sem tarthatjuk, hogy az élőlények úgy fejlődtek ki, ahogy a darwinizmus állítja.

Valószínűleg kevesen gondolják, hogy ezzel halálos csapás érte a darwinizmust. Mivel a logikai helyességgel kapcsolatos aggályok ugyanezt a sémát követik, a Lucas/Penrose érvet sincs okunk megkérdőjelezni pusztán amiatt, hogy szkeptikus kétely tárgya lehet. A szkepticizmus olcsó, bárhol működtetni lehet, de emiatt nem kell automatikusan kapitulálnunk előtte. A szkepticizmus – a szisztematikus és destruktív kétely értelmében – filozófiai betegség, amire nincs hatékony gyógyír. De attól, hogy egy betegségre nincs gyógyír, nem kell tudatosan arra törekednünk, hogy elkapjuk.

A Lucas/Penrose érv ellenfele erre azzal vághat vissza, hogy ő teljesen egészséges, és pusztán arra hívja fel a figyelmet, hogy ha a logikánk helytelen, akkor a 3. premissza hamis. Ez nem állítható párhuzamba a darwinizmussal kapcsolatos öncélú szkeptikus kekeckedéssel.

De miért ne lenne párhuzamba állítható vele? Lucas és Penrose szerint azonosítani tudunk Gödel-mondatokat és látjuk róluk, hogy igazak. Ez kétségkívül

csak akkor igaz, ha a logikánk helyes. De ebből nem következik, hogy igazolnunk kell egy további premisszát, miszerint a logikánk helyes. Ha Lucasnak és Penrose-nak szüksége van egy ilyen premisszára, akkor a darwinistának is szüksége is van egy hasonló premisszára. A darwinista szerint jó okunk van azt gondolni, hogy a mai élőlények évmilliókkal korábbi fajokból alakultak ki a természetes szelekció révén. De csak akkor van jó okunk ezt gondolni, ha a látásunk megbízható, ha nem hallucináljuk a fosszíliákat, nem hamisak a DNS-re vonatkozó elméleteink, nem vonunk le belőlük hamis következtetéseket és így tovább. A két eset ismeretelméleti szempontból teljesen analóg.

„A természettudományok esetében joggal hiszünk érzékszerveink és kognitív módszereink megbízhatóságában, mert ez a legegyszerűbb magyarázat arra, miért ilyen fejlett és sikeres a természettudomány."

De Lucas és Penrose is hivatkozhat ugyanerre. A matematika rendkívül sikeres, nem mellesleg az összes természettudomány alapja, tehát a legegyszerűbb hipotézis az, hogy a matematikai képességeink megbízhatóak. Ahogy Penrose írja:

> Nem érzem túlzottan hihetőnek azt a felvetést, hogy a matematikusok valójában egy helytelen formális rendszerre alapozzák a matematikai fogalmaikat és hiteiket. Talán az olvasó is egyetért velem abban, hogy ha ez elvben *lehetséges* is, a legkevésbé sem *hihető*. (Penrose 1994: 140)

Három rekontrát ismertettem arra az ellenérvre, hogy a nemteljességi demarkációhoz előbb igazolnunk kellene a logikánk helyességét. A transzcendentális rekontra szerint ez a racionális gondolkodás szükségszerű posztulátuma; a pragmatikus rekontra szerint a matematikai gyakorlat kiszűri a hibás következtetéseket; az antiszkeptikus rekontra szerint pedig nem jogos az az elvárás, hogy igazoljuk a kognitív mechanizmusaink megbízhatóságát, mielőtt használni kezdjük őket.

Zárásként egy olyan kifogást szeretnék megfogalmazni a Lucas/Penrose érv ellen, amivel nem találkoztam a szakirodalomban. Szerintem a kortárs MI-technológia fényében sokkal célravezetőbb (a demarkációellenes tábor számára) a 2. premisszát tagadni:

(2) Ha x egy MI, akkor létezik egy olyan $\mathcal{A}_x$ rekurzív formális axiómarendszer, hogy az x által igaznak tekintett (matematikai) állítások halmaza $\mathcal{A}_x$ logikai következményeinek részhalmaza.

Az LLM-ek esetében messze nem egyértelmű, hogy (2) igaz. (2) inkább a régimódi, szimbolikus MI paradigmáját idézi, ahol a cél egy-egy tudásterület formális reprezentációja volt, emiatt pedig a matematika esetében valóban konkrét axiómarendszerekről lehetett beszélni. De a mai MI, mint arról a 2. fejezetben szó volt, konnekcionista alapon működik, emiatt pedig nincs egy jól definiálható, axiomatikus tudásbázisa, amiből logikai szabályok segítségével generálná matematikai hiteit.

A Lucas/Penrose-érv híve persze válaszolhatja, hogy egy LLM esetében is értelmeznünk kell a modell által elfogadott állítások halmazát, amennyiben a modellt gondolkodó lénynek akarjuk tekinteni. De amiatt, hogy a modellek explicit reprezentációk helyett a paramétereikből eredő valószínűség-eloszlások segítségével ⌜gondolkodnak⌝, a (2)-es premissza ebben a formában mégiscsak tarthatatlan, vagy legalábbis az LLM-ek belső mechanizmusaival kapcsolatos szofisztikált elmefilozófiai vagy matematikai igazolásra szorul, ennek híján pedig a levegőben lóg.

A nemteljességi demarkációval kapcsolatban az a további kifogás is felmerülhet, hogy nem következik belőle a gépi ⌜gondolkodás⌝ és emberi gondolkodás lényegi különbsége. A vita kedvéért fogadjuk el, hogy mi képesek vagyunk olyan matematikai belátásokra, melyekre a gépek nem. Következik-e ebből, hogy a gépek nem gondolkodnak? Nem világos, miért következne ez. Ha x és y matematikai képességei radikálisan eltérnek, akkor legfeljebb az egyikük lehet gondolkodó lény? A nemteljességi demarkáció pusztán annyit igazol, hogy a matematikai megértés nem gépesíthető. Nem igazolja, hogy az MI-nek nem lehet intellektuális tapasztalata a holland tájképfestészettel (vagy bármilyen, matematikán kívüli témával) kapcsolatban.

Ahhoz, hogy a nemteljességi demarkációt kiterjeszthessük az intellektuális tapasztalat teljes spektrumára, igazolni kellene, hogy ha x intellektuális tapasztalattal rendelkezik valamivel kapcsolatban, akkor x elvben képes felismerni formális rendszerek Gödel-mondatairól, hogy igazak. Ezt a fajta összefonódást a gondolkodás különféle formái közt elég nehéznek tűnik igazolni, emiatt pedig a Lucas/Penrose érv a kognitív demarkáció szempontjából végeredményben haszontalan.

10.

Eldönthetetlenségi demarkáció

Egy kérdés számításelméleti értelemben akkor eldönthető (*decidable*), ha Turing-géppel (vagy általában valamilyen effektív algoritmussal) választ lehet adni rá. Eldönthető például az, hogy hány prímszám van 0 és n között, ahol n egy tetszőleges természetes szám. Vannak azonban eldönthetetlen kérdések is, melyek a kiszámíthatóság elvi korlátait jelzik.

Az eldönthetetlenség leghíresebb példája a megállási probléma (*Halting Problem*). Az első két alfejezetben a megállási problémára épülő demarkációs érvet ismertetem, ami szerintem nagyjából annyira sikeres, mint a nemteljességi demarkáció (vagyis kevéssé). Ezután rátérek az eldönthetetlenség egy ritkábban emlegetett, de a megállási problémánál sokkal érdekesebb esetére, a Rice-tételre, ami véleményem szerint kifejezetten erős demarkációs érvhez vezet. Az érv hatóköre korlátozott, de egy alapvető emberi kognitív képességet érint.

10.1. A megállási probléma

A számítógépes programok végeredményben betűk és más szimbólumok sorozatai. Át lehet őket másolni egyik gépről a másikra, fel lehet tölteni az internetre, ki lehet nyomtatni. Ha egy ilyen szöveget elmentünk, a gép vagy szerver nullák és egyesek sorozataként tárolja. Emiatt minden számítógépes program egy számsorozatnak feleltethető meg.

A programok tipikusan számokkal dolgoznak: számokat fogadnak inputként és számokat generálnak outputként, még akkor is, ha a felhasználók számára az inputok és outputok teljesen más alakot öltenek. Ha rákattintunk egy linkre, ettől pedig elindul egy videó, akkor felhasználói szemszögből kattintás az input és mozgókép az output, de a gép – és az interneten hozzá csatlakozó szerverek – végig számokat manipulálnak. Mikor a linkre kattintunk, a processzor számokként kódolt adatokat kap arról, hogy a képernyő melyik részén kattant az egér, majd ez alapján számokként kódolt üzeneteket küld az internetre, ahonnan szin-

tén számok formájában érkezik meg a hang és mozgókép. Tehát a kattintástól a videó lejátszásáig tartó folyamat számokon végzett műveletekből áll.

Mivel a programok számokkal dolgoznak és maguk is számok sorozatai, egy program inputja lehet egy másik program, vagy akár saját maga. Képzeljünk el egy programot, ami programokat fogad inputként és megállapítja, hogy milyen nyelven – C, JavaScript stb. – íródtak. Mivel a programnyelveknek precízen definiált szótára és szintaxisa van, semmi akadálya, hogy egy szövegfájlról egy programmal megállapítsuk, megfelel-e egy adott programnyelv szabályainak. Tehát a programok képesek más programokat vizsgálni.

Ha P egy program, akkor jelölje $P(i)$ az i input hatására P által generált outputot. $P(i)$ és i egyaránt (természetes) számok, a program tehát lényegében egy függvény, ami számokból számokat csinál. Feltehetnénk azt is, hogy P-nek egynél több inputja van – $P(i_1, i_2)$ stb. –; ez esetben két- vagy többváltozós függvényekről beszélnénk.

Ha P egy program, jelölje $\ulcorner P \urcorner$ a P program kódszámát, vagyis azt a számot, ami P-t mint betűk és szimbólumok sorozatát kódolja. A korábban említett programnyelv-felismerő Q programtól azt várjuk el, hogy bármely P program esetén a $Q(\ulcorner P \urcorner) = n$ input jelezze, hogy P melyik programnyelven íródott (pl. 1 = C, 2 = JavaScript stb.). Általában, ha egy programról el akarjuk dönteni, hogy rendelkezik-e a $\mathscr{T}$ tulajdonsággal, akkor az adott kérdést eldöntő T programtól valami olyasmit várunk, hogy $T(\ulcorner P \urcorner) = 1$ esetén P rendelkezzen $\mathscr{T}$-vel, és $T(\ulcorner P \urcorner) = 0$ esetén P ne rendelkezzen $\mathscr{T}$-vel.

Vajon a programok mely tulajdonságai eldönthetőek ebben az értelemben? Ezt a kérdést az 1930-as években kezdték vizsgálni a matematikusok, és több híres számítástudományi tétel forrása lett. Turing 1936-os cikke, ahonnan a Turing-gépek fogalma származik, két kérdésről is belátta, hogy eldönthetetlen: (i) P véges vagy végtelen sok különböző outputot tud-e produkálni, (ii) van-e olyan input, hogy $P(i)$ tartalmaz egy bizonyos szimbólumot (pl. a 9-es számot). A leghíresebb eldönthetetlen kérdés ezeknél drámaibb és közérthetőbb: lefagy-e P az i inputtól?

A programok mindenki számára ismert tulajdonsága, hogy néha összeomlanak, működésképtelenné válnak, lefagynak. Ennek sok oka lehet, beleértve a hardverhibát. Az egyik tipikus ok az, hogy a program olyan utasításlabirintusba kerül, amiben a végtelenségig kering. A programok általában a programozók figyelmetlensége miatt kerülnek végtelen ciklusba, de szándékosan is írhatunk ilyen programot. Íme egy Turing-gép, ami a $\{0, 1, \text{ÜRES}\}$ szimbólumkészlettel dolgozik és bekapcsoláskor azonnal végtelen ciklusba kerül:

$\mathbf{T}^{\infty}$

Utasítás	*Olvas*	$\langle$*Ír, Mozog*$\rangle$	*Következő utasítás*
1	0	$\langle$0, JOBBRA$\rangle$	1
1	1	$\langle$1, JOBBRA$\rangle$	1
1	ÜRES	$\langle$ÜRES, JOBBRA$\rangle$	1

Bármi szerepel a szalagon, $\mathbf{T}^\infty$ folyamatosan jobbra mozog rajta. Sosem produkál outputot, magától sosem áll le.

A számítástudomány leghíresebb eldönthetetlenségi tétele azt mondja ki, hogy eldönthetetlen, hogy egy tetszőleges P program egy tetszőleges i input hatására produkál-e outputot vagy a végtelenségig szöszmötöl, tehát megáll-e P az i inputon. Ez a *megállási probléma* (*Halting Problem*). A megállási problémát a Turing-gépek kontextusában szokták tárgyalni, de a számítások bármilyen (a Turing-gépekkel ekvivalens) matematikai modelljére érvényes, ezért a továbbiakban egyszerűen csak „programok"-ról fogok beszélni.

Tegyük fel, hogy M egy megállás-detektáló program: bármely P program és i input esetén M nullát ír ki outputként, ha P az i inputon sosem áll le, és 1-et, ha P előbb-utóbb megáll. M tehát egy kétváltozós függvénynek tekinthető:

$$M(\ulcorner P \urcorner, i) = \begin{cases} 0 & \text{ha } P \text{ örökké fut az } i \text{ inputon} \\ 1 & \text{ha } P \text{ egy idő után leáll, vagyis létezik } P(i) \end{cases}$$

Ha a megállási probléma eldönthetetlen, akkor ilyen M program nincs. Ennek kicsit pongyola bizonyításához tegyük fel, hogy M létezik. Ekkor M felhasználásával építhetünk egy unortodox U megállásdetektort, ami a következőt csinálja:

$$U(\ulcorner P \urcorner) = \begin{cases} 0 & \text{ha } P \text{ nem áll le } \ulcorner P \urcorner\text{-n} \\ (\text{végtelen ciklus}) & \text{ha } P \text{ leáll } \ulcorner P \urcorner\text{-n} \end{cases}$$

U a $\ulcorner P \urcorner$ input beolvasása után lefuttatja M-et a $(\ulcorner P \urcorner, \ulcorner P \urcorner)$ inputpáron, és ha nullát kap eredményül, akkor kiírja, hogy „0", ha pedig 1-et, akkor futtatni kezdi $\mathbf{T}^\infty$-t, azaz végtelen ciklusba taszítja magát. Egy ilyen programnak nyilván semmi értelme, de létrehozható, ha M létezik.

Vizsgáljuk meg, mi történik, ha U a saját kódját kapja inputnak:

$$U(\ulcorner U \urcorner) = \begin{cases} 0 & \text{ha } U \text{ nem áll le } \ulcorner U \urcorner\text{-n} \\ (\text{végtelen ciklus}) & \text{ha } U \text{ leáll } \ulcorner U \urcorner\text{-n} \end{cases}$$

Mint látható, U pontosan akkor kerül végtelen ciklusba, ha nem kerül végtelen ciklusba. Ez lehetetlen, tehát U nem létezik. Csakhogy ha M létezik, akkor U is létezik. Emiatt M sem létezik. A megállási probléma tehát eldönthetetlen: nem kiszámítható, hogy egy tetszőleges program egy adott inputon leáll-e.[1]

Felvethetnénk, hogy U léte nem vezet ellentmondásra, amennyiben a megállásdetektor, M, nem mindig áll meg. Ha például M végtelen ciklusba kerül

[1] Gyakori tévképzet, hogy a megállási problémát Turing 1936-os cikke bizonyította be. Turing ott két teljesen más probléma eldönthetetlenségét igazolta (véges sok inputja van-e P-nek, ill. van-e olyan i input, hogy $P(i)$ tartalmaz egy adott szimbólumot). A megállási probléma egy speciális esetét futólag belátta Kleene (1952: 382), de maga a „megállási probléma" mint kifejezés, valamint a tétel ma ismert változata Martin Davistől ered (Davis 1958: 70).

attól a kérdéstől, hogy U megáll-e $\ulcorner U \urcorner$-n, vagyis ha $M(\ulcorner U \urcorner, \ulcorner U \urcorner)$ nem 1 és nem is 0, hanem M ilyenkor végtelen ciklusba kerül, akkor az $U(\ulcorner U \urcorner)$ számításnak sem lesz eredménye.

A formális ellentmondást ezzel valóban kiküszöböljük, de a megállási probléma szempontjából ennek nincs jelentősége. Hiszen ez esetben M nem tudja *bármelyik* programról megállapítani, hogy egy adott inputtól megáll-e. Azt sem mondhatjuk, hogy U-n kívül viszont bármelyik program viselkedését megjósolja, mert végtelen sok U-hoz hasonló program létezik, ha M létezik.

10.2. Megállási demarkáció

Roger Penrose a Gödel-tétel mellett a megállási problémára is épített egy demarkációs érvet. Tegyük fel, hogy az M program helyesen modellezi a matematikusok tudását arról, hogy a programok mely inputokon állnak le és melyeken nem. Penrose arra hívja fel a figyelmet, hogy ekkor az előző alfejezetben szereplő U program esetében mi látjuk, hogy $M(\ulcorner U \urcorner, \ulcorner U \urcorner)$ végtelen ciklust eredményez, M ezen az inputon nem áll le, de M ezt nyilván nem tudja megállapítani. Tehát M nem modellezi helyesen a matematikusok tudását – a matematikai gondolkodás nem gépesíthető (Penrose 1994: 73–75).

Penrose itt hasonló stratégiát követ, mint a nemteljességi demarkáció esetében. Ha a gondolkodásunk gépesíthető, akkor léteznie kell egy olyan programnak, ami a matematikusokkal ekvivalens módon viselkedik. De ha lenne ilyen M program, akkor lenne olyan matematikai tény, amit mi felismernénk, M pedig nem.

Akárcsak a nemteljességi demarkáció esetében, itt is felmerül az a kifogás, hogy igazolnunk kellene logikai rendszerünk helyességét, mielőtt a programok megállásáról szóló ítéleteinkben, így az M megállásával kapcsolatos ítéletekben, megbízhatunk. Erre az ellenérvre ugyanolyan transzcendentális, pragmatikus és antiszkeptikus válaszok adhatóak, mint a 9. fejezetben.

Érdekes módon a megállási demarkációnál (szemben a nemteljességi demarkációval) az LLM és a szimbolikus MI különbsége nem jelent problémát, mert itt eleve programokról van szó. Teljesen mindegy, M hogyan reprezentálja a világot; az a fontos, hogy nem lát bizonyos tényeket, amiket mi látunk.

A megállási demarkációt azonban végeredményben ugyanaz a probléma fenyegeti, mint a nemteljességi demarkációt. A vita kedvéért tegyük fel, hogy a konklúzió igaz, mi emberek valóban látunk bizonyos matematikai tényeket, amiket a programok nem. Következik-e ebből, hogy az MI nem gondolkodik, nincs intellektuális tapasztalata? Nem világos, miért következne ez.

10.3. Rice-demarkáció

A fejezet hátralevő részében az eldönthetetlenségi demarkáció egy másik változatát dolgozom ki. Ez a változat a Rice-tételre épül, ami a számítástudományon kívül viszonylag ismeretlen, pedig tartalma jóval közérthetőbb és az MI szempontjából sokkal relevánsabbnak tűnik. A Rice-tétel kicsit pongyolán megfogalmazva azt mondja ki, hogy Turing-géppel nem lehet eldönteni, hogy egy Turing-gép mit csinál. Szabatosabban, nincs olyan Turing-gép, ami egy tetszőleges T Turing-gépről el tudja dönteni, hogy T az f függvényt számolja-e ki (ahol f a természetes számokon értelmezett tetszőleges parciális függvény). Nincs például olyan Turing-gép, ami el tudja dönteni, hogy az inputként kapott program az összeadásfüggvényt számítja-e ki. Mivel a programok végeredményben semmi mást nem csinálnak, mint függvényeket számítanak ki, ez azt jelenti, hogy programok képtelenek megállapítani, hogy a programok mit csinálnak. Ha *mi* képesek vagyunk erre, akkor az emberi elme nem egy program.

Ez az érv sem általában a gondolkodást vitatja el a gépektől, de a nemteljességi és megállási demarkációnál jóval alkalmasabb rá, hogy a szűk matematikai gondolkodáson túlmutató különbséget mutasson ki az emberi értelem és az MI között. Legalábbis emellett fogok érvelni. Először a Rice-tételt próbálom precízen kimondani.

T_1 és T_2 Turing-gépek *funkcionálisan ekvivalensek*, ha ugyanazt a függvényt számítják ki, vagyis bármely i inputra $T_1(i) = T_2(i)$.[2] Nevezzük funkcionálisan ekvivalens Turing-gépek bármely halmazát *egyszerű* funkcionális típusnak. Turing-gépek egy $\mathcal{F}$ halmaza akkor funkcionális típus, ha (i) $\mathcal{F}$ egy egyszerű funkcionális típus (vagyis van olyan φ parciális függvény, hogy $\mathcal{F}$ a φ-t kiszámító Turing-gépek halmaza), vagy (ii) $\mathcal{F}$ funkcionális típusok uniója. $\mathcal{F}$ akkor *nemtriviális*, ha nem üres és nem azonos az összes Turing-gép halmazával.

A Rice-tétel azt mondja ki, hogy a nemtriviális funkcionális típusokat nem lehet Turing-géppel azonosítani: ha $\mathcal{F}$ egy nemtriviális funkcionális típus, akkor nincs olyan Turing-gép, hogy $F(\ulcorner T \urcorner) = 1$ bármely $T \in \mathcal{F}$ esetén, és $F(\ulcorner T \urcorner) = 0$ egyébként.

A Rice-tétel bizonyításához tegyük fel, hogy F létezik. Mivel $\mathcal{F}$ nemtriviális funkcionális típus, tehát nem üres és a komplementere sem üres, létezik olyan A és B gép, hogy $A \in \mathcal{F}$ és $B \notin \mathcal{F}$. F felhasználásával építsünk egy paradox P gépet, ami a következőt csinálja: az i input észlelésekor P megkérdezi F-től, hogy ő maga (azaz P) $\mathcal{F}$ eleme-e. Ha F igennel felel (azaz $F(\ulcorner P \urcorner) = 1$), akkor P a $B(i)$ számot adja outputnak, ha pedig F nemmel felel ($F(\ulcorner P \urcorner) = 0$), akkor P az $A(i)$ számot adja outputnak. Vagyis ha F szerint P az $\mathcal{F}$ típusba tartozik, akkor P úgy viselkedik, mintha nem oda tartozna (mivel $B \notin \mathcal{F}$), ha pedig F szerint P *nem* az $\mathcal{F}$ típusba tartozik, akkor P úgy viselkedik, mintha

[2] Mivel a Turing-gépek nem mindig adnak outputot, $T_1(i) = T_2(i)$ szigorúan véve nem igaz, ha T_1 és T_2 nem áll le az i inputon. Hogy ezt elkerüljük, a funkcionális ekvivalenciát definiálhatjuk úgy is, hogy nincs olyan i, melyre $T_1(i) \neq T_2(i)$.

oda tartozna (mert $A \in \mathcal{F}$). Ez ellentmond annak, amit F-ről feltettünk (hogy felismeri $\mathcal{F}$ elemeit), tehát F nem létezik. Általában: a nemtriviális funkcionális típusok Turing-géppel azonosíthatatlanok.

A bizonyítás kihagyott egy fontos kérdést: hogyan tudja P megkérdezni F-et arról, hogy P eleme-e $\mathcal{F}$-nek? A legkézenfekvőbb megoldás az, ha P felírja a szalagjára a saját kódját, $\ulcorner P \urcorner$-t, elkezdi futtatni F-et, a végén ellenőrzi, hogy 0 vagy 1 az eredmény, majd ennek függvényében lefuttatja A-t vagy B-t a szalagon megőrzött eredeti i inputon. Azt, hogy létezik ilyen P, az úgynevezett rekurziótétel segítségével lehet igazolni.[3]

A Rice-tételre épülő *naiv* demarkációs érvhez vegyünk egy nemtriviális funkcionális típust, mondjuk az $f(x) = x + 1$ függvényt kiszámító programok halmazát.

(1) Mi emberek képesek vagyunk eldönteni, hogy egy tetszőleges Turing-gép az $f(x) = x + 1$ függvényt számítja-e ki.

(2) Nincs olyan, Turing-gépet implementáló fizikai rendszer, ami képes eldönteni, hogy egy tetszőleges Turing-gép az $f(x) = x+1$ függvényt számítja-e ki.

(3) Tehát mi nem vagyunk Turing-gépet implementáló fizikai rendszerek.

Mivel a 2. premissza a Rice-tétel következménye, az ellendrukkernek az 1. premisszát kell támadnia. A legkézenfekvőbb ellenérv szerint semmi sem igazolja, hogy *bármelyik* programról el tudnánk dönteni, hogy az $f(x) = x+1$ függvényt számítja-e ki.

Ennek az ellenérvnek van egy általános és egy konkrétabb változata. Az általános változat a korábban látott szkeptikus aggály aktuális inkarnációja: honnan tudjuk, hogy helyesen gondolkodunk? Lehet, hogy egy program az $x+1$ függvényt komputálja, de mi hibásan arra következtetünk, hogy valami mást csinál. Akárcsak a 9.2. és 10.2. alfejezetben, itt is adhatunk transzcendentális, pragmatikus vagy antiszkeptikus választ az aggályra.

[3] Ld. pl. Rogers (1987:180). A rekurziótétel azt mondja ki, hogy bármely f rekurzív (Turing-géppel kiszámítható) függvény esetében létezik olyan n szám, hogy $\varphi_n = \varphi_{f(n)}$, ahol φ_k a k kódszámú parciális rekurzív függvény. A saját kódját a szalagra író gép létezésének bizonyításához legyen f az a függvény, ami az n kódú géphez hozzárendeli annak a Turing-gépnek a kódját, ami az i inputot megkapva ráírja a saját szalagjára n-t, lefuttatja F-et, majd aszerint, hogy 0 vagy 1 az eredmény, $A(i)$-t vagy $B(i)$-t adja outputként.

A Rice-tételt Henry Gordon Rice eredetileg a rekurzív függvényekről szóló tételként mondta ki, konkrétan azt igazolta, hogy természetes számok rekurzívan felsorolható halmazainak egyetlen nemtriviális halmaza sem rekurzívan felsorolható (Rice 1953). A tételt modern tankönyvek általában úgy fogalmazzák meg, hogy Turing-gépek egyetlen „szemantikai tulajdonsága" sem eldönthető (pl. Hopcroft, Motwani & Ullman 2007: 398, Sipser 2013: 241). Én szemantikai tulajdonságok helyett funkcionális tulajdonságokról beszélek, mert a „szemantika" szó filozófiai kontextusban általában a természetes nyelvi jelentésre, nem a Turing-gépek által elfogadott inputokra vonatkozik.

Az ellenérv *konkrétabb* változata az emberi kognitív képességek gyakorlati korlátaira hívja fel a figyelmet. A szkepticizmust félretéve ismerjük el, hogy képesek vagyunk programokat elemezni, például egy 100 soros programról meg tudjuk állapítani, hogy az $x+1$ függvényt számítja-e ki (ilyen feladatokat rendszeresen kapnak informatikus hallgatók). Ha ez igaz is, nyilván nem tudjuk az *összes lehetséges* programról eldönteni, hogy melyik függvényt számítja ki. Képzeljünk el egy több kvadrillió utasításból álló programot, ami az i input beolvasása után elkezd iszonyatosan bonyolult differenciálegyenleteket számolni, majd felfoghatatlanul sok lépés után kiírja az $i+1$ számot. Tegyük fel, hogy ez annyira nyakatekerten következik az inputból, hogy a világ legjobb informatikusaiból álló konzorcium sem jönne rá, mit művel a program, még akkor sem, ha évezredeken át vizsgálnák. Mivel az emberi funkció-detektáló képességnek is van határa, az 1. premissza hamis.

Ezt az ellenérvet a Rice-tétel egy hosszkorlátozott változatával lehet semlegesíteni. Ha T egy Turing-gép, jelölje $|T|$ a gép hosszát. Ezt mérhetjük az instrukciók számával vagy a T kódolásához szükséges szöveg hosszával; teljesen mindegy. A kérdés az, hogy ha hosszkorlátozott funkciódetektorokról beszélünk, azaz olyan F gépekről, melyek csak $|T| < k$ esetén képesek eldönteni, hogy T az $\mathcal{F}$ halmaz eleme-e, akkor sikertelenné válik-e a Rice-demarkáció. A mögöttes intuíció az, hogy mivel egy ember nem képes tetszőlegesen hosszú programokat vizsgálni, csak hosszkorlátozott Turing-gépek implementációi lehetünk, emiatt pedig az eredeti (hosszkorlátozást nem tartalmazó) Rice-tételre nem lehet demarkációs érvet építeni olyan módon, ahogy (1)–(3) teszi.

A mögöttes intuíció nem alaptalan, mert ha F hosszkorlátozott, akkor a Rice-tétel eredeti bizonyítása érvénytelenné válik: a paradox P gép létéből többé nem az következik, hogy F nem létezik; csak az következik belőle, hogy vagy F nem létezik, vagy P meghaladja F számítási korlátait, azaz $|P| \geq k$.

Amellett fogok érvelni, hogy a demarkáció ennek ellenére hosszkorlátozott kontextusban is működik, ha olyan funkciódetektorokra koncentrálunk, melyek rendelkeznek az alábbi képességgel:

Moduláris Megértés Az F funkciódetektáló Turing-gép rendelkezik a moduláris megértés képességével, ha bármely T_1, T_2, T_3 gép esetében, melyekről F meg tudja állapítani, hogy $\mathcal{F}$-be tartoznak-e, F meg tudja ezt állapítani arról a T_* gépről is, ami a következő műveleteket végzi az i inputon:
1. T_* ráírja a szalagjára a saját kódját
2. T_* lefuttatja a saját kódján T_1-et
3. Ha az eredmény 0, akkor T_* a $T_2(i)$ outputot adja vissza
4. Ha az eredmény 1, akkor T_* a $T_3(i)$ outputot adja vissza

Nevezzünk egy $\mathcal{F}$ funkcionális típust *k-nemtriviális*nak, ha vannak k-nál rövidebb gépek $\mathcal{F}$-en belül és kívül is. A továbbiakban koncentráljunk k-nemtriviális funkcionális típusokra.

Könnyen belátható, hogy ekkor nem létezik olyan F, amely (i) bármely $|T| < k$ gépről el tudja, dönteni, hogy $\mathcal{F}$-be tartozik-e, (ii) rendelkezik a moduláris megértés képességével, és (iii) k-nál rövidebb, azaz $|F| < k$. Ha ugyanis létezik ilyen F, akkor ugyanúgy ellentmondásra jutunk, mint az eredeti Rice-tétel esetében. Legyen $T_1 = F$, $T_2 = A$, $T_3 = B$, ahol $A \in \mathcal{F}$ és $B \notin \mathcal{F}$, $|A| < k, |B| < k$. Mivel $|F| < k$, a moduláris megértés kritériuma vonatkozik erre a trióra, és azt kapjuk, hogy egyrészt $F(|T_*|) = 0$ vagy $F(|T_*|) = 1$, másrészt $F(|T_*|) = 0$ akkor és csak akkor, ha $F(|T_*|) = 1$, ami lehetetlen. Tehát nincs olyan hosszkorlátozott funkciódetektor, ami rendelkezik a moduláris megértés képességével és rövidebb a saját hosszkorlátjánál, k-nál.

Nevezzük (pusztán a rövidség kedvéért) „funkcionalistának" azt a filozófust, aki az emberi elmét implementált Turing-gépnek, vagy egy Turing-géppel ekvivalens számítási rendszernek tekinti. Legyen E a legrövidebb olyan Turing-gép, ami funkcionálisan ekvivalens az emberi elmével (vagy a legrövidebbek egyike, ha több minimális jelölt van). Az eredeti Rice-demarkáció, azaz (1)–(3) blokkolása érdekében feltesszük, hogy E egy hosszkorlátozott funkciódetektor, vagyis E csak k-nál rövidebb programok funkcionális tulajdonságairól tud megbízhatóan nyilatkozni.

Két eset lehetséges: (a) $|E| \geq k$, vagy (b) $|E| < k$. Az (a) esetben az emberi elme nem alkalmas az emberi elme funkcionális tulajdonságainak vizsgálatára, hiszen E hossza meghaladja E korlátját. A funkcionalista öngólt rúg ezzel a feltevéssel, hiszen ekkor azt állítja, hogy elménk megértése meghaladja kognitív képességeinket, tehát a funkcionalizmus funkcionalista alapon igazolhatatlan.

A (b) esetben a Rice-tétel hosszkorlátozott változatából következik (tetszőleges k-nemtriviális $\mathcal{F}$ funkcionális típust alapul véve), hogy nem rendelkezünk a moduláris megértés képességével. De nyilván rendelkezünk vele: ha értjük a T_1, T_2, T_3 programok működését, akkor értjük a „Ha $T_1(n) = 0$, akkor T_2, egyébként T_3" típusú program működését tetszőleges n-re, így a szóban forgó programot kódoló n-re is. A hosszkorlátozás tehát nem ássa alá a Rice-demarkációt, hanem arra kényszeríti a funkcionalistát, hogy válasszon két lehetőség közül: az elme vagy nem egy implementált Turing-gép, vagy nem tudhatjuk, hogy az.

10.4. A Rice-demarkáció előnyei

A Rice-tétel véleményem szerint sokkal sokkal hatékonyabb demarkációs eszköz, mint a Gödel-tétel vagy a megállási probléma, mert nem obskúrus matematikai belátásokat vitat el a gépektől, hanem egy mindenki által ismert és léptennyomon használt képességet, a funkcionális tulajdonságok felismerésének képességét. Az eszközöket input/output összefüggések jellemzik: a mikrohullámú sütő például hideg ételekből mint inputból felmelegített ételeket állít elő outputként. Mivel az eszközöket „ha/akkor" szabályokkal leírható oksági összefüggések definiálják, minden eszközt megfeleltethetnünk egyfajta algoritmusnak. Ha ez

a megfeleltetés kellően precízzé tehető, akkor a Rice-demarkáció alapján az MI általában az eszközöket sem képes felismerni, nem tudja, hogy a tárgyak mire valók.

Felvethetnénk, hogy a Rice-demarkáció túlságosan *absztrakt* egy ilyen tézis igazolásához, sőt, még a számítógépes programok informatikailag lényeges funkcionális típusaira sem vonatkozik. A Rice-tételből ugyanis nem következik, hogy a programok praktikusan fontos funkcionális tulajdonságai algoritmikusan felismerhetetlenek lennének. Elképzelhető, hogy egy megfelelően kalibrált LLM hatékonyan – és az embernél sokkal hatékonyabban – tudja a gyakorlatban fontos funkcionális típusokba sorolni a ténylegesen létező programokat. Mivel a Rice-tétel bizonyítása egy praktikusan haszontalan, paradox programra épül, ezt a lehetőséget nem zárja ki.

Elvben valóban nem kizárható, hogy létezhet mágikus funkciófelismerő MI, ami az összes praktikus igényt kielégíti. A pragmatikus ellenérvre csak pragmatikus választ lehet adni: jelenleg semmi jele, hogy számítógépes programok képesek lennének megbízhatóan azonosítani számítógépes programok funkcionális tulajdonságait, hiszen ha képesek lennének rá, akkor az appok és operációs rendszerek mindig azt csinálnák, amit készítőik szándékai szerint csinálniuk kellene, és nem omlanának össze időről időre programozási hibák miatt. Bár a Rice-tétel szigorú matematikai értelemben nem cáfolja, hogy létezhet gyakorlati célokra tökéletesen alkalmas funkciódetektáló MI, alapot ad az ezzel kapcsolatos kételyre annak fényében, hogy nincs (pedig rengeteg pénzt megérne) egy olyan algoritmus, ami hatékonyan detektálná a jelenleg használt programok működési szabályait. A pragmatikus ellenérv emiatt még pragmatikus szinten sem cáfolja a Rice-demarkációt.

Ami a Rice-tételnek az eszközökre való kiterjesztését illeti: valóban nincs szigorú matematikai kapcsolat a programok és általában az eszközök között, de nyilván van köztük analógia, és egyelőre annak sincs jele, hogy az MI-k hatékonyan felismernék fizikai tárgyak funkcionális tulajdonságait, ne pedig külső jegyek (például a tárgyak külalakjával kapcsolatos statisztikai adatok) alapján kategorizálnák őket.

A pragmatikus szempontoktól elszakadva itt is felvethető, hogy ha a szóban forgó demarkációs érv netán sikeres is, nem következik belőle, hogy az MI emberi értelemben nem gondolkodik, nem rendelkezik intellektuális tapasztalattal. Tehát a Rice-demarkáció nem kognitív demarkációs érv az 1. fejezetben definiált értelemben. Csupán annyit igazolhat (ha amúgy helyes), hogy az MI nem tud megoldani bizonyos feladatokat, melyeket mi meg tudunk.

Ez az ellenérv részben jogos, részben téves. Téves annyiban, hogy az emberek esetében a funkcionális tulajdonságok felismerése vélhetőleg sajátos intellektuálist tapasztalaton alapul: *látjuk* – a program struktúráját vizsgálva –, hogy a program mit csinál, és ez alapján ismerjük fel, hogy ezt vagy azt a függvényt számolja ki. Tehát az MI esetében a funkciódetektálás képességének hiánya egyben az intellektuális tapasztalat egy fajtájának a hiányát is jelen-

ti. Az ellenérv viszont annyiban jogos, hogy ettől függetlenül lehet az MI-nek (más kérdésekkel kapcsolatban) intellektuális tapasztalata; a Rice-demarkáció valóban korlátozott érvényű. De sokkal kevésbé korlátozott érvényű, mint a nemteljességi és megállási demarkáció, és összefüggésbe hozható az emberi értelem egyik egészen alapvető megnyilvánulásával, az eszközhasználattal.

11.

Az önreprezentáció elátkozott szentháromsága

A következő két fejezet a formális nyelvek önreprezentációs korlátait próbálja felhasználni annak igazolására, hogy az MI – szemben az emberrel – nem tudja reprezentálni önmagát, pontosabban saját reprezentációs tevékenységét. Az így keletkező demarkációs érv az önreflexió egyik lényeges aspektusát tagadja meg a gépektől; azt, hogy képesek lennének saját magukat gondolkodó lényként elgondolni. Először egy logikai állítást bizonyítok, ami azt mondja ki, hogy egyetlen (elsőrendű) nyelv sem tudja egyszerre reprezentálni saját szintaktikai tulajdonságait, szemantikai tulajdonságait, valamint azt a tényt, hogy reprezentálja saját tulajdonságait („metareprezentáció"). Ezt az inkompatibilitást nevezem az önreprezentáció elátkozott szentháromságának. Mielőtt rátérek a formális kifejtésre, megpróbálom felvázolni az állítás alapgondolatát.

Az emberi nyelv képes reprezentálni saját szintaktikai tulajdonságait. Szintaktikainak nevezek minden olyan tulajdonságot, ami kizárólag a mondatok mint jelsorozatok felépítésével kapcsolatos. Az írott magyar nyelv egyes mondatainak szintaktikai tulajdonsága például az, hogy 6-nál többször szerepel bennük az „e" betű. A magyar nyelv képes kifejezni az ilyen tulajdonságokat; például ki tudja fejezni azt a tényt, hogy az előző mondatban 6-nál többször szerepel az „e" betű. Általában, a magyar nyelv (az összes többi természetes nyelvhez hasonlóan) képes azonosítani a nyelv kifejezéseit alkotó szimbólumokat, leírást tud adni bármelyik mondat formális struktúrájáról. Ezt a jelenséget *szintaktikai önreprezentáció*nak hívom.

Mikor arról beszélek, hogy a magyar nyelv kifejez valamilyen tulajdonságot vagy képes leírást adni valamiről, nem arra gondolok, hogy a Nyelv mint olyan kifejez vagy reprezentál valamit. Ezek metaforák, melyek nyelvtanilag komplikált állításokat rövidítenek, azt az állítást például, hogy a magyar nyelvet használó írástudó felnőttek ki tudják fejezni azt a tényt, hogy egy magyar mondat írott változatában 6-nál többször szerepel az „e" betű. A nyelv „ké-

pességei" kapcsán tehát saját reprezentációs tevékenységünkről van szó. De ezt hosszadalmas lenne minden alkalommal precízen megfogalmazni, ezért a tömörség kedvéért végig metaforákat használok.

A nyelvek képesek *szemantikai* tulajdonságokat is kifejezni. A „szemantika" itt a mondatok és nyitott formulák extenziójával kapcsolatos tényeket jelenti, azaz mondatok esetében az igazságértékekre vonatkozó tényeket, nyitott formulák esetében pedig a formulát igazzá tevő dolgok halmazaival kapcsolatos tényeket. Például az „x egy alma" nyitott formula egyik szemantikai tulajdonsága, hogy az almákra vonatkozik. Később precízebben is definiálom majd a szemantikai tulajdonságokat (mint formulák koextenzivitásra zárt halmazait), de már az előbbi elnagyolt meghatározás is egyértelművé teszi, hogy a magyar nyelv (a többi természetes nyelvhez hasonlóan) ki tud fejezni szemantikai tulajdonságokat. Például ki tudja fejezni azt a tényt, hogy az „alma" az almákat jelöli, a „jerke" pedig fiatal nőstény bárányokat és kecskéket jelöl (precízen: az „x egy jerke" nyitott formula extenziója a fiatal nőstény bárányok és kecskék halmaza). Ezt a jelenséget *szemantikai önreprezentáció*nak nevezem.

A nyelv képes reprezentálni azt is, hogy reprezentálja saját magát. Az a magyar kifejezés például, hogy „a szó, ami a magyar ABC harmadik, első, harmadik és első betűjéből áll (ebben a sorrendben)" a „baba" szót jelöli. Az előző mondat pedig kifejezte a tényt, hogy ezt jelöli. Tehát a magyar nyelv (és általában minden természetes nyelv) képes reprezentálni azt, hogy reprezentálja saját magát. Ezt a jelenséget *metareprezentáció*nak hívom.

Az önreprezentáció elátkozott szentháromságának tétele szerint egyetlen logikailag koherens nyelv sem rendelkezhet a fenti három önreprezentációs tulajdonsággal. Ezt egy olyasfajta paradox kifejezés megszerkesztésével igazolom, mint „ha ez a kifejezés az almákra vonatkozik, akkor x nem egy alma, és ha nem az almákra vonatkozik, akkor x egy alma" – ami pontosan akkor vonatkozik az almákra, ha nem vonatkozik rájuk. A 12. fejezetben amellett érvelek majd, hogy a magyar nyelvet (a metaforát feloldva: a magyar nyelvet használó emberek gondolkodását) a „metaszintek" léte menti meg az inkoherenciától, és a metaszintek vezetnek majd kognitív demarkációhoz.

11.1. Az önreprezentáció szentháromsága

Legyen $\mathscr{F}$ egy (elsőrendű) formális nyelv, melynek szimbólumkészlete logikai konstansokat ($\neg$, $\exists$, $\forall$, $\wedge$, $\vee$, $\rightarrow$, $\equiv$), tagoló írásjeleket (zárójel, kettőspont), változójeleket ($x, y, z, v, z, w, x_1, y_1, z_1 \ldots$), individuális konstansokat (individuumokat jelölő primitív neveket) ($a, b, c \ldots$), függvényjeleket ($f, g \ldots$) és primitív predikátumokat ($F, G \ldots$) tartalmaz. $\mathscr{F}$ egy *kifejezése* („term") bármely (a) változójel, (b) individuális konstans, (c) $f(t_1 \ldots t_n)$ alakú szimbólumsorozat, ahol f egy n változós függvényjel, $t_1 \ldots t_n$ pedig kifejezések. $\mathscr{F}$ *formulája* bármely (i) $F(t_1 \ldots t_n)$ alakú jelsorozat, ahol F egy n változós primitív predikátumjel,

$t_1 \dots t_n$ pedig kifejezések, (ii) bármely $\forall x : \varphi$ vagy $\exists x : \varphi$ alakú jelsorozat, ahol x tetszőleges változójel, φ pedig egy formula (az így keletkező formulákban az x változó *kötött*), (iii) formulák tetszőleges (véges) konjunkciója ($\wedge$), diszjunkciója ($\vee$) vagy kondicionális összekapcsolása ($\rightarrow$, $\equiv$). Egy formulában az x változó akkor *szabad*, ha x nem kötött. $\mathscr{F}$ *mondatai* a szabad változót nem tartalmazó formulák.

Az $\mathscr{F}$ nyelvet használó *modellek* két komponensből, egy U halmazból és egy I interpretációból állnak. U a modell „univerzuma": azoknak a dolgoknak a halmaza, melyekről a modell (a nyelv szóban forgó interpretációja szerint) állításokat tesz. Az I interpretáció egy függvény, ami az individuális konstansokhoz az univerzum elemeit, az n változós függvényjelekhez U-n értelmezett n változós függvényeket, az n változós predikátumjelekhez pedig az U elemeiből álló n elemű sorozatok halmazait rendeli. I alapján rekurzívan definiálható a mondatok igazsága ($\Leftrightarrow$ = akkor és csak akkor, ha):

(1) Ha a t kifejezés az a individuális konstans, akkor $I(t) = I(a)$. Ha t „$f(t_1 \dots t_n)$" alakú, akkor $I(t) = I(f)(I(t_1) \dots I(t_n))$.

(2) Az $F(t)$ atomi mondat (ahol F egy predikátumjel és t egy kifejezés) akkor igaz az $\mathcal{M}(U, I)$ modellben (röviden: $\mathcal{M} \models F(t)$), ha $I(t) \in I(F)$. Hasonlóan a többváltozós predikátumjelekre.

(3) Ha a φ mondat „$\neg\psi$" alakú, akkor $\mathcal{M} \models \varphi \Leftrightarrow \mathcal{M} \not\models \psi$.

(4) Ha a φ mondat „$\psi \vee \chi$" alakú, akkor $\mathcal{M} \models \varphi \Leftrightarrow \mathcal{M} \models \psi$ vagy $\mathcal{M} \models \chi$.

(5) Ha a φ mondat „$\psi \wedge \chi$" alakú, akkor $\mathcal{M} \models \varphi \Leftrightarrow \mathcal{M} \models \psi$ és $\mathcal{M} \models \chi$.

(6) Ha a φ mondat „$\psi \rightarrow \chi$" alakú, akkor $\mathcal{M} \models \varphi \Leftrightarrow \mathcal{M} \models (\neg\psi) \vee \chi$.

(7) Ha a φ mondat „$\psi \equiv \chi$" alakú, akkor $\mathcal{M} \models \varphi \Leftrightarrow \mathcal{M} \models (\psi \rightarrow \chi) \wedge (\chi \rightarrow \psi)$.

A kvantifikált mondatok („valamely x-re …", „minden x-re …") szemantikájához szükség van a *kiértékelés*ek fogalmára. Kiértékelés bármely olyan függvény, ami a változójelek ($x, y, z \dots$) halmazáról az univerzum elemeire képez, vagyis lényegében individuális konstansokként kezeli a változójeleket. Ha az I interpretációt egy K kiértékeléssel kombináljuk (I^K), akkor a változókat tartalmazó kifejezéseknek és a nyílt formuláknak is szemantikai értéket tudunk tulajdonítani a fentiekkel analóg módon:

(8) Ha a t kifejezés az a individuális konstans, akkor $I^K(t) = I(a)$. Ha t az x változójel, akkor $I^K(t) = K(x)$. Ha t „$f(t_1 \dots t_n)$" alakú, akkor $I^K(t) = I^K(f)(I^K(t_1) \dots I^K(t_n))$.

(9) Az $F(t)$ atomi formula (ahol F egy predikátumjel és t egy kifejezés) akkor igaz $\mathcal{M}$-ben a K kiértékelés mellett ($\mathcal{M}^K \models F(t)$), ha $I^K(t) \in I(F)$. Hasonlóan a többváltozós predikátumjelekre.

(10) (... A 10–14. szabály megegyezik a 3–7. szabállyal, csak $\mathcal{M}$ helyett $\mathcal{M}^K$, „mondat" helyett pedig „formula" szerepel bennük...)

(15) Ha a φ formula „$\exists x : \psi(x)$" alakú, akkor $\mathcal{M}^K \models \varphi \Leftrightarrow$ van olyan L kiértékelés, hogy $\mathcal{M}^L \models \psi(x)$.

(16) Ha a φ formula „$\forall x : \psi(x)$" alakú, akkor $\mathcal{M}^K \models \varphi \Leftrightarrow$ bármely L kiértékelésre $\mathcal{M}^L \models \psi(x)$.

A kiértékelések segítségével definiálható a kvantifikált állítások igazsága az eredeti (nem egy adott kiértékelésre relativizált) modellben:

(17) Ha a φ mondat „$\exists x : \psi(x)$" alakú, akkor $\mathcal{M} \models \varphi \Leftrightarrow$ van olyan K kiértékelés, hogy $\mathcal{M}^K \models \psi(x)$.

(18) Ha a φ mondat „$\forall x : \psi(x)$" alakú, akkor $\mathcal{M} \models \varphi \Leftrightarrow$ bármely K kiértékelésre $\mathcal{M}^K \models \psi(x)$.

A továbbiakban kis görög betűkkel (α, β, φ stb.) $\mathscr{F}$ formuláit fogom jelölni. Ezek a szimbólumok nem részei $\mathscr{F}$-nek, csak a róla szóló metanyelvnek. Az $\alpha(x)$ jelölés arra utal, hogy az α formulában az x az egyetlen szabad változó. A görög betűk mellett időnként `ILYEN` nagybetűs szavakkal jelölöm $\mathscr{F}$ egyes formuláit. Ezek a szavak ugyanazt a metanyelvi funkciót töltik be, mint a görög kisbetűk, de utóbbiakkal ellentétben közvetlenül utalnak a szóban forgó formulák interpretációjára, tehát remélhetőleg segítik a megértést.

Legyen Σ az $\mathscr{F}$ nyelv szimbólumaiból képzett véges sorozatok halmaza és tegyük fel, hogy létezik egy injektív $R : \Sigma \to U$ függvény, azaz olyan függvény, melyre $s_1 \in \Sigma$, $s_2 \in \Sigma$, $s_1 \neq s_2$ esetén $R(s_1) \neq R(s_2)$. A kibővített $\mathcal{M}(U, I, R)$ modellt *impliciten önreprezentáló*nak nevezem, R outputjait pedig implicit jelölőknek: ha $u = R(\varphi)$ (ahol $u \in U$, $\varphi \in \Sigma$), akkor u implicit jelölete φ, és φ implicit jelölője u.

Az $\mathcal{M}(U, I, R)$ impliciten önreprezentáló modell *felismeri jelsorozatait*, ha minden $s \in \Sigma$-re létezik a nyelvben egy olyan $\texttt{ID}[s](x)$ formula (az s jelsorozatot azonosító leírás), amit kizárólag s implicit jelölője, azaz $R(s)$ tesz igazzá:

$$I(\texttt{ID}[s](x)) = \{R(s)\}$$

$\mathcal{M}(U, I, R)$ *kifejezi a konkatenációt*, ha minden $n \in \mathbb{N}$-re létezik $\mathcal{M}$-ben egy olyan $\texttt{CON}(x_1, \ldots x_n, y)$ formula, ami pontosan akkor igaz, ha y implicit jelölete az $x_1, \ldots x_n$ implicit jelöleteinek konkatenációja (a megadott sorrendben). Precízebben ($s_1 \oplus s_2$ jelöli az s_1 és s_2 szimbólumsorozatok konkatenációját):

$$I(\texttt{CON}(x_1, \ldots x_n, y)) = \\ = \{\langle u_1 \ldots u_n, u_{n+1}\rangle \in U^{n+1} : R^{-1}(u_{n+1}) = R^{-1}(u_1) \oplus \ldots \oplus R^{-1}(u_n)\}$$

Valójában formulák egész családjáról van itt szó, mivel minden $n \geq 2$-re kell egy külön $\texttt{CON}^n$ formula az n elemű konkatenáció kifejezéséhez. Az egyszerűség kedvéért ezt nem jelölöm külön.

$\mathcal{M}(U, I, R)$ *szintaktikailag önreprezentáló*, ha felismeri a jelsorozatait és kifejezi a konkatenációt. Egy szintaktikailag önreprezentáló nyelv képes leírni bármelyik szimbólum-sorozatának struktúráját.

A $\varphi(x_1 \dots x_n)$ és $\psi(x_1 \dots x_n)$ n változós formulák *koextenzívek* $\mathcal{M}$-ben, ha az univerzumnak ugyanazok az elemei teszik igazzá őket, azaz

$$\mathcal{M} \models \forall x_1 \dots \forall x_n : \varphi(x_1 \dots x_n) \equiv \psi(x_1 \dots x_n)$$

Az $S^{\mathcal{M}} \subseteq \Sigma$ halmaz n változós formulák *monadikus szemantikai tulajdonsága* (az $\mathcal{M}$ modellben), ha $S^{\mathcal{M}}$ olyan halmaz, ami (i) n változós formulákat tartalmaz és (ii) zárt a koextenzióra, azaz bármely két, $\mathcal{M}$-ben koextenzív $\varphi(x_1 \dots x_n)$ és $\psi(x_1 \dots x_n)$ formulára igaz, hogy vagy mindkettő eleme $S^{\mathcal{M}}$-nek, vagy egyik sem. Az n változós formulák k tagú szemantikai relációinak definíciója hasonló; ez esetben az $S^{\mathcal{M}}$ halmaz n változós formulák k tagú sorozatainak halmaza lesz, vagyis $S^{\mathcal{M}} \subseteq \Sigma^k$.

Egy szemantikai tulajdonság akkor *nemtriviális*, ha nem üres és nem az összes (n változós) formula vagy formula-sorozat halmaza.

$\mathcal{M}(U, I, R)$ *szemantikailag önreprezentáló*, ha képes kifejezni egy nemtriviális szemantikai tulajdonságot, tehát van olyan $S^{\mathcal{M}} \subseteq \Sigma^k$ $(k \geq 1)$ nemtriviális szemantikai tulajdonság és van $\mathcal{M}$-ben egy olyan $\mathsf{S}(x_1 \dots x_k)$ formula, hogy

$$I(\mathsf{S}(x_1 \dots x_k)) = \{\langle R(s_1) \dots R(s_k) \rangle : \langle s_1 \dots s_k \rangle \in S^{\mathcal{M}}\}$$

Végezetül az $\mathcal{M}(U, I, R)$ modell *metareprezentáló*, ha kifejezi a formulák és az őket azonosító leírások közti viszonyt, tehát létezik egy olyan $\texttt{META}(x, y)$ formula, melyre

$$I(\texttt{META}(x, y)) = \{\langle u_1, u_2 \rangle : \ u_1 = R(\texttt{ID}[s](x)) \ \wedge \ u_2 = R(s)\}$$

A metareprezentáló modellek reprezentálják azt a tényt, hogy egyes kifejezéseik a saját szimbólumsorozataikat reprezentálják.

11.2. Elátkozottság

1. Állítás. *Ha $\mathcal{M}(U, I, R)$ egy impliciten önreprezentáló modell, akkor $\mathcal{M}$ nem lehet egyszerre szintaktilag önreprezentáló, szemantikailag önreprezentáló és metareprezentáló.*

Bizonyítás. Tegyük fel, hogy $\mathcal{M}$ mindhárom tulajdonsággal rendelkezik. Az egyszerűség vegyük azt az esetet, amikor $\mathcal{M}$ egyváltozós formulák valamely monadikus szemantikai tulajdonságát fejezi ki. Az állítás könnyen kiterjeszthető

többváltozós formulák monadikus tulajdonságaira és formulák relációira. Van tehát egy $S^{\mathcal{M}}$ szemantikai tulajdonság, amit $\mathcal{M}$ az $\mathtt{S}(x)$ formulával fejez ki. Mivel $S^{\mathcal{M}}$ nemtriviális, létezik olyan $\varphi_0(y)$ és $\varphi_1(y)$, hogy $\varphi_1(y) \in S^{\mathcal{M}}$ és $\varphi_0(y) \notin S^{\mathcal{M}}$.

Tekintsük az alábbi formulákat:

$$\alpha(x) =_{df}$$
$$\exists y \exists z \exists v \exists w : \mathtt{ID}[\exists x :](y) \wedge \mathtt{META}(z, x) \wedge \mathtt{ID}[\wedge](v) \wedge \mathtt{CON}(y, z, v, x, w) \wedge \mathtt{S}(w)$$

$$\beta(x, y) =_{df} \quad (\alpha(x) \wedge \varphi_0(y)) \vee (\neg\alpha(x) \wedge \varphi_1(y))$$

$$\gamma(y) =_{df} \quad \exists x : \mathtt{ID}[\beta(x, y)](x) \wedge \beta(x, y)$$

Emlékeztető: $\mathtt{ID}[\ldots](x)$ azt fejezi ki, hogy x implicit jelölete a „…” szimbólumsorozat, precízen $I(\mathtt{ID}[\ldots](x)) = \{R(\ldots)\}$. A $\mathtt{META}(z, x)$ a szimbólumsorozatok és az őket azonosító leírások közti viszonyt reprezentálja, pontosabban az univerzum azon u_z, u_x elemei elégítik ki, melyekre $u_x = R(s)$ esetén $u_z = R(\mathtt{ID}[s](x))$. $\mathtt{CON}(y, z, v, x, w)$ pedig azt fejezi ki, hogy w implicit jelölete az y, z, v, x implicit jelöleteinek konkatenációja.

A bizonyítás alapötlete az, hogy a fenti $\alpha(x)$, $\beta(x, y)$ és $\gamma(x)$ formulák interpretációjából következik, hogy $\gamma(x) \in S^{\mathcal{M}}$ pontosan akkor igaz, ha $\gamma(x) \notin S^{\mathcal{M}}$, vagyis ellentmondásra jutunk.

Ennek intuitív igazolásához tegyük fel, hogy a c konstans interpretációja a $\beta(x, y)$ formula implicit jelölője, azaz $I(c) = R(\beta(x, y))$. Ekkor $\alpha(c)$ pontosan akkor igaz $\mathcal{M}$-ben, ha w implicit jelölete egyrészt eleme S-nek, másrészt azonos az alábbi négy szimbólumsorozat konkatenációjával: (1) $\exists x :$, (2) az $I(c)$ implicit jelöletét azonosító leírás, (3) $\wedge$, (4) $I(c)$ implicit jelölete. Ezek összefűzése éppen $\gamma(y)$. Tehát $\alpha(c)$ pontosan akkor igaz, ha $\gamma(y) \in S^{\mathcal{M}}$.

Ugyanakkor azt is tudjuk, hogy a $\beta(c, y)$ formula koextenzív $\gamma(y)$-nal, hiszen $\gamma(y)$ definíció szerint úgy áll elő, hogy $\beta(x, y)$-ba behelyettesítjük x helyére $\beta(x, y)$ implicit jelölőjét, a c konstans interpretációja viszont épp ez a jelölő.

Továbbá, $\beta(c, y)$ pontosan akkor *nem* eleme $S^{\mathcal{M}}$-nek, ha $\alpha(c)$ igaz. Ha ugyanis $\alpha(c)$ igaz, akkor $\beta(c, y)$ a β formula definíciója szerint koextenzív $\varphi_0(y)$-nal, márpedig kikötöttük, hogy $\varphi_0(y) \notin S^{\mathcal{M}}$. Ha pedig $\alpha(c)$ hamis, akkor $\beta(c, y)$ hasonló okokból koextenzív $\varphi_1(y)$-nal, márpedig kikötöttük, hogy $\varphi_1(y) \in S^{\mathcal{M}}$. Tehát:

- $\mathcal{M} \models \alpha(c)$ akkor és csak akkor, ha $\gamma(y) \in S^{\mathcal{M}}$
- $\beta(c, y)$ és $\gamma(y)$ koextenzív
- $\mathcal{M} \models \alpha(c)$ akkor és csak akkor, ha $\beta(c, y) \notin S^{\mathcal{M}}$
- $S^{\mathcal{M}}$ zárt a koextenzivitásra.

Ez a négy tény ellentmondásra vezet, mert az következik belőlük, hogy $\gamma(y) \in S^{\mathcal{M}}$ akkor és csak akkor, ha $\gamma(y) \notin S^{\mathcal{M}}$.

Általános esetben nem garantálható, hogy létezik olyan c konstans, melynek interpretációja a $\beta(x,y)$ formula implicit jelölője, de a bizonyítás ekkor is hasonlóan építhető fel, csak be kell vonnunk a kiértékeléseket. Az implicit önreprezentáció definíciójából tudjuk, hogy van olyan $u \in U$, melyre $u = R(\beta)$. Legyen K egy olyan kiértékelés, ahol $K(x) = u$. Ekkor a következők igazak:

(i) Ha $\mathcal{M}^K \models \alpha(x)$, akkor $\mathcal{M} \models \forall y : \gamma(y) \leftrightarrow \varphi_0(y)$, ha pedig $\mathcal{M}^K \not\models \alpha(x)$, akkor $\mathcal{M} \models \forall y : \gamma(y) \leftrightarrow \varphi_1(y)$.

(ii) Mivel $\varphi_0 \notin S^{\mathcal{M}}$, $\varphi_1 \in S^{\mathcal{M}}$, és $S^{\mathcal{M}}$ zárt a koextenzivitásra, (i)-ből következik, hogy $\mathcal{M}^K \models \alpha(x) \;\Leftrightarrow\; \gamma \notin S^{\mathcal{M}}$.

(iii) A K kiértékelésnél az α, β, γ formulák definíciója miatt $\alpha(x)$ azt fejezi ki, hogy γ eleme $S^{\mathcal{M}}$-nek, így $\mathcal{M}^K \models \alpha(x) \;\Leftrightarrow\; \gamma \in S^{\mathcal{M}}$.

(iv) Tehát, (ii) és (iii) miatt, $\gamma \in S^{\mathcal{M}} \Leftrightarrow \gamma \notin S^{\mathcal{M}}$. □

Az 1. állításból levezethető Gödel első nemteljességi tétele. Ehhez előbb alkalmazni kell az állítást az igazságpredikátumokra:

2. Állítás. *Ha $\mathcal{M}(U, I, R)$ szintaktikailag önreprezentáló és metareprezentáló, akkor nem fejezheti ki a saját igazságpredikátumát.*

Bizonyítás. Az igazság a nulla változós formulák (azaz a mondatok) szemantikai tulajdonsága, hiszen ha $S^{\mathcal{M}}$ a modell igaz mondatainak halmaza, akkor $\varphi \in S^{\mathcal{M}}$ és $\mathcal{M} \models \varphi \equiv \psi$ esetén nyilván $\psi \in S^{\mathcal{M}}$. Mivel $S^{\mathcal{M}}$ nemtriviális (nem üres és nem az összes mondat halmaza), vonatkozik rá az 1. állítás. □

Legyen $\mathcal{M}(\mathbb{N}, I)$ az elemi aritmetika egy standard modellje. Ezt Gödel-számozás segítségével kibővíthetjük egy impliciten önreprezentáló $\mathcal{M}(\mathbb{N}, I, G)$ modellé, ami szintaktikailag önreprezentáló lesz, mert felismeri a jelsorozatait és kifejezi a konkatenációt. Metareprezentáló is lesz, mert van olyan kétváltozós $\varphi(m,n)$ formula, amit az m, n természetes számok pontosan akkor elégítenek ki, ha $n = G(\psi)$ és $m = G(x = n)$.

Azt is tudjuk továbbá, hogy az $\mathcal{M}(\mathbb{N}, I, G)$ modellnek van olyan $\mathsf{BIZ}(m,n)$ formulája, melyet az m, n természetes számok pontosan akkor elégítenek ki, ha $m = G(\varphi)$, $n = G(\psi)$ és ψ bizonyítja φ-t.

Tegyük fel, hogy az elemi aritmetika helyes, azaz nem lehet benne hamis állítást bizonyítani. Ha minden aritmetikai igazság bizonyítható, azaz ha minden $\mathcal{M} \models \varphi$ esetén van olyan m és n, hogy $m = G(\varphi)$ és $\mathcal{M} \models \mathsf{BIZ}(m,n)$, akkor $\mathcal{M}$ kifejezi a saját igazságpredikátumát, méghozzá a $\exists y : \mathsf{BIZ}(x,y)$ formulával. Ez azonban a 2. állítás miatt lehetetlen. Tehát ha az elemi aritmetika helyes, akkor egyes aritmetikai igazságok bizonyíthatatlanok.

11.3. Értékrések és értéktorlódások

Az 1. állítás kiterjeszthető olyan modellekre is, ahol az igazság és hamisság mellett van egy harmadik, „szürkezónás” kategória is. Ez többféleképpen is előfordulhat; a két leggyakoribb feltevés szerint a szürkezónás állításoknak nincs

igazságértéke („igazságérték-rés", *truth value gap*), vagy egyszerre igazak és hamisak („igazságérték-torlódás", *truth value glut*). Modellelméleti szinten ez azt jelenti, hogy az I interpretációs függvény nem az univerzum (sorozatainak) részhalmazait rendeli a nyelv predikátumjeleihez, hanem mindegyikhez két részhalmazt, egy extenziót és egy anti-extenziót rendel.

Egy egyváltozós F predikátumjel esetében ez a következőképpen néz ki:

$$I(F) = \langle H^F_{ext}, H^F_{antiext} \rangle \subseteq \mathscr{P}(U) \times \mathscr{P}(U)$$

$I(a) \in H^F_{ext}$ esetén $\mathcal{M} \models F(a)$, míg $I(a) \in H^F_{antiext}$ esetén $\mathcal{M} \models \neg F(a)$, egyébként pedig $F(a)$ szürkezónás (vö. Priest 2006: 5. fejezet).

Az értékréses modellekben egyes F-ek esetében egyszerre teljesül az, hogy $H^F_{ext} \cup H^F_{antiext} \neq U$ és az, hogy $H^F_{ext} \cap H^F_{antiext} = \emptyset$, tehát az extenzió és az antiextenzió együttesen nem fedik le a modell teljes univerzumát. Az értéktorlódásos modellek azok, ahol ennek fordítottja igaz, vagyis egyes predikátumok esetében az extenziónak és antiextenziónak van közös eleme.

Mivel a koextenzivitás korábbi definíciója ezekben a modellekben nem alkalmazható, a továbbiakban a koextenzivitás jelentse azt, hogy a formulák extenziója *és* antiextenziója is azonos.

Ha az önreprezentáció szentháromságához kapcsolódó összes speciális formula ((CONCAT$(x_1 \ldots x_n)$ stb.) klasszikus módon viselkedik, vagyis ha az interpretációjuk nem szürkezónás, akkor az 1. állítás értékrések és értéktorlódás esetén is igaz marad. Ez esetben ugyanis az 1. állítás bizonyításában az $\alpha(c)$ formula klasszikus interpretációt kap, ami oda vezet, hogy $\gamma(y)$ ismét paradox módon viselkedik, egyszerre eleme és nem eleme $S^{\mathcal{M}}$-nek. E feltételek tehát – az önreprezentációs formulák klasszikussága – az értékréses és értéktorlódásos modelleknél is garantálják az elátkozottságot.

Nyilván az az eset érdekes, amikor az önreprezentáció egyes elemei is szürkezónásak. Rengeteg kombináció lehetséges, de most csak azokkal a modellekkel foglalkozom, ahol az önreprezentációt biztosító formulák közül egyedül az $S^{\mathcal{M}}$ szemantikai tulajdonságot kifejező $\mathsf{S}(x)$ formula szürkezónás. Ezeket „szemantikailag szürkezónás" modelleknek nevezem. A többi formula esetében nehezebb elhinni, hogy nemklasszikus módon viselkednek. Mikor lenne például sem igaz, sem hamis (vagy egyszerre igaz és hamis), hogy az $F(x) \wedge G(x)$ formula az $F(x)$ és $\wedge G(x)$ jelsorozatok konkatenációja? Hasonló a helyzet a jelsorozatok felismerésével és a metareprezentációval.

A szemantikai predikátum esetében viszont kifejezetten ismerős a szürkezónás felvetés: a hazug-paradoxonnak azokra a megoldásaira hasonlít, melyekben a paradox mondat („Ez a mondat hamis") nem rendelkezik igazságértékkel, vagy egyszerre igaz és hamis. Az 1. állítás bizonyításában szemantikailag szürkezónás modelleknél nem ellentmondás adódik, hanem arra jutunk, hogy $\alpha(c)$ a szürkezónába esik, vagyis modellünk nem mindig képes koherens igen/nem választ adni arra a kérdésre, hogy egy formula rendelkezik-e az $S^{\mathcal{M}}$ szemanti-

kai tulajdonsággal. Az 1. állítás emiatt átalakul azzá az állítássá, hogy ha egy modellben a szemantikai önreprezentáció és metareprezentáció nem szürkezónás, akkor a modell csak szürkezónás módon fejezheti ki bármelyik szemantikai tulajdonságát.

Ez a tény akkor különösen érdekes, és egyben bizarr, ha a koextenzivitásra koncentrálunk. A koextenzivitás szemantikai tulajdonság, konkrétan egy bináris szemantikai viszony formulák között, hiszen triviálisan igaz, hogy bármely két koextenzív formula koextenzív. A szürkezónás modellekben (ha a szemantikai és szintaktikai önreprezentáció terén klasszikusak) a koextenzivitást kifejező formula (ha létezik) garantáltan a szürkezónába esik, emiatt pedig, ha egy ilyen modellt használunk hiteink kifejezésére, sosem lehetünk biztosak benne, hogy a „$\varphi(x_1 \dots x_n)$ és $\psi(x_1 \dots x_n)$ koextenzív" alakú állítások koherensek és van igazságértékük. Tehát a szürkezónás modellek formálisan igazolnak egy quine-i hangulatú tézist, miszerint a jelentésazonosság meghatározatlan, defektes viszony, vagy legalábbis csak így fejezhető ki.

11.4. A modális verzió

Az 1. állítás a standard kvantifikált modális logikára is kiterjeszthető, ha a szemantikai tulajdonságok a *szükségszerű* koextenzivitásra zártak. Az egyszerűség kedvéért csak a konstans univerzumú (*constant domain*) esetekkel foglalkozom. Rögzítsünk egy elsőrendű nyelvet, és adjuk hozzá a $\Box$ és $\Diamond$ logikai konstansokat (a szükségszerűség ill. lehetőség operátorát). Egy $\mathcal{K}(U, W, I, E)$ Kripke-modell egy U univerzumból, a világok W halmazából, egy $E \subseteq W \times W$ elérhetőségi relációból és egy olyan I interpretációból áll, ami a konstansjelekhez U egy-egy elemét, a predikátumjelek és világok párosaihoz pedig az univerzum (sorozatainak) részhalmazait rendeli, azaz ha F egy egyváltozós predikátumjel, akkor $I(\langle F, w\rangle) \subseteq U$ minden $w \in W$-re. I azt rögzíti, hogy a w világon U mely elemei rendelkeznek az F tulajdonsággal. A szemantika többi része hasonló a standard elsőrendű esethez, de világokra relativizálva, például $\mathcal{K} \models_w F(a)$ ($\mathcal{K}$ szerint a w világban igaz $F(a)$), ha $I(a) \in I(\langle F, w\rangle)$. Az egyetlen újdonságot a modális operátorok jelentik:

- $\mathcal{K} \models_w \Box\varphi \;\Leftrightarrow$ ha $v \in W$ és $E(w, v)$, akkor $\mathcal{K} \models_v \varphi$ (valami akkor szükségszerű w-ben, ha igaz minden w-ből elérhető világban)
- $\mathcal{K} \models_w \Diamond\varphi \;\Leftrightarrow$ van olyan $v \in W$, hogy $E(w, v)$ és $\mathcal{K} \models_v \varphi$ (valami akkor lehetséges w-ben, ha igaz egy w-ből elérhető világban)

Az ilyen modellek a korábban jelzett módon impliciten önreprezentálóvá tehetők egy injektív $R : \Sigma \to U$, függvénnyel. Az önreprezentáció különböző formáinak definíciója hasonló, de a vonatkozó formuláktól (CONCAT stb.) elvárjuk, hogy minden lehetséges világban ugyanaz legyen az extenziójuk (minden

világban az univerzumnak ugyanazok az elemei tegyék igazzá őket). A szemantikai tulajdonságok definíciója annyiban módosul, hogy minden szemantikai tulajdonság zárt a *szükségszerű* koextenzióra. Például, ha $S^{\mathcal{M}}$ n változós formulák egy monadikus szemantikai tulajdonsága, akkor bármely $\varphi(x_1 \dots x_n)$ és $\psi(x_1 \dots x_n)$ formulákra igaz, hogy ha minden w-re $I(\langle\varphi, w\rangle) = I(\langle\psi, w\rangle)$, azaz ha φ és ψ szükségszerűen koextenzívek, akkor vagy mindkettő eleme $S^{\mathcal{M}}$-nek, vagy egyik sem.[1]

E feltételek mellett az 1. állítás releváns változata („egy impliciten önreprezentáló Kripke-modell nem lehet egyszerre szemantikailag önreprezentáló, szintaktikailag önreprezentáló és metareprezentáló") ugyanúgy bizonyítható, mint az 1. állítás. Az ott szereplő $\mathsf{S}(x)$ formulának a fenti feltételek mellett minden világon ugyanaz lesz az extenziója, azaz bármely c-re vagy szükségszerűen igaz $\mathsf{S}(c)$ (minden világban), vagy szükségszerűen hamis (minden világban). Az, hogy két formula szükségszerűen koextenzív-e, világfüggetlen tény, és a konstansok jelölete sem változik világok között. Mivel $\mathsf{S}(c)$ igazságértéke világfüggetlen, $\alpha(c)$ is vagy szükségszerű, vagy lehetetlen, ez pedig garantálja, hogy ha $I(c) = \beta(x, y)$, akkor $\beta(c, y)$ a korábbiakhoz hasonló paradox módon egyszerre eleme és nem eleme $S^{\mathcal{M}}$-nek.

[1] A $\Box\forall x_1 \dots \forall x_n : \varphi(x_1 \dots x_n) \equiv \psi(x_1 \dots x_n)$ tárgynyelvi formulával csak akkor lehet definiálni a szükségszerű koextenzivitást, ha minden világ minden világból elérhető.

12.

Önreprezentációs demarkáció

Az előző fejezet az elsőrendű predikátumlogika modelljeiről látta be, hogy nem rendelkezhetnek egyszerre három önreprezentációs tulajdonsággal: a szemantikai önreprezentáció, a szintaktikai önreprezentáció és a metareprezentáció képességével. Ezt a tételt a fejezet végén kiterjesztettem az értékréses és értéktorlódásos, illetve modális logikára is. Bár a bizonyítások formális nyelvekre vonatkoztak, az alábbiakban a természetes nyelvre vonatkoztatom őket, abból a meggondolásból, hogy az elsőrendű predikátumlogika része a természetes nyelv (racionális ismeretszerzés érdekében precízzé tett) logikájának, a modell-elméleti apparátus pedig azt a tényt tükrözi vissza, hogy mondataink jelentése a bennük szereplő kifejezések jelentésének rekurzív szabályokon alapuló kombinációja (Button & Walsh 2018: 19–21). A modell-elmélet tehát értelmezésem szerint nem platonista porhintés és a matematika belügye, hanem nyelvileg megformált gondolataink egyik alapvető strukturális tulajdonságának vizsgálata. Az egyszerű mondatok – „$F(a)$" – igazsága az alany és az állítmány jelöleteinek függvénye, a komplex mondatok és kifejezések jelentése pedig rekurzív módon adódik az egyszerűbb mondatok és kifejezések jelentéséből. Mivel ez igaz a (logikailag letisztogatott) természetes nyelvre, a modell-elméleti tételek is alkalmazhatóak rá. Emiatt pedig a magyar nyelv (és általában egyik természetes nyelv) sem lehet egyszerre szemantikailag önreprezentáló, szintaktikailag önreprezentáló és metareprezentáló. De mint az előző fejezetben láttuk, a magyar nyelv a jelek szerint mindhárom tulajdonsággal rendelkezik. Nyelvünk eszerint inkonzisztens?

A paradoxon feloldására véleményem szerint Tarski (1944) klasszikus javaslata, a metaszintek bevezetése a leghatékonyabb eszköz. Tarski a hazugparadoxon – „Ez a mondat hamis" – kapcsán vetette fel, hogy a nyelvi önreprezentáció elemzésekor el kell különítenünk a tárgynyelvet és a metanyelvet. Az a kijelentés, hogy „»A hó fehér« magyar mondat igaz", bár magyar mondatnak tűnik, Tarski szerint egy olyan nyelvhez, a *metanyelv*hez tartozik, ami hasonló a mindennapokban használt magyar nyelvhez, de nem azonos vele. (Modell-elméleti fogalmakkal: kifejezve egy másik modellről van szó.) A meta-

nyelv segítségével vizsgáljuk a tárgynyelvet, az „alapszintű” magyar nyelvet. A metanyelv gazdagabb, mint a tárgynyelv. Az „alapszintű” magyar nyelv nem tartalmazza például a saját igazságpredikátumát, mert ekkor inkonzisztens lenne. Az igazságpredikátum a metanyelv szintjén jelenik meg, ha pedig a metanyelv mondatairól beszélünk, akkor egy meta-metanyelven, és így tovább. A paradoxnak tűnő kijelentések rosszul formált mondatok, mert keverik a tárgynyelv és metanyelv szintjét.

Tarski a formális logika kontextusában beszélt metanyelvekről, én azonban a gondolkodás és az ezt kifejező természetes nyelv kapcsán szeretnék bevezetni egy analóg fogalmat, melynek segítségével feloldható az önreprezentáció szentháromságának átka. Hangsúlyozom, hogy nem Tarski elméletét veszem át és szigorúan véve nem is az ő fogalomkészletét használom, hanem egy hasonló ötletet próbálok kidolgozni, ami az emberi gondolkodásra, nem pedig (elsősorban) a formális logikára vonatkozik.

A megoldási javaslat a következő: a magyar nyelvet – a többi természetes nyelvhez hasonlóan – azért nem teszi inkonzisztenssé az önreprezentáció elátkozott szentháromsága (a metaforát feloldva: a magyar nyelv használóinak gondolkodása azért nem inkonzisztens), mert a reprezentációs tevékenységünk reprezentációja, a metareprezentáció más gondolati szinten, „metaszinten” zajlik, mint a szemantikai és szintaktikai önreprezentáció. (Lehet, hogy ez utóbbiak is egymástól eltérő szinteken zajlanak, de ez most mindegy.) Amikor nyelvileg kifejezzük azt a tényt, hogy a nyelv leírja saját (szintaktikai vagy szemantikai) önreprezentációs tevékenységét – a metaforát feloldva: mikor arról gondolkodunk, hogyan reprezentálják a mondataink a mondatainkat – , akkor egy metaszintről nézünk vissza az alapszintre, a szintaktikai és szemantikai tulajdonságokat kifejező mondatainkra, és ez a szintkülönbség nem nyomható össze egyetlen monolitikus kognitív rendszerré, ami egyszerre végezne szemantikai önreprezentációt, szintaktikai önreprezentációt és metareprezentációt. A természetes nyelv pragmatikus rugalmassága elleplezi azt a tényt, hogy a nyelv által kifejezett gondolatok néha hierarchikus viszonyban állnak egymással; a gondolataink egy részének a gondolataink másik része a tárgya és emiatt előbbiek, ha úgy tetszik, nem ugyanannak a kognitív univerzumnak a részei. Ezért nem logikailag inkoherens a természetes nyelv. A metareprezentáció más reprezentációs szinten történik, mint a szemantikai és szintaktikai önreprezentáció; metaszintre emelkedve egy tágabb keretben, mintegy kívülről reprezentáljuk a reprezentációinkat.

Ha elfogadjuk ezt a megoldást, akkor segítségével az alábbi demarkációs érv konstruálható:

(1) Szükségszerű, hogy ha x nem képes metaszintű gondolkodásra, akkor x vagy inkonzisztens, vagy nem képes szintaktikai önreprezentációra, vagy nem képes szemantikai önreprezentációra, vagy nem képes metareprezentációra.

(2) Aki gondolkodik, az potenciálisan konzisztens, képes szemantikai önreprezentációra, szintaktikai önreprezentációra és metareprezentációra.

(3) Az MI szükségszerűen nem képes metaszintű gondolkodásra.

(4) Tehát az MI nem gondolkodik.

Az érv logikai szerkezete a következő ($\Box$ = szükségszerű, $\Diamond$ = lehetséges):

(1) $\Box\forall x : \neg M_1(x) \rightarrow (\neg K(x) \vee \neg S_1(x) \vee \neg S_2(x) \vee \neg M_2(x))$

(2) $\forall x : G(x) \rightarrow \Diamond(K(x) \wedge S_1(x) \wedge S_2(x) \wedge M_2(x))$

(3) $\forall x : A(x) \rightarrow \Box\neg M_1(x)$

(4) $\forall x : A(x) \rightarrow \neg G(x)$

$A(x)$	x egy MI
$G(x)$	x gondolkodik
$K(x)$	x konzisztens
$M_1(x)$	x képes metaszintű gondolkodásra
$M_2(x)$	x képes metareprezentációra
$S_1(x)$	x képes szintaktikai önreprezentációra
$S_2(x)$	x képes szemantikai önreprezentációra

A standard (normális) elsőrendű modális logikában a fenti levezetés érvényes, tehát az érvet a premisszák szintjén kell támadni.

12.1. Az 1. premissza védelmében

Az 1. premissza melletti fő érv az, hogy a metaszintű gondolkodás fogalma gyors és elegáns megoldást kínál az önreprezentáció elátkozott szentháromságából eredő problémára (arra, hogy a 11.2.1. állítás fényében a magyar nyelv logikailag inkoherensnek tűnik). Ezt a premisszát elvben támadni lehet a probléma olyan alternatív megoldásai alapján, melyek nem posztulálnak metaszintű gondolkodást. Amennyire látom, hasonló alternatívák merülnek fel, mint a hazug-paradoxon esetében, ezért a továbbiakban ezeket tekintem át Beall, Glanzberg és Ripley (2025) cikke alapján. A hazug-paradoxon megoldásainak részletes rekonstrukciója, adaptálása és kritikája messze túlmutat e könyv keretein. Célom mindössze annyi, hogy hihetővé tegyem, hogy a metaszint feltételezése abduktív alapon a legjobb megoldás lehet.

A hazug-paradoxon egyes megoldásai nem egyértelműen relevánsak jelen kontextusban, ezeket gyorsan zárójelbe is teszem. Ide tartoznak a nemkontraktív, nemtranzitív és nemreflexív logikák, valamint a Kripke-féle fixpont-alapú megoldás (Beall, Glanzberg & Ripley 2025: 4.2, 4.3.2). Nem világos (legalábbis

számomra), hogy ezek a bonyolult formális konstrukciók hogyan adaptálhatóak a 11.2.1. állításra, ahol nem konkrétan az igazságpredikátumról van szó, hanem tetszőleges szemantikai tulajdonságról.

A hazug-paradoxon kontextualista megoldása szerint (i.m. 4.4) az igazságpredikátum jelentése kontextusfüggő és a látszólag paradox igazságértékek valójában eltérő kontextusokhoz tartoznak. Ahhoz, hogy ez az ötlet alkalmazható legyen a 11.2.1. állításra, azt kellene állítanunk, hogy az I interpretációs függvény kontextusfüggő. Ezzel azonban nem cáfoljuk a 11.2.1. állítást, hiszen továbbra is elismerjük, hogy egy *adott* kontextusban lehetetlen a háromféle önreprezentáció együttes jelenléte. A fenti demarkációs érv tehát újrafogalmazható a kontextusváltás képességére hivatkozva. Az így keletkező alternatív 2. és 3. premissza érzésem szerint nagyjából ugyanúgy és ugyanannyira lesz védhető, mint a premisszák aktuális változata. Tehát ha a kontextualizmus netán a metaszinteknél kielégítőbben oldja is fel az önreprezentáció szentháromságának átkát, a szóban forgó demarkációs stratégiát nem veszélyezteti.

Mint arról a 11.3. alfejezetben szó volt, a szemantikai paradoxonok értékrésekre vagy értéktorlódásokra épülő megoldásai (Beall et al. 2025: 4.1.1–4.1.2) szintén képesek feloldani az önreprezentáció szentháromságának átkát, cserébe viszont el kell fogadnunk egy quine-i jellegű tézist, miszerint a szemantikai fogalmaink defektesek. A szemantikai fogalmakat tartalmazó állításokról – például a kifejezések koextenzivitásával kapcsolatos állításokról – nem garantálható, hogy koherensek és van igazságértékük.[1] De felvethető, hogy ez az ár túl magas. Ha van más működőképes megoldás, akkor nincs okunk szükségszerű igazságnak tekinti, hogy a szemantikai fogalmaink defektesek. A koextenzivitás például – az, hogy két állítmány ugyanazokra a dolgokra vonatkozik – teljesen érthető, világosan kifejezhető tulajdonságnak tűnik, melyet lépten-nyomon használnak a nyelvészetben, filozófiában és logikában, illetve általában bárhol, ahol definíciókat mondanak ki és ezzel expliciten a koextenzivitásról állítanak valamit. Tehát a metaszintű gondolkodás vonzóbb megoldást kínál, mint az értékréses és értéktorlódásos modellek.

12.2. A 2. premissza védelmében

A 2. premissza azt fejezi ki, hogy egy gondolkodó lény képes konzisztens módon szemantikai önreprezentációt, szintaktikai önreprezentációt és metareprezentációt végezni. Ez ránk emberekre láthatólag igaz, hiszen a nyelv elemzésekor rendszeresen használjuk az önreprezentáció mindhárom formáját és ha ezt a metaszintek segítségével tesszük, akkor konzisztensek vagyunk. A 2. premissza tehát elfogadható, ha „gondolkodó lény”-en a velünk azonos kognitív képességekkel rendelkező létezőket értjük.

[1] Pontosabban ez a helyzet, ha a szintaktikai önreflexió és a metareflexió klasszikus módon viselkedik, ami a természetes nyelv esetében hihetőnek tűnik (vö. 11.3. alfejezet).

12.3. A 3. premissza védelmében

A 3. premissza szerint az MI természeténél fogva nem képes metaszintű gondolkodásra, nem tudja saját ⌜mentális⌝ állapotait saját ⌜mentális⌝ állapotai tárgyává tenni. Erre nagyon egyszerű érvem van. Mint arról a 2. fejezetben szó volt, az „MI" jelen kontextusban a nagy nyelvi modellek különféle változatait jelöli, melyek az alábbi mechanizmuson alapulnak:

szöveg (input)
↓
tokenizáció
↓
tokenek lecserélésére vektorokra
↓
... (vektoralgebrai műveletek) ...
↓
output generálása egy vektorként adódó valószínűség-eloszlás alapján

Ennek a folyamatnak nincs metaszintje. Ha visszatöltünk bele valamit, ami kijött belőle, ugyanaz történik vele, mint bármi mással. Saját ⌜gondolatai⌝ ugyanúgy tokenekre hasadnak és vektorok sorozatává szétesve ugyanazokon a vektoralgebrai operációkon mennek keresztül, mint bármely más input. Az MI ugyanazon a szinten kezeli, ugyanabban a szemantikai térben helyezi el a „Nyílik a pitypang" betűsorozatot, mint az „Egyes gondolataim a pitypangokra vonatkoznak" betűsorozatot. Nem képes reflektálni saját magára, tárggyá tenni saját magát, metaszintre emelkedve vizsgálni saját reprezentációit, azok tartalmát és hordozóit.

Az önreflexív MI hívei erre azt felelhetik, hogy az inputokon végzett matematikai műveletek típusazonosságából nem következik, hogy az MI gondolkodása is homogén, metaszint-mentes lenne, hiszen a metaszint ebben a formában a mi neuronjainkban sincs jelen, vagy legalábbis semmi okunk azt gondolni, hogy jelen van. Nem tudunk olyan speciális agyterületről, ami csak akkor aktiválódik, ha más agyterületeink reprezentációs tevékenységét elemzi.

Ez a válasz véleményem szerint hamis analógiával dolgozik, mert az agyat nem értjük annyira, hogy a fenti folyamatábrához hasonló képünk lenne arról, mit csinál. A folyamatábra persze sokkal egyszerűbb, mint bármelyik működő MI, de egy adekvát absztrakciós szinten az MI tényleg csak azt csinálja, amit a folyamatábra mutat. Az agyról nem áll rendelkezésre a fentihez hasonló kompakt leírás, ami összefoglalná a benne zajló folyamat lényegét. Az agyban neuronok tüzelnek és ettől más neuronok is tüzelnek, a tüzelések valószínűségét pedig hormonok, neurotranszmitterek és az agy-ér gáton átjutó különféle molekulák befolyásolják. Ennek az elképesztően bonyolult folyamatnak nagyon

sok lokális részletét ismerjük. De ez nem teljes leírása az emberi agy output-generáló tevékenységének olyan értelemben, ahogy a fenti folyamatábra teljes leírása a transzformer-architektúra – és ezáltal a mai MI – output-generáló tevékenységének.

További probléma, hogy az ellenérv előfeltételezi a fizikalizmust, pontosabban azt, hogy az emberi önreprezentációs tevékenység teljes egészében az agyban zajlik. Ha a fizikalizmus hamis és létezik anyagtalan elme vagy lélek, vagy vannak nemfizikai intencionális állapotaink, akkor abból, hogy a fenti demarkációs érv egy változata esetleg az emberi agyra is igaz, nem következik, hogy a demarkáció sikertelen. Ha a fizikalizmus hamis, akkor nem meglepő, hogy az agy sem gondolkodik. Tehát még ha működne is az ellenérv, akkor is csak annyit igazolna (ha az (1)–(4) érv amúgy helyes), hogy a 3. premissza és a fizikalizmus közül az egyik hamis.

Miért ne mondhatnánk, hogy az MI elvben megfigyelheti a saját belső tevékenységét, és ezáltal reprezentálja saját magát (Walker et al. 2025)? Elképzelhetünk ilyen önfigyelő folyamatokat, és az MI-vel vezérelt robotokban vélhetőleg zajlanak is ilyen folyamatok, de ez nem az a fajta önreprezentáció, amiről a metaszintek kapcsán szó van, hanem puszta visszacsatolási mechanizmus: a rendszer jelenlegi állapotáról érkező információ befolyásolja a rendszer későbbi állapotait. Ilyen mechanizmust a termosztátok is tartalmaznak, de ez nem teszi őket metaszinten gondolkodó rendszerekké.

Millière és Buckner (2024: 10) szerint „átcímkézési hibát" (*redescription fallacy*) követünk el, ha az MI intellektuális képességeivel kapcsolatban az MI architektúrájának informatikai vagy matematikai tulajdonságaira hivatkozunk. Szerintük a demarkációs vitának arra kell koncentrálnia, hogy milyen *empirikus* bizonyítékaink vannak az MI kognitív képességeivel kapcsolatban. Mivel a fenti demarkációs érv nem empirikusan tesztelhető kérdésekkel, hanem az MI belső mechanizmusával foglalkozik, az érv (a szóban forgó megközelítés szerint) az átcímkézési hiba paradigmatikus esete.

Ez a reakció szerintem inkább ideológiai pajzs, mint racionális ellenérv. Azt a módszertani elvet, hogy az MI matematikai leírásából tilos az MI képességeivel kapcsolatos következtetéseket levonni (mert ehhez csak az MI empirikus megfigyelése alapján van jogunk) kizárólag akkor van okunk elfogadni, ha elköteleztük magunkat az intelligencia funkcionalista felfogása mellett. Ha kijelentjük, hogy az MI értelmi képességeinek megítélésekor kizárólag arra hagyatkozhatunk, amit *csinál*, arra nem, ami a belsejében történik, akkor – Turinghoz hasonlóan – egy verbális manőverrel, egy fogalmi újítás bevezetésével rövidre zárjuk a demarkációs vitát.

Szofisztikáltabb ellenérvet konstruálhatunk, ha azt állítjuk, hogy az MI architektúrájának matematikai tulajdonságairól nem lehet leolvasni, hogy az MI ⌜gondolatainak⌝ mi a tartalma, emiatt pedig azt sem lehet leolvasni róla, hogy az MI tud-e metaszinten gondolkodni. Az illusztráció kedvéért tegyük fel, hogy a hitek tartalma oksági előtörténetük függvénye: egy hit (mint agyi állapot

vagy processzorállapot) akkor vonatkozik az F-ekre, ha F-ek okozzák. (Ez a naiv oksági jelentéselmélet könnyen cáfolható (ld. 6.5.2. alfejezet), de ha a naturalizált szemantika egy védhetőbb, komplikált változatát vennénk alapul, az ellenérv szempontjából ugyanoda jutnánk.) Az oksági szemantika alapján az MI egyes ⌜gondolatai⌝ nyugodtan irányulhatnak az MI egyéb ⌜gondolataira⌝, ha előbbieket az utóbbiak okozzák. Mivel az oksági szemantikát nem cáfoltuk, a 3. premissza hamis.

Ezzel az ellenérvvel az a gond, hogy az önreprezentáció szentháromságát sújtó átoknak az oksági szemantika is áldozatul esik. Ha ugyanis a hitek tartalma puszta oksági összefüggéseken múlik, akkor a szintaktikai önreflexió, szemantikai önreflexió és a metareprezentáció ugyanazon a nyelvi szinten zajlik, hiszen az okságnak metafizikai értelemben nincsenek különböző „szintjei"; nincs olyan, hogy meta-okság.[2] Ha tehát az oksági szemantika igaz és az oksági szemantika szerint az MI metareprezentál, akkor az MI inkonzisztens. Az oksági szemantika híve esetleg segítségül hívhatja az értékréses vagy értéktorlódásos logikát a paradoxon feloldására; ekkor az MI ⌜gondolkodása⌝ nem lesz inkoherens, csak kicsit meghatározatlan vagy helyenként zavaros. De ha az MI csak így menthető meg az inkoherenciától, akkor továbbra is van lehetőség demarkációra, mert az emberi gondolkodás esetében értékrések és értéktorlódások nélkül is el lehet kerülni a paradoxont.

Felmerülhet az a technikai jellegű ellenérv is, hogy az LLM-ek eleve nem klasszikus kétértékű logikával dolgoznak, tehát nem vonatkozik rájuk a 11.2.1. állítás. Ha igen/nem típusú kérdést teszünk fel egy MI-nek, akkor a program nem az „igen" vagy „nem" választ generálja magában, hanem egy valószínűségeloszlást, ahol az opciók egyike sem kap 0 vagy 1 valószínűséget.[3] Mivel a 11.2.1. állítás klasszikus kétértékű logikával dolgozik, nem alkalmazható az MI-re.

Fontos tisztázni, hogy a vitapartner itt pontosan mit állít. Ha azt állítja, hogy a transzformer-architektúra sajátosságai miatt az MI összevissza tippelget és emiatt ⌜hitei⌝ csak véletlenül alkothatnak koherens rendszert, akkor a demarkációs érv valóban hatályát veszti (mert az MI hitrendszerére nem illeszkednek a klasszikus elsőrendű modellek), de adódik egy új, még hatékonyabb demarkációs érv:

(5) Az MI ⌜hitrendszere⌝ csak véletlenül alkothat koherens rendszer.

[2] Beszélhetünk esetleg a másodrendű tulajdonságok – a tulajdonságok tulajdonságainak – oksági szerepéről és az ezekből eredő másodrendű oksági szintről. De az önreflexió és metareprezentáció esetében szintek *közötti* viszonyokról van szó, ezeket tehát az első- és másodrendű tulajdonságok közti oksági viszonyoknak kellene megfeleltetnünk. Még ha léteznének is ilyenek, hogyan lehetnének alapvetően más természetűek, mint az elsőrendű oksági viszonyok? Miért nem homogenizálja az okság a tulajdonságok közti összes oksági viszonyt egyetlen szintté?

[3] Szigorúan véve az LLM-ek a következő tokenhez rendelnek valószínűséget, de ez utóbbiak alapján definiálhatóak a teljes válaszokra vonatkozó valószínűségek.

(6) Ha x gondolkodó lény, akkor x hitrendszere nem csak véletlenül alkothat koherens rendszert (hanem reflexió, revízió és fogalmi tisztázás révén is).

(7) Ha az MI gondolkodik, akkor a ⌜hitrendszere⌝ a hitrendszere.

(8) Tehát az MI nem gondolkodik.

Az ellenérvnek az a változata tehát, ami elfogadja az 5. állítást, pürrhoszi győzelmet jelent.

Az ellenérv másik értelmezése szerint az MI bayesiánus módon rendelkezik hitekkel, és ez az oka, hogy az állításokhoz valószínűségeket, nem pedig klasszikus igazságértékeket rendel. A bayesiánus ellenérv sikere azon múlik, hogy bayesi keretben reprodukálható-e a 11.2.1. állítás. Ennek tisztázására itt nem vállalkozom; csak két rövid rekontrát fogalmazok meg.

Egyrészt, ha a 11.2.1. állítás bayesiánus kontextusban netán reprodukálhatatlan is, a demarkációs érv továbbra is védhető marad, ha az emberi hitrendszer a bayesiánus bizonytalanságok ellenére is *elvben* kétértékűvé tehető egyes (a 11.2.1. állítás feltételeinek megfelelő) univerzumokon (például tudományos kontextusokban). Ha az MI ⌜hitrendszere⌝ szintén elvben kétértékűvé tehető, akkor az 1. állítás újra relevánssá válik és a 3. premissza újra védhető. Ha pedig az MI ⌜hitrendszere⌝ még elvben sem tehető kétértékűvé, akkor kapunk egy újabb demarkációs érvet: az emberi gondolkodás koherens módon kétértékűvé tehető, az MI ⌜gondolkodása⌝ viszont lényegileg bayesiánus.

Ugyanakkor bayesiánus bonyodalmak nélkül is meggyőzően illusztrálható, hogy az 1. állítástól a klasszikus elsőrendű logika feladásakor sem feltétlenül szabadulunk meg. Mint a 11.4. alfejezet megmutatta, a standard modális logikában, emiatt pedig a standard episztemikus logikában is igaz a 11.2.1. állítás egy változata. Ha az MI-nek bayesi episztemológiát tulajdonítunk, akkor vélhetőleg elismerjük, hogy a standard episztemikus logika egy modelljét is realizálja, emiatt pedig továbbra is vonatkozik rá az (1)–(4) érv egy változata.

Végezetül, akárcsak a többi matematikai alapú demarkációs stratégiánál, itt is felvethető, hogy ha netán sikeres is a szóban forgó demarkációs érv, akkor sem igazolja, hogy az MI emberi értelemben ne gondolkodna, ne lenne belső intellektuális tapasztalata. Lehet, hogy a fent tárgyalt érv helyes, de az MI ettől függetlenül rendelkezik intellektuális tapasztalattal, ugyanúgy töpreng, spekulál és tudatosan következtet, mint mi, csak az önreprezentációs képességei szerényebbek.

Ez a felvetés igaz, de az önreprezentációs demarkációnak nem az a célja, hogy általában a gondolkodást vitassa el az MI-től, hanem az, hogy a gondolkodás egy speciális formáját, a metaszintű gondolkodást, a gondolkodásról való gondolkodást vitassa el tőle. Metaszintű reflexióra minden normális ember képes, tehát ha az önreprezentációs demarkáció helyes, akkor az MI kognitív szempontból lényegileg különbözik tőlünk.

Ráadásul ha gondolataink lényegi tulajdonsága, hogy reflektálni tudunk rájuk – szintaktikai és szemantikai szempontból, valamint metaszintről –, akkor az önreprezentációs demarkációra egy általános demarkációs érv is építhető. Ekkor az emberi gondolatokra szükségszerűen igaz lesz, hogy tulajdonosuk reflektálni képes rájuk (szemantikai és szintaktikai szempontból, valamint metaszinten), az MI ⌜gondolataira⌝ viszont ez nem lesz igaz. Ha így van, akkor a gondolkodás általában különbözik a ⌜gondolkodástól⌝. Az önreprezentációs demarkációtól tehát egyenes út vezet az általános kognitív demarkációig, ha gondolataink lényegileg feltételezik a komplex önreprezentáció lehetőségét.

II. rész

Számításoldali demarkáció

13.

Implementáció

A számítógépek programokat futtatnak, és az MI maga is egy program, ami számítógépeken fut. Milyen típusú fizikai jelenség ez? Mit kell tennie egy fizikai tárgynak, hogy egy programot futtasson?

A programok futtatását a továbbiakban „implementáció"-nak nevezem, a „programot implementál", „programot futtat", „számítást végez" szókapcsolatokat pedig szinonimaként használom. A „program" és az „absztrakt számítási modell" kifejezések ugyanazt jelentik majd. „Program"-on nem a számítógép processzorában ténylegesen jelen levő programokat értem, hanem egy absztrakt struktúrát vagy működési elvet, melynek paradigmatikus esete a Turing-gép. „Számítógép"-nek általában számítást végző fizikai rendszereket nevezem annak ellenére, hogy az implementáció egyes felfogásai szerint nem minden számítást végző rendszer *gép* abban az értelemben, hogy emberek építették volna számítások elvégzésére. De ennyi pontatlanságot megér, hogy ne kelljen a „számítást végző fizikai rendszer" kifejezést ismételgetni.

13.1. Az Implementációs Tézis

Milyen fizikai jelenség az implementáció? Mit kell tennie egy fizikai tárgynak ahhoz, hogy számítást végezzen? Egy olyan $\mathscr{I}(x, \mathbf{P})$ relációt keresünk, ami egy x fizikai tárgy és egy $\mathbf{P}$ program (mint absztrakt számítási modell) között pontosan akkor áll fenn, ha x implementálja $\mathbf{P}$-t.

Az MI metafizikájának egyik kulcskérdése ez, noha a demarkációs viták-ban szinte sosem esik róla szó. Minden MI egy program implementációja, a mentális képességeit pedig (ha vannak neki) vélhetőleg ennek – a program implementálásának – köszönheti, nem pedig az implementációt kísérő egyéb fizikai jelenségeknek. Ennek igazolására tegyük fel, pusztán a példa kedvéért, hogy a ChatGPT 7-nek már lesz tudata, de továbbra is a maiakhoz hasonló (bár náluk nagyobb teljesítményű) számítógépeken fut, a maiakhoz hasonló (bár komp-

lexebb) nagy nyelvi modell formájában. Az alábbi állítások közül melyik igaz ebben az elképzelt jövőbeli szituációban?

A A ChatGPT 7-nek amiatt van tudata, hogy egy számítógép egy programot (egy nagy nyelvi modellt) implementál.

B A ChatGPT 7-nek a szóban forgó program implementálásától független fizikai tények miatt van tudata, például a processzorokat hűtő ventilátorok pörgése, a tápegység működése, a vezetékek anyaga és formája, a processzorban levő félvezetők kémiai felépítése vagy más hasonló tények miatt.

Összesen négy elvi opciónk van: csak *A* igaz, csak *B* igaz, *A* és *B* is igaz, *A* és *B* is hamis. Ez utóbbit kizárhatjuk, hiszen fizikailag megmagyarázhatatlanná tenné, hogy a ChatGPT 7-nek mitől van tudata.

A „csak *B* igaz" opciót szintén kizárhatjuk. Ha a ChatGPT 7-nek nincs anyagtalan lelke, akkor tudatának nyilván szoros köze van a futtatott programhoz. Ennek tagadása ahhoz hasonló álláspontho vezetne, mint ha valaki fizikalista lenne (elutasítaná a testtől független elme létét), ugyanakkor tagadná, hogy az agynak köze van az öntudathoz, és azt feltételezné, hogy az öntudat a gyomorból, tüdőből, belekből és egyéb, koponyán kívüli szervekből ered. Ez teljesen implauzibilis, és analóg okokból a „csak *B* igaz" opció is az.

Az „*A* és *B* is igaz" opció logikailag koherens, ha az „amiatt *P*, hogy *Q*" típusú állításokban a *P* és *Q* közti magyarázati viszony részleges lehet, vagyis az „amiatt" szót a „részben vagy egészében amiatt" értelmében használjuk. Ekkor *A* és *B* együttes felvállalásával azt állítjuk, hogy a ChatGPT 7 tudata *részben* a számítógép által implementált programnak, *részben* a gép egyéb fizikai tulajdonságainak, vagyis a szoftver és a hardver együttműködésének köszönhető. Ez a nézet az arisztoteliánus elmefilozófiára hasonlít, ahol az ember mint gondolkodó lény test és lélek elválaszthatatlan fúziója. Az MI esetében egy ilyen tézisből az következik, hogy a ChatGPT 7 programja bizonyos számítógépeken implementálva tudatossággal jár együtt, másfajta számítógépeken implementálva pedig nem. De ez elég különös lenne. Az illusztráció kedvéért tegyük fel, hogy a ChatGPT 7 kódja szilikon-chipeken futtatva tudatossággal jár együtt, kvantumszámítógépeken futtatva viszont nem. Milyen racionális magyarázat lehet erre? Miért generál ugyanaz a program tudatot, ha tranzisztorokon fut, miközben *nem* generál tudatot, ha ionokon és lézer-impulzusokon (vagy egyéb kvantum-architektúrán) fut? Az MI-ről intuitíve feltételezzük, hogy sokféleképpen realizálható és realizálása azon múlik, hogy a megfelelő programot implementálja-e egy fizikai rendszer. Aki azt állítja, hogy a ChatGPT 7 csak akkor okoz tudatosságot, ha tranzisztorokon (vagy csak akkor, ha lézerrel manipulált ionokon stb.) fut, hasonló álláspontot képvisel, mint az elmefilozófiában John Searle (1980), aki szerint az elmét csak biológiai organizmusokon lehet realizálni, számítógépeken nem, hiába futtatja egy számítógép ugyanazt

a szoftvert, mint az elme. A többszörös realizálhatóság Searle szerint a karteziánus dualizmus rejtett továbbélése. De ez a felvetés az MI esetében teljesen *ad hoc*. (Talán az emberi tudat esetében is az, de ez most mindegy.)

Jó okunk van tehát azt feltételezni, hogy csak A igaz, vagyis ha az MI-nek van tudata, akkor tudata egy program implementációjából ered. Általában, az MI kognitív tevékenysége, függetlenül attól, hogy mennyire hasonlít az emberére, egy program futtatásának köszönhető:

Implementációs Tézis (IT) Az MI amiatt ⌜gondolkodik⌝, hogy egy fizikai rendszer egy absztrakt számítási modellt implementál.

Fontos hangsúlyozni, hogy **IT** nem zárja ki a járulékos fizikai körülmények fontosságát. Ahhoz, hogy egy programot implementáljunk, rengeteg olyan dologra van szükség, ami szigorúan véve nem része az implementációnak. A processzornak áramhoz kell jutnia, nem szabad túlmelegednie, kommunikálnia kell a gép egyéb komponenseivel, és így tovább. Ilyen értelemben az MI tudata (ha létezik) a program implementációjától konstitutív értelemben független fizikai tényeknek is köszönheti a létét. De az MI tudata és a hűtőventilátor, tápegység stb. közti viszony eltér attól a viszonytól, amire **IT** utal. **IT** egy olyan „amiatt P, hogy Q" típusú viszonyra utal, ami csak akkor áll fenn, ha Q igazsága *garantálja* és *magyarázza* P igazságát. Ez nem érvényes az MI tudata és a hűtőventilátor vagy tápegység közti viszonyra, hiszen a ventilátor és a tápegység tevékenysége nem garantálja egy MI jelenlétét. Tehát **IT** összefér azzal, hogy az MI létezésének vannak szükségszerű fizikai feltételei.

13.2. Az Implementációs Tézis metafizikája

IT a metafizikai eredettel (*metaphysical grounding*) kapcsolatos állítást fejez ki. Viszonylag elterjedt nézet, hogy vannak alapvető (fundamentális) és derivatív (függő) létezők (tények, tulajdonságok), és utóbbiak (a derivatív létezők) metafizikai értelemben az előbbiekből (a fundamentális létezőkből) erednek. Az asztal tulajdonságai például nyilván az asztalt alkotó atomok tulajdonságaiból erednek: az asztal derivatív létező, metafizikailag függ az atomjaitól. Az atomjai pedig metafizikailag függnek az őket alkotó kvarkoktól és leptonoktól. A derivatív létezők nem igényelnek valódi ontológiai elköteleződést, mert létüket garantálják a fundamentális létezők. Ha Isten megteremti az asztal összes atomját (asztal alakban elrendezve), akkor nem kell ezek után magát az asztalt is megteremtenie. Ahhoz tehát, hogy kimerítő leírást adjunk a világról, elég a fundamentális létezőket (és az eredet törvényeit) leírni.

A metafizikai eredet elmélete éles szakítást jelent a quine-i „lapos" ontológiával (Schaffer 2009). Quine (1948) számára az ontológia alapkérdése az volt, hogy mi létezik, mik azok a dolgok, melyek felett kvantifikálunk. Ezzel kapcsolatban Quine óvatosságra intett, és azt javasolta, a legjobb tudományos elméleteinknek az elsőrendű logika segítségével ráncba szedett, precíz változatát vegyük

alapul. Ha a metafizika alapkérdésének ehelyett azt tekintjük, hogy mi miből *ered*, akkor a Quine számára lelki terhet jelentő problémák egy része megoldódik. Többé nem az lesz a kérdés, hogy tényleg *léteznek-e* asztalok (vagy csak asztal alakban elrendezett elemi részecskék léteznek), ami a quine-i módszertant szentírásként kezelő Peter van Inwagen (1990) és vitapartnerei számára súlyos metafizikai dilemmának tűnt. A kérdés az, hogy az asztalok léte miből *ered*. Ha meg tudjuk nevezni metafizikai alapjukat, akkor a létük nem igényel további (az alapjukon túlmutató) elköteleződést. A létezők kanonikus listája helyett tehát a *fundamentális* létezők kanonikus listáját keressük, és nyugtalanságot csak akkor kell éreznünk, ha ebben a kontextusban kényszerülünk ontológiailag kényelmetlen vállalásokra. Az, hogy valaki elfogadja az asztalok, tulajdonságok vagy elmék létét, nem feltétlenül jelent ontológiai eltévelyedést, hiszen tekintheti őket derivatív létezőknek.

A metafizikai eredet elmélete az utóbbi évtizedek egyik legfontosabb metafizikai fejleménye, és hívei szerint sok korábbi filozófiai vitát újraértelmez. Raven (2020: 1) úgy véli, lényegében a filozófia összes alapkérdése az eredetre vonatkozik. Mitől igazságos egy társadalom? Mitől van egy mondatnak jelentése? Mitől rendelkezik valaki tudással? Mitől jó vagy rossz egy cselekedet? Mi miatt van egy lénynek tudata? Mitől szép egy műalkotás? Az ilyen „mitől", „mi miatt" (*in virtue of what*) kérdések (az eredethívők szerint) egytől egyig a metafizikai eredettel kapcsolatos kérdések.

A metafizikai eredet *posztmodális* jelenség (Sider 2020), vagyis (állítólag) nem artikulálható modális fogalmak (a szükségszerűség és lehetőség fogalmának) segítségével. Az 1980-as és 1990-es években tipikusan a szükségszerűségre, konkrétabban a szupervenienciára hivatkozva fogalmaztak meg az analitikus filozófusok olyan állításokat, melyek a metafizikai eredetről szóló mai tézisekre hasonlítanak. Az elmefilozófiai fizikalizmus standard megfogalmazása például az volt, hogy az elme az agyon (vagy általában a fizikai valóságon) szuperveniál. Az eredethívők szerint ezek a tézisek valójában az eredetről szólnak, és nem a szupervenienciára hivatkozva kell tárgyalni őket, mert az eredet és a szupervenienciа nem azonos (Schaffer 2009: 364). Az eredet egyrészt azért nem azonos a szupervenienciával, mert a szupervenienciа lehet szimmetrikus (pl. az asztal létén szuperveniál az asztalt alkotó atomok léte), de az eredet nem. Másrészt a szükségszerű dolgok definíció szerint tetszőleges kombinációban szuperveniálnak egymáson, az eredethívők szerint viszont nemtriviális eredetviszonyokban állhatnak. A számok például (egyes felfogások szerint) részben az üreshalmazból erednek, nem pedig fordítva, viszont a számok és az üreshalmaz léte egymáson szuperveniál amiatt, hogy szükségszerűek. Fine (1994) legendássá vált példájával élve Szókratész léte szuperveniál a Szókratészt tartalmazó egyelemű halmaz, {Szókratész} létén (és megfordítva), de emiatt nyilván nincs köztük kölcsönös ontológiai függés. A metafizikai eredet tehát egy sajátos, a modalitásra nem redukálható metafizikai jelenség.

Azért időztem hosszabban az eredetnél, hogy jelezzem, az Implementációs Tézisnek komolyan vehető metafizikai tartalma van. Ezzel kapcsolatban a következő terminológiai kikötéseket teszem. Az eredet metafizikai értelemben „mindenevő" viszony, mely tárgyak, állapotok és események között is fennállhat. Ha G eredete F, akkor F metafizikai értelemben *magyarázza* G-t: G azért létezik (vagy azért olyan, amilyen), mert F létezik (vagy mert F olyan, amilyen). Az eredet *relatív* fundamentalitást jelez: valaminek a metafizikai eredete maga is eredhet valamiből. Az eredet továbbá lehet teljes vagy részleges: ha G részben F-ből ered, akkor G létét részben F magyarázza, de magyarázhatja még valami más is. Az előző alfejezetben szereplő A és B téziseknél az $A \wedge \neg B$ opció annak az állításnak felel meg, hogy az MI tudata metafizikailag teljes egészében a szoftverből ered, a $\neg A \wedge B$ opció annak, hogy teljes egészében a hardverből, az $A \wedge B$ opció annak, hogy részben a szoftverből és részben a hardverből, a bizarr $\neg A \wedge \neg B$ opció pedig annak, hogy az MI tudata nem ered sem a szoftverből, sem a hardverből.

13.3. Definíciós kritériumok

Tegyük fel, hogy egy egyszerű programot (mondjuk a természetes számokon végzett összeadást) több különböző fizikai rendszeren implementálunk: fémrudakon és fogaskerekeken (mint Leibniz tette), régimódi vákuumcsöveken, egy modern mikrochipen, valamint egy hidraulikus szerkezeten, ami gátakat emel fel és ereszt le a kalkuláció során. Mi a közös ezekben a fizikailag teljesen eltérő folyamatokban – a fogaskerekek mozgásában, a vákuumcsövek villódzásában, a mikrochip tranzisztoraiban, a vízáramlásban –, amitől mindegyik ugyanazt a számítást végzi? Objektív, fizikai tulajdonságuk, hogy ugyanazt a programot futtatják? Ha igen, pontosan *melyik* tulajdonságuk ez? Ahhoz, hogy megértsük az MI természetét, erre a kérdésre kell választ adnunk. Definiálnunk kell, mit jelent implementálni egy programot.

Mikor lesz egy ilyen definíció adekvát? A továbbiakban a Gualtiero Piccinini (2015: 11–14) által javasolt kritériumrendszer egy változatát fogom alkalmazni a rivális hipotézisek összehasonlítására. Eszerint az implementáció természetéről szóló ideális $\mathscr{I}$ elméletnek a következő kritériumokat kell kielégítenie:

Objektivitás $\mathscr{I}$ alapján az implementáció legyen objektív jelenség. Ne következzen például $\mathscr{I}$-ből, hogy a szoba fala egy szövegszerkesztőt implementál.

Magyarázat Ha egy fizikai rendszer egy absztrakt számítási modellt implementál, akkor a modell adjon (legalább részleges) *magyarázatot* a rendszer viselkedésére (nem a metafizikai eredet értelmében, hanem számítástechnikai értelemben). Ha $\mathscr{I}$ szerint az implementált számítás nem magyarázza az implementáló rendszer fizikai viselkedését, akkor jó okunk van alternatív definíciókat preferálni.

Visszaigazolás $\mathscr{I}$ igazolja vissza a köznapi és informatikai ítéleteket azzal kapcsolatban, hogy mely fizikai rendszerek végeznek számításokat. Ha $\mathscr{I}$-ből például az következik, hogy az okostelefonok nem végeznek számításokat, akkor $\mathscr{I}$ nem megfelelő definíció.

Kizárás $\mathscr{I}$ igazolja vissza a köznapi és informatikai ítéleteket azzal kapcsolatban, hogy mely fizikai rendszerek *nem* végeznek számításokat. Ha például $\mathscr{I}$ szerint a hurrikánok és naprendszerek számításokat végeznek, akkor $\mathscr{I}$ nem megfelelő definíció.

Piccinini két további kritériumot is kidolgoz: a definíció engedje meg számítási hibák (*miscomputation*) lehetőségét, továbbá igazolja vissza a számítási kapacitásokkal kapcsolatos tényeket. Ezeket a kritériumokat nem használom, mert szerintem nem változtatnak a kialakuló összképen.

Az **Objektivitás**, **Magyarázat**, **Visszaigazolás** és **Kizárás** nem megkérdőjelezhetetlenek, és van olyan implementáció-felfogás – a formalizmus –, ami szinte egyiknek sem felel meg. A formalizmus tárgyalásakor (15. fejezet) hosszabban kitérek majd rá, miért érdemes jobban bíznunk a kritériumokban, mint az őket megsértő elméletekben.

A következő fejezetekben sorra veszem az implementáció ma létező definícióit, és amellett fogok érvelni, hogy a legmeggyőzőbb definícióból kézenfekvően adódik egy számításoldali demarkációs érv. Szó lesz *a szemantikai* felfogásról, ami a fizikai számítást információfeldolgozásnak tekinti (14. fejezet), a *formalista* felfogásról, ami puszta formális analógiát követel meg a program és a számítógép működése között (15. fejezet), a *strukturalista* felfogásról, ami a számítógép műszaki felépítésére vagy a programból kiolvasható kontrafaktuális összefüggésekre hivatkozik (16. fejezet), az *eszközfelfogásról*, ami a számítógépeket emberi célok eléréséhez köti (17. fejezet), végül az *idealista* felfogásról (18–19. fejezet), ami az implementációt elmefüggőnek, külső megfigyelők interpretációiból eredő jelenségnek tekinti. A 20. fejezet végül bemutatja, hogyan adódik az implementációs idealizmusból egy erős demarkációs érv.

14.

A szemantikai felfogás

Közhely, hogy a számítógépek adatokkal dolgoznak. A számítógéppel foglalkozó iparág neve „informatika", és a számítógépek kapcsán lépten-nyomon információról beszélünk. Emiatt magától értetődőnek tűnhet, hogy a programok implementálása valamiféle információfeldolgozást jelent. Ez az álláspont – az implementáció *szemantikai* felfogása – az elmefilozófusok között különösen népszerű. Jerry Fodor (1975: 34) egyik aranyköpése szerint „nincs számítás reprezentáció nélkül" (*no computation without representation*). De nem kell elmefilozófiával foglalkoznunk, hogy a szemantikai felfogás vonzónak tűnjön. A számítógépekkel kapcsolatos köznapi és szakértői diskurzus is ezt sugallja.

A szemantikai felfogás elemzése előtt hadd vázoljak fel két egyszerű példát, melyek a szemantikai felfogás fő vonzerejét és fő hátulütőjét illusztrálják. Az első Rescorla (2013: 694–695) egyik gondolatkísérletén alapul, a másik pedig Piccinini (2015: 35) egyik érvén.

Legyen **T** egy véges szalagú Turing-gép, ami bináris számrendszerben egyet hozzáad az inputhoz, vagyis az $f(x) = x + 1$ függvényt számítja ki. **T** ábécéje a $\{0, 1, \text{ÜRES}\}$ halmaz, felépítése pedig a 2. fejezetben bemutatott $\mathbf{T}^{+1}$ géphez hasonlít. Az alábbi input/output összefüggések jellemzik:

Input	Output
0	1
1	10
11	100
100	101
...	...

Tegyük fel, hogy ténylegesen megépítjük **T**-t. Legyen fizikailag realizált változatának neve **TR**. Tegyük fel azt is, hogy **TR** egy olyan felhasználóhoz kerül, aki csak a tízes számrendszert ismeri és nem látta **TR** programját. Elkezd kísérletezni **TR**-el, és azt tapasztalja, hogy a masina az 1-nél nagyobb számjegyeket nem hajlandó felismerni, a nullákból és egyesekből álló számok esetén

pedig fura dolgokat csinál. Ha 0-t kap inputnak, hozzáad egyet, de ha 1-et, akkor megszorozza tízzel. A 100-hoz szintén hozzáad egyet, a 11-et viszont megszorozza tízzel és levon belőle tízet. Emberünk tehát arra jut, hogy **TR** egy bonyolult és a gyakorlatban teljesen haszontalan függvényt számít ki, ami időnként az $f(x) = 10x$ függvénnyel esik egybe, máskor az $f(x) = x + 1$ vagy $f(x) = 10x - 10$ függvénnyel, 1-nél nagyobb számjegyeket tartalmazó inputok esetében pedig definiálatlan.

TR tehát különböző szemantikák alapján különböző függvényeket számít ki. Ha kettes számrendszerben értelmezzük, akkor **TR** az $f(x) = x+1$ függvényt számítja ki; ha tízes számrendszerben, akkor egy sokkal bonyolultabb (parciális) függvényt számít ki. Az tehát, hogy egy fizikai tárgy mint számítógép milyen programot implementál, részben az általa manipulált szimbólumok jelentésétől függ. Nincs számítás reprezentáció nélkül

Következzen egy példa, ami ezzel ellentétes felfogást sugall. Képzeljük el, hogy eljutunk a Marsra, és egy barlangrendszerben felfedezzük egy rég kihalt, időszámításunk előtti marsi civilizáció írásos emlékeit. A betűk teljesen idegenek számunkra és fogalmunk sincs, melyik szó mit jelent (talán nem is szavakból állnak a jelsorozatok). Találunk egy gépet is, nevezzük **M**-nek, ami marsi szimbólumok rövid sorozatait olvassa be, a szóközöknél szétválasztja őket, és átrendezi a sorozatokat. Az átrendezés logikája hasonló ahhoz, mintha egy földi számítógép ABC-sorrendbe rakna szavakat. Noha a marsi betűk teljesen idegenek számunkra és nem hordoznak semmi jelentést, nehéz tagadni, hogy **M** számításokat végez: egy szabály alapján rendez át sorozatokat. Matematikai értelemben azt a függvényt számítja ki, melynek bemenete természetes számok véges sorozata, kimenete pedig ugyanazoknak a számoknak nagyság szerint rendezett sorozata. De **M** anélkül végzi ezeket számításokat, hogy az inputoknak és outputoknak jelentése lenne (mivel a marslakók kihaltak). Tehát egy program implementációjához nincs szükség szemantikára.

A két gondolatkísérlet egymással ellentétes intuíciókat mozgósít: **TR** esetében mintha szükség lenne a számításhoz szemantikára, **M** esetében pedig nem. Melyik nézőpont a helyes? Felvethetnénk, hogy az első, mert azt kell figyelembe vennünk, hogy egy gép milyen *matematikai* függvényt számít ki, ez pedig az interpretációtól függ. (Még **M** esetében is interpretációt igényel, hogy ABC-sorrendről vagy fordított ABC-sorrendről van szó.) Ezzel két probléma van. Egyrészt nem teljesen egyértelmű, hol húzódik a matematika határa. Ha $\mathfrak{M}$ a marsi betűk halmaza, $\overline{\mathfrak{M}}$ pedig a belőlük képzett véges sorozatok halmaza, akkor egy $\overline{\mathfrak{M}}$-ből $\overline{\mathfrak{M}}$-be képző függvény *matematikai* értelemben tökéletesen definiált objektum ($\overline{\mathfrak{M}}$ elemeiből képzett párok halmaza, ahol $\overline{\mathfrak{M}}$ minden elemének pontosan egy párja van). Az **M** gép tehát egy matematikai függvényt számít ki, annak ellenére, hogy nincs szemantikája.

De lehet, hogy „matematikai függvény"-en szigorúan csak *aritmetikai* függvényeket akarunk érteni, melyek számokból számokat állítanak elő. Ekkor **M** valóban nem egy matematikai függvényt számít ki. De ha ezt állítjuk, impliciten

feltételezzük a szemantikai felfogást, hiszen a számítógép belsejében szaladgáló elektromos jelek önmagukban nem szimbolizálnak számokat, interpretálni kell őket. Az intuíciók csatáját nem tudjuk azzal eldönteni, hogy a számítást aritmetikai függvények kiszámításával azonosítjuk, hiszen ezzel lényegében elismételjük a szemantikai felfogást támogató intuíciót anélkül, hogy igazolnánk. **M** példája azt mutatja, hogy nem egyértelmű, miért kellene a számításokat aritmetikai függvények kiszámítására korlátoznunk.

A konfliktus feloldására a filozófiában megszokott módon bevezethetünk két különböző fogalmat, a tág és szűk számítások fogalmát. Egy *tág* értelemben vett számításnak része a gép által manipulált fizikai mintázatok információtartalma. **TR** különböző számításokat végez aszerint, hogy kettes vagy tízes számrendszerben értelmezik a szalagjára írt jeleket.

A *szűk* értelemben vett számítás az a fizikai folyamat, ami a számítógépben attól függetlenül zajlik, hogy milyen interpretáció kapcsolódik outputjaihoz a külvilágban. **TR** ugyanazt a szűk értelemben vett számítást végzi, akár kettes, akár tízes számrendszerben értelmezik a szalagjára írt jeleket. **M** pedig csak szűk értelemben vett számítást végez, mióta a marslakók kihaltak.

E megkülönböztetés segítségével megmagyarázhatjuk, miért vezet a két gondolatkísérlet látszólag ellentétes eredményre. **TR** a tág értelemben vett, **M** pedig a szűk értelemben vett számítás fogalmát illusztrálja. Tehát a „számítás" szó eltérő jelentései szerint igaz, hogy **M** számításokat végez, **TR** pedig különböző felhasználók számára különböző számításokat végez. Emiatt a gondolatkísérletek nem mondanak ellent egymásnak; két különböző jelenségről, a szűk ill. tág számításokról van szó.

Bár a gondolatkísérletek konfliktusát feloldja ez a megkülönböztetés, a szemantikai felfogás körüli vitát nem dönti el, csak átfogalmazza az alapkérdést. A kérdés most az, hogy az implementációhoz *szűk* vagy *tág* számításokra van-e szükség. A tág számításoknak definíció része a szemantika, a szűk számításoknak pedig, mint **M** példája mutatja, nem.

Ahhoz, hogy a szemantikai felfogásról vitázni tudjunk, tisztázni kell, mit értünk „szemantikán". Három opciót veszek szemügyre: a természetes információt, a Shannon-információt és a naturalizált nyelvi jelentést.

14.1. Természetes információ

Az információ talán legkézenfekvőbb formája a Paul Grice (1989) által tárgyalt *természetes jelentés*, amire a hétköznapokban „x az y-t jelzi" vagy „x jele y" típusú kifejezésekkel utalunk. A kiütés a bárányhimlő jele, az évgyűrűk a fa életkorát jelzik, a füst tüzet jelez, és így tovább. A jel és jelölet közti viszony itt kontrafaktuális „ha/akkor" összefüggéseken alapul: ha a gyerek nem lenne bárányhimlős, nem volnának kiütései, ha a fa fiatalabb lenne, kevesebb évgyűrűje volna, ha nem égne valami, nem lenne füst. A természetes jelentés tehát egy-

fajta korreláció, amit többféleképpen is definiálhatunk. Mondhatjuk, hogy F_2 fizikai folyamat vagy állapot az F_1 fizikai folyamat vagy állapot jele, ha (a) F_2 nem lenne jelen, ha F_1 nem lenne jelen, vagy (b) F_1 és F_2 együttes fennállása valószínűbb, mint F_1 fennállása F_2 nélkül, vagy (c) F_2 tulajdonságai szisztematikusan összefüggnek F_1 tulajdonságaival (pl. ha nagyobb lenne a tűz, több lenne a füst).

A számítógépek által manipulált jelsorozatoknak többnyire van grice-i értelemben vett természetes jelentése. A billentyűzethez kapcsolódó adatbuszban keletkező elektromos jeleknek például az a természetes jelentése, hogy valaki nyomogatja a billentyűzetet, mert ez a két esemény tipikusan együttjár és szisztematikusan összefügg. De a természetes jelentés nem tűnik szükségszerűnek egy program implementálásához. Ha a billentyűzet adatbuszában elektromos zavar miatt keletkeznek jelek, és a szövegszerkesztő vagy a ChatGPT ezeket inputként értelmezi, akkor számítás zajlik a gépben annak ellenére, hogy a jeleknek nincs természetes jelentése. De nem feltétlenül kell a véletlenre apellálnunk ennek kimutatására. Tegyük fel, hogy valaki filmet néz a laptopján. A bináris számok pixelekké és hanghullámokká való átalakítása rengeteg számítást igényel, de ezeknek a számoknak nincs természetes jelentése. Felvethetnénk, hogy a képernyő egy adott pixelének színe és fényereje kontrafaktuális értelemben összefügg a pixel tartalmát kódoló számokkal, utóbbiak tehát természetes információt hordoznak. De ez a korreláció teljességgel konvencionális, és különböző videoformátumok esetében eltérhet, ráadásul a legtöbb mai videoformátum tömörítő algoritmusokat is használ, emiatt pedig a fájlban tárolt számokat eleve nem is lehet egy az egyben pixeleknek megfeleltetni. Nincs „természetes" korreláció a pixelek és bitek között, és ha lenne, akkor is lehetne ettől eltérő módon kódolni és dekódolni filmeket. Tehát a természetes információ nem feltétele a számításoknak.

14.2. Shannon-információ

A Shannon-információ (Shannon 1948) a természetes információ kifinomultabb, a kriptográfiában, számítástechnikában és távközlésben használt definíciója, ami az információ matematikai elméletnek is alapja (Cover & Thomas 2006). Piccinini (2015: 227-229) intuitív megfogalmazása szerint a Shannon-információ azt méri, hogy az „üzenet" – ez szemantikai fogalomnak hangzik, de lehet a sebességmérő mutatójának állása is – mennyire *valószínűtlen*. Egy valószínűtlen eseménynek nagyobb az információtartalma, shannoni értelemben, mint egy valószínű eseménynek. Egy lottóötösnek nagyobb a Shannon-információtartalma, mint a nulla találatnak, mert sokkal valószínűtlenebb.

Az információ továbbítása a shannoni keretben korrelációt jelent. A sebességmérő akkor továbbítja az információt, ha a kocsi sebessége és a sebességmérő mutatójának állása erősen korrelál. Az információ mennyiségének és az infor-

mációtovábbítás hatékonyságának számszerűsítésével Shannon érdekes matematika tételeket tudott bizonyítani. A „zajos csatornák alaptétele" például azt mondja ki, hogy ha egy csatorna adásába véletlen zaj keveredik, akkor az információközvetítés hiba-aránya tetszőlegesen kicsire szorítható, amennyiben a közvetített információ mennyisége kisebb a csatorna kapacitásánál.

Bár a Shannon-információ mérnöki szempontból hasznos fogalom, programok implementációjához nem elégséges és nem is szükséges a Shannon-információ jelenléte. Nem elégséges, mert egy régi szalagos magnetofon shannoni értelemben információt közvetít, mégsem végez számításokat (Rescorla 2024: 6.1). Másrészt a Shannon-információ jelenléte nem is szükséges a számításokhoz (Piccinini 2015: 238). Egy teljesen determinisztikus, külső inputokra érzéketlen számítógép (például egy olyan gép, ami bekapcsoláskor prímtényezőire bont egy előre meghatározott számot, majd leáll) számításokat végez, mégsem hordoz Shannon-információt. További probléma – a szemantikai felfogás szempontjából –, hogy eleve nem világos, mit jelentene Shannon-információt *feldolgozni*. Ha például egy számítógép szenzorok adatai alapján kiszámítja, mekkora az átlagos páratartalom, akkor Shannon-információt közvetít (mert az inputjai és outputjai korrelálnak), de a Shannon-információ mentén nehéznek tűnik artikulálni, milyen értelemben *végez* számításokat. A számítás nem puszta korrelációt jelent a kommunikációs csatorna két vége között, hanem ahhoz van köze, ahogy az input a rendszer belsejében átváltozik outputtá.

14.3. Naturalizált szemantika

A szemantikai felfogást elvben felépíthetjük a *nyelvi* értelemben vett jelentés fogalmára is: akkor végez egy fizikai rendszer számításokat, ha egyes állapotainak nyelvi (vagy általában kognitív) tartalma van. Nevezzük ezt *szigorú* szemantikai felfogásnak.

A szigorú szemantikai felfogás nem egy jól definiált, monolitszerű elmélet, hanem egymástól élesen eltérő elméletek családja, melynek tagjait a nyelvi jelentés alternatív definíciói különböztetik meg egymástól. Mivel a nyelvi jelentés természete erősen vitatott kérdés, a szigorú szemantikai felfogás változatai radikálisan eltérhetnek. Csak az illusztráció kedvéért tekintsük azt a dualista elméletet, hogy egy fizikai tárgy vagy folyamat akkor hordoz jelentést, ha egy mentális szubsztancia jelentést tulajdonít neki. Ez a jelentéselmélet a szigorú szemantikai felfogással kombinálva elmefüggővé teszi a számításokat, hiszen az következik belőle, hogy egy fizikai rendszer csak akkor végez számításokat, ha egy megfigyelő jelentést tulajdonít az állapotainak. Ez a tézis radikálisan eltér a szemantikai elmélet olyan variánsaitól, ahol egy fizikai rendszer nem a külső megfigyelők mentális állapotai miatt hordoz jelentést. Jelen alfejezetben az

utóbbi variánsokra koncentrálok, a számításokat elmefüggőnek tekintő nézeteket pedig az idealizmus címkéje alá sorolom (19. fejezet).[1]

Ahhoz, hogy a szigorú szemantikai felfogás ne váljon idealistává, a nyelvi jelentés *naturalizált* elméletére van szükség, ami a jelentést a gépekben vagy gép-világ viszonyokban horgonyozza le és nem hivatkozik a felhasználók interpretációira. Mivel a szemantika naturalizálásával kapcsolatos vita minimum fél évszázada zajlik, egyetlen alfejezet keretében nehéz alaposan áttekinteni a releváns opciókat. Az alábbiakban két leegyszerűsített példán próbálom bemutatni, miért nem tűnik működőképesnek a naturalizált szemantikára épülő implementáció-felfogás.

A naturalizált szemantika egyes változatai (pl. Fodor 1990) a külvilág és a jelentéshordozók közti oksági összefüggésekből eredeztetik a jelentést. Az alapvető intuíció szerint egy szó jelentése a szó használatát okozó dolgoktól függ: a „cica" például akkor jelenti a cicákat, ha használatát cicák okozzák. A történet persze ennél bonyolultabb, hiszen ha egy gyerek cicás rajzok alapján tanulja meg a „cica" jelentését, a „cica" akkor is a cicákat, nem pedig a cicás rajzokat jelenti számára (Loewer 2017: 178). De ezeket a komplikációkat most tegyük félre. Tételezzük fel, hogy rögzíteni lehet egy kellően bonyolult oksági kritériummal, hogy az „F" szó mikor jelöli az F-eket. A kulcskérdés jelen kontextusban nem az, hogy létezik-e ilyen kritérium, hanem az, hogy amennyiben létezik, kellenek-e számítások implementációjához ilyen oksági összefüggések.

Mielőtt ezt tisztáznánk, érdemes megemlíteni, hogy az oksági szemantika alapján az MI outputjai ugyanolyan értelemben rendelkezhetnek jelentéssel, mint a mi mondataink (Cappelen és Dever 2021). Ha pedig az oksági szemantikát a gondolati tartalmakra is vonatkoztatjuk, akkor az MI-nek ugyanúgy lehetnek intencionális állapotai, mint nekünk. Az oksági szemantika tehát kétségkívül fontos antidemarkációs fegyver. De a szigorú szemantikai felfogás szempontjából ez irreleváns, mert utóbbi nem specifikusan az MI belső folyamatairól szól, hanem általában *bármilyen* számításról, beleértve a mosógépek és mikrohullámú sütők által végzett számításokat. Ebben az általános kontextusban pedig nem egyértelműen relevánsak az oksági szemantika által feltételezett összefüggések. Tegyük fel, hogy egy mikrohullámú sütő visszaszámol n másodpercet és a nullát elérve kikapcsol. Egy ilyen mechanizmus nyilván programot futtat, még ha viszonylag primitívet is. De milyen (naturalizált) *nyelvi* tartalom található meg benne? A sütésre, melegítésre, teakészítésre stb. vonatkozó tartalom aligha van benne jelen, hiszen a visszaszámláló mechanizmus pontosan ugyanazt csinálná, ha a mikró teljesen üres lenne, vagy ha csak visszaszámolásra és kikapcsolásra lenne képes, melegítésre nem. Az egyetlen nyelvi tartalom, ami ehhez a számításhoz kapcsolható, a másodperc fogalma, hiszen a gép lényegében semmi mást nem csinál, mint másodpercenként eggyel csökkenti egy változó

[1] A szemantikai nézet egyes képviselői, például Shagrir (2022), nem hajlandóak elárulni, milyen szemantikai elmélettel dolgoznak („szemantikai pluralizmus"). Ezeket a felfogásokat metafizikailag értelmezhetetlennek tekintem.

értékét és ellenőrzi, nagyobb-e nullánál.[2] De nem világos, hogy a sütő egy belső állapotának miért kellene *nyelvi* értelemben azt jelentenie, hogy „Még n másodperc van hátra", és még kevésbé világos, hogy ha lenne is a sütőnek ilyen belső állapota, hogyan hordozná ez az állapot a szóban forgó jelentést az *oksági szemantika* alapján. Az, hogy a mikró n másodperc múlva ki fog kapcsolni, nyers fizikai törvényekből ered, nem pedig annak a következménye, hogy n másodperces múltbeli események bonyolult oksági viszonyban állnak a sütő jelenlegi állapotával. Ahhoz tehát, hogy egy fizikai tárgy számításokat végezzen, nincs szükség az oksági szemantika értelmében vett jelentésre.

Második példa. A naturalizált szemantika jelenlegi legnépszerűbb változata a *teleoszemantika*, ami a mentális állapotok tartalmát (és ebből eredően a nyelvi jelentést) az evolúciós előtörténetre vezeti vissza (Millikan 1984, Neander 2017, Papineau 1993, Schulte és Neander 2022: 3.2). Az alapötlet durván leegyszerűsítve az, hogy pl. egy béka A agyi állapota akkor hordozza azt a kognitív tartalmat, hogy „Ha kinyújtom a nyelvemet, el fogok kapni egy legyet", ha A hatására a béka megfelelő időpontban nyújtja ki a nyelvét, mikor egy légy repked a közelében, és ennek a reakciónak a kiváltása az A állapot biológiai értelemben vett funkciója (Loewer 2017: 181). A biológiai funkciókat a teleoszemantika hívei az evolúciós előtörténet alapján definiálják: az x organizmus T tulajdonságnak akkor a φ testi folyamat a funkciója, ha x elődei számára evolúciós előnyt jelentett, hogy T a φ folyamatot okozza (bővebben ld. Garson 2016: 3. fejezet, ill. alább 17.1. alfejezet).

A teleoszemantikai felfogás elsősorban a mentális tartalmakat akarja definiálni; a szavak jelentése a mentális tartalmak függvénye. „A cica alszik" mondat jelentését például arra vezetjük vissza, hogy ez a mondat fejezi ki az alvó cicákat reprezentáló mentális állapotainkat. Az átkötés részleteivel most nem foglalkozom. Az egyszerűség kedvéért azt feltételezem, hogy az implementáció teleoszemantikai felfogása ugyanolyan értelemben társít szemantikai tartalmat a számításokhoz, mint a teleoszemantika a mentális állapotokhoz.

Az implementáció teleoszemantikai felfogását Davidson (1987) Mocsáremberével analóg példákkal lehet cáfolni. Tegyük fel, hogy villám csap egy hulladéktelepbe, és rendkívül valószínűtlen kvantummechanikai véletlenek miatt összeáll a fémdarabokból egy vadonatújnak tűnő, működőképes mikrohullámú sütő, M_1. Legyen M_2 egy vele kvalitatívan azonos gép, ami viszont gyárban készült, gondos műszaki tervek alapján. Feltehető, hogy M_2-nek van egyfajta evolúciós előtörténete; az M_2-t létrehozó egyéneknek kétségkívül volt. A vita kedvéért azt is tételezzük fel, hogy ennek az evolúciós előtörténetnek köszönhetően hordoznak M_2 egyes állapotai – teleoszemantikai értelemben – jelentést, és részben e jelentéseknek köszönhető, hogy M_2 számításokat végez, például visszaszámol n másodpercet a kikapcsolásig. Nehéz elképzelni, hogyan ruházná

[2] Elvben felvethető, hogy a természetes számok fogalma is jelen van a rendszerben. Mivel a matematika értelmezése kihívást jelent az oksági szemantika számára (Benacerraf 1973), ezt az opciót félreteszem.

fel az evolúciós előtörténet M_2 állapotait ilyen tartalommal, de ez most mindegy. Az implementáció teleoszemantikai felfogása amiatt bukik el, hogy M_1 – a hulladéktelepen keletkezett mikró – nyilván szintén számításokat végez, amikor visszaszámol nulláig. M_1-et pontosan ugyanazokra a célokra lehet felhasználni, mint M_2-t, ugyanazok a fizikai folyamatok játszódnak le benne, és egy mérnök pontosan ugyanazokat a számítási folyamatokat tulajdonítaná neki. De M_1-nek nincs evolúciós előtörténete, emiatt pedig teleoszemantikai értelemben nem hordoznak jelentést a belső állapotai. Tehát a számításokhoz nincs szükség teleoszemantikai értelemben vett jelentésre.

Felvethetnénk, hogy M_1 folyamatait a felhasználók mentális állapotai ruházzák fel jelentéssel, utóbbiaknak viszont teleoszemantikai alapon nagyon is van tartalma. Ez kétségtelen (amennyiben a teleoszemantika igaz), de ezen az úton az implementáció idealista elméletéhez jutunk, mert külső megfigyelők mentális állapotaira vezetjük vissza, hogy egy fizikai rendszer (például M_1 vagy M_2) végez-e számításokat. Az idealizmus nem ördögtől való (sőt, véleményem szerint ez a helyes megközelítés), de ahogy korábban jeleztem, az ilyen nézeteket a 19. fejezetre halasztom és a szemantikai felfogás kapcsán csak a nem-idealista nézeteket vizsgálom.

A fenti két egyszerű példa nyilván nem zárja le végérvényesen a naturalizált szemantikára épülő implementáció-definíciókkal kapcsolatos vitát, már csak azért sem, mert maga a naturalizált szemantika bonyolult és szerteágazó téma. De én személy szerint nem látom, hogy a példákban jelzett elvi problémákat hogyan lehetne feloldani anélkül, hogy az implementációs idealizmus egy változatához jutnánk.

15.

A formalista felfogás

Ahhoz, hogy egy fizikai tárgy egy programot implementáljon, vélhetőleg a tárgy és a program (mint absztrakt számítási modell) közti strukturális hasonlóságra van szükség. Az implementáció formalista felfogásának lényege az, hogy valójában *csak* erre van szükség. Ha megfeleltethetőek egymásnak a gép működésének fázisai és a számítási modell absztrakt szabályai, akkor a gép az adott programot futtatja. Piccinini (2015) ezt *mapping account*-nak, leképezési felfogásnak hívja. Én formalistának nevezem, hogy hangsúlyozzam, mennyire elvont jelenségnek tekintik e felfogás hívei az implementációt.

Nyilvánvalónak tűnik, hogy egy gép akkor futtat egy programot, ha a gép működése bizonyos értelemben tükrözi a program struktúráját vagy engedelmeskedik a program által definiált szabályoknak. De mint látni fogjuk, a formalizmus ezt az alapötletet annyira liberálisan értelmezi, hogy trivializálja a számítás fogalmát. A formalizmus elleni standard kifogás emiatt az, hogy *pánkomputacionalizmus*hoz vezet: az következik belőle, hogy minden fizikai tárgy, még egy vödör víz is, számításokat végez. Az alábbiakban három rövid és egy hosszabb példán keresztül mutatom be, hogyan trivializálja a formalizmus az implementáció fogalmát, majd amellett érvelek, hogy a formalizmus emiatt elfogadhatatlan.

A formalizmus buktatóinak gyors illusztrálására tekintsük az alábbi egyszerű formalista elméletet:

> Számítógép minden olyan fizikai rendszer, melynek dinamikai fejlődése „input" állapotok valamely halmazából „output" állapotok valamely halmazához vezet. Az állapotokat egy meghatározott módon felcímkézzük, a gépet preparáljuk az egyik input-címkéhez tartozó állapotban, majd megmérjük a kimeneti állapotot... [A] mért kimeneti állapot a preparált bementi címke specifikus f függvénye lesz [...]. [A] gépről ilyenkor azt mondjuk, hogy „kiszámítja" az f függvény értékeit. (Deutsch 1985: 97)

Ez a definíció a számítások implementációját lényegében az input/output összefüggésekre redukálja. Egy $f(x)$ matematikai függvényt bármely fizikai rendszer komputál, ami az x címkével ellátott kezdőállapotból az $f(x)$ címkével rendelkező végállapotba fejlődik.

Íme annak bizonyítása, hogy e nézet alapján egy szabadrúgás minden, véges intervallumokon definiált függvényt kiszámít. Tegyük fel, hogy az f függvény az $[a, b]$ szakaszról képez a $[c, d]$ szakaszra, ahol a, b, c, d valós számok ($a < b$, $c < d$). Tekintsünk egy tetszőleges szabadrúgást és tegyük fel, hogy a csatár lábának e energiájától függ, hogy a labda hány másodpercet tölt a levegőben. Az e érték 0 és E között mozoghat, a labda pedig az elrúgást követően $T(e)$ másodperc múlva ütközik valamivel (a földdel, kapufával stb.), ahol $T(e)$ a newtoni mozgástörvényeken és a szabadrúgás körülményein alapuló függvény.[1] Képezzük le a $[0, E]$ intervallumot az $[a, b]$ intervallumra ($e \mapsto a + e \cdot \frac{(b-a)}{E}$), a $[T(0), T(E)]$ intervallumot hasonló módon a $[c, d]$ intervallumra, $T(e)$-t pedig feleltessük meg $f(x)$-nek. Ezzel máris beláttuk, hogy a szabadrúgás az $f(x)$ függvényt komputálja. Mivel $f(x)$ bármely (intervallumról intervallumra képző) függvény lehet, minden szabadrúgás minden ilyen függvényt komputál. És mivel ez nemcsak a szabadrúgásokra, hanem tetszőleges fizikai folyamatra igaz, minden fizikai folyamat minden (intervallumról intervallumra képző) függvényt komputál. A formalizmus fenti egyszerű definíciójától tehát pár lépésben eljutottunk a pánkomputacionalizmus egy formájához.

Felvethetnénk, hogy csalás történt, mert a labda röptének időtartama valójában nem azonos $f(x)$-el a fenti transzformáció szerint. A szabadrúgás csak akkor komputálná mondjuk a $\sin(x)$ függvényt, ha igaz lenne, hogy a labda röptének időtartama $\sin(a + e \cdot \frac{(b-a)}{E})$, de ezt nem garantálja, hogy a $[T(0), T(E)]$ intervallumot lineárisan ráképeztük a $[c, d]$ intervallumra. Deutsch pont ezért beszél a fenti idézetben *mérések*ről. De nyilván használhatunk olyan leképezést $[T(0), T(E)]$ és $[c, d]$ között, hogy a $T(e)$ értékhez a $\sin(a + e \cdot \frac{(b-a)}{E})$ érték tartozzon, és építhetünk egy órát, ami ezt az értéket „méri". Mivel Deutsch semmi kikötést nem tesz a címkézés módszertanával kapcsolatban, egy ilyen címkézés érvényesnek minősül, és ilyen címkézés mellett a szabadrúgás már a szinuszfüggvényt komputálja. Ha a formalista arra hivatkozik, hogy a szóban forgó mérőeszköznek semmi haszna, ezért használata nem megengedett, vagy felpanaszolja, hogy a mérőeszközöknek valódi fizikai mennyiségeket kell mérniük, akkor túllépünk a formalista felfogáson és hasznosságot vagy valódi fizikai mennyiségekkel való kapcsolatot követelünk az implementációhoz. Ez a kiegészítés az implementáció egyéb – strukturalista, szemantikai, teleológiai vagy idealista – felfogásaihoz vezet, tehát a formalizmus feladását követeli.

Deutsch formalizmusa azért trivializálható könnyen, mert nem vesz tudomást a számítások belső struktúrájáról. Emiatt joggal vádolható azzal is, hogy

[1] E függvénynek szigorúan véve vektor-bemenettel kellene dolgoznia (ami reprezentálja, hogy a csatár lába melyik irányban mennyi energiát ad át a labdának), de ez mellékes. Tegyük fel, hogy a vektor első két komponensét fixáljuk, e pedig a harmadik komponens.

elvéti az implementáció lényegét. Egy szám négyzetének kiszámítása nem abban merül ki, hogy az x címkéjű bemenet hatására előjön az x^2 címkéjű kimenet. Ahhoz, hogy a kezdő- és végállapot egy fizikai számítás kezdő- és végpontja legyen, valami másnak is történnie kell. Például meg kell szoroznunk x-et önmagával valamilyen algoritmus szerint. Tehát nemcsak a számítás inputjait és outputjait kell a fizikai rendszer kezdő- és végállapotaihoz rendelnünk, hanem a köztes fázisokat is meg kell feleltetnünk egy absztrakt számítási modell fázisainak. Az alábbi négy gondolatkísérlet azt mutatja be, hogy a formalizmusnak ez a szofisztikáltabb változata is pánkomputacionalizmushoz vezet.

15.1. Négy trivializáló érv a formalizmus ellen

A formalista felfogás alapgondolata az, hogy a számítógép működése valamilyen formális értelemben megfeleltethető a program műveleteinek (Chalmers 1996a: 311, Piccinini 2015: 17, Sprevak 2018: 5):

Formalizmus Az F fizikai rendszer akkor implementálja az A absztrakt számítási modellt, ha van olyan, F állapotait és A fázisait összekapcsoló α függvény, hogy ha az F_i fizikai állapotot az F_{i+1} fizikai állapot követi, akkor az absztrakt számítási modellben az $\alpha(F_i)$ fázist az $\alpha(F_{i+1})$ fázis követi.

Ez a definíció akkor lenne precíz, ha megmondanánk, mik egy absztrakt számítási modell fázisai (és eleve mi egy absztrakt számítási modell). A precíz változat nem fogalmazható meg anélkül, hogy kiválasztanánk a számítások egy specifikus modelljét. Ha például a Turing-gépeket tekintjük a számítások absztrakt modelljeinek, akkor a „fázisok" a szalag állapotainak és a parancsok sorszámainak lehetséges kombinációi („számítási lépés", ld. Monk 1976: 15). Ezt az absztrakt struktúrát – a számítási lépések sorát – kell megfeleltetni a **Formalizmus** alapján a Turing-gépet implementáló fizikai rendszer állapotainak. Copeland érve kapcsán (15.1.4) előkerül majd egy ennél sokkal realisztikusabb számítási modell is. Egyelőre maradjunk a fenti leírás által sugallt intuitív képnél: az implementált program és az implementáló fizikai rendszer működése párhuzamba állítható abban a tisztán formális értelemben, hogy a rendszer átmenetei megfeleltethetőek a program lépéseinek.

Az alábbiakban előbb Sprevak (2018) nyomán három rövid trivializációs érvet mutatok be, melyek azt próbálják igazolni, hogy formalista alapon gyakorlatilag minden fizikai rendszer, még egy vödör víz is, számításokat végez. Az első két érv egyszerű intuíció-pumpa; a harmadik, Putnamtól eredő gondolatmenet szintén kicsit elnagyolt, de történetileg fontos. Copeland (1996) trivializációs érve (16.1.4) azonban kellően alapos ahhoz, hogy ezeket az intuíciókat igazolja.

15.1.1 Lycan vödre

> Képzeljük el egy forrásvízzel teli, átlátszó műanyagvödröt, amit kitettek a napra. Mikroszkopikus szinten a vödör gigászi komplexitásforgatag: tele van konvekciós áramlásokkal, ész nélkül szaporodnak benne baktériumok és egyéb mikroorganizmusok, és így tovább. Molekuláris szinten ez még frenetikusabb aktivitást jelent. Miért ne lehetne ez a kavalkád annyira komplex, hogy egy rövid időre véletlenül implementál valami emberalkotta programot abban az értelemben, hogy precíz korreláció áll fenn a vödörben zajló mikroszintű események és a program input-, output- és állapotváltozói között? (Lycan 1981: 39)[2]

Lycan az elmefilozófiai funkcionalizmus lehetséges cáfolataként hozza elő a vödröt. Ha az elme lényegében egy program, a programok implementációja pedig formális összefüggésekben merül ki, akkor egy vödör forrásvíz is öntudatra ébredhet egy-egy pillanatra. Ez a példa nemcsak trivializálja az implementációt, de azt is érzékelteti, mi a formalizmus körüli vita tétje az MI kontextusában. Ha a formalista felfogás igaz, akkor egy vödör víz is futtatja néha a ChatGPT-t, sőt, annak jövőbeli verzióit is, melyeket még piacra sem dobtak. Ha pedig az MI gondolkodik, akkor időnként egy vödör víz is gondolkodik. A formalizmus tehát nemcsak a pánkomputacionalizmus rémét idézi meg (minden fizikai tárgy számításokat végez), hanem egyfajta funkcionalista alapú pánpszichizmushoz is vezet, ahol a fizikai világ minden darabkájának van (időnként) tudata. Vagy legalábbis ez a helyzet, ha a formalizmus igaz és Lycan vödre valóban a formalizmus egyik szükségszerű következményét illusztrálja.

15.1.2 Searle fala

> Elkerülhetetlennek látszanak az alábbi következtetések: 1. Bármely tárgyról adható olyan leírás, mely alapján digitális számítógépnek minősül. 2. Bármely program és bármely kellően komplex tárgy esetében adható olyan leírás, mely alapján a tárgy a szóban forgó programot implementálja. A hátam mögött levő fal például jelenleg a Wordstar programot [a Word egyik elődjét] futtatja, mert a molekulák rezgésének egyes mintázatai izomorfak a Wordstar formális struktúrájával. De ha a fal a Wordstart futtatja, akkor egy kellően nagy fal bármelyik programot futtatja, beleértve az agy által implementált programot, akármi is az... (Searle 1992: 208–209)

Searle szintén egy elmefilozófiai vita kapcsán hozza elő ezt a példát, amit a Kínai Szoba (6. fejezet) kiegészítésének szán. Ha a falon fut az emberi agy által (állítólag) implementált program, a falnak pedig nincs elméje, akkor a

[2] Lycan szerint a példa eredetileg Ian Hinckfuss-tól ered.

számítógépeken futtatott MI-nek sincs elméje, mert egy program futása (a fal példája alapján) nem elegendő ahhoz, hogy jelen legyen az értelem. Bár Searle érve trivializálja az implementációt, maga Searle nem a formalizmust hibáztatja az abszurd következményekért, hanem az erős MI koncepcióját. A formalista felfogás szerinte demarkációs célra használható fel: ha a formalizmus igaz, akkor az MI nem gondolkodik, hiszen a fal nyilván nem gondolkodik.

15.1.3 Putnam köve

Putnam szintén az elmefilozófiai funkcionalizmus kritikájához használta fel a formalista felfogást és az abból eredő pánkomputacionalizmust. Érvének eredeti változata viszonylag hosszú (Putnam 1988: 120–125), de még így sem eléggé általános, mert a Turing-gépeknél jóval primitívebb számítási modellre épül. Csak egy tömör rekonstrukciót idézek Sprevak (2018) alapján.

Putnam az inputmentes véges állapotú automaták (*Inputless Finite State Automaton*, IFSA) működését elemzi formalista alapon. Az IFSA-k szalagmentes Turing-gépekre hasonlítanak: csak belső állapotaik vannak, melyek egy átmenettábla szerint változnak. Tekintsük egy IFSA-t, ami az A és B állapotok között oszcillál, tehát a működési logikáját az alábbi táblázat írja le:

$\Rightarrow$	
A	B
B	A

Putnam szerint ezt a programot egy kődarab is futtatja, ha zajlik benne valamilyen fizikai változás, például felmelegszik a napon:

> Tekintsük a kődarab pályáját a fázistérben a t_0 és t_n időpontok között. [...] Putnam a fázisteret négy régióra bontja (r_1, r_2, r_3, r_4), melyeken a kő t_0 és t_n között áthalad. A régiók a kő állapotát írják le a (t_0, t_n) intervallum négy egyforma hosszú szakaszán. Minden régió fizikai tulajdonságok egy halmaza, a rendszer egy lehetséges állapota. A kő tehát az intervallum első negyedében az r_1 régió által definiált fizikai állapotban van; a második negyedében az r_2 által definiált fizikai állapotban [és így tovább].
>
> Miközben a kő áthalad a fázistér négy szomszédos régióján (r_1, r_2, r_3, r_4), egyben a fázistér két össze nem függő régiója ($r_1 \vee r_3$, illetve $r_2 \vee r_4$) között oszcillál. Putnam az $r_1 \vee r_3$ régiót az A számítási fázisnak, $r_2 \vee r_4$-et a B fázisnak felelteti meg. Így izomorfizmus jön létre a kő fizikai állapotai és azok változásai, illetve az FSA absztrakt állapotai és azok átmenetei között. (Sprevak 2018: 10).

Ez a trivializációs stratégia hasonlít a szabadrúgással kapcsolatos korábbi példához, de radikalizálja azt, mert a program fázisait a kőállapotok *diszjunkcióinak*

felelteti meg, tehát nem a kő természetes tulajdonságait azonosítja a számítás fázisaival. Putnam arra mutat rá, hogy a formalista felfogásba az is belefér, ha tetszőlegesen összekutyult fizikai „állapotokat" tekintünk számítási lépésnek. Nem szükségszerű (legalábbis a formalizmus definíciója alapján), hogy az F_i-k mérnökileg értelmes vagy számunkra jól felismerhető állapotok legyenek.

Felvethetnénk, hogy Putnam érve életidegen. A találomra definiált belső állapotokon realizált, input és output nélküli „programok" távolról sem hasonlítanak egy tipikus számítógépes programra, főleg nem az olyan komplex alkalmazásokra, mint az MI. De ez nem feltétlenül teszi Putnam stratégiáját hiteltelenné. Chalmers (1995) például kidolgozta az érv egy általánosabb változatát, ami Turing-gépekre is érvényes.

Mielőtt rátérnék egy realisztikusabb trivializációs érvre, a történeti hitelesség érdekében hangsúlyozom, hogy Lycan, Searle és Putnam a fenti példákat nem a formalizmus *cáfolatának* szánták, hanem arra akarták felhívni a figyelmet, hogy a formalizmus – ami az implementáció korrekt elméletének tűnt számukra – olyan kontraintuitív következményekhez vezet, melyek aláássák az elmefilozófiai funkcionalizmust, vagy legalábbis annak egyes változatait. De a kontraintuitív következmények ugyanannyi alapot adnak a formalizmus megkérdőjelezésére. Ezért nevezem ezeket az érveket – talán szerzőik szándékaival ellentétben – az implementációt trivializáló érveknek.

15.1.4 Copeland trivialitási érve

Képzeljünk el egy egyszerű processzort, aminek három 8 bites regisztere van (I, A, D). Az I regiszter az aktuális instrukciót tárolja, D az inputként kapott adatot, A pedig szabadon írható-olvasható. A regiszterek tartalmát jelezze felülvonás, például $\overline{I}$ az I regiszter tartalma, az aktuális instrukciót kódoló nyolc bit. Maguknak a regisztereknek is van egy nyolcbites címe, melynek segítségével hivatkozni lehet a tartalmukra. Ezt kétszeres felülvonás jelzi: $\overline{\overline{I}}$ annak a regiszternek a tartalma, melynek címe az I regiszterben jelenleg tárolt szám. I lehetséges tartalmai ilyesfajta számítási műveleteknek felelnek meg:

	$\overline{I}$	Instrukció
(I1)	00000001	„A tartalmát tedd egyenlővé D tartalmával" $(\overline{A} \Rightarrow \overline{D})$
(I2)	00000010	„A tartalmát tedd egyenlővé A és D tartalmának összegével" $(\overline{A} \Rightarrow \overline{A} + \overline{D})$
(I3)	00000011	„A tartalmát tedd egyenlővé A és D tartalmának szorzatával" $(\overline{A} \Rightarrow \overline{A} \times \overline{D})$
(I4)	00000100	„A tartalmát tedd egyenlővé A és a $\overline{D}$ címen levő regiszter tartalmának összegével" $(\overline{A} \Rightarrow \overline{A} + \overline{\overline{D}})$
...	...	...

A processzornak van egy belső órajele, ami az utasítások végrehajtásának ritmusát diktálja. Ha a processzor ciklusonként egy utasítást hajt végre, akkor egy program futását az alábbihoz hasonló táblázat jellemzi majd ($i_1, d_1 \ldots$ mindegyike egy 8 bites szám):

Ciklus	$\overline{I}$	$\overline{D}$	$\overline{A}$	$\overline{\overline{D}}$
1	i_1	d_1	a_1	x_1
2	i_2	d_2	a_2	x_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	i_n	d_n	a_n	x_n

Copeland (1996) bemutatja, hogyan tehetjük precízzé Searle állítását a szövegszerkesztőt futtató falakról ennek az egyszerű (de nem valóságidegen) számítási modellnek a segítségével. Tegyük fel, hogy a táblázatban jelzett absztrakt programfutás a Wordstar szövegszerkesztő gépi kódú programjának felel meg (egy bizonyos időintervallumon). Ez persze irreális feltevés, hiszen még egy primitív szövegszerkesztő futtatásához is sokkal bonyolultabb processzorra van szükség. De az érven semmit sem változtatna, ha ezerszer ekkora táblázattal dolgoznánk.

Annak kimutatásához, hogy egy tetszőleges fal a $t_1 \ldots t_n$ intervallumban a fenti számítási folyamatot implementálja, minden időpontban feleltessük meg a fal egy-egy területét (nem feltétlenül mindig ugyanazt) valamelyik processzor-regiszternek, vagyis tekintsük a fal egy-egy (akár ciklusonként változó) területén levő molekulák egy t_k-beli állapotát a regiszter k. ciklusbeli realizálásának:

Ciklus	$\overline{I}$ realizálói	$\overline{D}$ realizálói	$\overline{A}$ realizálói	$\overline{\overline{D}}$ realizálói
t_1	F_1^1	F_1^2	F_1^3	F_1^4
t_2	F_2^1	F_2^2	F_3^3	F_4^4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	F_n^1	F_n^2	F_n^3	F_n^4

F_j^i a fal valamelyik régiójának j. ciklusbeli fizikai állapota. Egyedül az a fontos, hogy különbözzön az összes többi többi F_l^k ($k \neq i$, $l \neq j$) állapottól.

A fenti táblázatról leolvasható az instrukció-regiszter falban realizált állapotait definiáló $\overline{\mathbf{I}}$ függvény, ami a $\{t_1 \ldots t_n\}$ halmazról képez a fal állapotainak halmazára oly módon, hogy tetszőleges k-ra $\overline{\mathbf{I}}(t_k) = F_k^1$. Hasonlóan adódnak a másik két regiszter realizálóit kijelölő $\overline{\mathbf{A}}$ és $\overline{\mathbf{D}}$ függvények és a $\overline{\overline{D}}$ hivatkozások realizálóit kijelölő $\overline{\overline{\mathbf{D}}}$ függvény. Létezik továbbá – a két fenti táblázat közti egyértelmű megfeleltetés alapján – egy olyan ν függvény, hogy bármely F_j^i esetén $\nu(F_j^i)$ az F_j^i által realizált bináris szám: $\nu(F_1^1) = t_1$, $\nu(F_1^2) = d_1$ stb.

Ezután definiáljuk a „+” és „×” műveleteket, valamint az „=” relációt oly módon, hogy bármely $t_i, t_j, t_k \in \{t_1 \dots t_n\}$ és $\mathbf{X}, \mathbf{Y} \in \{\overline{\mathbf{I}}, \overline{\mathbf{A}}, \overline{\mathbf{D}}, \overline{\overline{\mathbf{D}}}\}$ esetén akkor és csak akkor teljesüljön, hogy (a) $\mathbf{X}(t_i) = \mathbf{Y}(t_j)$, (b) $\mathbf{X}(t_i) = \mathbf{Y}(t_j) + \mathbf{Z}(t_k)$, illetve (c) $\mathbf{X}(t_i) = \mathbf{Y}(t_j) \times \mathbf{Z}(t_k)$, ha igaz, hogy (a) $\nu(\mathbf{X}(t_i)) = \nu(\mathbf{Y}(t_j))$, (b) $\nu(\mathbf{X}(t_i)) = \nu(\mathbf{Y}(t_j)) + \nu(\mathbf{Z}(t_k))$, illetve (c) $\nu(\mathbf{X}(t_i)) = \nu(\mathbf{Y}(t_j)) \times \nu(\mathbf{Z}(t_k))$. Jogunk van így interpretálni a műveleti jeleket és az egyenlőségjelet, mert formális szinten ezt semmi sem tiltja. Egy ilyen interpretáció alapján pedig a realizációs függvények kielégítik a processzor-instrukcióknak megfelelő formális axiómákat:

„I1” $\forall t : \overline{\mathbf{I}}(t) = |00000001| \rightarrow \overline{\mathbf{A}}(t) = \overline{\mathbf{D}}(t)$ („Ha az I regisztert t-ben realizáló falterület a 00000001 számnak megfelelő mikroszkopikus állapotban van, akkor az A regisztert t-ben realizáló falterület mikroszkopikus állapota ugyanannak a számnak felel meg, mint a D regisztert t-ben realizáló falterület mikroszkopikus állapota.”)

„I2” $\forall t_k : \overline{\mathbf{I}}(t_k) = |00000010| \rightarrow \overline{\mathbf{A}}(t_k) = \overline{\mathbf{A}}(t_{k-1}) + \overline{\mathbf{D}}(t_k)$

„I3” $\forall t_k : \overline{\mathbf{I}}(t_k) = |00000011| \rightarrow \overline{\mathbf{A}}(t_k) = \overline{\mathbf{A}}(t_{k-1}) \times \overline{\mathbf{D}}(t_k)$

„I4” $\forall t_k : \overline{\mathbf{I}}(t_k) = |00000100| \rightarrow \overline{\mathbf{A}}(t_k) = \overline{\mathbf{A}}(t_{k-1}) + \overline{\overline{\mathbf{D}}}(t_k)$

... ...

A nyíl a materiális kondicionálist jelöli, vagyis $P \rightarrow Q$ pontosan akkor igaz, ha P hamis vagy Q igaz ($P \rightarrow Q \equiv \neg P \vee Q$).

Ezzel készen is vagyunk: definiáltunk olyan függvényeket, melyek alapján a fal ugyanazoknak a szabályoknak engedelmeskedik, mint az absztrakt számítási modell.

Copeland érve nagyon primitív absztrakt architektúrát feltételez, de a falak mikroszkopikus állapotainak végtelen sokfélesége miatt bonyolultabb architektúrákra is konstruálhatunk hasonló érvet.

Felvethetnénk, hogy ez az érv puszta szemfényvesztés, mert nem garantáltuk, hogy a fal egy adott területe tényleg *realizál* egy adott regisztert. A falnak nincsenek funkcionális részei, pedig az absztrakt modell egyértelműen ezt sugallja. A fal területei teljesen *ad hoc* módon „reprezentálják” a regiszterek tartalmát is. Például semmi garancia rá, hogy a 00000001 szám két különböző reprezentációja fizikai szempontból hasonló. Vélhetőleg semmi közös nincs bennük azon kívül, hogy mindegyiket a 00000001 szám reprezentációjának *tekintjük*. A processzor működésének szabályai sem érvényesek a szó valódi értelmében. Aligha igaz például, hogy ha t_1-ben a fal az aktuálistól eltérő „utasítást” kapott volna, akkor az annak megfelelő „tevékenységet” végezte volna, például $\overline{\mathbf{I}}(t_2) = 00000010$ esetén az $\overline{\mathbf{A}}(t_1) = \overline{\mathbf{A}}(t_1) + \overline{\mathbf{D}}(t_2)$ egyenlet igaz lett volna. Sem a fal fizikai dinamikája, sem a fenti interpretáció nem garantálja ezt. Logikai

értelemben az jelenti a problémát, hogy az implementált szabályok csak a materiális kondicionálisok („→”) szintjén tükrözik az absztrakt szabályrendszert, nem kontrafaktuálisok szintjén.

Ezek a kifogások jogosak, de nem cáfolják a trivializációs érvet. A kifogások emiatt a *formalizmus* elleni kifogásoknak tekinthetőek, akárcsak maga az érv. A következő fejezetben tárgyalt strukturalista felfogás pont arra tesz kísérletet, hogy az implementációt a fizikai rendszer struktúrája vagy diszpozíciói alapján definiálja, amihez a fenti $\overline{\mathbf{I}}(t)$, $\overline{\mathbf{D}}(t)$, $\overline{\mathbf{A}}(t)$ és $\overline{\overline{\mathbf{D}}}(t)$ függvények nem elegendőek.

15.2. Pánkomputacionalizmus

A formalista felfogás alapján minden fizikai rendszer számításokat végez. Ez a tézis a *pánkomputacionalizmus*. Ahogy a pánpszichista minden anyagi létezőt tudatosnak vél, a pánkomputacionalista a természet minden részét számítógépnek nézi.

Elvben háromféleképpen lehet reagálni arra a tényre, hogy a formalizmus pánkomputacionalizmushoz vezet. Akinek a formalizmus és a pánkomputacionalizmus egyaránt szimpatikus, elégedetten nyugtázhatja, hogy a két tézis szorosan összefügg. Ezt tette például Putnam és Searle, akik a pánkomputacionalizmusra támaszkodva támadták azt az elméletet, hogy az elme egyfajta számítási folyamat lenne (*computational theory of mind*).

Akinek mindkét nézet antipatikus, a trivialitási érveket a formalista felfogás cáfolatának tekintheti. Az implementáció metafizikájával foglalkozó filozófusok körében ez a hozzáállás tűnik tipikusnak.

Akinek a formalizmus szimpatikus, a pánkomputacionalizmus viszont nem, megpróbálhatja úgy módosítani a formalista felfogást, hogy a trivializációs érvek működésképtelenné váljanak. Amennyire látom, ezek a kísérletek a strukturalista felfogáshoz vezetnek, amit a következő fejezet tárgyal. Jelen fejezet lezárásaként a többségi állásponthoz csatlakozva a pánkomputacionalizmus ellen érvelek, és a könyv további részében azt feltételezem, hogy a pánkomputacionalizmus hamis.

A vita előtt fontos tisztázni, hogy a „pánkomputacionalizmus” legalább három különböző elméletet jelölhet. A pánkomputacionalizmus első lehetséges változata szerint minden fizikai folyamat egyben számítási folyamat is. Ezt az „esse est computare” (**EEC**) elméletnek fogom hívni, Berkeley püspök után szabadon. Az EEC enyhébb formában is megfogalmazható: nem szó szerint *minden*, csak minden *kellően komplex* fizikai folyamat számítási folyamat. A Wordstar futtatásához (talán) nem elég egy hidrogénatom, de egy fal már igen. Ez a komplexitási kitétel nem teszi EEC-t kevésbé abszurddá, tehát dialektikai szempontból a leszűkítésnek nincs jelentősége.

A pánkomputacionalizmus másik lehetséges formája két univerzális kvantort alkalmaz: *minden* fizikai folyamat *minden* programot implementál. Ezt

*túlburjánzó pánkomputacionalizmus*nak (**TPK**) fogom hívni. TPK-ból következik EEC, de EEC-ből nem következik TPK.

EEC és TPK mellett létezik egy további elmélet, amit néha pánkomputacionalizmusnak neveznek: az a hipotézis, hogy a világ alapvetően *információból* áll. Ezt a bizarr ontológiai feltevést „It From Bit" (**IFB**) elméletnek fogom hívni (Wolfram 2002, Lloyd 2006).

A pánkomputacionalizmus mindhárom formája – EEC, TPK és IFB – hamisnak vagy irrelevánsnak tekinthető a demarkációs vita kontextusában, de eltérő okokból. IFB véleményem szerint egyszerűen értelmetlen. Semmi értelme azt állítani, hogy a szék információból áll. Aki azt állítja, hogy a szék információból áll, az vagy nincs tisztában az „információ" és a „szék" szavak jelentésével, vagy sajátos, új jelentésben használja ezeket a szavakat, de anélkül, hogy ezt a futurisztikus újbeszélt világossá tudná tenni. Lehet persze, hogy ebben tévedek. Lehet, hogy pallérozatlanságom miatt képtelen vagyok felfogni a modern fizikai és informatika mély lételméleti tanulságait. De akár értelmetlen, akár mély és értelmes IFB, nem következik a fenti trivialitási érvekből, sem általában az implementáció formalista felfogásából, sem (amennyire látom) a formalizmus később tárgyalt alternatíváiból. Ezért jelen kontextusban – az implementáció természete kapcsán – figyelmen kívül hagyható.

EEC és TPK határozottan nem értelmetlenek, de annyira abszurdak, hogy lehetetlen komolyan venni őket. A számítás az, amit a számítógépek csinálnak. Több tízezer évnek kellett eltelnie, amíg sikerült építenünk ilyen szerkezeteket, és igazi lendületet csak az utóbbi fél évszázadban vett a fejlődésük. A számítás természetével kapcsolatos metafizikai kérdés arra vonatkozik, hogy ezek a szerkezetek – melyek kontingens, emberalkotta tárgyak, és a bolygón található fizikai rendszerek elenyészően kis részét képezik – pontosan *mitől* számítógépek, struktúrájuk vagy tevékenységük mely aspektusa a számítás mint absztrakt algoritmus fizikai realizációja. Ha erre a kérdésre azt válaszoljuk, hogy voltaképpen a vödör és a fal is számítógép, akkor gyanítható, hogy valamit figyelmen kívül hagytunk. Képzeljük el, hogy a biológia még gyerekcipőben jár és az ősemberek egy csoportja megkéri a helyi bölcset, hogy derítse ki, miféle lények a sündisznók. A bölcs elvonul elmélkedni, majd közli, hogy a sünök természetét vizsgálva ráébredt, valójában *minden* sün: a barlang is sün, a madár is sün, a kövek is sünök; sőt, nemcsak, hogy minden sün, hanem *minden* dolog *mindegyik* sün. Felvethető, hogy ez a bölcs nem egy eladdig ismeretlen metafizikai igazságot tárt fel, hanem mély felfedezésnek nézte a pánsünizmust cáfoló tényeket. Hasonló a helyzet EEC-vel és TPK-val.

EEC és TPK módszeresebb kritikáját az implementáció-elmélet módszertani kritériumaira lehet alapozni (13.3. alfejezet). EEC és TPK a négy kritérium közül kettőt súlyosan megsért: a **Kizárás** elvét, mely azt írja elő, hogy ne minősüljenek számítógépnek olyan fizikai rendszerek, melyeket az informatikusok és átlagemberek nem tartanak annak, és a **Magyarázat** elvét, mely azt írja elő, hogy a számítást végző rendszerek viselkedését az általuk implementált

program magyarázza. A **Kizárás** azért sérül, mert EEC és TPK vödröket, falakat és köveket is számítógépnek állít be, a **Magyarázat** pedig azért, mert az implementáció feltételeként azonosított absztrakt összefüggések lényegében bármelyik (kellően komplex) fizikai rendszerben megtalálhatóak, tehát semmiféle értelmes korlátot vagy vezérelvet nem jelentenek a rendszer működésével kapcsolatban. A **Magyarázat** elvét talán megmentheti az EEC egy enyhített változata, ami minden fizikai rendszerhez pontosan 1 db. implementált programot rendel, mégpedig azt a szabályt, ami az adott rendszer működését a természeti törvények szintjén jellemzi. Ez az „EEC Light" elmélet azonban inkább csak formális, nem pedig szubsztantív értelemben elégíti ki a **Magyarázat** kritériumát, hiszen a számítástechnikai magyarázat fogalmát lecseréli a fizikai magyarázat fogalmára és gyakorlatilag felszámolja a számítási magyarázatot mint sajátos jelenséget.

EEC és TPK továbbá csak formális, nem pedig szubsztantív értelemben elégíti ki a **Visszaigazolás** elvét, mely azt követeli, hogy minősüljenek számítógépnek azok a dolgok, melyeket eleve annak tartunk. EEC és TPK alapján a preteoretikusan számítógépnek tartott tárgyak nem valami sajátos, a többi tárgytól eltérő tulajdonságuk miatt minősülnek számítógépnek, hanem azért, mert *minden* tárgy valamilyen programot implementál. EEC és TPK ráadásul még ilyen gyenge értelemben sem feltétlenül elégíti ki a **Visszaigazolás** elvét, mert EEC és TPK is logikailag kompatibilis azzal a nézettel, hogy laptopok, asztali számítógépek, mobiltelefonok és szerverek valójában nem léteznek (a példa kedvéért tegyük fel, hogy a Peter van Inwagen (1990) érveihez hasonló okokból). Ez a fajta eszköz-nihilizmus nem mond ellent EEC-nek és TPK-nak, hiszen ha laptopok stb. nem léteznek, viszont molekulák és naprendszerek (valamiért) igen, akkor EEC és TPK alapján léteznek számítógépek. Tehát EEC és TPK kizárólag azért elégíti ki a **Visszaigazolás** elvét, mert EEC-től és TPK-tól függetlenül eleve hiszünk a **Visszaigazolás** szempontjából fontos tárgyak létében.

Végezetül felvethető, hogy EEC és TPK az **Objektivitás** elvét is megsérti, mert trivializálja az implementációt. Analógia: ha minden sün, akkor a sünség *bizonyos* értelemben objektív jelenség (hiszen minden, ami létezik, az létezik), azonban az a kategória, amit a pánsünizmus felfedezése előtt a „sün" jelölt, nem feltétlenül lesz objektív, hiszen ha *minden* sün, akkor az ennél jóval szűkebb prefilozófiai sün-fogalmunk nem feltétlenül a valóság egyik illesztékét jelöli ki. Hasonlóan, EEC és TPK alapján az informatikusok és átlagemberek implementáció-fogalma nem feltétlenül ragad meg egy valós fizikai tulajdonságot. A helyzet analóg a mereológiai univerzalizmus esetével a komplex tárgyak metafizikájában (Wasserman 2021). A mereológiai univerzalista szerint fizikai tárgyak bármely halmaza egy további fizikai tárgyat alkot, létezik például egy tárgy, ami a Holdból és a ma élő emberek füleiből áll. A mereológiai univerzalizmus egyesek szerint visszaigazolja azt a prefilozófiai meggyőződésünket, hogy léteznek komplex fizikai tárgyak – asztalok, székek, házak, fák –, de

megszabadít attól a kellemetlen feladattól, hogy definiáljuk, elemi részecskék pontosan mely halmazával azonos egy adott makroszkopikus tárgy. Pont emiatt azonban univerzalista alapon csak félszívvel tekinthetjük objektívnek az olyan kategóriáinkat, mint „asztal", „szék" stb., mert metafizikai értelemben semmi nem választja el a szék alakú atomfelhőket az őket tartalmazó nagyobb atomfelhőktől. A pánkomputacionalizmus esetében hasonló módon nincs alapvető különbség egy vödör víz és egy laptop között.

A pánkomputacionalisták felpanaszolhatják, hogy a fenti négy kritérium diszkriminatív, mert eleve kizárja az ő nézetüket. Ez a fajta patthelyzet gyakori a filozófiában, mert a filozófiai szempontból lényeges fogalmak esetében legitimnek minősülhetnek a revizionizmus különböző formái; a metafizikában például nem muszáj ugyanúgy értelmezni a „tárgy" szót, ahogy az átlagember értelmezi. A filozófus ilyenkor nem tehet mást, mint hogy rámutat paradigmatikus esetekre és feltételezi, hogy a rájuk épülő intuíciók alkalmasak a szóban forgó kategória rögzítésére. A pánkomputacionalizmus azonban pont erre nem képes, és a négy módszertani kritérium erre a hiányosságára világít rá.

16.

A strukturalista felfogás

Egy program implementálásához vélhetőleg a program struktúrája és a számítógép működése közti analógiára van szükség; a formalistáknak ebben igazuk van. De az analógiának, szemben a puszta formális megfeleléssel, *valósnak* kell lennie; a fizikai rendszer működésének hűen kell tükröznie a számítás struktúráját, bármit is jelent ez. Az ilyen erős megfelelésekre épülő implementációfelfogásokat Rescorla (2013) nyomán „strukturalistá"-nak fogom hívni.

Tegyük fel, hogy az F fizikai rendszer az A absztrakt számítási modellt implementálja. Jelölje $\Phi(A_i)$ a számítás A_i részének fizikai realizálóit, vagyis F-nek azokat a részeit vagy állapotait, melyek az A_i számítási fázisnak vagy komponensnek felelnek meg. A_i lehet például a Turing-gép szalagja egy adott pillanatban, vagy az, hogy a Turing-gép ráír valamit a szalagra és átvált egy másik utasításra. A strukturalista arra törekszik, hogy a formalizmusnál gazdagabb kritériumokat támasszon Φ-vel kapcsolatban. A szakirodalomban alapvetően kétféle kritériumot találunk:

I Kategoriális hasonlóság A $\Phi(A_i)$ halmaz elemei, azaz A_i fizikai realizálói, legyenek objektíve hasonlóak egymáshoz, és ha $A_i \neq A_j$, akkor $\Phi(A_i)$ elemei objektíve különbözzenek $\Phi(A_k)$ elemeitől. Azonos jelsorozatokat például azonos vagy nagyon hasonlói fizikai mintázatok realizáljanak, különbözőeket pedig különbözőek. Ez a kritérium egyértelműen sérül a formalista felfogásnál, ahol tetszőleges fizikai állapotok (például a fal különböző régióinak molekuláris szintű tulajdonságai) realizálhatják ugyanazt a regiszter-állapotot.

II Kontrafaktuális hasonlóság F viselkedése képezze le A logikáját: ha az absztrakt számítási modellben az A_i fázis után mindig az A_j fázis következik, akkor a $\Phi(A_i)$ fizikai állapotokat mindig (vagy legalábbis nagy valószínűséggel) a $\Phi(A_j)$ állapotok valamelyike kövesse.

Az I. kritériumra épülő felfogást *kategoriális strukturalizmus*nak, a II. kritériumra épülő felfogást *kontrafaktuális strukturalizmus*nak fogom hívni.[1] A két megközelítés nem ellentétes, sőt, elválaszthatatlanok, amennyiben egy fizikai rendszer kategoriális tulajdonságai szorosan összefüggnek a viselkedésével. A kategoriális strukturalizmust többek között Godfrey-Smith (2009) képviseli, a kontrafaktuális strukturalizmust Chalmers (1995, 1996a), a kettőt kombinálja Copeland (1996). Az alábbiakban amellett érvelek, hogy egyik variáns sem kielégítő.

16.1. Kategoriális strukturalizmus

A kategoriális strukturalista szerint az F fizikai rendszer akkor realizálja az A programot, ha A részeinek fizikai realizálói objektíve hasonlítanak egymásra és objektíve különböznek A más részeinek realizálóitól: $\Phi(A_i)$ elemei objektíve hasonlóak, bármely $A_i \neq A_k$ esetén pedig $\Phi(A_i)$ és $\Phi(A_k)$ elemei objektíve különböznek. Ez a kritérium különbözteti meg a kategoriális strukturalistát a formalistától. Copeland trivializációs érvében (15.1.4. alfejezet) a hasonlósági kritérium azért nem teljesül, mert a processzor-regiszterek egy-egy állapotát véletlenszerűen kiválasztott mikroszkopikus mintázatok „realizálják", $\Phi(A_i)$ elemeit nem köti össze semmi azon túl, hogy mindegyiket A_i realizálójának tekintjük. Egy valódi számítógép esetében egy regisztert (mint absztrakt komponenst) tranzisztorokból álló logikai kapuk realizálnak, és a program absztrakt modelljében szereplő állapot – például egy bizonyos jelsorozat – objektíve hasonló fizikai állapotoknak felel meg: a „00000001" jelsorozat realizálásakor az első hét bitet realizáló tranzisztorokban alacsony, a nyolcadik bitet realizáló tranzisztorban pedig magas a feszültség. Ilyen kategoriális hasonlóságokat nem találunk a program absztrakt modellje és a falak mikroszintű állapotai között. Ahogy Godfrey-Smith fogalmaz a formalizmusról a strukturalizmusra való átmenet kapcsán:

> A [formalizmus] eredeti koncepcióját egészítsük ki a természetes(ebb) és természetellenes(ebb) realizációk közti különbséggel. A természetességet azokon a kategóriákon belül keressük, melyek alapján az absztrakt állapotokat fizikai állapotoknak feleltetjük meg. A hasonlóságnak ez a fogalma [...] nem általános és univerzális, hanem a gép működésének alacsonyszintű leírásától függ. (Godfrey-Smith 2009: 289–290)

[1] Rescorla (2013) szűkebb értelemben használja a „strukturalizmus" címkét és kizárólag a II-es típusú felfogást jelöli vele. Az implementációs strukturalizmusnak természetesen semmi köze a tudományfilozófiai (Worall 1989, Ladyman & Ross 2007), matematikafilozófiai (Benacerraf 1965), nyelvészeti (Saussure), irodalomelméleti (Jacobson), antropológiai (Lévi-Strauss), pszichoanalitikai (Lacan) vagy marxista (Althusser) strukturalizmushoz.

Hasonló elvet fogalmaz meg Copeland:

> A megfeleltetés nem lehet *ex post facto*. Ezáltal tudjuk elkülöníteni a valódi implementációk halmazát az olyan imposztoroktól, melyek nem tartják tiszteletben az axiomatikus elmélet [az absztrakt számítási modell] szándékolt jelentését. (Copeland 1996: 350)

Természetes és természetellenes realizációkra azonban csak akkor hivatkozhatunk, ha világos értelmet tudunk adni annak a kijelentésnek, hogy egy fizikai rendszer $F_1 \dots F_n$ állapotai objektíve hasonlítanak vagy különböznek. Ehhez a természetes tulajdonságok (*natural properties*) fogalmát használhatjuk fel, ami David Lewistól ered:

> A tulajdonságokról néha úgy gondolkodunk, mintha *szaporák* (*abundant*), néha pedig úgy, mintha *ritkák* (*sparse*) lennének. A szapora tulajdonságok bármennyire extrinzikusak lehetnek, akármilyen nyakatekerten tákolhatjuk össze őket. Nem törődnek a valóság kvalitatív illesztékeivel, hanem sokszor teljesen önkényesen osztják fel a világot. Ha két tárgy ugyanazzal a szapora tulajdonsággal rendelkezik, nem biztos, hogy hasonlóak. Két tökéletesen egyforma tárgy számtalan szapora tulajdonságban eltér; ahogy teljesen különböző tárgyak számtalan szapora tulajdonságon osztoznak. Sok szapora tulajdonság kifejezhetetlen bármilyen, számunkra akár csak elvben elsajátítható nyelven, és még akkor is kifejezhetetlen maradna, ha képesek lennénk végtelen mondatokat leírni és meg tudnánk nevezni olyan dolgokat, melyek a megismerhetetlenség örök homályába burkolóznak. A szapora tulajdonságokból voltaképpen ugyanolyan sok van, mint halmazokból, mert minden halmaz esetében létezik az a tulajdonság, hogy valami annak a halmaznak az eleme. [...]
>
> A ritka tulajdonságokkal egész más a helyzet. Ha két dolog rendelkezik ugyanazzal a ritka tulajdonsággal, akkor kvalitatívan hasonlóak. A ritka tulajdonságok a valóság illesztékeit jelölik ki, intrinzikusak, specifikusak, pont elegen vannak ahhoz, hogy a dolgokat teljes mértékben, de redundancia-mentesen jellemezzék, és példányaik halmazai *ipso facto* nem szedett-vedettek. [...] Ha egy tulajdonság ehhez a szűk kisebbséghez tartozik, *természetes* tulajdonságnak nevezem. (Lewis 1986: 59–60)

A kategoriális strukturalista lényegében azt követeli, hogy a $\Phi(A_i)$ fizikai realizálókat kösse össze valamilyen természetes tulajdonság, $A_i \neq A_k$ esetén pedig $\Phi(A_i)$ és $\Phi(A_k)$ elemei között legyen valami természetes különbség. Az absztrakt program és a fizikai rendszer közti párhuzamoknak tehát a fizikai rendszer lewisi értelemben „ritka” tulajdonságain kell alapulnia.

David Chalmers, aki a kontrafaktuális strukturalizmus híve, szkeptikus ezzel a stratégiával kapcsolatban:

> Egyesek amiatt kritizálják a fenti [Searle vödréhez hasonló] érvet, hogy önkényes diszjunkciókkal definiált, „természetellenes" fizikai állapotokra épül. De ezt a kifogást nehéz precízen kidolgozni, mert nehéz olyan objektív különbséget felfedezni a „természetes" és „természetellenes" tulajdonságok között, ami a kifogást megalapozottá tenné. (Chalmers 1996a: 312)

Chalmers nem részletezi, miért nincs objektív különbség a természetes és nem természetes tulajdonságok között. Érvelhetne például azzal, hogy a természetes tulajdonságok az aktuális tudományos paradigmától függő társadalmi konstrukciók, melyek történetileg változékony módon tesznek kitüntetetté egyes (szapora) fizikai tulajdonságokat. De ez a hipotézis valójában nem alkalmas a kategorikus strukturalizmus kritikájára, mert hiába intünk búcsút a társadalmi konvencióktól független objektivitásnak, a kategoriális hasonlóságra épülő implementáció-felfogás védhető marad mint az aktuális konvenció kontextusában megfogalmazott kritérium. Ha történetileg kontingens is, hogy mi számít természetes tulajdonságnak, az aktuális történeti kontextusban egyes tulajdonságok annak számítanak, és ezekre felépíthető kategoriális strukturalista alapon az implementáció fogalma.

Ha Chalmerséhez hasonló kritikát akarunk megfogalmazni a kategoriális strukturalizmussal szemben, több értelme van azt állítani, hogy létezhetnek ugyan természetes tulajdonságok – a tömegről, töltésről és egyéb alapvető fizikai tulajdonságokról például nyugodtan feltételezhetjük, hogy természetesek –, az absztrakt programokat (prefilozófiai ítéleteink szerint) implementáló rendszerekben azonban nem találunk olyan természetes tulajdonságokat, melyek alapján egy mobiltelefon objektíve számítógépnek minősülne, egy vödör víz viszont nem.

Ez a felvetés őrültségnek tűnhet. A számítógép belsejében vezetékek, merevlemezek, memóriakártyák, processzorok és egyéb, célzottan a számítógép működtetésére gyártott alkatrészek találhatóak. Egy ilyen rendszernek *nyilván* olyan a felépítése, hogy egyes részei az általa futtatott programok absztrakt struktúráját képezik le. A falra ez nem igaz. Tehát van olyan objektív különbség a számítógép és a fal között, ami miatt a számítógép számításokat végez, a fal pedig nem.

Ez a rekontra azonban nem működik a chalmersi kifogás szóban forgó változata ellen. „Objektív"-nek, mint arról a 18. fejezetben részletesebben is szó lesz, két teljesen eltérő jelenséget is nevezhetünk. Jelölhetjük vele az interszubjektív invarianciát: x akkor objektíve F, ha nem ízlés-, kultúra- vagy nézőpontfüggő, hogy x F-e, ha x minden normális megfigyelő számára F. Ha a Himalája léte objektív (ebben az értelemben), akkor egyetlen normális megfigyelő sem ítéli úgy, hogy a Himalája nem létezik. Nevezzük ezt objektivitás$_1$-nek. Kétségkívül vannak objektív$_1$ különbségek a számítógépek és a falak között, melyek alapján a számítógép számításokat végez, a fal pedig nem. Nem ízlés-, kultúra-

vagy nézőpontfüggő, hogy a számítógépben processzorok és adattároló eszközök vannak, a falban pedig nincsenek.

Ugyanakkor nevezhetjük „objektív"-nak a külső tárgyak reprezentációinkról leválasztott tulajdonságait is. Nevezzük ezt objektivitás$_2$-nek. Az objektivitás$_2$ és az objektivitás$_1$ két különböző jelenség. Szivárvány idején minden normális megfigyelő számára úgy tűnik, hogy színes csíkok vannak az égen, de ezzel a tulajdonsággal az ég a reprezentációinkról függetlenül nem rendelkezik: az égen valójában nincsenek színes csíkok, az agyunk az égből érkező fotonok egyes tulajdonságai miatt generálja csíkok látványát. Tehát a szivárvány mint az égen húzódó csíkos mintázat objektív$_1$, de nem objektív$_2$. Csak az objektív$_2$ tény, hogy fent az égben vízpermet található, ami a szóban forgó objektív$_1$ jelenséget (a színes csíkok látványát) okozza.

A kétféle objektivitás különbségét jól illusztrálja a hétköznapi tárgyak létéről szóló metafizikai vita (van Inwagen 1990, Wasserman 2021). A hétköznapi tárgyak – asztalok, székek, autók stb. – objektíve$_1$ léteznek, hiszen bármely két normális megfigyelő egyet fog érteni abban, léteznek-e asztalok, székek, autók és hasonló tárgyak. Ugyanakkor messze nem egyértelmű, hogy az ilyen tárgyak léte objektív$_2$ lenne. A fizikai és kémiai egyenletekben nem szerepelnek székek, asztalok és autók, és aligha lehet pusztán fizikai vagy kémiai fogalmakkal definiálni ezeket a tárgyakat. Egy asztal elemezhető kémiai és fizikai elméletek segítségével, de nincs olyan komplex kémiai vagy fizikai leírás, ami atomi vagy molekuláris szinten definiálná egy asztal létének szükségszerű és elégséges feltételeit. Mikroszkopikus szinten a számunkra ismerős fizikai tárgyak inkább felhőkre hasonlítanak, nincsenek éles körvonalaik (Unger 1980), ráadásul sokszorosan realizálhatóak, mert nem az atomi szintű felépítésük, hanem a funkciójuk definiálja őket. Tehát felvethető, hogy a hétköznapi tárgyak léte nem objektív$_2$, annak ellenére, hogy kétségkívül objektív$_1$, és létüknek van objektív$_2$ alapja.

A kategoriális strukturalizmus ennek megfelelően két teljesen eltérő változatban fogalmazható meg:

KS1 Az implementáció objektív$_1$ hasonlóságokon és különbségeken alapul: $\Phi(A_i)$ minden eleme objektíve$_1$ hasonló és ha $A_j \neq A_k$, akkor $\Phi(A_j)$ és $\Phi(A_k)$ elemei objektíve$_1$ különböznek.

KS2 Az implementáció objektív$_2$ hasonlóságokon és különbségeken alapul: $\Phi(A_i)$ minden eleme objektíve$_2$ hasonló és ha $A_j \neq A_k$, akkor $\Phi(A_j)$ és $\Phi(A_k)$ elemei objektíve$_2$ különböznek.

Ahhoz, hogy természetes tulajdonságokra alapozzuk az implementációt, a **KS2** tézist kell védenünk, ahhoz viszont, hogy a szemmel látható, számunkra adódó tulajdonságok alapján különítsük el laptopot és a vödröt, a **KS1** tézisre van szükségünk. Emiatt a kategoriális strukturalista nem apellálhat a természeti tulajdonságokra a laptopok és falak megkülönböztetésekor.

Felvethetnénk, hogy mégiscsak vannak objektív$_2$ különbségek a laptopok és falak között, emiatt pedig **KS2** is igaz, ha **KS1** igaz. De ez elhamarkodott következtetés. A természetes tulajdonságok azok a tulajdonságok, melyek az igaz természettudományos elméletekben szerepelnek. Miattuk hasonlítanak egymásra a vízmolekulák és különbözik a víz a metántól. Ezek a tulajdonságok nem csak a reprezentációinkban vannak jelen, hanem a reprezentációinkról leválasztott külvilágban is, legalábbis a standard realista felfogás szerint, amit a kategoriális strukturalista nem kérdőjelez meg. A számítógépek strukturális tulajdonságai azonban nem objektívak$_2$. A fizika és kémia nem használja az alaplap vagy tranzisztor fogalmát, sem a számítógépek leírására szolgáló egyéb műszaki fogalmakat. Chipek, adatbuszok és memóriakártyák nem nőnek a földből, létüket kizárólag olyan folyamatokra lehet visszavezetni, melyekben alapvető szerepet kap a célirányos emberi tevékenység és az erre jellemző absztrakció. Egy tipikus memóriacella tranzisztorokkal realizált logikai kapuk bonyolult hálózata és minden memóriacella, ami ugyanazt az architektúrát használja, objektíve$_1$ hasonló, mert ugyanarra az absztrakt modellre épül. De ez az objektív$_1$ hasonlóság nem felel meg egyértelműen objektív$_2$ hasonlóságoknak. Atomi szinten nem teljesen egyformák a tranzisztorok, nem hajszálpontosan ugyanakkora feszültségek tárolják bennük a bináris számokat. A memóriacellák fizikailag és kémiailag csak annyira hasonlítanak egymásra, mint két bankjegy: számunkra, vagyis objektíve$_1$, egyformák, de a természeti tulajdonságok szintén, objektíve$_2$, különböznek, különbségeik pedig csak az emberi fogalmak absztrakciós szintjén tűnnek el (vö. Fodor 1974). Emiatt **KS1**-ből nem következik **KS2**.

Mivel a számítógépek és alkatrészeik számítástechnikailag releváns tulajdonságai nem objektívak$_2$, a kategoriális strukturalista nem hivatkozhat a természetes tulajdonságokra, mikor azt állítja, hogy a laptop a természetes tulajdonságok alapján programokat implementál, a fal pedig a természetes tulajdonságok alapján nem implementál semmit. Ez csak az objektivitás$_1$ értelmében, nem pedig a természetes tulajdonságokra jellemző objektivitás$_2$ értelmében igaz. A strukturalistának **KS2**-re lenne szüksége, de csak **KS1** igazolható.

Ellenérvként felvethetnénk, hogy a számítógépek struktúrája objektív$_2$: atomokból állnak, elektromágneses jelenségek zajlanak bennük, minden komponensükről és belső folyamatukról számot tud adni a fizika és a kémia.

Ez igaz, de nem következik belőle, hogy a számítógépek műszaki jellemzői természetes tulajdonságok lennének. Egy asztal is atomokból áll, és a fizika és kémia teljes egészében számot tud adni a viselkedéséről, de ebből nem következik, hogy az asztalság természetes tulajdonság. Az asztal fogalma nem szerepel a fizikai és kémiai elméletekben. Atomi szinten definiálhatatlan, hogy egy asztal micsoda és egyáltalán hol kezdődik és hol ér véget. Az asztalokat lehetetlen anélkül definiálni, hogy célirányos emberi tevékenységről beszélnénk és tekintetbe vennénk az emberi érzékelés jellegzetességeit; létük objektív$_1$, de nem objektív$_2$. Hasonló a helyzet a számítógépekkel.

A fenti gondolatmenet részben módosítja, részben alátámasztja a kategorikus strukturalizmus chalmersi kritikáját. Chalmers téved, mikor azt állítja, hogy a természetes tulajdonságok kiléte homályos, de abban igaza van, hogy a strukturalista nem definiálhatja rájuk hivatkozva az implementációt. Azok az objektív tulajdonságok, melyek összekötik a számítógépeket és megkülönböztetik őket a falaktól, objektívak$_1$, de nem objektívak$_2$. Mivel a természetes tulajdonságok objektívak$_2$, nem a természetes tulajdonságok teszik hasonlóvá egymáshoz a számítógépeket és különböztetik meg őket a falaktól.

Átválthat-e a strukturalista **KS1**-re, vagyis állíthatja-e azt, hogy az implementáció objektív$_1$ (de nem objektív$_2$) jelenség? Elvben igen, de ekkor a számítás hasonlóvá válik a szivárványhoz: olyan jelenség lesz, ami nem leválasztható a reprezentációinkról, nincs „odakint" a világban, hanem a fizikai létezőkről adott interpretációink része. Ezzel a felfogással – amit a kategoriális strukturalisták általában elutasítanak, és emiatt nem is tekinthető nézetük egyik variánsának – a 19. fejezetben lesz szó.

A fentiek miatt az implementáció elméletével kapcsolatos módszertani elvárások közül (13.3. alfejezet) a kategoriális strukturalizmus (a **KS2** változatban) megsérti a **Visszaigazolás** és a **Kizárás** elvét, vagyis nem tudja biztosítani, hogy a prefilozófiai ítéleteink szerint számítógépnek minősülő tárgyak strukturalista alapon is azok, a prefilozófiai ítéleteink szerint nem számítógépnek minősülő tárgyak pedig nem azok. Mivel hiányoznak a kategoriális strukturalista által keresett objektív$_2$ tulajdonságok, **KS2** alapján meghatározhatatlan, hogy mely tárgyak végeznek számításokat és melyek nem. A kategoriális strukturalizmus a **Magyarázat** elvét is megsérti, még akkor is, ha a vita kedvéért feltételezzük, hogy rendelkezésre állnak a szükséges objektív$_2$ tulajdonságok. A kategoriális strukturalista ugyanis semmit nem mond arról, hogy ezekből a fizikai tulajdonságokból miért következne, hogy a rendszer által implementált program magyarázza a rendszer működését.

16.2. Kontrafaktuális strukturalizmus

A kontrafaktuális strukturalista nem a számítógépek belső felépítésére, hanem a hajlamaira, diszpozícióira koncentrál. Mivel a számítások szabályokon alapulnak, egy program futtatása szabályok alkalmazását jelenti. Egy A absztrakt számítási modellt implementáló rendszernek tehát mindenekelőtt az A által definiált szabályoknak megfelelően kell viselkednie.

Az alapötlet illusztrálásához legyen T egy Turing-gép, ami a $\{0 \ldots 9, \textsc{üres}\}$ szimbólumkészlettel dolgozik és bekapcsoláskor az írófej alatt levő számot megnöveli 1-gyel. (Ha a cellában 9 szerepel, T egy nullát ír a helyére, ha pedig a cella üres, akkor 1-et.) A kontrafaktuális strukturalista azt követeli, hogy ha az F fizikai rendszer T-t implementálja, akkor F-nek legyenek olyan $F_0 \ldots F_9, F_{\textsc{üres}}$ állapotai, melyek igazzá teszik az alábbi állításokat:

(1) Ha F az F_0 állapotban van, akkor az F_1 állapotba kerül.

(2) Ha F az F_1 állapotban van, akkor az F_2 állapotba kerül.

(3) Ha F az F_2 állapotban van, akkor az F_4 állapotba kerül.

(4) ... stb.

Fontos, hogy ezek a „ha-akkor" állítások *nem* materiális kondicionálisok. A $P \rightarrow Q$ materiális kondicionális akkor igaz, ha P hamis vagy Q igaz. Emiatt minden olyan materiális kondicionális igaz, melynek előtagja (P) hamis. Az a materiális kondicionális például, hogy „Ha a Hold sajtból van, akkor vasból van", igaz, mivel a Hold nem sajtból van ($\neg P$). De ha a fenti állítások csak ilyen értelemben igazak, akkor nagyon könnyű T-t implementálni. (1) például materiális kondicionálisként igaz lesz, ha F sosem kerül az F_0 állapotba. Extrém esetben a szabályokat egy olyan „gép" is kielégítheti, ami sosem kerül az $F_0, F_1, \ldots F_{\text{üres}}$ állapotok valamelyikébe.

A kontrafaktuális strukturalista szerint a fentiekhez hasonló „ha-akkor" összefüggések *kontrafaktuális* kondicionálisok, melyek azt fejezik ki, hogy mi történne vagy történt volna bizonyos körülmények között. (1) kontrafaktuális olvasatban például csak akkor teljesül, ha F aktuális állapotától függetlenül igaz, hogy ha F az F_1 állapotba kerülne, akkor utána az F_2 állapotba fejlődne.

Általánosabban, ha az A absztrakt számítási modellnek része az a szabály, hogy az A_j számítási fázis után az A_k fázis következik, akkor ennek az A-t implementáló F rendszerben egy kontrafaktuális formájában kell megjelennie: ha a rendszer a $\Phi(A_j)$ állapotok valamelyikében lenne, akkor utána a $\Phi(A_k)$ állapotok valamelyikébe kerülne (Chalmers 1996a, Maudlin 1989).

Egyes strukturalisták ezt a kritériumot nem specifikusan kontrafaktuálisokra hivatkozva fogalmazzák meg, hanem oksági összefüggéseket (Chrisley 1995, Scheutz 1999, 2001) vagy diszpozíciókat (Klein 2008) követelnek a fizikai rendszertől: a $\Phi(A_j)$-állapotoknak $\Phi(A_k)$-állapotokat kell okozniuk, vagy F-nek rendelkeznie kell azzal a diszpozícióval, hogy egy $\Phi(A_j)$-állapotba kerülve egy $\Phi(A_k)$-állapotba fejlődik tovább. A kauzális és diszpozicionális strukturalizmusból is következik a szóban forgó kontrafaktuálisok igazsága, ezért a továbbiakban a „kontrafaktuális strukturalizmus" címke alá sorolom ezeket az elméleteket is.

A kontrafaktuális strukturalizmus jobban muzsikál a módszertani kritériumok terén, mint kategoriális rokona. A fizikai rendszerek viselkedési szabályai vélhetőleg objektívak$_2$. Az, hogy egy processzor x inputra az y outputot adja-e, a reprezentációnktól független jelenség (még ha maga a processzor mint fizikai tárgy nem is az). Az **Objektivitás** kritériuma tehát az interpretációinktól függetlenül teljesül. A **Visszaigazolás** kritériuma is teljesül, mert prefilozófiai ítéleteink a számítógépek diszpozícióiról vélhetőleg tükrözik a gépek által aktuálisan implementált programok logikáját; ha nem tükröznék, akkor fura lenne azt állítani, hogy a gép az adott programot futtatja.

Kevésbé egyértelmű, hogy a **Magyarázat** elve érvényesül. Egy mobiltelefon működését a kontrafaktuális strukturalizmus fényében kevéssé *magyarázza* az, hogy a mobiltelefon egy programot implementál, hiszen ez utóbbi állítás (a strukturalista szerint) csak annyit jelent, hogy a programban szereplő $A_i \Rightarrow A_j$ átmeneteknek megfeleltethető egy $\Phi(A_i) \Rightarrow \Phi(A_k)$ alakú kontrafaktuális. De ezek a kontrafaktuálisok nem az $A_i \Rightarrow A_j$ absztrakt szabály *miatt* igazak, és nem is ezzel a szabállyal magyarázzuk őket. A mobiltelefon tranzisztorainak hajlamait a villamosmérnökök által használt fogalmakkal és végeredményben a fundamentális fizikával magyarázzuk. Még rosszabb a helyzet a **Kizárás**sal. Ez a kritérium akkor teljesül, ha nem nyilvánítunk számítógépnek olyan dolgokat, melyek prefilozófiai ítéleteink szerint nem azok. A kontrafaktuális strukturalizmus alapján viszont bármilyen, számítógéppel modellezhető fizikai folyamat számítási folyamatnak minősül majd:

> Algoritmikus leírásokat olyan folyamatokról is lehet adni, melyek paradigmatikusan nem minősülnek számításnak – a galaxisok létrejöttéről, meteorológiai jelenségekről, a verejtékezésről [...]. Erre elegendő bizonyíték a számítógépes modellek széles körű használata az empirikus tudomány összes területén. Emiatt pedig az implementációnak azok a felfogásai, melyek az algoritmusok és a fizikai struktúrák közti [kontrafaktuális] analógiákra apellálnak, minden ilyen folyamatról (és az azt tartalmazó rendszerről) azt fogják állítani, hogy számításokat végez. (Piccinini 2015: 24)

A kontrafaktuális strukturalizmus tehát a pánkomputacionalizmus egyik változatához, konkrétan az *Esse Est Computare* (EEC) elmélethez vezet, miszerint minden fizikai rendszer komputál valamit (15.2. alfejezet). A kontrafaktuális strukturalista annyiban tompítja ezt a következtetést, hogy csak a számítógéppel (elvben) modellezhető folyamatokat tekinti számításoknak.

Chalmers szerint (aki kontrafaktuális strukturalista) ez nem probléma, mert még ha minden fizikai rendszer számításokat végez is, nem következik, hogy *minden* fizikai rendszer *minden* programot implementál:

> Továbbra is objektív kérdés, hogy egy adott számítási leírás egy adott rendszer adekvát leírása-e. A semmitmondás problémája csak akkor merülne fel, ha minden rendszer minden [absztrakt számítási modellt] implementálna, de nem ez a helyzet. Egy adott komplex [számítási modellt] csak a fizikai rendszerek egy töredéke fog implementálni. (Chalmers 1996a: 331)

Chalmers szerint tehát a kontrafaktuális strukturalizmussal azért nincs semmi gond, mert (a 15.2. alfejezet terminológiáját használva) csak EEC következik belőle, TPK nem. De ez az enyhébb következmény is rendelkezik a pánkomputacionalizmus összes lényeges negatív tulajdonságával. Ha a pánkomputacionalizmust elutasítjuk, akkor el kell utasítanunk a kontrafaktuális strukturalizmust

is. Amellett, hogy számítógépeknek tekinti a falakat és csak triviális értelemben tekinti számítógépnek a laptopokat, a **Magyarázat** kritériumát sem elégíti ki.

Összefoglalva: a természetes tulajdonságokra építő kategoriális strukturalizmus (**KS2**) működésképtelen, az objektív$_1$ tulajdonságokra építő kategoriális strukturalizmus (**KS1**) a strukturalizmus szellemét meghazudtoló idealista elmélet (melyről később, a 19. fejezetben részletesen szó lesz), a kontrafaktuális strukturalizmus pedig pánkomputacionalizmushoz vezet, ami elfogadhatatlan. A strukturalista felfogás tehát nem alkalmas az implementáció definiálására.

17.

Az eszközfelfogás

Az eszközfelfogás szerint egy F fizikai tárgy akkor implementálja a P programot, ha F olyan emberalkotta eszköz (*artifact*), melynek az a funkciója, hogy a P által definiált módon viselkedjen. Az eszközfelfogásnak elvben sok változata létezhet, attól függően, hogy mit gondolunk az eszközök metafizikájáról. „Az" eszközfelfogásról beszélni emiatt kissé félrevezető, hiszen az eszközök természetével kapcsolatban élesen eltérő elméletek keringenek a társadalomontológiában (*social ontology*). Az a radikális álláspont például, hogy az „eszköz" teoretikus tartalommal nem rendelkező, zavaros kifejezés, melyet a társadalom elemzésekor kerülnünk kell (Sperber 2007) az implementáció eszközfelfogásával kombinálva arra a konklúzióra vezetne, hogy számítógépek voltaképpen nem is léteznek.

Az alábbiakban az eszközfelfogás egy specifikus változatára koncentrálok, melyet Piccinini (2015) képvisel. Piccinini maga „mechanikai felfogás"-nak (*mechanical account*) hívja, de valójában mechanisztikus és teleológiai kritériumok sajátos keverékéről van szó, amit a teleológiai komponens tesz igazán egyedivé. Annak érdekében, hogy ne mossam össze az eszközfelfogás elvben lehetséges összes formáját Piccinini elméletével, utóbbit *teleológiai eszközfelfogás*nak vagy röviden *teleológia felfogás*nak fogom hívni.

A teleológiai felfogás három fő komponensből áll. Első komponense egy strukturalista jellegű elképzelés az implementált számítások alaptermészetéről: a számítás eszerint fizikai mintázatok szabályszerű manipulációját jelenti. Ilyen mintázat lehet például a magasabb és alacsonyabb feszültségek eloszlása egy processzor tranzisztoraiban. A második, „mechanisztikus" kritérium szerint a fizikai mintázatokat egy mechanizmus – térben elkülönült részek törvényszerű interakcióiból álló rendszer – manipulálja. A harmadik kritérium a teleológiai komponens: egy fizikai mintázatokat szabályszerűen manipuláló mechanizmus csak akkor végez számításokat, ha az a *funkciója*, hogy a szóban forgó szabályoknak megfelelően viselkedjen. Mindjárt rátérek arra, hogyan értelmezi Piccinini a funkciókat.

Piccinini két tudományfilozófiai hagyományt kombinál, a mechanizmusok elméletét (Bechtel & Richardson 1993, Machamer, Darden & Craver 2000, Craver 2013) és a teleológiai magyarázatok elméletét. A mechanisztikus és teleologikus magyarázatokat sokszor élesen szembeállítják. A mechanikus magyarázat egy rendszer működését a rendszer részeiből vezeti le (egy számítógép esetében például a processzorok, tápegység, ventilátorok stb. tulajdonságaiból). A teleológiai magyarázatok viszont – a közkeletű felfogás szerint – a rendszer működésének következményeire koncentrálnak és nem törődnek azzal, hogy ezeknek milyen mechanizmus okozta. A karburátor funkciója az, hogy levegővel keverje az üzemanyagot és befecskendezze az égéstérbe; ennek pontos mechanizmusa a motor mint funkcionális rendszer szempontjából mellékes.

Piccinini (2015: 5. fejezet) megkérdőjelezi ezt a szembeállítást. Szerinte a mechanisztikus és teleologikus magyarázatok szorosan összefüggnek, mert minden funkcionális magyarázatnak tartalmaznia kell mechanisztikus komponenseket. Az üzemanyag-porlasztásnak olyan fizikai feltételei vannak, melyek alapján egy karburátort csak nagyon speciális mechanizmusok képesek realizálni; nem lehet bármiből karburátor. Számításokat agyszövet segítségével éppúgy lehet realizálni, mint fogaskerekek, vákuumcsövek vagy szilikonlapok felhasználásával, de az MI futtatásához szükséges, másodpercenként több tízmilliárd aritmetikai műveletet igénylő programokat csak nagyon különleges mechanizmusokon lehet realizálni.

A mechanisztikus és funkcionális magyarázatok persze nem *azonosak*, hiszen ha valaminek funkciót tulajdonítunk, akkor tevékenységének eredményét és ennek tágabb összefüggéseit hangsúlyozzuk, nem az adott tárgy belső struktúráját. Abból pedig, hogy egy mechanizmus valamit okoz, nem következik, hogy ennek okozása a funkciója (a számítógépnek például nem funkciója, hogy hőt termeljen, sőt, ez műszaki szempontból inkább problémát jelent).

Mitől lesz egy kupac drótnak, műanyagnak, mágnesnek és szilikonnak az a *funkciója*, hogy fizikai mintázatokat manipuláljon algoritmikus szabályok alapján? Vélhetőleg nem attól, hogy erre *használjuk*, hiszen a teleológiai felfogás a számítást objektív (az előző fejezet terminológiája szerint objektív$_2$), a külvilágban zajló, a mi emberi reprezentációinktól független folyamatnak szeretné tekinteni. Piccinininek tehát az objektív (objektív$_2$) teleológia fogalmára van szüksége. A 17.1. alfejezet röviden ismerteti a biológia filozófiája által kínált opciókat, a 17.2. alfejezet pedig bemutatja, hogyan keveri Piccinini az egyik ilyen opciót a tudatos (propozicionális attitűdökből eredő) célok fogalmával. Amellett fogok érvelni, hogy a teleológiai felfogás egyik változata sem képes megragadni az implementáció lényegét.

17.1. Naturalizált teleológia

A teleológia a premodern filozófia egyik központi témája volt, de a modern tudomány fellendülése óta általában gúny és megvetés tárgya. A standard elképzelés szerint Arisztotelész természetfilozófiája – a modern tudományt megelőző domináns paradigma – „vég-okokat", objektív célokat tulajdonított az élőlényeknek, sőt az élettelen tárgyaknak is, a természetben zajló változást pedig ezekre az objektív célokra vezette vissza. Ahogy Hobbes epésen megjegyezi, Arisztotelész szerint a kő azért esik le, mert a földön szeretne lenni (Hobbes 1651/1998: 451). A modern természettudomány látszólag elveti a természetes célok fogalmát, a fizikai rendszereket alkotórészeikre bontja fel, az alkotórészek viselkedését pedig teleológiától mentes, matematikai törvényszerűségekkel modellezi.

A sors iróniája, hogy miközben a modern természettudomány bizonyos értelemben az arisztotelészi világkép felszámolásából született, továbbra is szüksége van egyfajta célfogalomra az organizmusok megértéséhez. A szívnek az a funkciója, hogy keringesse a vért, nem pedig az, hogy dobogó hangot hallasson, noha mindkettő következik fizikai tulajdonságaiból. A biológiai funkció tehát több a puszta fizikai következménynél.

A funkciók persze nem azonosak a célokkal. Felvethetnénk, hogy a modern tudománynak csak funkciókra van szüksége, teleológiára nincs. De cél és funkció egymás közeli rokonai:

> [N]oha van különbség funkció és cél között, ezek a jelenségek szorosan összefüggnek. [...] Ha egy organizmusnak valamilyen célt tulajdonítunk, azt feltételezzük, hogy az organizmus hajlamos létrehozni egy bizonyos állapotot. A funkciókkal kapcsolatos elméleteknek pedig tipikus előfeltevése, hogy a következmények alapján tulajdonítunk funkciókat. A célokra hivatkozó magyarázatok és a funkcionális magyarázatok tehát ugyanannak a *genus*nak az alfajai. [...] De a kapcsolat talán ennél is szorosabb. [...] [H]a azt mondjuk, a szív funkciója a vér keringetése, mi mást állítunk, mint hogy a vérkeringés elősegíti az organizmus egyik célját, a túlélést? És ha azt mondjuk, hogy a levélnehezék funkciója a papírköteg egyben tartása, mi mást állítunk, mint hogy a levélnehezék készítőinek és használóinak az volt a célja, hogy a papírok ne szóródjanak szét? (Garson 2016: 18)

A funkciókat és a célokat is irányulás jellemzi. Ha x funkciója a φ folyamat, akkor x tevékenysége azt *segíti elő*, hogy φ beinduljon vagy fennmaradjon. Ha y célja φ, akkor y arra *törekszik*, hogy φ fennálljon. Mindkét esetben *valamiért*, valaminek érdekében történik valami. A célok és funkciók abból a szempontból is szorosan összefüggnek, hogy ha x funkciója φ, akkor vélhetőleg létezik olyan y, hogy y célja φ, vagyis a funkció teleológiát feltételez. Ez vélhetőleg fordítva is igaz: y aligha törekedhet φ-re anélkül, hogy ehhez felhasználna valamit, ami-

nek így φ elősegítése lesz a funkciója. A modern tudomány kontextusában is beszélhetünk tehát teleológiáról, ha ezen nem többet (de nem is kevesebbet) értünk, mint a funkciók jelenlétét. Ez nem az arisztotelészi metafizika felélesztése, hanem a funkcionális összefüggések elismerése.[1]

A Piccinini-féle implementáció-felfogás nem légüres térben jött létre, hanem a biológia filozófiájának egyik kortárs témájához, a biológiai funkciók természetéről szóló vitához kapcsolódik. Ezen belül öt markáns álláspontot szeretnék bemutatni.[2] Mindegyik esetben megvizsgálom, hogy kiterjeszthető-e az adott felfogás az eszközök (mint funkcióval rendelkező emberalkotta tárgyak) definiálására.

Szelekció A funkciók szelekcióalapú (*selected effects*) vagy etiológiai felfogása az evolúciós előtörténetből eredezteti a funkciókat. Az x élőlény y részének akkor a φ tevékenység a funkciója (pl. az ember (x) szemének (y) akkor a látás (φ) a funkciója), ha x ősei számára szelekciós előnyt jelentett, hogy y a φ tevékenységet végzi.

Fitnesz A fitnesz-alapú (*fitness-contribution*) felfogás szerint a T biológiai tulajdonság akkor rendelkezik funkcióval, ha T tipikusan – nem történeti távlatban, hanem az aktuális kontextusban – szignifikánsan növeli az organizmus hozzájárulását a következő generáció génállományához.

Makacs Plaszticitás A kibernetikai vagy behaviorista felfogás néven is ismert elmélet szerint a teleológia robusztus hajlam egy adott állapot elérésére. Precízebben: x-nek akkor célja a C állapot, ha x plasztikusan és makacsul irányul C-re. x akkor irányul plasztikusan C-re, ha hajlamos sok különféle kezdőállapotból elérni C-t, és x akkor irányul makacsul C-re, ha x a pályáját érintő perturbációk hatására (pl. ha akadály kerül az útjába) úgy módosítja a viselkedését, hogy továbbra is elérje C-t. Funkcióval azok a dolgok rendelkeznek, melyek elősegítik a makacs és plasztikus viselkedést.

Autonómia Ezt néha organizációs felfogásnak vagy *autopoiészisz*-alapú felfogásnak hívják, és elsősorban a túlélésre való törekvést, vagyis az élőlények legalapvetőbb célját szokás definiálni vele. Az x fizikai rendszernek akkor célja a túlélés, ha x-ben olyan körkörös oksági folyamat zajlik, melynek egyes fázisai egymás létrejöttét segítik elő. Ilyen például a táplálkozás, emésztés és táplálékkeresés ciklusa: a táplálkozás lehetővé teszi az emésztést, az emésztés energiát termel és lehetővé teszi a táplálékkeresést. Az élőlények működését biokémiai szinten is ilyen körfolyamatok jellemzik: az enzimek lehetővé teszik a fehérjék szintézisét, ami a degradálódó enzimek újratermeléséhez vezet. A funkciók a körfolyamat egyes részeinek

[1] Vannak persze, akik magát az arisztotelészi teleológiát akarják feltámasztani a funkciók kapcsán (Feser 2019).

[2] Az itt következő áttekintés alapja Kodaj, Bernáth és Pickup (2025). A kimaradó elméleteket Garson (2016) 6. fejezete tárgyalja.

a körfolyamat fenntartásában játszott szerepéből erednek. A táplálkozás funkciója például az energiatermelés; általában, x funkciója az a tevékenység, ami x-et a körfolyamat részévé teszi.

Oksági Szerep Az oksági szerepekre épülő felfogás (*causal role account*) a funkciókat a funkcionális magyarázatokban, nem pedig magukban a biológiai rendszerekben lokalizálja. A funkcionális magyarázatok a rendszer viselkedését a részek tulajdonságaira vezetik vissza. Mivel ugyanazt a rendszert több különböző absztrakciós szinten is le lehet írni, a funkciók léte részben a rendszerről adott interpretációinktól függ.

A fentiek közül egyik felfogás sem tekinthető uralkodónak a biológia filozófiájában, és többel kapcsolatban felmerült megsemmisítő erejűnek vélt ellenérv. A **Szelekció**t például sokak szerint cáfolja a Donald Davidson (1987) által (más célra) kidolgozott gondolatkísérlet Mocsáremberről, egy embernek tűnő lényről, aki villámcsapás hatására áll össze a mocsárban lebegő szerves anyagból. Mocsárember szívének vélhetőleg az a funkciója, hogy keringesse a vért, de a szelekcionista megközelítés szerint nincs funkciója, mert Mocsárembernek nem voltak ősei. A szelekcionisták természetesen tagadják, hogy Mocsárember cáfolná az elméletüket.[3]

A továbbiakban a vita kedvéért azt feltételezem, hogy mindegyik megközelítés védhető, és versenyeztetésük a biológia filozófiájának belügye. A biológia filozófiája jelen esetben csak inspiráció: a kérdés az, hogy kiterjeszthető-e valamelyik funkciófelfogás az eszközökre, ezáltal pedig a számítógépekre. Ennek megfelelően a felfogások alapötleteire koncentrálok: a **Szelekció** a felmenők evolúciós sikereiből eredezteti a funkciókat, a **Fitnesz** az egyedek reprodukciós sikereiből, a **Makacs Plaszticitás** bizonyos diszpozíciókból, az **Autonómia** az önújratermelő körfolyamatokból, a **Szerep** pedig a magyarázó tevékenységünkből. Vajon e megközelítések valamelyike alkalmas-e arra, hogy azt állítsuk, a számítógépeket a funkciójuk teszi számítógépekké?

A **Szelekció** alapú eszközfelfogás alapján egy tárgy akkor rendelkezik funkcióval, ha „ősei" evolúciós előnyökkel rendelkeztek (vagy azzal ruházták fel használóikat). Ez a megközelítés véleményem szerint két okból is reménytelen. Egyrészt a fizikai tárgyaknak nincs olyan egyértelmű, reprodukción és leszármazáson alapuló előtörténete, mint az élőlényeknek. A laptopomnak nincs családfája, nincsenek szülei és nagyszülei. Létrejött valahogy, egy gyárban rakták össze, ennyiben tehát van története, ahogy az architektúrájának is van. De ez alapján nem definiálhatóak a laptop *ősei*, így az ősei által nyújtott szelekciós előnyök sem. Másrészt az eszközhasználat (legalábbis számítógépek esetében) kizárólag az emberi társadalom kontextusában értelmezhető, az emberi társadalmat viszont vélhetőleg nem a darwini értelemben vett evolúciós folyamatok,

[3] Ld. Papineau (2022). A Mocsárembernél szofisztikáltabb anti-szelekcionista érveket ismertet Kingsbury (2023).

vagy nem csak azok, vezérlik. A laptopomnak nemcsak jól definiálható ősei nincsenek, de eleve tisztázatlan, hogy őseinek miféle létharcban kellett volna fennmaradniuk és továbbörökíteniük a „génjeiket". Tekinthetjük a piaci versenyt a tárgyak közti létharcnak, de az eszközfunkciókat lehetetlen pusztán a piaci versenyből eredeztetni, hiszen egy piacilag sikertelen vagy eleve piacra sem dobott tárgy is lehet jól funkcionáló eszköz. Nem állítom, hogy a technológia evolúciójáról beszélni értelmetlen. Azt állítom, hogy az eszközfunkciókat nem lehet a darwini típusú evolúciós előtörténetre visszavezetni.

A **Fitnesz** esetében nem kell azonosítanunk a laptop szüleit és nagyszüleit, mert az aktuális populációból, nem pedig az evolúciós előtörténetből erednek a funkciók. Tárgyak szaporodásáról sem kell beszélnünk, mert az eredeti felfogást módosítva mondhatjuk azt, hogy x-nek akkor a φ folyamat fenntartása a funkciója, ha φ fenntartása valószínűbbé teszi x fennmaradását. Ez lefedi az olyan eseteket is, amikor egy tárgyat a készítők szándékaitól eltérő célokra használnak, például ha a Covid-lezárásokat nyilvántartó telefonos alkalmazás a társadalmi elnyomás eszköze lesz (Hurshman 2023). Az így keletkező teleológiai implementáció-felfogás szerint x akkor végez számításokat, ha x fennmaradását valószínűbbé teszi, hogy x fizikai mintázatokat manipulál szabályszerű módon. De ez a definíció hibás, mert a tartós fennmaradás nem szükséges ahhoz, hogy valami számításokat végezzen. Egy egyszer használatos számológép is számításokat végez, akárcsak egy elhagyatott számítógép, ami bekapcsolva maradt. Egy ilyen számítógép nyilván nem *eszköz* többé, de fura lenne azt mondani, hogy már nem futtat programokat.

A **Makacs Plaszticitás** alapján egy tárgy akkor rendelkezik funkciókkal, ha elősegíti a makacs és plasztikus viselkedést. Ez általában igaz az eszközökre és tipikusan nem igaz az eszköznek nem használt tárgyakra. Az implementáció definíciója azonban nem építhető a makacs plaszticitásra, mert ahhoz, hogy egy fizikai rendszer implementáljon egy programot, nem kell elősegítenie semmit. Egy bekapcsolva maradt, elhagyatott számítógép programokat implementál, de nem segíti a makacs és plasztikus viselkedést. Felvethetnénk, hogy maga a programfuttatás makacs és plasztikus jelenség, mivel a számítógép hajlamos ugyanazokat a fizikai mintázatokat generálni különböző kezdőállapotok vagy akár zavaró körülmények esetén, de ez vélhetőleg csak a számítógépek egy speciális, szűk halmazára (pl. az önvezérlő drónokra vagy bombakereső robotkutyákra) igaz.

Az **Autonómia**-felfogás ön-újratermelő körfolyamatokból eredezteti a funkciókat. Az emberek esetében az önújratermelésben lényeges szerepet kapnak fizikai tárgyak is – aligha lennénk képesek ugyanebben a formában újratermelni magunkat, ha nem lenne telefonunk, ruhánk, lakóhelyünk stb. Az **Autonómiá**ra épülő funkciófogalom tehát elegánsan kiterjeszthető az eszközökre. Az így keletkező implementáció-felfogás azonban a korábbihoz hasonló problémákba ütközik. Olyan rendszerek is implementálhatnak programokat, melyek egyetlen organizmus önújratermelő körfolyamataiban sem kapnak szerepet és

maguk sem rendelkeznek ilyen körfolyamatokkal. A magára hagyott számítógép fennmaradásához nem feltétlenül járul hozzá az, hogy programot futtat. Tehát az **Autonómia** nem használható az implementáció definiálására.

Utolsó versenyzőnk az **Oksági Szerep**. Ha ezt az eszközökre hangszereljük át, azt kapjuk, hogy x akkor eszköz, ha x része egy olyan rendszernek (vélhetőleg egy ágens és különféle tárgyak összességének), melynek tulajdonságait mi mint külső megfigyelők részben x tulajdonságaira vezetjük vissza. Ha ez netán extenzionálisan adekvát implementáció-definícióhoz vezet is, a definíció az idealizmus egy változata lesz amiatt, hogy az oksági szerepek (a szóban forgó funkció-felfogás szerint) a megfigyelők reprezentációitól függnek.

17.2. A Piccinini-féle eszközfelfogás

Piccinini a **Fitnesz**-felfogást kombinálja a propozíciós attitűdökön alapuló teleológiával, vagyis a tudatos célkövetéssel, ami érthető okokból nem szerepel a fenti listán. Piccinini megkülönbözteti az *objektív* célokat, melyek az evolúciós fitneszen alapulnak (2015: 112), a *szubjektív* céloktól, melyek mentális állapotokból erednek és akár a célok közti konfliktusokhoz is vezethetnek:

> Érző organizmusok esetében a jólét, a fájdalom hiánya és a gyönyör akkor is célnak tekinthető, ha akadályozza az objektív célokat, a túlélést és inkluzív fitneszt. Az intelligenciából – az érzékenységgel társulva vagy attól függetlenül – szintén eredhet olyan cél (pl. a tudásvágy), ami zavarja más célok megvalósítását. Az értelem és érzékenység közös eredője lehet az etika és morál a maga sajátos céljaival, például a nem vérrokon fajtársak túlélése iránti igénnyel. (Piccinini 2015: 116)

A célokból erednek a funkciók mint olyan tulajdonságok, melyek az objektív vagy szubjektív célok megvalósulását segítik elő (uo.) A funkciók természete emiatt nemcsak elméleti szinten nem egységes, de a gyakorlatban komoly zavarok forrása lehet. A szubjektív célok akadályozhatják az objektív célok megvalósulását, egyes szubjektív célok pedig más szubjektív célokét.

Piccinini implementáció-elmélete végeredményben tehát így foglalható össze:

> **Teleológiai Felfogás** Egy F fizikai rendszer akkor végez számítást, ha (i) F objektív vagy szubjektív célok megvalósulását elősegítő mechanizmus, és (ii) F egyik funkciója fizikai mintázatok manipulálása olyan szabályok szerint, melyek kizárólag a mintázatok fizikai alkotórészeinek különbségeire érzékenyek. (Piccinini 2015: 121 alapján)

Számítógép az, ami fizikai mintázatokat alakít át fizikai (nem pedig szemantikai) tulajdonságok alapján, egy algoritmusnak megfelelően, és ezáltal objektív vagy szubjektív célokat szolgál.

Piccinini (2015: 141–148) szerint a teleológiai felfogás az összes vonatkozó módszertani kritériumnak megfelel. A számítás ez alapján objektív (**Objektivitás**), a számítógép viselkedését részben az általa implementált programok magyarázzák (**Magyarázat**), a prefilozófiai ítéleteink alapján számítógépnek minősülő rendszerek továbbra is annak minősülnek (**Visszaigazolás**), a prefilozófiai ítéleteink szerint számításokat nem végző rendszerek, mint a falak vagy hurrikánok, nem minősülnek annak (**Kizárás**).

Piccininivel ellentétben úgy vélem, a teleológiai felfogás e kritériumok egy részének egyáltalán nem felel meg. Mielőtt ezt részletezném, egy általános problémát szeretnék jelezni. Piccinini egyik alapfeltevése az, hogy a számítógépek által végzett mintázat-manipuláció elősegíti szubjektív és objektív céljaink megvalósítását. Csakhogy a számítógépekben zajló fizikai folyamatoknak szűk értelemben nem, csak bizonyos társadalmi kontextusokban van ilyen funkciója. A számítások elemi fázisai – például amikor a processzor lefuttat egy „while" ciklust – egyáltalán nem rendelkeznek teleológiai funkciókkal, és a mechanizmus egésze is csak akkor rendelkezik teleológiai funkciókkal, ha meghatározott (súlyosan kontingens) módokon használják. Tegyük fel, hogy az iPhone semmilyen objektív vagy szubjektív célom megvalósulását nem segíti elő; nincs rá szükségem, nincs rá pénzem, és eleve irritál mint termék. Ez esetben az iPhone az én szempontomból nem végez számításokat, mert semmi köze a céljaimhoz. Piccinini erre talán azt felelheti, hogy nem *egyes* eszközök és *egyes* emberek közti viszonyokról van szó, hanem arról, hogy egy mechanizmus kielégít-e szubjektív vagy objektív célokat a populáción belül. De ez sem sokat javít a helyzeten. Ha egy kivételével az összes iPhone megsemmisülne és az utolsó példány egy garázsban felejtve futtatna valamilyen (semmilyen külső eszközzel nem kommunikáló) programot, akkor az iPhone-ok nem járulnának hozzá az emberi populáció szintjén szubjektív vagy objektív célok megvalósításához, de utolsó példányuk továbbra is végezne számításokat (amíg akkumulátora engedi).

A fentiek miatt megkérdőjelezhető, hogy a teleológiai felfogás megfelel-e az **Objektivitás** kritériumának. A számítógépek felhasználási módjai annyira sokrétűek, hogy lehetetlen minden egyes számítógépet szubjektív és objektív célok jól meghatározható halmazához kapcsolni, akár egyéni, akár populációs szinten. Vannak haszontalan számítógépek, melyek semmilyen célt nem szolgálnak, mégis számításokat végeznek. Egy számítógépet célmentes kontextusba helyezhetünk anélkül, hogy ettől megszűnne számításokat végezni. Ezt az invarianciát a teleológiai felfogás képtelen megmagyarázni, sőt, definíció szerint kizárja. Hasonló okokból a teleológia felfogás a **Visszaigazolás** kritériumát sem teljesíti, hiszen prefilozófiai ítéleteink szerint egy bekapcsolva hagyott, elfeledett iPhone továbbra is végez számításokat.

Hasonló a helyzet a **Magyarázat** kritériumával. Szubjektív és objektív céljaink nem magyarázzák, hogy a Windows futtatásakor miért viselkedik másképp a számítógép, mint a Linux futtatásakor. Tehát a Piccinini-féle felfogás teleológiai komponense semmiféle magyarázati szerepet nem kap. A magyarázatot a

felfogás strukturalista komponense szolgáltatja: a számítógép azért viselkedik máshogy, mikor más operációs rendszert tesznek rá, mert az operációs rendszerek eltérései miatt más kontrafaktuális szabályszerűségek vonatkoznak rá. A magyarázat terhét végeredményben a strukturalista komponens hordozza, a strukturalizmus viszont egyéb okokból tarthatatlan.

18.

Antirealizmus

A következő két fejezet az implementáció idealista felfogását tárgyalja. Eszerint egy fizikai rendszer akkor végez számításokat, ha mi megfigyelők számításként (mintázatok szabályszerű manipulációjaként) értelmezzük a tevékenységét. Ez egy antirealista nézet. Mivel azonban az „antirealizmus" egymástól radikálisan eltérő metafizikai álláspontokat jelölhet, fontos előbb elkülöníteni ezek főbb típusait.

Az antirealizmus a modern metafizika vissza-visszatérő témája. Valaki akkor antirealista az F-ekkel kapcsolatban, ha úgy véli, F-ek nem részei az interpretációinktól független valóságnak, nem léteznek „odakint", a reprezentációinkon kívül. A morális antirealisták szerint a morális jó és rossz fogalma puszta emberi interpretáció olyan nyers fizikai tényekről, melyekben valójában nincsenek jelen értékek. A matematikai antirealisták szerint a számok és egyéb matematikai struktúrák „odakint" nem, csak a matematikusok képzeletében léteznek. A tudományos antirealisták szerint a természettudomány által posztulált, közvetlenül nem érzékelhető dolgok (pl. az elemi részecskék) csak a valóságról alkotott elméleteinknek részei, nem a külvilágnak. És így tovább. Általában, az F-ekkel kapcsolatos antirealizmus az a tézis, hogy F-ek objektíven nem, csak a mi interpretációinkban vagy reprezentációinkban léteznek.

Fontos azonban megkülönböztetni az antirealizmus gyenge és erős változatát. *Gyenge* értelemben akkor antirealista valaki az F-ekkel kapcsolatban, ha tagadja az F-ek létét, de nem állítja, hogy az F-ek léte ellentmond a természet törvényeinek. Tipikus gyenge antirealizmus a jetitagadás. Aki nem hisz a gorillaszerű hegyi óriások létében, az többnyire nem tagadja, hogy ilyen lények *elvileg* létrejöhettek volna. Nem a nomologikus lehetőségüket zárja ki a világból, hanem a tényleges létüket.

Erős értelemben akkor antirealista valaki, ha nem egyszerűen tagadja az F-ek létét, hanem úgy véli, F-ek a természet törvényei alapján eleve csak a reprezentációinkban létezhetnek. Az erős és gyenge antirealizmus ellentmond egymásnak, hiszen ha valaki gyenge antirealista az F-ekkel kapcsolatban, akkor nomologikusan lehetségesnek, míg ha erős antirealista, akkor nomologikusan

lehetetlennek tartja az F-ek létét. Ezt a különbséget a boszorkányok és a jetik különbségével illusztrálhatjuk.

A valóságban sem boszorkányok, sem jetik nincsenek (tegyük fel), tehát a boszorkányokkal és jetikkel kapcsolatos antirealizmus egyformán megalapozott. Az is igaz, hogy a boszorkányok és jetik csak az emberek képzeletében léteznek (akármit is jelent ez). De boszorkányok nemcsak, hogy nem léteznek az emberek képzeletén kívül, nem is létez*hetnek* azon kívül a világunkban érvényes természeti törvények szerint. A mágia nomológiailag lehetetlen, emiatt pedig a varázserővel rendelkező nők is azok. A boszorkányokkal kapcsolatos antirealizmus tehát az erős antirealizmus egy formája, ellentétben a jetikkel vagy ufókkal kapcsolatos antirealizmussal. Hegyi óriásmajmok és űrutazó idegen létformák létét nem zárják ki a természet törvényei, boszorkányokét igen.

A továbbiakban az F fajtát vagy tulajdonságot („boszorkány", „morálisan helyes" stb.) *reprezentációba zárt* fajtának vagy tulajdonságnak nevezem, ha F-ek léte elgondolható, de nomológiailag lehetetlen. Az F-ekkel kapcsolatos erős antirealizmus eszerint az a tézis, hogy az F reprezentációba zárt fajta vagy tulajdonság. A reprezentációba zárt fajtákat és tulajdonságokat „puszta fenomén"-nek nevezem majd.

Az erős antirealizmus sem egységes nézet. Két markáns alfaját különböztethetjük meg, az *ingoványos* és a *fundált* antirealizmust. Ez nem bevett terminológia (ahogy a „gyenge" és „erős" sem), de fontos különbségre utal.

Az ingoványos és fundált antirealizmus különbségét a boszorkányok és a szivárványok különbségével illusztrálhatjuk. Ha a „szivárvány" szó az égen vihar után végighúzódó színes sávokat jelenti, akkor a szivárványokkal kapcsolatban az erős antirealizmus a helyes álláspont. A szivárványok – mint színes égi sávok – nem részei a valóságnak, hiszen az égen nem találhatóak színes sávok vihar után. Sem az angyalok, sem az ufók nem szoktak speciális filctollal csíkokat rajzolni az égre. Szivárványt amiatt látunk, hogy az odafent lebegő vízpermeten megtörik a napfény, a fotonok pedig a látóidegünkbe csapódva színes csíkok látványát okozzák. Ennek az optikai illúziónak objektív alapja van, és megfelelő körülmények között minden normális megfigyelő tapasztalja. A szivárvány nem *hallucináció*, annak ellenére, hogy amit látunk (a színes sávok összessége) valójában nincs jelen az égen, csak az égről alkotott reprezentációnkban. Az agyunk által színes csíkokként interpretált jelenség fizikai leírásában nem szerepel a „piros", „lila", „kék" stb. szó, vagy legfeljebb fényhullámhosszok nevének rövidítéseként szerepel benne.[1] Emiatt a szivárvány nem is *létezhet* (a természet aktuális törvényei szerint) máshol, mint a reprezentációinkban.

[1] A piros szín vizuális élménye nem azonosítható egyetlen konkrét fényhullámhosszal sem, tehát a „piros fény" és hasonló kifejezések eleve pontatlanok. Egy adott színélményt fényhullámok teljesen eltérő kombinációi válthatják ki. Ez a jelenség a *metamerizmus* (Gunther 2012: 420), amit Newton 1704-ben írt le *Optika* c. művében (Levkowitz 1997: 6). A metamerizmusnak ipari jelentősége is van (pl. a nyomdászatban és a számítógépes grafikában).

A szivárvány tehát puszta fenomén, ugyanakkor a szivárványok léte nem puszta kitaláció, szemben a boszorkányok létével. Akik régen boszorkánynak néztek valakit, azok tévedtek vagy hallucináltak, aki viszont szivárványt lát az égen (megfelelő légköri viszonyok között), az nem téved vagy hallucinál: tényleg van az égen valami, amit a vizuális rendszerünk törvényszerűen színes csíkokként interpretál. Ez az élmény a természet törvényeinek megfelelően és általuk magyarázható módon adódik az interpretált fizikai jelenségből. Ezt nevezem *fundált* (erős) antirealizmusnak.

A boszorkányok esetében nyomát sem találjuk annak a törvényszerű összefüggésnek, ami a szivárvány látványa és az azt kiváltó légköri jelenség között fennáll. Egy feltűnően szép nő vagy egy baljós öregasszony látványa nem váltja ki törvényszerűen minden normális megfigyelőben azt az ítéletet, hogy a szóban forgó nő a Sátán szolgája. Csak abnormális megfigyelőkben keletkeznek ilyen ítéletek, és még esetükben is bizonytalan és kontingens módon. Boszorkányreprezentációk nem törvényszerűen, hanem földrajzilag-történetileg eltérő körülmények között keletkeznek, melyek nem alkotnak egységes típust. Ezt nevezem *ingoványos* antirealizmusnak. Nincsenek jól megragadható, törvényszerű összefüggések az F-ek élménye vagy feltételezése és a külvilág (nomológiailag szükségszerűen F-mentes) tényei között.

Az ingoványos és a fundált antirealizmus alighanem egy spektrum két végét képezi. Elképzelhető, hogy egyes reprezentációba zárt fajták vagy tulajdonságok esetében köztes álláspontot kell felvennünk. Jó példa erre a társadalmi nem (*gender*). Az, hogy mi számít nőiesnek vagy férfiasnak, történetileg és földrajzilag erősen változékony, tehát a „nő" és „férfi" mint társadalmi (nem pedig pusztán biológiai) fajta puszta fenomén, mely a reprezentációinkról leválasztva nem létezik, hiszen valaki attól lesz nő vagy férfi (társadalmi, nem pusztán biológiai értelemben), hogy bizonyos módon reprezentálják.

Nem azt állítom, hogy a nemekkel kapcsolatos szociálkonstruktivizmus garantáltan igaz; azt próbálom illusztrálni, hogy az erős antirealizmusnak létezhetnek az ingoványos és fundált pólusok közti formái. A gender mint társadalmi konstrukció egyáltalán nem olyan zavaros jelenség, mint a boszorkányhit, hiszen egy adott társadalomban többnyire robusztus kritériumok szabályozzák, ki melyik nemhez tartozik. Ugyanakkor a gender nem is annyira fundált jelenség, mint a szivárvány, hiszen a nemi hovatartozás kritériumai történetileg és földrajzilag erősen változékonyak, sőt, elvben tetszés szerint alakíthatóak.

Talán vannak további példák is köztes esetekre; ezt a kérdést nem tudom és nem is akarom itt tisztázni. Az egyszerűség kedvéért a pólus két szélére koncentrálok, mert az implementációs idealizmus szempontjából az ingoványos és fundált antirealizmus elkülönítése a legfontosabb.

Az implementációs idealizmus a fundált erős antirealizmus egy formája. Azért fontos hangsúlyozni ezt, mert amikor az idealista azt állítja, hogy számítások a valóságban nem léteznek, akkor metafizikai értelemben teljesen mást

állít, mint az, aki a jetik vagy a boszorkányok létét tagadja Hasonlítsuk össze a következő két állítást:

A1 Ott egy boszorkány!

A2 Ott egy szivárvány!

A1 garantáltan hamis, bárkire mutatunk is rá. **A2** viszont gyakran igazságot fejez ki. A szivárványokra rá lehet mutatni, lokalizálni lehet őket. Ha egy puszta fenomén, mint a szivárvány, fundált, akkor bizonyos értelemben tényleg „ott" van, bizonyos értelemben objektív; minden normális megfigyelő tapasztalni fogja a külvilág adott szegletében. De a szivárvány – mint gigászi színes csíkok sorozata – *nem* objektív abban az értelemben, hogy a reprezentációinkon kívül is létezne. Tehát kétféle objektivitásról beszélhetünk (vö. 16.1. alfejezet):

O1 Az F-ek valóságosak$_1$ (létük objektív$_1$) $=_{df}$ (a) Van olyan, a külvilágban előforduló körülmény, melynek hatására minden normális megfigyelő törvényszerűen F-eket tapasztal vagy feltételez, de (b) F-ek nem létezhetnek (a természet törvényei alapján) a reprezentációinkon kívül.

O2 Az F-ek valóságosak$_2$ (létük objektív$_2$) $=_{df}$ F-ek léteznek és a reprezentációinkról leválasztott külvilág részei, nem csak a reprezentációink részei (az F nem reprezentációba zárt fajta vagy tulajdonság, nem puszta fenomén).

Az objektivitás fenti két típusát definiálhatjuk tömörebben is:

O1 Az F-ek valóságosak$_1$ (létük objektív$_1$) $=_{df}$ F fundált és reprezentációba zárt tulajdonság

O2 Az F-ek valóságosak$_2$ (létük objektív$_2$) $=_{df}$ F fundált és *nem* reprezentációba zárt tulajdonság

Fundáltság Az F tulajdonság fundált $=_{df}$ Van olyan, a külvilágban előforduló körülmény, melynek hatására minden normális megfigyelő törvényszerűen F-eket tapasztal vagy feltételez.

Reprezentációba zártság Az F tulajdonság reprezentációba zárt $=_{df}$ F-ek léte elgondolható, de nomológiailag (a természet törvényei szerint) lehetetlen, hogy F-ek létezzenek.

Az antirealizmus három fő formája (és a realizmus) az alábbi fogalmi kombinációknak felel meg:

	F reprezentációba zárt tulajdonság	F *nem* reprezentációba zárt tulajdonság
F fundált	fundált erős antirealizmus	realizmus
F nem fundált	ingoványos erős antirealizmus	gyenge antirealizmus

Ez a táblázat nem a tulajdonságok általános kategóriarendszere. Csak az elgondolható tulajdonságokat fedi le, azokat, melyek elvben részei lehetnek a reprezentációinknak. Ha F elgondolhatatlan, felfoghatatlan, akkor nem lehet fundált és nem alkalmazható rá (sem pozitív, sem negatív értelemben) a reprezentációba zártság fenti fogalma, emiatt pedig kívül esik a fenti kategóriákon. De az antirealizmus jellemzése szempontjából ez lényegtelen, mert az antirealizmus filozófiailag életképes fajtái elgondolható tulajdonságokra vonatkoznak.

Az antirealizmus eltérő fajtáinak összemosásából súlyos félreértések eredhetnek. Az állatorvosi ló ezen a téren a filozófiatörténet leghírhedtebb antirealista elmélete, George Berkeley idealizmusa. Berkeley szerint kizárólag az anyagtalan tudatok valóságosak$_2$, az anyagi tárgyak csupán az elménkben Isten parancsai alapján keletkező ideák (vagy legalábbis ezekre redukálhatóak). Berkeley tehát az anyagi világgal kapcsolatos fundált erős antirealizmus (és a mentális szubsztanciákkal kapcsolatos realizmus) híve volt. Egyik kortársa, az irodalmár Samuel Johnson egy anekdota szerint úgy „érvelt" Berkeley ellen, hogy belerúgott egy kőbe és kijelentette: „Tessék, megcáfoltam" (*I refute it thus*). Ezt a dialektikai manővert manapság is sokan hiszik az idealizmus elleni letaglózó erejű érvnek. Érdemes azonban észrevenni, hogy kövek rugdosásával nem a fizikai tárgyakkal kapcsolatos *fundált* antirealizmust cáfoljuk, hanem a fizikai tárgyakkal kapcsolatos *ingoványos* antirealizmust, amit viszont Berkeley sosem állított. Az F-ekkel kapcsolatos ingoványos antirealizmust valóban cáfolni lehet azzal, ha rámutatunk az F-tapasztalatok törvényszerű jelenlétére, de a fundált erős antirealizmust ez a manőver definíció szerint képtelen cáfolni.

Az implementáció idealista felfogása szerint a számítások létmódjukat tekintve olyanok, mint a szivárványok, vagy mint Berkeley világképében a fizikai tárgyak. Egyes fizikai folyamatokat törvényszerűen számításként interpretálunk (ha jól megfigyeljük őket), ennyiben tehát az „Ott egy számítógép!" mondat ugyanolyan módon igaz (megfelelő körülmények között), mint az „Ott egy szivárvány!". De a számítások nincsenek *ott* a gépekben, ugyanúgy, ahogy az égen sincsenek gigantikus színes csíkok. A számítások puszta fenomének: csak a reprezentációinkban léteznek, a reprezentációinkról leválasztott külvilágban nem létezhetnek. A 19. fejezetben amellett érvelek, hogy ez a felfogás az implementáció legjobb modellje, a 20. fejezetben pedig amellett, hogy ha igaz, akkor a demarkációs tézis is igaz, mert az emberi gondolkodás (szemben a számításokkal) nem puszta fenomén.

19.

Az idealista felfogás

Az implementáció idealista felfogása szerint egy fizikai rendszer akkor végez számításokat, ha az adott rendszert (alapos vizsgálatok után) törvényszerűen számításként *interpretáljuk*, valamilyen absztrakt számítási modellt vetítünk rá. Az előző fejezetben megpróbáltam elkülöníteni ezt az antirealizmus egyéb formáitól. Most amellett fogok érvelni, hogy az implementációs idealizmus az implementáció legjobb definíciója.

Az idealista felfogás a (kombinált) strukturalizmus fundált erős antirealista változatának tekinthető, mely a program és a programot implementáló fizikai rendszer közti analógiákat a reprezentációinkban lokalizálja. A *kategoriális* strukturalisták szerint egy A absztrakt számítási modellt akkor realizálja az F fizikai rendszer, ha van olyan Φ megfeleltetés A részei és F részeinek halmazai között, hogy (a) $\Phi(A_i)$ elemei objektíve hasonlóak egymáshoz, és (b) $A_i \neq A_j$ esetén $\Phi(A_i)$ és $\Phi(A_j)$ elemei objektíven különböznek. A *kontrafaktuális* strukturalisták szerint az implementációhoz az kell, hogy az A program minden $A_i \Rightarrow A_j$ átmenet-szabályra igaz legyen (kontrafaktuális értelemben), hogy ha F a $\Phi(A_i)$ állapotok valamelyikében van, akkor a $\Phi(A_j)$ állapotok valamelyikébe fejlődik. A kombinált strukturalizmus a kategoriális és kontrafaktuális összefüggéseket egyaránt megköveteli.

Az idealista annyit tesz hozzá a kombinált strukturalizmushoz, hogy $\Phi(A_i)$ elemeinek objektíve$_1$ hasonlónak lennie, és $A_i \neq A_j$ esetén $\Phi(A_i)$ és $\Phi(A_j)$ elemeinek objektíve$_1$ különbözniük kell. Az idealista ezzel megoldja a kategoriális strukturalizmussal kapcsolatos legfontosabb problémát, amit az okoz, hogy nem állnak rendelkezésre megfelelő objektív$_2$ hasonlóságok és különbségek a természetes tulajdonságok szintjén (16.1. alfejezet). Ugyanakkor az idealizmus megoldja a kontrafaktuális strukturalizmussal kapcsolatos fő problémát is, ami abból eredt, hogy minden fizikai rendszer, ami számítógéppel modellezhető (hurrikán, naprendszer, verejtékmirigy stb.) számítógépnek minősül a program és a rendszer közti kontrafaktuális párhuzamok miatt (16.2. alfejezet). Az idealistát ez a probléma azért nem fenyegeti, mert ő a kontrafaktuális összefüggések mellett objektív$_1$ megfelelést követel a kategoriális struktúra terén, ez

pedig nincs jelen a hurrikánok és hasonló rendszerek esetében. Egy hurrikánnak nincs szalagja és írófeje, mint egy Turing-gépnek, nincsenek processzorai és regiszterei stb., emiatt pedig az idealista által előírt Φ függvény az ilyen esetekben nem létezik.

Azt a fajta összefüggést, amit az idealista ír elő a programok és a programokat futtató gépek között, *stabil projektálhatóság*nak fogom nevezni. Ha a fent jellemzett Φ függvény létezik, akkor A *stabilan projektálható* F-re. Az idealista szerint az implementáció stabil projektálhatóság.[1]

Az idealista felfogás *nem* feltételezi, hogy *mindenki* projektálja a szóban forgó számítási modellt. Amikor azt mondom, hogy az F fizikai rendszer tevékenységét a megfigyelők törvényszerűen az A típusú számításként interpretálják, vagyis stabilan projektálják F-re A-t egy megfelelő Φ közvetítésével, akkor olyan megfigyelőkre gondolok, akik tanulmányoztak F-hez nagyon hasonló rendszereket és magyarázni akarják F működését. Az számít, hogy Φ *elvben* rendelkezésre áll, nem az, hogy Φ-t mindenki felismeri, aki F-re ránéz. Az idealista azt sem követeli meg, hogy A-t valaha ténylegesen projektálják F-re. Képzeljük el, hogy egy laptop leesik a gyártósorról és véletlenül bekapcsolódik, anélkül, hogy bárki észrevenné. Az idealista szerint ez a tárgy számításokat végez, mert azok a megfigyelők, akik tanulmányoztak a laptophoz hasonló rendszereket és magyarázni akarják a laptop működését, törvényszerűen programokra hivatkoznának a laptop leírásakor, vagyis létezik egy olyan absztrakt számítási modell, ami stabilan projektálható a laptopra, függetlenül attól, hogy soha senki nem projektálja rá.

Felvethetnénk, hogy ez a felfogás *relativizmushoz* vezet a számítások létével kapcsolatban. Tegyük fel, hogy a marslakók belelátnak a processzorokba, közvetlenül érzékelik az ott zajló kvantummechanikai folyamatokat, emiatt pedig a mi laptopjainkat nagyon primitív kvantumszámítógépeknek tekintik és teljesen más számítási modelleket projektálnak rájuk, mint mi. Ekkor a marslakók számára a laptop nem azokat a programokat futtatja, mint számunkra, tehát a számítás relatívvá válik. Még a pánkomputacionalizmus egy változata is igaz lehet. Tegyük fel, hogy a vénusziak közvetlenül érzékelnek *minden* mikroszkopikus folyamatot, látják például, hogy a naprendszer hogyan elégíti ki a kvantumtér-elmélet egyenleteit. Emiatt a vénusziak számításnak látják a bolygók mozgását is, tehát számukra a naprendszer – illetve általában minden fizikai rendszer – számításokat végez.

Ez a fajta relativizmus valóban következik az idealizmusból, mert az idealista relációs jelenségnek, megfigyelők és fizikai rendszerek közti viszonynak tekinti a számítást. Radikálisan más megfigyelők radikálisan más számításokat projektálhatnak ugyanarra a rendszerre. De ez egybevág prefilozófiai ítélete-

[1] Az implementációs idealizmust nagyon kevesen védik. Két általam ismert kortárs képviselője Dewhurst (2018), aki a mechanisztikus magyarázatok elmélete alapján (Craver 2013) dolgoz ki egy hasonló felfogást, illetve Schweizer (2019), aki a Dennett-féle „intencionális hozzáállás" mintájára értelmezi a fizikai számításokat.

inkkel, tehát nem okoz a pánkomputacionalizmushoz hasonló problémát. Ha a vénusziak ellátogatnának hozzánk és elmesélnék, hogy számukra a naprendszer egy számítógép, azt gondolnánk, furcsa kognitív képességeik miatt teljesen mást értenek „számítógép"-en; a földi számítógép-fogalom és a vénuszi-számítógép fogalom csak nevében hasonlít egymásra, tartalmuk teljesen eltér. Tehát az a fajta relativizmus, amit az implementációs idealizmus okoz, nem abszurd. Visszaigazolja azt a prefilozófiai ítéletünket, hogy ha léteznének tőlünk teljesen eltérő kognitív képességekkel rendelkező megfigyelők, akkor fajfüggő lehetne, mi minősül számítógépnek.

Szemben a korábbi fejezetekben tárgyalt alternatív felfogásokkal, az idealizmus kielégíti az implementáció definíciójával kapcsolatos összes kritériumot (ld. 13.3.). Kielégíti a **Magyarázat** kritériumát, hiszen ha A projektálható F-re, akkor a $\Phi(A_i) \Rightarrow \Phi(A_j)$ típusú kontrafaktuálisok miatt a projekció segítségével megjósolható F viselkedése.[2] A számítási modell persze nem fog *mindent* magyarázni F viselkedésével kapcsolatban. Ha a processzor meghibásodik, akkor A alapján nem lehet majd megjósolni F viselkedését, és A eleve nem fogja magyarázni F viselkedésének minden aspektusát (például a processzor melegedését). Az idealista felfogás azonban pontosan azt adja, amit a **Magyarázat** kritériuma követel: specifikusan informatikai jellegű magyarázatokat, melyek a számítási szempontból lényeges fizikai változásokat az implementált program struktúrájára vezetik vissza.

Az idealista felfogás a **Visszaigazolás** kritériumát is kielégíti, mert számítógépnek minősíti a preteoretikus ítéleteink szerint számítást végző fizikai rendszereket. A preteoretikus ítéleteink ugyanis nyilván a fent jellemzett, stabilan projektálható Φ realizációs viszonyokon alapulnak.

Az idealista felfogás a **Kizárás** kritériumát is kielégíti, azaz nem minősít számítógépnek olyan fizikai rendszereket, melyek prefilozófiai ítéleteink alapján nem azok. Az idealizmus egyértelműen kizárja például a 15.1. alfejezetben szereplő perverz implementációkat. A idealista szerint a fal nem futtatja a Wordstart, mert nem tudjuk projektálni rá a Wordstar programját. Az idealista felfogás, a kontrafaktuális strukturalizmussal szemben, a hurrikánokat, galaxisokat és egyéb, számítógéppel modellezhető rendszereket sem tekinti számítási folyamatoknak. Erről korábban már szó volt, de talán érdemes ismét

[2] Ez nem lesz igaz az olyan degenerált esetekben, mint egy 1 állapotú Turing-gép, ami egyetlen jelet képes felismerni és semmilyen kontrafaktuális összefüggést nem elégít ki (folyamatosan az A_0 állapotban van). De ilyen programok esetében eleve nehezen alkalmazható a **Magyarázat** kritériuma, hiszen segítségükkel aligha lehet egy rendszer viselkedését magyarázni. A ténylegesen létező számítógépekre szigorú értelemben véve a nem-degenerált Turing-gépek sem projektálhatóak, mert a ténylegesen létező számítógép tárolókapacitása véges, tehát a Turing-gépnek lesznek olyan átmenetei, melyek nem felelnek meg a fizikai rendszer lehetséges átmeneteinek. De az idealistának nem kell azt állítania, hogy minden számítógépre egy Turing-gép projektálható. A számítógépekre tipikusan gépi kódú vagy emberek számára olvasható nyelven – C-ben, C++-ban, Javában stb. – írt programokat projektálunk.

hangsúlyozni. Legyen H egy hurrikán, P_H pedig egy program, ami a hurrikánt modellezi. Feltehetjük, hogy P_H valamilyen differenciálegyenlet alapján ad becsléseket a hurrikán fizikai paramétereiről. Az idealista felfogás szerint H nem implementálja P_H-t, mert P_H mint *számítógépes program* nem projektálható H-ra: a hurrikánban nincsenek változók, ciklusok és hasonló struktúrák. Egyedül P_H kimeneteit tudjuk megfeleltetni H tulajdonságainak. A Φ realizációs viszonyhoz szükséges objektív$_1$ kategoriális hasonlóságok teljesen hiányoznak.

Az idealista felfogás végezetül az **Objektivitás** kritériumát is kielégíti: az idealista szerint a számítás objektív$_1$ jelenség. Az, hogy egy számítási modell projektálható-e egy fizikai rendszerre, nem ízléskérdés, nem vélemény vagy kulturális hovatartozás kérdése. Egy Android telefonról minden jól informált megfigyelő úgy ítéli, hogy a Linux kernelt futtatja. A számítás tehát – az idealista felfogás alapján – reprezentációba zárt, de fundált, vagyis objektív$_1$ jelenség.

Két ellenérv merülhet fel. Az egyik az, hogy az idealista csak trükkösen teljesíti az **Objektivitás** kritériumát, mert a kritérium igazából azt követelné, hogy a számítások legyen objektívak$_2$. Az idealista egy antirealista objektivitásfogalmat használ fel, de ezzel csalást követ el.

Ha az implementáció adekvát elméletétől elvárhatnánk, hogy objektívnak$_2$ mutassa be az implementációt, akkor ez a kifogás jogos lenne. De nem várhatjuk el, hiszen nem zárhatjuk ki *a priori*, hogy a számítás puszta fenomén, a reprezentációinkról nem leválasztható jelenség. A prefilozófiai ítéleteink ezzel kapcsolatban nem mondanak semmit, mert a prefilozófiai ítéleteinkből hiányzik az objektivitás$_1$ és objektivitás$_2$ megkülönböztetése. Ha az implementáció adekvát elméletével kapcsolatos elvárásaink fő célja az, hogy a prefilozófiai ítéleteinkben megbújó számítás-fogalmat minél kevesebb torzítással explicitté tegyük, akkor nem írhatjuk elő, hogy a számítás legyen objektív$_2$, mert a prefilozófiai ítéleteinkből csak annyi következik, hogy a számítás objektív$_1$ vagy objektív$_2$.

Felvethetnénk, hogy ugyanilyen alapon a formalizmust is védeni lehet a pánkomputacionalizmus vádjával szemben (15.2. alfejezet), arra hivatkozva, hogy a **Kizárás** kritériuma (miszerint falak, hurrikánok stb. ne minősüljenek számítógépnek) ugyanúgy diszkvalifikálja a formalistát, ahogy az objektivitás$_2$ előírása diszkriminálná az idealistát. De ez a panasz alaptalan, vagy legalábbis nem alkalmazható az **Objektivitás** idealista kezelésére, mert prefilozófiai ítéleteinkből igenis következik, hogy a fal nem végez számításokat, az viszont nem következik belőlük (az ehhez szükséges filozófiai distinkciók hiánya miatt), hogy a számítások objektívak$_2$, nem pedig objektívak$_1$.

A másik lehetséges ellenérv az, hogy az idealista földrajzilag-történetileg relatívvá teszi az implementációt, még akkor is, ha elfeledkezünk a marslakókról és vénusziakról. Egy őserdei törzs, mely a modern civilizációtól teljesen elzártan él, nem képes egy Android telefonra projektálni a Linux kernelt, mert sem a mobiltelefonokról, sem az operációs rendszerekről nincs fogalma. Tehát (az idealista felfogás szerint) a törzs számára a mobiltelefon nem végez számí-

tásokat. Hasonlóan, egy 1946-ban tevékenykedő mérnök nem lett volna képes számításokat projektálni tranzisztorok milliárdjaiból álló processzorokra, mert a tranzisztorokat akkoriban még nem találták fel. Tehát az 1946-ben élő mérnök számára egy korabeli számítógép, például az ENIAC, számításokat végez, egy mobiltelefon viszont nem. És így tovább. Emiatt a számítás az idealizmusban még csak nem is objektív$_1$, hiszen különböző humán szemlélők nem ugyanazokat a programokat projektálják ugyanazokra a tárgyakra.

Az idealista erre azt felelheti, hogy a számítások valóban kicsit más módon objektívak$_1$, mint a szivárványok, hiszen a szivárványokat minden (nem látássérült) megfigyelő látja, a programok implementálását viszont csak az „látja", aki valamennyire ért az algoritmusokhoz és a számítógépekhez. Az absztrakt számítások projektálásának vannak bizonyos episztemikus előfeltételei, de ettől nem lesz az implementáció kevésbé objektív$_1$. Hasonlóan, egy Beethoven-versenyműi sajátosságai sem lesznek kevésbé objektívak$_1$ attól, hogy a felismerésükhöz zenei előképzettségre van szükség. A szivárványok esetében az számít normális szemlélőnek, akinek van szeme, tiszta a tudata stb.; a számítások esetében pedig az, akinek megfelelő informatikai és műszaki előismeretei vannak. Ha az elzárt törzs tagjainak vagy az ENIAC programozójának elmagyaráznánk, hogyan működik az Android telefon és mi a Linux kernel, akkor egyetértenének velünk, hogy előbbi az utóbbit futtatja.

Felmerülhet az a kérdés is, hogy az idealista képes-e értelmezni programok emberi beavatkozástól független keletkezését. Képzeljünk el egy robotot, ami módosítani tudja saját programját vagy új programokat hoz létre és lefuttatja őket. Felmerülhet a gyanú, hogy a kódok keletkezése és implementációja ekkor nem pusztán objektív$_1$ jelenség, hiszen bármiféle emberi interpretáció, emberi beavatkozás nélkül zajlik. De idealista szemszögből nézve itt sincs semmi metafizikai újdonság: a robotra stabilan projektálható egy számítási struktúra, és ennek a projekciónak a segítségével meg tudjuk magyarázni, hogy miért projektálhatóak a robotra vagy az általa kezelt számítógépekre újabb programok. Az idealistát talán azzal az próbálhatjuk meg itt zavarba hozni, ha elképzeljük, hogy a robot – mondjuk az MI által kipusztított emberi civilizáció romjain – létrehoz egy teljes robot-társadalmat, a robotok pedig városokat és űrhajókat építenek, olyan technológiai újításokat hoznak létre saját programjaik folyamatos módosítása és új programok telepítése révén, melyekről az emberek még csak nem is álmodtak. Egy magasan fejlett robottársadalomra fura lenne azt mondani, hogy puszta projekció. De véleményem szerint az ilyen radikális példák sem ingatják meg az idealista felfogást, mert az idealista nem tagadja, hogy a szóban forgó szituációban a robotok tényleg városokat és űrhajókat építenek. Egyedül azt fogja tagadni, hogy a robotok által implementált programok a reprezentációinkon kívül is léteznek. Ezeket a programokat projekcióknak fogja tekinteni, melyekkel hatékonyan magyarázható a robotok tevékenysége és ennek eredménye. A robotok processzorai fizikailag nem tartalmazzák a programokat, csak alapot adnak a projekcióra.

Szintén érdekes ellenpéldának tűnhet, ha valaki fejben vagy papíron, tudatosan futtat egy programot, mint a Kínai Szoba lakója. A Kínai Szoba lakója egy programot implementál (chatbotként viselkedik egy fix algoritmus alapján), de ez a tevékenysége nem merül ki abban, hogy stabilan projektálható rá egy számítási modell. A szobára nem csak mi külső megfigyelők projektáljuk a chatprogramot, hanem a szoba lakója a mi reprezentációinktól függetlenül (objektíve$_2$) a chatprogramot futtatja. A 6.4. alfejezet terminológiáját használva az algoritmusok *szubjektív* futtatása tehát nem tűnik projekciónak.[3] Mivel az algoritmusok szubjektív futtatása is az implementáció egy esete, az implementációs idealizmus hamis.

Az idealista erre az ellenérvre azt felelheti, hogy a Kínai Szoba lakója esetében – illetve általában az algoritmusok szubjektív futtatásakor – nem az algoritmus jelenléte objektív$_2$, hanem az algoritmust futtató személy reprezentációi saját tevékenységéről. Ha Searle a Kínai Szobában ül, lázasan lapozgatva a szabálykönyvet, akkor reprezentációi („papír", „toll", „vektorszorzat" stb.) nem reprezentációba zárt tulajdonságok, nem pusztán a Searle-t megfigyelő szubjektumok reprezentációinak a részei, hanem Searle saját mentális állapotainak a részei. Tehát Searle gondolatai, miközben a chatprogramot futtatja, objektívak$_2$, és gondolataiban objektívan$_2$ jelen vannak a programot (mint absztrakt számítási modellt) definiáló szabályok. A programok szubjektív futtatása nem az implementáció egy *fajtája*, hanem az implementáció egyik lehetséges metafizikai vagy oksági eredete. Az implementáció metafizikai vagy oksági eredete objektív$_2$ lehet akkor is, ha az implementáció pusztán objektív$_1$ és a stabil projektálhatóságban merül ki. Az idealista valósnak$_2$ ismerheti el Searle kognitív folyamatait ugyanúgy, ahogy valósnak$_2$ ismeri el (amennyiben nem berkeleyánus) a számítógépekben zajló atomfizikai és kémiai folyamatokat mint a digitális számítások metafizikai alapját.

[3] A 6.4. alfejezet a szubjektív és a konstitutív futtatás között tett különbséget (75. old.) Konstitutív futtatás idealista alapon lehetetlen, legalábbis ha a gondolkodással kapcsolatban realisták vagyunk.

20.

Implementációs demarkáció

Az implementáció idealista felfogása szerint egy fizikai rendszer akkor implementál egy programot, ha a program mint absztrakt számítási modell stabilan projektálható rá. Ha ez igaz, akkor a demarkációs tézis is igaz, vagy legalábbis emellett fogok most, a II. rész lezárásaként, érvelni.

Az implementációs demarkáció alapötlete az, hogy az implementációs idealizmus alapján a számítások jelenléte a külső megfigyelőkhöz fűződő viszony függvénye, míg a gondolkodás nem a külső megfigyelőkhöz fűződő viszony függvénye. Tehát a számítógépek nem gondolkodnak. Szabatosabban:

(1) Ha az MI gondolkodik, akkor ez a tény (metafizikai értelemben) abból ered, hogy egy fizikai rendszer egy programot implementál.

(2) Ha egy F fizikai rendszer a P programot implementálja, akkor ez a tény abból ered, hogy P stabilan projektálható F-re.

(3) Tehát, (1) és (2) miatt, ha az MI gondolkodik, akkor ez a tény abból ered, hogy egy program stabilan projektálható egy fizikai rendszerre.

(4) Ha x gondolkodik, akkor ez a tény nem abból ered, hogy egy program stabilan projektálható egy fizikai rendszerre.

(5) Tehát az MI nem gondolkodik.

Az érv logikailag érvényes, amennyiben az eredet tranzitív, vagyis ha a T_1 tény T_3-ból ered, valahányszor T_2 a T_3-ból, T_1 pedig T_2-ből ered. Mivel jelen esetben T_3 leírása (a 2. premissza utolsó tagmondata) egy definíció behelyettesítésével adódik T_2 leírásából (az 1. premissza utolsó tagmondatából), a tranzitivitás itt aligha megkérdőjelezhető.

Az érv három premisszája (1), (2) és (4). Az 1. premissza véleményem szerint magától értetődő, de érveltem is mellette a 13.1. alfejezetben. A 2. premissza az implementációs idealizmusból adódik, amit a 14–19. fejezet próbált igazolni. Tehát egyedül a 4. állítás szorul külön indoklásra.

Számomra (4) is egyértelműen igaznak tűnik. A kognitív folyamataim metafizikai értelemben nem abból erednek, hogy mások hogyan hajlamosak reprezentálni engem. Általában, az emberi gondolkodás nem külső megfigyelők reprezentációin, nem stabil projektálhatóságon múlik. A gondolkodás *képességének* kialakulásához nyilván szoros köze van annak, hogyan reprezentálnak valakit (ha egy csecsemőt téglának reprezentálnak, sose lesz belőle gondolkodó ember). De ha valaki már elkezdett gondolkodni, nem a külső projektálhatóságon múlik, hogy vannak-e intellektuális élményei.

Felvethetnénk, hogy a számítógép sem a külső megfigyelők reprezentációi miatt futtat programokat. Ha bekapcsolunk egy számítógépet és magára hagyjuk, anélkül fut rajta az operációs rendszer, hogy a gépet bárki reprezentálná. De ez az ellenérv figyelmen kívül hagyja, hogy az implementációs idealizmus milyen szerepet tulajdonít a külső szemlélőknek. Az implementációs idealizmus szerint egy fizikai rendszer akkor végez számítást, ha stabilan projektál*ható* rá egy program. A magára hagyott számítógépre stabilan projektálható az operációs rendszer (mint absztrakt számítási modell). Tehát az, hogy a gép megfigyelők nélkül is számításokat végez, nem mond ellent az idealista tézisnek, miszerint a projektálhatóságban merül ki az a tény, hogy számításokat végez. Esetünkben viszont nem a projektálhatóságban merül ki, hogy gondolkodunk.

Négy elmefilozófiai álláspontról tudok, melyek alapján (4) kikezdhető: Wittgenstein privátnyelv-érve, a dennetti „intencionális hozzáállás", az illuzionizmus (Frankish 2016) és az eliminatív materializmus (Churchland 1981).

Wittgenstein privátnyelv-érve – ami szigorúan véve nem egy érv, hanem talányos aforizmák szövevénye – egyes értelmezések szerint azt igazolja, hogy nincsenek abszolút értelemben privát mentális tartalmaink, melyekhez kizárólag mi férünk hozzá és melyeket csak mi tudunk leírni. Ezt esetleg felhasználhatnánk (4) cáfolatára, arra hivatkozva, hogy a mentális állapotainknak publikusnak (ennyiben tehát megfigyelhetőnek, projektálhatónak) kell lenniük. Sajnos senki sem tudja, hogy Wittgenstein érvének pontosan mik a premisszái és mi a logikai szerkezete; az érvnek ezek a mozzanatai privátak maradtak. De ha félre is tesszük ezt a problémát és kizárólag a konklúzióra koncentrálunk, akkor sem tűnik nehéznek megvédeni (4)-et, mert a privátnyelv-érv (ha létezik) nem azt mutatja ki, hogy csak az gondolkodik, akire valamilyen absztrakt számítási modell projektálható, hanem azt, hogy a mentális tartalmaink lényegéhez tartozik a megoszthatóság. Tehát az idealista demarkáció a privátnyelv-érv fényében is sikeres lehet, mert továbbra is lényegi eltérés lesz a gondolatok és a számítások között.

A 4. állítást hatékonyabban lehet támadni a Daniel Dennett-féle „intencionális hozzáállás" (*intentional stance*) alapján. Dennett (1987) úgy véli, azért tulajdonítunk mentális tartalmakat (hiteket, gondolatokat, érzéseket) más embereknek, mert ennek segítségével meg tudjuk jósolni a viselkedésüket. Dennett szerint a mentális tartalmak léte *ki is merül* ebben az interpretációs tevékenységben, vagyis az, hogy valaki gondolkodik, pusztán annyit jelent, hogy

gyakorlati szempontból hasznos gondolatokat tulajdonítani neki. Ebből az elméletből valóban következik, hogy a gondolkodás egyfajta projektálhatóságban merül ki, tehát (4) az intencionális hozzáállás alapján hamis. Dennett nemcsak az idealista demarkációt blokkolja, hanem segít egy *anti*demarkációs érv kidolgozásában is, hiszen a gondolkodást teljesen hasonló (kívülről projektált) jelenségnek tekinti, mint az idealista a számítást.

Dennett elméletével legalább három probléma van. A legfőbb probléma az, hogy mindennapos, alapvető és univerzális tapasztalatunk, hogy gondolkodunk (a belső intellektuális tapasztalat értelmében), és puszta introspekcióval meg tudjuk állapítani, hogy ez nem a külső szemlélők által pragmatikus okokból ránk projektált tulajdonság. Tehát az intencionális hozzáállás elmélete eleve nehezen hihető.

Mint arra Howard Robinson rámutatott, Dennett elméleteivel kapcsolatban általános probléma, hogy igazolásuk láthatólag a „Jerikó-módszerre" épül: Dennett „azt hiszi, hogy ha elég sokáig masírozik egy filozófiai probléma körül relevánsnak tűnő tudományos eredményeket harsogva, akkor a probléma spontán megszűnik" (Robinson 1993: 174). Dennett a tudományosan megalapozott fizikalizmust szeretné összeegyeztetni a mentális tartalmakkal, erre pedig kézenfekvő megoldás, ha a mentális tartalmakat az intelligens organizmusok által végzett predikciós tevékenységre redukálja, hiszen ez utóbbihoz elég a nyers fizika: predikcióra bizonyos értelemben egy barométer is képes. De még ha fizikalista alapon kifogástalan is ez a megoldás, egy ilyen elmélet *puszta léte* nem érv amellett, hogy az elmélet *igaz*, vagy csak akkor érv mellette, ha a Jerikó-módszert legitimnek fogadjuk el.

A harmadik probléma az, hogy az intencionális hozzáállás ideológiája szerint a mentális tartalmak léte kimerül a mentális tartalmak (mentális tartalmaktól mentes) tulajdonításában, ehhez viszont a mentális tartalmak tulajdonításának olyan elemzésére lenne szükség, mely kizárólag az emberek fizikai viselkedésére és/vagy agyi állapotaira hivatkozik. Mivel az agyi állapotoknak sem tulajdoníthatunk *sui generis* tartalmakat, legfeljebb a külvilággal kapcsolatos oksági összefüggéseket, az elmélet végeredményben oda fut ki, hogy a mentális tartalmak tulajdonítása bizonyos (társas) viselkedési diszpozíciókra redukálható. De *a priori* hamisnak tűnik a felvetés, hogy mások gondolkodó lényként való interpretálása viselkedési hajlamokban merül ki. Ahogy Robinson írja:

> [Dennett] nem magyarázza meg, hogyan transzmutálódnak a fizikai viselkedésmódok interpretációs aktusokká. Persze ha egy viselkedés kellően komplex, akkor *interpretálhatjuk* szemantikai vagy intencionális jelenségként, de ezzel csak visszajutunk a vita kezdetére, ahhoz a kérdéshez, hogy hogyan lehet egy bizonyos fizikai mozgás *maga* az, hogy interpretálunk valamit. (Robinson 2016: 216)

Ha Dennett elmélete netán igaz lenne, a fenti demarkációs érv egy változata akkor is védhető maradna, csak nem a belső életünkkel, hanem a predikcióra

használt fogalmainkkal kapcsolatban. Felvethető ugyanis, hogy ha (4) dennetti okok miatt hamis, a gondolkodás akkor is *másképp* függ a külső interpretációktól, mint ahogy az MI processzoraiban zajló ⌜gondolkodás⌝ függ a mi projekcióinktól. Az MI-re egy absztrakt számítási modellt projektálunk, amikor felfogjuk a működését, más emberekre viszont nem egy absztrakt számítási modellt projektálunk, amikor gondolkodó lényekként interpretáljuk őket. Tehát az emberek gondolatai és az MI ⌜gondolatai⌝ között még dennetti alapon is lehet különbség, ha (1) és (2) igaz.

Két opció maradt (4) tagadására, az illuzionizmus és az eliminatív materializmus. Az illuzionisták szerint mentális életünk, beleértve a gondolkodást, puszta illúzió: azt hisszük, hogy vannak privát érzéseink, gondolataink, szándékaink stb., de ez puszta látszat. Az eliminatív materialisták szerint pedig még csak nem is látszat, hanem egy ódivatú, téves ideológia. A fenomenális és intencionális tartalmak a „népi pszichológia" mitikus képződményei, melyekkel a modern természettudomány ugyanúgy leszámolt, mint a boszorkányhittel.

> Az emberi értelmet [...] szinte közhely olyan állapotok táncának tekinteni, melyek mondatokat vagy propozíciókat tartalmaznak [...]. Ezek a feltevések alapvető részei az emberi kognitív tevékenységgel kapcsolatos standard koncepciónak, amit „népi pszichológiának" szoktak nevezni abból a meggondolásból, hogy a nem-szakemberek körében is általánosan elterjedt. Bármennyire is általánosak ezek az alapfeltevések, valószínűleg tévesek. Ma már világos, hogy az emberekben (illetve általában az állatokban) a kogníció alapja az aktivációs vektor. Ma már egyértelműnek tűnik, hogy a kognitív számítások alapja a vektortranszformáció. Az pedig tagadhatatlan, hogy az emlékezet alapja a szinaptikus súlyok konfigurációja. Ezek a dolgok nem állnak semmilyen lényegi kapcsolatban mondatokkal vagy propozíciókkal vagy köztük levő logikai viszonyokkal. A kogníció hagyományos nyelvközpontú modelljével ma az agyközpontú modell áll szemben, mely nem tulajdonít semmilyen szerepet a nyelvnek. (Churchland 1995: 322–323)

Az illuzionista szerint tehát a gondolkodás illúzió, az eliminativista szerint pedig egy bukott elmélet téves posztulátuma.

A filozófusok többsége számára (köztük számomra is) hihetetlennek tűnnek ezek a tézisek. Az eliminativizmussal kapcsolatos fő probléma (meglátásom szerint) az, hogy senkinek sem sikerült még leváltania az érzésekkel, tapasztalatokkal, hitekkel és célokkal kapcsolatos állításokat és elméleteket az idegsejtek biokémiai folyamataival és ezek algoritmikus struktúrájával kapcsolatos állításokra és elméletekre. A pszichológia és kognitív tudomány jelentős része az eliminativisták által ódivatú mítosznak minősített állapotok vizsgálatára épül és a „népi pszichológia" hagyományos fogalmait használja vagy ezekre épít precízebb tudományos fogalmakat. A társadalomtudományok és bölcsészettu-

dományok számára axiómának számít, hogy az embereknek vannak érzései, gondolatai és céljai; e feltevés nélkül aligha lehet vizsgálni a piac működését, a társadalmi rétegződést, a történelmet vagy a művészetet. Emellett az összes tudomány számára, köztük a természettudományok számára is, módszertanilag elengedhetetlen feltevés, hogy vannak gondolataink, melyek segítségével tényeket és elméleteket tudunk elemezni. Tudományos viták során a vélekedéseinket fejezzük ki és próbáljuk igazolni, nem pedig a neuronjaink vektoralgebrai számításaira hivatkozunk.

Az illuzionizmus számomra szintén hamisnak tűnik. Az a tény, hogy az illuzionizmus számomra hamisnak tűnik, mintha már önmagában cáfolná az illuzionizmust, hiszen ha valami valamilyennek tűnik számomra, akkor a tudatom létezik és nem puszta káprázat. Az „illúzió" szó standard jelentésének része, hogy valaki számára tűnik valami (tévesen) valamilyennek, tehát ha az értelem a szó standard jelentése szerint illúzió lenne, akkor mégiscsak valóságos lenne, mert léteznének kognitív állapotok. Az illuzionistának emiatt nonstandard jelentésben kell használnia az „illúzió" szót, mikor arról beszél, hogy a gondolkodás illúzió. De mi lenne ez az alternatív jelentés? Ha az illuzionista egyszerűen annyit állít, hogy tévesen tartjuk gondolkodó lényeknek magunkat, akkor álláspontja az eliminativizmusra redukálódik.

Az eliminativistákat és illuzionistákat aligha győzik meg ezek a kifogások. De jelen kontextusban nem is feladatom meggyőzni őket, mert ha az eliminativizmus vagy az illuzionizmus igaz, akkor a gondolkodás – mint sajátos belső folyamat, intellektuális élmény – valójában nem létezik, emiatt pedig a demarkációs kérdés érvényét veszti. Triviálisan igaz lesz, hogy az MI nem gondolkodik, hiszen mi sem gondolkodunk; tévedünk vagy puszta illúzió áldozatai vagyunk, mikor gondolkodó lénynek tekintjük magunkat.

Érdekes módon azonban a demarkációs vita akkor sem teljesen tét nélküli, ha az eliminativizmus vagy illuzionizmus igaz. Ez esetben a vita a népi pszichológiával kapcsolatos vizsgálódás, mely arra a kérdésre keresi a választ, hogy *ha* a népi pszichológia igaz lenne, akkor az MI gondolkodna-e. Az idealista demarkációs érv ebben a szűk – tisztán fogalmi vagy kontrafaktuális – kontextusban továbbra is védhető marad.

Konklúzió

A korábbi fejezetek a demarkáció alábbi formáit vizsgálták:

(I) testi

(II) egzisztenciális

(III) referenciális

(IV) szemantikai

(V) fenomenális

(VI) introspekciós

(VII) nemteljességi

(VIII) eldönthetetlenségi

(IX) önreprezentációs

(X) implementációs demarkáció.

Az első kilenc érv az *elmeoldali* demarkáció területére tartozik, vagyis az emberi elme egy lényegi tulajdonságáról próbálja kimutatni, hogy az MI-ből hiányzik. A X. érv a *számításoldali* demarkáció egy formája, mely a számítógépek egyik lényegi tulajdonságáról állítja, hogy az emberekből hiányzik. Amellett érveltem, hogy hat érv – I, II, III, IV, VI, VII – sikertelen vagy erősen korlátozott érvényű. A maradék négy stratégia (V, VIII, IX, X) számomra sikeresnek tűnik, már amennyire egy filozófiai érv sikeres lehet. Racionálisan megalapozottá teszik a demarkációs tézist, de nyilván nem teszik racionálisan *megkérdőjelezhetetlenné*; igazoló erejük nem kiugró, de nem is elhanyagolható. A könyv lezárásaként szeretném áttekinteni, hogyan illeszkednek ezek az érvek az MI-vel kapcsolatos tágabb filozófiai kérdésekhez.

Kudarctipológia

Héder Mihály (2020) az MI filozófiai kritikájának a demarkáció szempontjából releváns formái között két fő típust különböztet meg: a *fenomenológiai kudarc*ra és a *performatív kudarc*ra alapuló kritikát. A performatív kudarc azt jelenti, ha az MI nem tud elvégezni egy olyan feladatot, amit az ember el tud, vagyis konkrétan a teljesítőképesség terén jelentkezik (empirikusan ellenőrizhető) különbség. „Fenomenológia kudarc"-on Héder nem pusztán a fenomenológiai alapú demarkációt érti, hanem általában azt a tézist, hogy az MI puszta imitáció – kívülről úgy tűnik, hogy az M típusú mentális állapotban van, de M igazából nincs benne jelen.

> A *puszta imitáció* érvben nem a sikeres működés a kérdés, hanem az, hogy az ágens „belülről", tehát saját maga szempontjából (ha egyáltalán ilyen szempont felvethető) hasonló élményt él-e át, mint egy intelligensen viselkedő ember? (Héder 2020: 18)

Héder szerint a performatív kritika mára talajvesztetté vált, mert az MI kognitív teljesítményével kapcsolatos kifogásokat csak a „nem igazi skót" (*no true Scotsman*) típusú manőverrel lehet fenntartani:

> Ezt az érvelési hibát Anthony Flew (1975) nevezte el így. A következőről van szó:
>
> „– Egy skót sohasem tenne ilyet!
>
> – XY, aki történetesen skót, épp ezt tette.
>
> – Egy *igazi* skót sosem tenne ilyet!"
>
> Az MI esetében az érvelési hibát ebben a formában látjuk viszont: „Az X feladatra, amely összetett intelligenciát/kreativitást/-bölcsességet igényel, a gépek soha nem lesznek képesek", jelenti ki valaki. Néhány hónappal, évvel később az MI képessé válik megoldani a feladatot úgy, hogy az minden konvencionálisan elvárt paraméternek megfelel. „Az *igazi* X-re azonban most sem képesek a gépek". Egy példa kedvéért az X helyére behelyettesíthetjük a rajzolást vagy a zenék komponálását; ekkor megtudjuk a fenti érvelőtől, hogy míg a gép csak korábban látott zenékből összegyúr egy statisztikailag megfelelő új változatot, addig az *igazi* zeneszerző valami többet csinál, például kifejezi a lelkét, az inspiráltság állapotában alkot stb. Ez a stratégia a végtelenségig alkalmazható, hiszen eljuthatunk oda, hogy az *igazi* egyenlő lesz az *emberi*vel, az *emberi*nek viszont mindig lehet olyan definíciója, ami az MI-t kizárja. (Héder 2024: 13)

Az általam sikeresnek ítélt négy demarkációs érv közül kettőből, a fenomenális és az implementációs demarkációból nyilvánvalóan következik a fenomenológiai kudarc. A fenomenális demarkáció szerint az MI-nek nincsenek kváléi (tehát „belül" nincsenek olyan érzéki tapasztalatai és ezekből adódó gondolatai, mint nekünk), az implementációs demarkáció szerint pedig az MI által végzett ⌜gondolkodás⌝, ami a számítás egy formája, a stabil projektálhatóságból ered, míg az emberi gondolkodás (mint megélt folyamat) nem a projektálhatóságból ered. Tehát az V. és a X. demarkációs stratégiai alapján az MI-ből egyértelműen hiányzik valami belső tapasztalat, ami bennünk megvan, ez pedig definíció szerint fenomenológiai kudarc. A VIII. és IX. érv esetében kevésbé nyilvánvaló a fenomenológiai kudarc, de (mint azt kifejtésükkor jeleztem) számomra itt is megalapozottnak tűnik ez a következtetés.

Érdekesebb kérdés a performatív kudarc. Héder szerint ennek kimutatására a technológia haladása miatt a filozófia nem képes többé, vagy csak csaló

retorikai manőverekkel képes. Véleményem szerint azonban van két olyan demarkációs érv (IV és VIII), melyek performatív kudarcot is jeleznek, vagy legalábbis koherens alapot adnak a performatív kudarc feltételezésére.

Performatív kudarc

Az MI kognitív képességeivel kapcsolatban erősen megoszlanak a vélemények. A ChatGPT-t fejlesztő cég vezetője, Sam Altman úgy véli, a jelenlegi MI lényegében ugyanannyit tud, mint egy doktori fokozattal rendelkező szakértő.[1] Mások (például Terence Tao matematikus) szerint a jelenlegi csúcsmodellek, bár az első-másodéves hallgatóktól elvárt feladatokat valóban képesek megoldani, egy kutató számára csak segédeszköznek jók, mert eléggé megbízhatatlanok.[2] Ez a vita a filozófián kívüli demarkációs érvek zavaros területére vezet, amit a korábban jelzett okokból (1.3. alfejezet) szeretnék elkerülni. Tehát nem fogok amellett érvelni, hogy a performatív kudarcra van empirikus bizonyíték. Ugyanakkor azt sem ismerem el, hogy a performatív kudarcra *nincs* empirikus bizonyíték. Azt feltételezem – de nem próbálom igazolni, csak a Taóhoz hasonló kutatókra mutogatok –, hogy ez a kérdés jelenleg eldöntetlen, mert a szakértők nem értenek egyet abban, mi minősül kudarcnak, sem abban, hogy jelenleg tapasztalható-e kudarc.

Ha jelenleg empirikus alapon nem igazolható és nem is cáfolható a performatív kudarc, akkor joggal merül fel a kérdés, hogy a performatív kudarc mint elvi lehetőség következik-e bármilyen *filozófiai* demarkációs érvből. Szerintem a fentiek közül két érvből is következik ez a lehetőség. Nem azt állítom, hogy ha a két érv valamelyike helyes, akkor a performatív kudarc *garantált*, hanem azt, hogy a két érv megmagyarázza, miből ered*het* performatív kudarc.

A fenomenális demarkációból (7.3. alfejezet) akkor eredhet performatív kudarc, ha a fenomenális tartalmaink annyira finoman szemcsézettek („analóg" jellegűek), hogy a fenomenális fogalmakat igénylő gyakorlati feladatok esetében nem lehet emberi szintű teljesítményt elérni a kválék közvetlen ismerete nélkül. Az illusztráció kedvéért tegyük fel, hogy a térbeli tájékozódás és a tárgyfelismerés kválékon alapul, és az ezzel kapcsolatos ítéleteink fenomenális fogalmakat igényelnek (olyan fogalmakat, melyek tartalmát részben kválék alkotják). Ha a szóban forgó kválék természete és viszonyrendszere annyira komplex, hogy számítások implementálásával nem megragadható, annak performatív kudarc lesz a vége; az MI nem fogja tudni tökéletesen imitálni a térben mozgó és tárgyakat használó emberek viselkedését. Lehet, hogy erre a hipotézisre ugyanúgy rácáfolnak majd a mérnökök, mint Dreyfus jóslataira, de az empirikus cáfolat

[1] „OpenAI GPT-5 is here: Company unveils 'PhD-level experts in your pocket'". *Fox Business*, 2025. aug. 7.

[2] „Can AI help us solve the hardest problems in mathematics? (ft. Terry Tao)" *YouTube*, 2025. dec. 30.

elvi lehetősége nem veszély, hanem annak a jele, hogy a filozófia is képes falszifikálható hipotéziseket megfogalmazni. Ha az MI filozófia kritikája empirikusan cáfolható, annál jobb: a mérnöki csodák segíthetnek *a priori* érvek hibáinak feltárásban. Jelen esetben az empirikus cáfolatból arra következtethetnénk, hogy a kválék nem túl finoman szemcsézettek vagy az MI olyan funkcionális komplexitást ért el, hogy a pszichofizikai törvények miatt már kváléi is vannak. Természetesen nem állítom, hogy a fenomenális demarkációból akár csak abduktív alapon levezettem a performatív kudarcot. Pusztán annyit akartam jelezni, hogy a kválék esetleges „analóg" természete megnyitja a performatív kudarc elvi lehetőségét.

Hasonló a helyzet a Rice-demarkációval (10. fejezet). Ennek egyik lehetséges értelmezése az, hogy az MI nem képes funkciókat azonosítani, nem tudja megállapítani, egy tárgy mire való. Magából a Rice-tételből ez logikailag nyilván nem következik, és az MI-optimista joggal spekulálhat arról, hogy minden praktikusan lényeges funkciófelismerő képességünk algoritmizálható. Nem állítom, hogy a Rice-demarkáció *kizárja* az algoritmizálhatóságot. Akárcsak az előző esetben, itt is csak annyit állítok, hogy *ha* az MI képtelennek bizonyul az emberihez hasonló célirányos gondolkodásmódra, funkciófelismerésre, akkor az ebből adódó performatív kudarc pontosan az a fajta hiányosság lesz, amire a Rice-demarkáció alapján gyanakodhatunk. Akárcsak a kválék esetében, a demarkációs érv itt sem a performatív kudarc igazolása, hanem koherens magyarázat a performatív kudarc egy lehetséges formájára.

A fenti két probléma – a kválékra épülő motorikus és perceptuális képességek problémája, illetve a teleologikus ítéletek problémája – elsősorban az MI-fejlesztők jelenlegi Szent Grálja, a Mesterséges Általános Intelligencia (AGI) kapcsán releváns. Az AGI olyan MI, ami képes minden olyan gyakorlati és intellektuális feladat megoldására, amire általában egy ember képes. Aligha lehet kizárni, hogy specifikus, leszűkített feladatok megoldására, kellően sok adat birtokában, be lehet tanítani már a mai MI-k robottesttel felszerelt változatait is. Egy LLM-et vélhetőleg be lehet paraméterezni olyan módon, hogy képes legyen tányérokat pakolni a mosogatógépbe. De ma még egyáltalán nem világos, meg lehet-e ezt tanítani az MI-nek olyan formában, hogy ne csak *bizonyos* típusú tányérokat legyen képes *bizonyos* típusú mosogatógépekbe bepakolni, hanem ezt a feladatot – vagy ami még jobb: általában az edények és evőeszközök megtisztítását – pontosan olyan általános értelemben tudja elvégezni, mint egy ember. (Steve Wozniak példáját használva: a cél az lenne, hogy ha az MI-t beküldik egy házba, ahol sosem járt, akkor magától találja meg a konyhát és főzzön egy kávét („Wozniak-teszt").) Az AGI-nak ilyen gyakorlati képességekre lenne szüksége, és pont az ilyen képességek algoritmizálhatatlanságát magyarázhatja a fenomenális demarkáció és a Rice-demarkáció. Kétségtelen, hogy jelenleg nincsenek olyan (védhető) filozófiai érvek, melyekből *következne* a performatív kudarc az ilyen területeken, de a két érv legalábbis a gyanú árnyékát veti az MI fejlődésével kapcsolatos vérmes reményekre.

Konstitutív kudarc

Nem tűnik alaptalannak az a felvetés, hogy a fenomenológiai és performatív kudarc mellett létezik egy harmadik fajta defekt is, a *konstitutív* kudarc. A konklúzió konklúziójaként ennek fogalmát próbálom kidolgozni a X. érv, az implementációs demarkáció segítségével.

Az MI kritikusai által azonosítani vélt defekteknek tipikusan az a lényege, hogy az ember rendelkezik egy intrinzikus F tulajdonsággal, amivel az MI nem rendelkezik. Vagyis ezek a defektek az elmeoldali demarkáció változatai. A Héder által azonosított fenomenális és performatív kudarc abban tér el egymástól, hogy előbbi esetben egy kategoriális tulajdonságról (kválé, belső tapasztalat stb.), míg utóbbi esetben egy diszpozicionális (funkcionális) tulajdonságról van szó (nem tud mosogatni stb.). De ha a számításoldali demarkációt, konkrétan az implementációs idealizmust is figyelembe vesszük, akkor adódik egy további lehetőség, mégpedig az, hogy a demarkáció alapját jelentő F tulajdonság extrinzikus, nem magában a hordozójában gyökerezik, hanem a hordozójától különböző létezők állapotaitól függ. Épp ez a helyzet az implementációs idealizmus szerint a számításokkal.[3] A számítás, ezáltal pedig a gépi ⌜gondolkodás⌝, objektív$_1$ jelenség, de emiatt nem magában a gépekben gyökerezik, hanem a projektálhatóságból ered. Az emberi gondolkodás viszont nem a projektálhatóságból ered (ld. 20. fejezet). Ez azt jelenti, hogy az ember mint gondolkodó lény ontológia státusza eltér a gépek mint számítást végző rendszerek ontológiai státuszától; mi intrinzikusan rendelkezünk kognitív állapotokkal, a gépek viszont csak extrinzikusan. Ezt nevezem *konstitutív kudarc*nak. A számítógép létmódja annyira eltér az emberétől, hogy eleve nem vonatkozhatnak rá azok a kategóriák, melyek az emberre mint cselekvő, érző, morálisan felelős lényre vonatkoznak.

Ez az állítás kicsit ezoterikusnak tűnhet, de egészen konkrétan illusztrálható. Tekintsük az MI-kreativitás kérdését. Egyetértek Héderrel abban, hogy a performatív kudarcra egyes kreatív iparágakban egyre inkább csak „Nem igazi skót!" típusú retorikai manőverrel hivatkozhatunk. Az MI már ma képes kreatív festményeket és látványterveket létrehozni (2022-ben egy MI megnyert egy képzőművészeti pályázatot[4]), és az MI-vel generált zenék is egyre népszerűbbek. Levonhatjuk tehát a következtetést, hogy az MI mint festő és zeneszerző utolérte (funkcionális értelemben) az embert. De ezt a következtetést a számításoldali demarkáció alapján könnyű aláásni, láthatóvá téve a konstitutív kudarcot. Gondoljuk át, pontosan *ki* is kreatív, amikor az MI létrehoz egy új festményt vagy zeneművet.

[3] Az extrinzikusság definícióját ld. a 19. oldalon. Az implementációs idealizmus szerint a programok futtatása azért extrinzikus jelenség, mert x attól futtatja a P programot, hogy P stabilan projektálható (x-től különböző megfigyelők által) x-re.

[4] *Théâtre D'opéra Spatial*. Wikipedia.

Tegyük fel, hogy Ancsa beírja egy MI appba a P promptot, és ennek hatására megjelenik egy elképesztően kreatív digitális festmény, D. Kinek a kreatív teljesítménye D? Felvethető, hogy D létrejötte (a technológiai infrastruktúrát – áramellátást, szerverparkot stb. – leszámítva, amit pusztán oksági előfeltételnek tekinthetünk) az alábbi tényezőkre vezethető vissza:

- az MI betanításához (az LLM paramétereinek kialakításához) használt fényképek és festmények készítőinek kreativitása
- a betanításhoz használt képeket szemantikailag felcímkéző segédmunkások kemény szellemi munkája
- az MI architektúráját kialakító és a betanítást végző informatikusok mérnöki kreativitása
- Ancsa kreativitása (P megfogalmazása).

D esztétikai értékei tehát egy egész hadseregnyi ember kreatív tevékenységéből és intellektuális erőfeszítéseiből erednek. Nem arról van szó, hogy az MI *mint entitás* kreatív lenne, hanem arról, hogy létezik egy elképesztően bonyolult számítógépes architektúra, amiben találkozik a humán képzőművészek, mérnökök és felhasználók kreativitása. E tényezők összjátéka teljesen váratlan és újszerű eredményeket hoz, de ez mit sem változtat azon, hogy továbbra is az emberi kreativitás egy – radikálisan hálózatosodott, kollektív, történetileg teljesen újszerű – formájáról van szó. Az MI nem kreál semmit, mert mint ágens eleve csak projekció.

A fenti négy – fenomenális, Rice-alapú, önreflexiós és implementációs – demarkáció együttes üzenete az, hogy az ember érző, teleologikus, önreflexióra képes ágens, aki intrinzikus módon rendelkezik ezekkel a tulajdonságokkal, az MI viszont nemcsak nem egy érző, teleologikus vagy önreflexióra képes rendszer, hanem eleve csak a mi reprezentációs tevékenységünk miatt rendelkezik (extrinzikus módon) kognitív tulajdonságokkal. Az MI olyan kísérteties jelenés, melyben saját képmásunkat véljük felfedezni, de közelebbről megnézve emberek ezrei által generált illúzióvá porlad szét.

Matematikai és logikai szimbólumok

Jelölés	Jelentés
$\wedge$	és
$\vee$	vagy
$\neg$	nem
$\equiv$	logikailag ekvivalens
$\rightarrow$	logikai „ha...akkor” (a materiális kondicionális értelmében: $A \rightarrow B \equiv \neg A \vee B$)
$\exists$	létezik
$\forall$	minden
$F(x)$	x rendelkezik az F tulajdonsággal
$\forall x : F(x)$	minden rendelkezik az F tulajdonsággal
$\Box P$	szükségszerű, hogy P
$\Diamond P$	lehetséges, hogy P ($\equiv \neg\Box\neg P$)
$\in$	halmaz eleme
$\mathbb{N}$	a természetes számok halmaza $(0, 1, 2 \ldots)$
$\{x \in A : F(x)\}$	az a halmaz, melynek elemei az A halmaznak is elemei és rendelkeznek az F tulajdonsággal
$\langle x_1, x_2 \ldots x_n \rangle$	az $x_1 \ldots x_n$ számokból (vagy egyéb matematikai objektumokból) álló sorozat
$A \times B$	az A és B halmazok elemeiből álló, $\langle a, b \rangle$ alakú rendezett párok halmaza
$A \cup B$	az A és B halmazok uniója (az a halmaz, melynek A és B összes eleme az eleme, és minden eleme A vagy B eleme)
$A \cap B$	az A és B halmaz metszete (az a halmaz, melynek minden eleme A-nak és B-nek is eleme)
$\emptyset$	az üreshalmaz (az a halmaz, melynek nincs eleme)
$f : A \rightarrow B$	az f függvény az A halmazból a B halmazba képez (minden $a \in A$ elemhez hozzárendel pontosan egy $f(a) \in B$ elemet)
$\mathscr{P}(A)$	az A halmaz részhalmazainak halmaza („hatványhalmaz”)
$A \subseteq B$	az A halmaz része a B halmaznak (A minden eleme B-nek is eleme)

Irodalomjegyzék

Aaronson, S. (2011) Why philosophers should care about computational complexity. arxiv.org

Anderson, P. W. (1974) More Is Different. *Science* 177: 4047, 393–396.

Avramides, A. (2023) Other Minds. *The Stanford Encyclopedia of Philosophy*, 2023. téli kiadás.

Barsalou, L. W. (1999) Perceptual symbol systems. *Behavioral and Brain Sciences* 22, 577–660.

Beall, J.C., M. Glanzberg & E. Ripley (2025) Liar Paradox. *The Stanford Encyclopedia of Philosophy* (2025. őszi kiadás).

Bechtel, W. & R. C. Richardson (1993) *Discovering Complexity: Decomposition and Localization as Scientific Research Strategies*. Princeton University Press.

Bell, J. L. (2023) Infinitary Logic. *The Stanford Encyclopedia of Philosophy* (2023. őszi kiadás).

Benacerraf, P. (1965) What Numbers Could Not Be. *The Philosophical Review* 74: 1, 47–73.

Benacerraf, P. (1967) God, the Devil, and Gödel. *Monist* 51, 9–32.

Benacerraf, P. (1973) Mathematical Truth. *The Journal of Philosophy* 70: 19, 661–679.

Block, N. (1978) Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9, 261–325.

Block, N. (1981) Psychologism and Behaviorism. *The Philosophical Review* 90:1, 5–43.

Block, N. (2002) Searle's Arguments Against Cognitive Science. In J. Preston & M. Bishop (szerk.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, 70–79.

Boolos, G. (1993) *The Logic of Provability*. Cambridge University Press.

Boolos, S., J. P. Burgess & R. C. Jeffrey (2007) *Computability and Logic* (5. kiadás). Cambridge University Press.

Borghi, A. M., & F. Binkofski (2014) *Words as Social Tools: An Embodied View on Abstract Concepts*. Springer

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

Bringsjord, S. (1992) *What Robots Can and Can't Be*. Springer.

Bringsjord, S. & K. Arkoudas (2004) The modal argument for hypercomputing minds. *Theoretical Computer Science* 317, 167–190.

Brooks, R. A. (1991) Intelligence without Representation. *Artificial Intelligence* 47: 139–159.

Butlin, P. et al. (2023) Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arxiv.org

Button, T. & S. Walsh (2018) *Philosophy and Model Theory*. Oxford University Press.

Cantwell Smith, B. (2019) *The Promise of Artificial Intelligence: Reckoning and Judgment*. MIT Press.

Cappelen, H. & J. Dever (2021) *Making AI Intelligible: Philosophical Foundations*. Oxford University Press.

Chalmers, D. (1995) On implementing a computation. *Minds and Machines* 4, 391–402.

Chalmers, D. (1996a) Does a rock implement every finite-state automaton? *Synthese* 108, 309–333.

Chalmers, D. (1996b) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. (1996c). Minds, Machines, and Mathematics. *Psyche* 2, 11–20.

Chalmers, D. (2023) Could a Large Language Model be Conscious? *Boston Review*, 2023. augusztus 9.

Chang, C.C. & H.J. Keisler (1973) *Model Theory*. North-Holland.

Chollet, F. (2019) On the Measure of Intelligence. arxiv.org

Chollet, F. et al. (2025) ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems. arxiv.org

Chrisley, R. L. (1995) Why everything doesn't realize every computation. *Minds and Machines* 4, 403–430.

Chudnoff, E. (2015) *Cognitive Phenomenology*. Routledge.

Churchland, P. M., (1981) Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* 78, 67–90.

Churchland, P. M. (1984) *Matter and Consciousness*. MIT Press.

Churchland, P. M. (1995) *The Engine of Reason, the Seat of the Soul*. The MIT Press.

Churchland, P. M. & P. Churchland (1990) Could a machine think? *Scientific American* 262:1, 32–37.

Coder, D. (1969) Gödel's Theorem and Mechanism. *Philosophy* 44, 234–237.

Cole, D. (2024) The Chinese Room Argument. *The Stanford Encyclopedia of Philosophy* (2024 téli kiadás).

Copeland, J. (1996) What is computation? *Synthese* 108, 335–359.

Copeland, J. (2002) The Chinese Room from a Logical Point of View. In J. Preston & M. Bishop (szerk.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, 104–122.

Cover, T. & J. Thomas (2006) *Elements of Information Theory*. Wiley.

Crane, T. & C. French (2021) The Problem of Perception. *The Stanford Encyclopedia of Philosophy* (2021. őszi kiadás).

Craver, C. F. (2013) Functions and Mechanisms: A Perspectivalist View. In P. Huneman (szerk.), *Functions: Selection and Mechanisms*, Springer, 133–158.

Davidson, D. (1987) Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association* 60, 441–458.

Davis, Martin (1958) *Computability & Unsolvability*. McGraw-Hill Book Company.

DeLong, H. (2004) *A Profile of Mathematical Logic*. Dover Publications.

Dennett, D. C. (1980) The milk of human intentionality. *Behavioral and Brain Sciences* 3, 428–430.

Dennett, D. C. (1984) *Elbow Room: The Varieties of Free Will Worth Having.* Clarendon Press.

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

Deutsch, D. (1985) Quantum Theory, the Church-Turing Principle and the Universal Quan- tum Computer. *Proceedings of the Royal Society,* Series A, 400, 97–117.

Dewhurst, J. (2018) Computing Mechanisms without Proper Functions. *Minds and Machines* 28, 569–588.

Dijkstra, E. W. (1984) The threats to computing science. *ACM South Central Regional Conference* (előadás).

Dnyeprov, A. (1969) Játék. In D. Bilenkin & A. Dnyeprov, *Hullámverés a Marson / Bíbormúmia*, Kozmosz Könyvek, 173–182.

Dove, G. (2009) Beyond perceptual symbols: A call for representational pluralism. *Cognition* 110, 412–431.

Dove, G. (2011) On the need for embodied and dis-embodied cognition. *Frontiers in Cognition* 1, 1–13.

Dreyfus, H. (1972) *What Computers Can't Do.* Harper & Row.

Dreyfus, H. (1974) Artificial intelligence. *The Annals of the American Academy of Political and Social Science* 412, 21–33.

Dreyfus, H. (1991) *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I.* The MIT Press.

Dreyfus, H. (2014a) Heidegger's Critique of the Husserl/Searle Account of Intentionality. In M. Wrathall (szerk.), *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*, Oxford University Press, 76–91.

Dreyfus, H. (2014b) Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian. In M. Wrathall (szerk.), *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*, Oxford University Press, 249–273.

Dummett, M. (1991) *The Logical Basis of Metaphysics.* Harvard University Press.

Feser, E. (2019) *Aristotle's Revenge: The Metaphysical Foundations of Physical and Biological Science.* editiones scholasticae.

Fine, K. (1994) Essence and modality. *Philosophical Perspectives* 8, 1–16

Fischer, M. H. & R. A. Zwaan (2008). Embodied language: A review of the motor system in language comprehension. *Quarterly Journal of Experimental Psychology* 6, 825–850.

Fitch, G. & M. Nelson (2024) Singular Propositions. *The Stanford Encyclopedia of Philosophy* (2024 tavaszi kiadás).

Flew, A. (1975) *Thinking About Thinking*. HarperCollins.

Fodor, J. (1974) Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese* 28:2, 97–115.

Fodor, J. (1975) *The Language of Thought*. Thomas Y. Crowell.

Fodor, J. (1990) *A Theory of Content and Other Essays*. The MIT Press.

Fodor, J. (2001) *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Bradford Books.

Frankish, J. (2016) Illusionism as a theory of consciousness. *Journal of Consciousness Studies* 23, 11–39.

Gans, J.S. (2026) A Model of Artificial Jagged Intelligence. arxiv.org.

Garson, J. (2016) *A Critical Overview of Biological Functions*. Synthese.

Gendler, T. S. & J. Hawthorne (2002) *Conceivability and Possibility*. Oxford University Press.

Godfrey-Smith, P. (2009) Triviality arguments against functionalism. *Philosophical Studies* 145, 273–295.

Gödel, K. (1951/1995) Some basic theorems on the foundations of mathematics and their implications. In uő, *Collected Works III*, Oxford University Press, 304–323.

Grice, P. (1989) *Studies in the Ways of Words*. Harvard University Press.

Gunther, K. (2012) *The Physics of Music and Color*. Springer.

Hamkins, J.D. (2020) *Lectures on the Philosophy of Mathematics*. The MIT Press.

Harnad, S. (1990) The symbol grounding problem. *Physica* D 42, 335–346.

Haugeland, J. (1985) *Artificial Intelligence: The Very Idea*. MIT Press.

Haugeland, J. (1998) Understanding natural language. In uő, *Having Thought: Essays in the Metaphysics of Mind*, Harvard University Press, 47–61.

Haugeland, J. (2002a) Authentic intentionality. In M. Scheutz (szerk.), *Computationalism: New Directions*, MIT Press, 159–174.

Haugeland, J. (2002b) Syntax, semantics, physics. In Preston, J. & M. Bishop (szerk.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, 379–392.

Hawley, K. & F. Macpherson (2011) *The Admissible Contents of Experience*. Wiley-Blackwell.

Heavey, C. K. & R.T. Hurlburt (2008) The phenomena of inner experience. *Consciousness and Cognition* 17:3, 798–810.

Hegel, G.W.F. (1979) *A logika tudománya I.* Akadémiai Kiadó.

Heidegger, M. (2010) *Logic: The Question of Truth*. Indiana University Press.

Héder, M. (2020) *Mesterséges intelligencia: Filozófiai kérdések, gyakorlati válaszok*. Gondolat Kiadó.

Héder, M. (2024) *Az MI etika új hullámai*. MTA Doktori értekezés.

Hobbes, T. (1651/1998) *Leviathan*. Oxford University Press.

Hopcroft, J. E., R. Motwani & J. D. Ullman (2007) *Introduction to Automata Theory, Languages, and Computation* (3. kiadás). Pearson Addison Wesley.

Hurlburt, R. T. & S. A. Akhter (2008) Unsymbolized thinking. *Consciousness and Cognition* 17:4, 1364-1374.

Hurshman, C. (2023) Artifacts and intervention: a persistence theory of artifact functions. *Synthese* 202: 128.

Hutto, D. D. & E. Myin (2013) *Radicalizing Enactivism*. The MIT Press.

Hutton, A. (1976) This Gödel is Killing Me. *Philosophia* 3, 135–144.

Jackson, F. (1982) Epiphenomenal Qualia. *Philosophical Quarterly* 32, 127–136.

James, W. (1961) *Psychology: The Briefer Course*. Harper and Row.

Kahneman, D. (2011) *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kaplan, D. (1989) Demonstratives. In J. Almog, J. Perry, & H. Wettstein (szerk.), *Themes from Kaplan*, Oxford University Press, 481–504.

Kingsbury, J. (2023) New challenges to the selected effects account of biological function. *Synthese* 202: 177.

Kirk, R. (1974) Zombies v. Materialists. *Proceedings of the Aristotelian Society* 48, 135–152.

Kirk, R. (2023) Zombies. *The Stanford Encyclopedia of Philosophy*, 2023. őszi kiadás.

Kleene, S. C. (1952) *Introduction to Metamatemathics*. North-Holland.

Klein, C. (2008) Dispositional implementation solves the superfluous structure problem. *Synthese* 165, 141–153.

Kodaj, D., L. Bernáth & M. Pickup (2025) Teleology for the twenty-first century: editorial introduction. *Synthese* 205:78.

Koellner, P. (2018a) On the question of whether the mind can be mechanized I: From Gödel to Penrose. *The Journal of Philosophy* 115:7, 337–360.

Koellner, P. (2018b) On the question of whether the mind can be mechanized II: Penrose's New Argument. *The Journal of Philosophy* 115:9, 453–484.

Kozlowski, A.C., C. Dai & A. Boutyline (2025) Semantic Structure in Large Language Model Embeddings. arxiv.org

Kriegel, U. (2007) Self-Consciousness. *Internet Encyclopedia of Philosophy*.

Kurzweil, J. (2002) Locked in his Chinese Room. In J.W. Richards (szerk.), *Are We Spiritual Machines: Ray Kurzweil vs. the Critics of Strong AI*, Discovery Institute, 128–171.

Ladyman, J. & D. Ross (2007) *Every Thing Must Go: Metaphysics Naturalized*. Oxford University Press.

Levkowitz, H. (1997) *Color Theory and Modeling for Computer Graphics, Visualisation, and Multimedia Applications*. Kluwer Academic Publishers.

Lewis, D. (1969) Lucas against Mechanism. *Philosophy* 44, 231–233.

Lewis, D. (1983) Extrinsic Properties. *Philosophical Studies* 44, 197–200.

Lewis, D. (1986) *On the Plurality of Worlds*. Blackwell.

Lipshitz, L. & L. A. Rubel (1987) A Differentially Algebraic Replacement Theorem, and Analog Computability. *Proceedings of the American Mathematical Society* 99:2, 367–372.

Lloyd, S. (2006) *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos*. Knopf.

Loewer, B. (2017) A guide to naturalizing semantics. In B. Hale, C. Wright & A. Miller (szerk.), *A Companion to the Philosophy of Language* (2. kiadás), I. kötet, Wiley Blackwell, 174–189.

Lucas, J. R. (1961) Minds, machines and Gödel. *Philosophy* 36, 112–127.

Lucas, J. R. (1970) Mechanism: A Rejoinder. *Philosophy* 45, 149–151.

Lucas, J. R. (1976) This Gödel is killing me: A rejoinder. *Philosophia* 6, 145–148.

Ludlow, P., Y. Nagasawa & D. Stoljar (2004) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. The MIT Press.

Lycan, W. G. (1981) Form, function, and feel. *Journal of Philosophy* 78, 24-50.

Machamer, P. K., L. Darden & C. F. Craver (2000) Thinking About Mechanisms. *Philosophy of Science* 67, 1–25.

Marcus, G. (2018) Deep Learning: A Critical Appraisal. arxiv.org.

Marshall, D. (2015) Intrinsicality and Grounding. *Philosophy and Phenomenological Research* 90, 1–19.

Marshall, D. & B. Weatherson (2023) Intrinsic vs. Extrinsic Properties. *The Stanford Encyclopedia of Philosophy*, 2023. őszi kiadás.

Maudlin, T. (1989) Computation and consciousness. *Journal of Philosophy* 86:8, 407–432.

McCullough, D. (1996) Can Humans Escape Gödel? *Psyche* 2, 57–65.

McDermott, D. (1996) *Penrose is wrong. *Psyche* 2, 66–82.

McDowell, J. (1998) Singular thought and the extent of inner space. In uő, *Meaning, Knowledge, and Reality*, Harvard University Press, 228–259.

Megill, J. (2011) The Lucas-Penrose Argument about Gödel's Theorem. J. Fieser & B. Dowden (szerk.), *Internet Encyclopedia of Philosophy*.

Millière, R. & C. Buckner (2024) A philosophical introduction to lanuage models I. archiv.org (2401.03910v1).

Millikan, R. (1984) *Language, Thought and Other Biological Categories*. MIT Press.

Monk, J. D. (1976) *Mathematical Logic*. Springer.

Neander, K. (2017) *The Mark of the Mental: In Defense of Informational Teleosemantics*. MIT Press.

Noë, A. (2004) *Action in Perception*. The MIT Press.

Noë, A. (2009) *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness*. Hill and Wang.

Opus, C. & A. Lawsen (2025) The Illusion of the Illusion of Thinking. arxiv.org

Papineau, D. (1993) *Philosophical Naturalism*. Blackwell.

Papineau, D. (2022) Swampman, teleosemantics and kind essences. *Synthese* 200: 509.

Patterson, D.A. & J.L. Hennessy (2014) *Computer Organization and Design* (5. kiadás). Morgan Kaufmann.

Penrose, R. (1994) *Shadows of the Mind: A Search for the Missing Theory of Consciousness*. Oxford University Press.

Penrose, R. (1999) *The Emperor's New Mind: Concerning Computers, Minds, and The Laws of Physics*. Oxford University Press.

Piccinini, G. (2015) *Physical Computation*. Oxford University Press.

Pinker, S. (1997) *How the Mind Works*. Norton.

Pitt, D. (2004) The Phenomenology of Cognition: Or What Is It Like to Think That P? *Philosophy and Phenomenological Research* 69:1, 1–36.

Plantinga, A. (2022) Aktualizmus és lehetsége világok. In Kocsis L. & Tuboly Á. (szerk.), *Lehetséges világok: Modális metafizikai tanulmányok*, Typotex, 54-77.

Pour-El, M. B. (1974) Abstract Computability and Its Relation to the General Purpose Analog Computer (Some Connections Between Logic, Differential Equations and Analog Computers). *Transactions of the American Mathematical Society* 199, 1–28.

Priest, G. (2006) *In Contradiction* (2. kiadás). Clarendon Press.

Prince, S. J. D. (2023) *Understanding Deep Learning*. The MIT Press.

Prinz, J. (2011) The Sensory Basis of Cognitive Phenomenology. In T. Bayne & M. Montague (szerk.), *Cognitive Phenomenology*, Oxford University Press, 174–196.

Putnam, H. (1981) *Reason, Truth, and History*. Cambridge University Press.

Putnam, H. (1988) *Representation and Reality*. MIT Press.

Quine, W. V. (1948) On what there is. *Review of Metaphysics* 2:5, 21–38.

Raatikainen, P. (2020) Gödel's incompleteness theorems. *The Stanford Encyclopedia of Philosophy* (2022 tavaszi kiadás).

Ramsey, W., S. Stich & J. Garon (1990) Connectionism, Eliminativism and The Future of Folk Psychology. *Philosophical Perspectives* 4, 499-533.

Raven, M. (2020) Introduction. In uő (szerk.), *The Routledge Handbook of Metaphysical Grounding*, Routledge, 1–14.

Reinhardt, W. (1986) Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems. *Journal of Philosophical Logic* 15:4, 427–474.

Rescorla, M. (2013) Against Structuralist Theories of Computational Implementation. *The British Journal for the Philosophy of Science* 64:4, 681–707.

Rescorla, M. (2024) The Computational Theory of Mind. *The Stanford Encyclopedia of Philosophy* (2024 téli kiadás).

Rey, G. (2002) Searle's misunderstandings of functionalism and strong AI. In Preston, J. & M. Bishop (szerk.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, 201–225.

Rice, H. G. (1953) Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society* 74: 2, 358–366.

Robinson, H. (1982) *Matter and Sense*. Cambridge University Press.

Robinson, H. (1993) Dennett on the Knowledge Argument. Analysis 53:3, 174–177.

Robinson, H. (2016) *From the Knowledge Argument to Mental Substance: Resurrecting the Mind*. Cambridge University Press.

Robinson, H. (2022) Why there cannot be a materialist theory of thought. Kézirat.

Robinson, W. S. (2011) A frugal view of cognitive phenomenology. In T. Bayne & M. Montague (szerk.), *Cognitive Phenomenology*, Oxford University Press, 197–214.

Rogers, H. (1987) *Theory of Recursive Functions and Effective Computatibility*. McGraw-Hill Book Company.

Rosser, B. (1936) Extensions of some theorems of Gödel and Church. *The Journal of Symbolic Logic* 1:3, 87-95.

Rubel, L. A. & M. F. Singer (1985) A Differentially Algebraic Elimination Theorem with Application to Analog Computability in the Calculus of Variations. *Proceedings of the American Mathematical Society* 94:4, 653–658.

Russell, B. (1910) Knowledge by Acquaintance and Knowledge by Description. *Proceedings of the Aristotelian Society* 11, 108–128.

Russell, B. (1912) *The Problems of Philosophy*. Oxford University Press.

Russell, B. (1927) *The Analysis of Matter*. Kegan Paul.

Russell, S. J. & P. Norvig (2022) *Artificial Ingelligence: A Modern Approach*. Pearson.

Sawyer, S. (2007) There is no viable notion of narrow content. In B.P. McLaughlin & J. Cohen (szerk.), *Contemporary Debates in Philosophy of Mind*, Blackwell, 20–34.

Schaffer, J. (2009) On what grounds what. In R. Wasserman, D. Manley & D. Chalmers (szerk.), *Metametaphysics*, Oxford University Press, 347–383.

Scheutz, M. (1999) „When physical systems realize functions..." *Minds and Machines* 9, 161–196.

Scheutz, M. (2001) Causal versus computational complexity. *Minds and Machines* 11, 534–66.

Schulte, P. & K. Neander (2022) Teleological Theories of Mental Content. *The Stanford Encyclopedia of Philosophy* (2022. nyári kiadás).

Schweizer, P. (2019) Computation in Physical Systems: A Normative Mapping Account. In D. Berkich & M. V. d'Alfonso (szerk.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Springer, 27–47.

Schwitzgebel, E. (2015) If materialism is true, the United States is probably conscious. *Philosophical Studies* 172: 7, 1697–1721.

Schwitzgebel, E. (2024) Introspection. *The Stanford Encyclopedia of Philosophy* (2024 őszi kiadás).

Searle, J. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 417–424.

Searle, J. (1989) Artificial Intelligence and the Chinese Room: An Exchange. *New York Review of Books* 36: 2.

Searle, J. (1992) *The Rediscovery of the Mind*. MIT Press.

Searle. J. (1996) Az elme, az agy és a programok világa. In Pléh Cs. (szerk.), *Kognitív tudomány*, Osiris Kiadó, 136–151.

Seth, A. (2025) Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences* (megjelenés alatt).

Shagrir, O. (2022) *The Nature of Physical Computation*. Oxford University Press.

Shank, R. & R. Abelson (1977) *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates.

Shannon, C. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 379–423, 623–656.

Shapiro, L. & S. Spaulding (2025) Embodied Cognition. *The Stanford Encyclopedia of Philosophy*, 2025. nyári kiadás.

Shapiro, S. (1997) *Philosophy of Mathematics: Structure and Ontology*. Oxford University Press.

Shojaee, P. et al. (2025) The Illusion of Thinking. arxiv.org.

Sider, T. (2020) *The Tools of Metaphysics and the Metaphysics of Science*. Clarendon Press.

Siewert, C. (2011) Phenomenal thought. In T. Bayne & M. Montague (szerk.), *Cognitive Phenomenology*, Oxford University Press, 236–267.

Simon, H.A. & A. Newell (1958) Heuristic problem solving: The next advance in operations research. *Operations Research* 6:1, 1–10.

Sipser, M. (2013) *Introduction to the Theory of Computation* (3. kiadás). Cengage Learning.

Sloman, A. & M. Croucher (1980) How to turn an information processor into an understanding. *Brain and Behavioral Sciences* 3: 447–448.

Soteriou, M. (2020) The Disjunctive Theory of Perception. *The Stanford Encyclopedia of Philosophy* (2020 nyári kiadás).

Sperber, D. (2007) Seedless Grapes: Nature and Culture. In E. Margolis & S. Laurence (szerk.), *Creations of the Mind: Theories of Artifacts and Their Representation*, Oxford University Press, 124–137.

Sprevak, M. (2018) Triviality arguments about computational implementation. In M. Sprevak & M. Colombo (szerk.), *Routledge Handbook of the Computational Mind*, Routledge, 175–191.

Stern, J. (2018) Proving that the mind is not a machine? *Thought* 7:2, 81–90.

Tarski, A. (1944) The semantic conception of truth: and the foundations of semantics. *Philosophy and Phenomenological Research* 4(3), 341–375.

Tőzsér, J. (2009) *Metafizika*. Akadémiai Kiadó.

Turing, A.M. (1936) On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* 58, 230–265.

Turing, A.M. (1948/2004) Intelligent machinery. In J. Copeland (szerk.), *The Essential Turing*, Clarendon Press, 410–432.

Turing, A. (1950) Computing Machinery and Intelligence. *Mind* 59:236, 433–460.

Tye, M. (2025) Qualia. *The Stanford Encyclopedia of Philosophy*, 2025. őszi kiadás.

Unger, P. (1980) The Problem of the Many. *Midwest Studies in Philosophy* 5, 411–467.

van Inwagen, P. (1990) *Material Beings*. Cornell University Press.

van Inwagen, P. (2009) *Metaphysics* (3.kiad.). Westview Press.

Végh J. (2024a) A kognitív fenomenológia körüli vita. Kézirat.

Végh, J. (2024b) Miért vitatkozunk a kognitív fenomenológián? *Magyar Filozódiai Szemle* 68: 4, 140–156.

Viswami, A. et al. (2017) Attention Is All You Need. arxiv.org

Walker, P. N. et al. (2025) Harnessing Metacognition for Safe and Responsible AI. *Technologies* 13: 107.

Wang, H. (1995) On computabilism and physicalism: some sub-problems. In J. Cornwell (szerk.), *Nature's Imagination: The Frontiers of Scientific Vision*, Oxford University Press, 161–189.

Wasserman, R. (2021) Material Constitution. *The Stanford Encyclopedia of Philosophy* (2021. őszi kiadás).

Williamson, T. (2002) Necessary existents. *Royal Institute of Philosophy Supplement* 51, 223–251.

Williamson, T. (2013) *Modal Logic as Metaphysics*. Oxford University Press.

Wolfram, S. (2002) *A New Kind of Science*. Wolfram Media.

Woolridge, D. (1963) *The Machinery of the Brain*. McGraw Hill.

Worrall, J. (1989) Structural Realism: The Best of Both Worlds?, *Dialectica*, 43:1–2, 99–124.

Wrathall, M. A. (2014) Introduction: Hubert Dreyfus and the Phenomenology of Human Intelligence. In M. Wrathall (szerk.), *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*, Oxford University Press, 1–22.

Wright, C. (1995) Intuitionists are not Turing Machines. *Philosophia Mathematica* 3:1, 86–102.

Zach, R. (2023) Hilbert's program. *The Stanford Encyclopedia of Philosophy* (2023 téli kiadás).

Már a mesterséges intelligenciával kapcsolatos kutatások kezdete, az 1960-as évek óta – sőt, tulajdonképpen már a XVII. század óta – léteznek filozófiai érvek amellett, hogy a gondolkodás gépesíthetetlen jelenség és az emberi elme radikálisan különbözik a programokat mechanikusan végrehajtó szerkezetektől. Az ilyen demarkációs érvek közül manapság John Searle „Kínai Szoba" gondolatkísérlete és Roger Penrose gödeli spekulációi a legismertebbek, de a potenciális demarkációs kísérletek spektruma ennél jóval szélesebb. A nyelvi jelentés (Searle) illetve a matematikai megértés (Penrose) mellett az introspekció, a referencia, a fenomenális tudat, a megtestesültség és az egzisztenciális szituáltság is alapot adhat rá, hogy megpróbáljuk elkülöníteni az emberi elmét a számítógépektől.

Ez a könyv tíz modern demarkációs érvet mutat be, a mai – nagy nyelvi modellekre épülő – MI-technológia kontextusába helyezve őket. Végkövetkeztetése szerint értelmünknek legalább három gépesíthetetlen aspektusa van: a fenomenális fogalmak használata, a funkciók felismerésének képessége és az önreflexió. Emellett a gépi számítások metafizikája is arra utal, hogy az MI-re csak rávetítjük a gondolkodást, intrinzikusan nincs benne jelen.

Kodaj Dániel az ELTE BTK Filozófiai Intézetének oktatója.
Szakterülete a metafizika, ismeretelmélet és vallásfilozófia.

ELTE | EÖTVÖS KIADÓ
Az ELTE hivatalos könyvkiadója

ELTE | BTK
BÖLCSÉSZETTUDOMÁNYI KAR

ISBN 978-963-489-958-7
9 789634 899587