**Válasz az MTA Doktora dolgozatomra kapott opponensi kérdésekre és kritikákra**

Mindenekelőtt szeretném megköszönni az MTA Tudomány- és technikatörténet osztályközi doktori bizottságának munkám befogadását, és külön három bírálómnak, Z. Karvalics Lászlónak, Márton Miklósnak és E. Szabó Lászlónak a sok és értékes észrevételt és kritikát, amit a munkámhoz kaptam.

Válaszomat a könnyű navigáció érdekében pontokba szedem és áttekintő táblázatban adom meg, hogy melyik pont kinek, mely kritikájára szól. Kitérek egyik állításom egy további kritikusára is, Kodaj Dánielre is. Az MTA doktora dolgozatom 2024-es leadása óta Kodaj kiváló könyvet írt a témában, amelynek egyik szakmai lektora lehettem és amelyben összességében kb. 5 oldalon át mutatja be és illeti kritikával a dolgozatban és korábbi könyvemben bemutatott álláspontomat. Természetesen mindez nem hivatalos része a védésnek, de a könyv olyan jól sikerült és a fogalomrendszer annyira hasznos, hogy fel tudom használni ebben a vitában is; továbbá értelmezésem szerint az adott kérdésben Kodaj Mártonnal azonos álláspontot foglal el.

Az alábbi táblázat áttekintő képet arra, hogy kinek mely kritikájára hol válaszolok, az átfedéseket is jelezve. Válaszomban szisztematikusan haladtam, és úgy gondolom, hogy minden pontra kitértem, ám ennek köszönhetően a válasz meglehetősen hosszú lett (bár így is jóval rövidebb, mint a kritikák összesen). A védésen rövidített előadásban tervezem összefoglalni ezek lényegét, és ahol szükséges ott referálhatunk erre a hosszabb verzióra is.

## Z. Karvalics László



## Márton Miklós



## E. Szabó László



## Kodaj Dániel

A bírálatokat beérkezési sorrendjükben dolgoztam fel, első Z. Karvalics László bírálata:

1)

Z. Karvalics észrevételével a dolgozat megközelítését illetően teljesen egyetértek, jó megfogalmazásnak tartom, hogy a bölcsészeti reflexiónak az MI „valóságátalakító" hatásához való felzárkóztatása a célja a dolgozatnak, a releváns technikatörténeti tényezők figyelembe vételével. És való igaz az is, hogy az *MI etika* cím jóval többet takar, mint a mesterséges ágens viselkedésének etikusságát.

2)

Készséggel elfogadom a Z. Karvalics címre vonatkozó kritikáját – az MI Etika valóban nem eléggé világos és egyértelmű kifejezés. De úgy gondolom, hogy érdemben jobb javaslat eddig az első felére nem érkezett, különösen nem az Donath Z. Karvalics által kedvelt „mesterséges létezők" („artificial entities") kifejezése, még akkor sem, ha a valóban mérföldkőnek számító *Oxford Handbook of Ethics of AI*-ban jelent meg.

2.1)

Úgy hiszem, hogy joggal feltételezem, hogy a „mesterséges entitás" alatt, kontextus hiányában egyesek klónokat, épületeket, hidakat, építőmérnöki tárgyakat vagy külső erővel létrehozott politikai entitásokat is érthet – azaz csupa olyan dolgot, ami nem tárgya vizsgálatunknak, így címként nem lenne szerencsés a javasolt „Mesterséges entitások etikai kérdései", hiszen egy címnek manapság feladata lehetőleg a kontextus ismerete nélkül is kijelölni a mű területét. Természetesen lehetünk megengedőbbek és megismerhetjük Donath definícióját (Donath 2020, p.52): „*This chapter is about the ethics of our relationships with artificial entities—bots, robots, and other computational systems created to interact with us as if they were sentient and autonomous individuals. They may be embodied as robots or exist only in software;"* A *Handbook* komoly problémája, hogy nem határolja le eléggé vizsgálata tárgyát, de ez érthető a diverz szerzőség ismeretében. De Donath fenti definíciója ezen túl is ellentétes dolgozatom bizonyos kulcsüzeneteivel: az intellektuális tisztesség megköveteli tőlünk, hogy nem dönthetjük el bizonyíték nélkül a mesterséges intelligencia státuszát, például azt, hogy pusztán imitáció-e („*interact with us* ***as if they were*** *sentient and autonomous individuals")* vagy hogy ha valamit nem tartunk megtestesültnek, az milyen természetű („*may be* ***embodied*** *as robots* ***or exist only in software***"**).** Nem értek egyet Z Karvalics javaslatának irányával sem, miszerint érdemes lenne az „*MI helyett is egy tágabb kategóriát használni*" – a dolgozatomra kapott bírálatok abban erősítenek meg, hogy az MI, mint kategória már most is olyan tág, hogy feldolgozása igen nehéz. Úgy gondolom, hogy a címben az MI-nek szerepelnie kell azért, hogy az olvasót ne vezessük félre a vizsgálat tárgyát illetően, de készséggel elismerem a lehetőséget, hogy található „Az MI Etika új hullámai"-nál jobb variáció. Úgy hiszem, hogy Z. Karvalics másik észrevétele fog időtállónak bizonyulni, amely az al-problémák önálló területté válását jelzi előre (a tervezési folyamat etikája, stb.). E. Szabó László ugyancsak kritizálja a címet, de erről később (67. pont).

3)

Jogos észrevétele Z. Karvalicsnak ugyanakkor, hogy az MI társadalomszervezési szinten is izgalmas kihívásokat és lehetőségeket rejt, amelyek általam alig érintett normatív kérdéseket is felvetnek – ez is egyike annak a sok területnek amelynek mélyebb feldolgozását bírálóim számonkérik a dolog jelentősége és aktualitása miatt – ám véleményem szerint egyetlen monográfia sem lesz képes minden

nagy jelentőséggel és aktualitással bíró és a témához kapcsolódó kérdést feldolgozni, ahogy erre a limitációra bevezetőmben utaltam is. Egyetértek ugyanakkor Z. Karvalics felhívásával a rigorózus fogalmi tisztázás fontosságára, viszont ezt sokkal nehézkesebbnek látom: a fogalomválasztás hat a dolgozat később tárgyalt kérdéseire, például azáltal, hogy reflektálatlan előfeltevéseket csempész a vizsgálatainkba, mint Donath a fenti példában. Továbbá attól félek, hogy Z. Karvalics stratégiája, amely szerint ha a mesterséges ágens aktivitást mutat, és erre egy igét használunk (pl. „dönt") akkor azt metonimikusan tesszük, zsákutcába fog jutni. Nem kétséges ugyanakkor, hogy a kívánatos megoldás az volna, hogy a rigorózusan tisztázott fogalomtól kiindulva térünk rá a minket érdeklő kérdésekre, de attól tartok, a fogalom tisztázása és a fogalomra építő kérdések tisztázása párhuzamosan, oda-vissza hatva valósítható meg csupán. Ha úgy tetszik, az egy bölcsész szövegekre vonatkozó konstruktőri limitáció.

4)

Z. Karvalics ezután nagyobb fogalmi felbontású, pontos distinkciókat és részletesebb leírásokat tartalmazó érvelés rekonstrukciót kívánna meg az általam bemutatott, MI-vel kapcsolatos szkepszis kritikájánál. Elismerve, hogy az ilyesmi mindig kívánatos, úgy gondolom, hogy a kimenetelen nem változtatna érdemben. Z. Karvalics jogosan veti fel, hogy megkülönböztetendő, hogy az MI-vel kapcsolatos előrejelzések között melyek voltak azok, amelyek arra vonatkoztak, hogy a performatív siker nem fog összejönni, és azok amelyek tartalmaztak a működési módra vonatkozó predikátumot is, tehát felmenthetők, mivel az MI végül más módon érte el a működést, mint amire a predikátum vonatkozott.

5)

Ez az észrevétel lehetőséget teremt egy fontos tisztázó lépésre amelynek a jelentőségét ez az kritika tudatosította bennem, amiért hálás vagyok. Ez a lépés pedig abból áll, hogy a GOFAI korában elképzelt számítások jellegét és a jelenkori MI hullám során kialakított számítási folyamatokat komoly alapossággal vetjük össze. Ez az összevetés pedig azt eredményezi, hogy ezen számítások jellege a kvantitatív jellemzőket leszámítva azonos, ahogyan a Neumann számítási architektúra is – ismét eltekintve a mennyiségi dimenziótól – azonos.

Létezhettek olyan kritikusok, amelyek olyasféléket állítottak, hogy „ilyen kicsi memóriával nem fog menni" vagy „masszív párhuzamosítás nélkül nem fog menni" esetleg az „órajelnek több nagyságrenddel gyorsabbnak kell lennie az intelligenciához", ők valóban nem tévedtek. Z. Karvalics nem említ konkrét szerzőket, így nem lehetek biztos benne, hogy kikre gondol, de el kell ismerni, hogy ez is része volt a korabeli diskurzusoknak, például a Lighthill (1973) jelentés az MI működése kapcsán felmerülő, akkori számítógépekkel nem kezelhető kombinatorikai robbanásra hivatkozott a támogatás leállításakor, így kellő jóindulattal ide sorolhatjuk. E válasz keretein belül nem tudom pótolni minden ilyen álláspont átfogó leírását, de számomra legérdekesebb példaként megemlítem a „Knowledge Acquisition Bottleneck" MI-kritikát az 1980-as évekből, amelyeket például Lenat és társai (1985) írtak le kiválóan. Itt valóban az volt az állítás, hogy az adott mérnöki módszerrel nem lesz megvalósítható a kitűzött cél, és ez nem cáfolódott.

Az általam vizsgált kritikusok, mint Dreyfus, Searle, Collins, Penrose, Larson vagy másutt Weizenbaum, Rodney Brooks azonban nem ilyen állításokat tettek. Ahogyan egy egészségügyi dolgozó egy adott eset kapcsán határozottan kijelenti, hogy az élettel összeegyeztethetetlen sérüléseket lát, úgy ők az MI performatív sikerével összeegyeztethetetlennek tartották a számítógépeket működési elvét, amelyre „szabálykövetés"-Larson, „szimbólumfeldogozás"-Searle, „formális rendszer"-Dreyfus és nem

ugyanúgy, de Penrose, „digitális explicit rendszer” – Collins, „központi feldolgozással, CPU-val vezérelt rendszer”-Brooks-ként hivatkoztak. Ebben ténylegesen tévedtek és úgy gondolom, hogy a helyes történeti rekonstrukció része ennek a rögzítése, nem elhallgatva, hogy talán elgondolhatatlan lehetett a mai számítástechnika jellege akkoriban. Kodaj (2026) könyve kiválóan összeszedi ezeket az általa demarkációsnak nevezett érveket (p.9).

Az MI születésében döntő szerepe volt a korabeli matematikusoknak és más, matematikával jól ismerő mérnököknek. Ennek köszönhetjük a Turing gépet, a Church-Turing tézissel leírt számítás fogalmat, a véges automaták definícióját, az input-output függvényeket, a keresés formális definícióját, stb. A legmodernebb MI pontos leírásához is elégségesek ezek az eszközök, még akkor is, ha olykor meglepi a befogadót, hogy matematikai leírás szempontjából lényegtelen mennyiségi kérdéseket leszámítva a ChatGPT „csak” egy ugyanolyan program mint bármely más program vagy hogy a következő LLM token előállítása ugyanolyan keresés mint minden más keresés, a rendszer pedig input-output függvényt valósít meg digitálisan. Az általam említett kritikusok eltérő mélységben, de ismerték és megpróbálták saját érvelésük hasznára fordítani ezt a matematikai eszköztárat, és épp emiatt tudok magabiztosan érvelni amellett, hogy tévedtek. Míg a Knowledge Acquisition Bottleneck miatt az MI kudarcát jövendölőkre igaz Z. Karvalics kifejezése, miszerint „máshogy” érte el az MI a performatív sikert, a fentebb felsoroltak kritikusokra ez nem áll fenn: az MI pontosan úgy érte el a performatív sikert ahogy szerintük lehetetlen volt, és ezzel cáfolta a performatív mesterséges intelligencia általuk felvázolt limitációit.[1] Ugyanakkor ez alátámasztja Z. Karvalicsnak azt az érvét, miszerint a „baj a teljes problématér elégtelen felépítéséből fakad”.

6)

Roppant megvilágítónak tartom Z. Karvalics gondolatmenetét a gépi fordítás során arról, hogy például irodalmi alkotások esetén felvethető, hogy egyáltalán nem lefordíthatók (sem ember, sem gép által), így nem lehet pontos semmilyen kijelentés, amely a feladatosztály MI általi megoldottságára utal. Ezt készséggel elfogadom; a *no true scotsman* érvelési hibára vonatkozó hivatkozásomat kérem, hogy tekintsük úgy, hogy az emberi, azaz a valóságban is megvalósult intelligencia szintjéhez képest vizsgáljuk az MI-t.

Z Karvalics felvetésére az elmúlt 4-5 év szövegterméséről: ezt természetesen csak mintavételezni tudjuk, tekintve, hogy bizonyosan százezrekben mérhető azok száma, amelyek valamilyen ide vágó érvet vagy véleményt fogalmaztak meg az MI-ről.

---

1 Félek, hogy maga Z. Karvalics is besorolható ebbe a kategóriába a Turing teszen sikeresen teljesítő LLM-ek kapcsán. Így fogalmaz „*Másrészt azzal, ahogyan a párbeszédképes MI gigantikus emberi szövegkorpuszokból használ fel tartalmat, s ezt ötvözi a grammatikai perfekcionizmussal, már valójában nem az eredeti kihívást teljesíti, amikor még szövegkonstrukciós ügyességre volt szüksége.*” Az eredeti kihívás az Alan Turing-ra jellemző pontos definícióval rendelkezik, és abban nem szerepel olyan, hogy ne használjon a rendszer hatalmas adatbázist (és a sorok között sem sejthető ilyen, a Turing-gép végtelen szalag koncepciója (legkésőbb) 1937-es, míg a Turing teszt 1950-es – a végtelen szalag lehetősége már akkor is elindította a gondolkodást a hatalmas adathalmazokról) és nem tartalmaz olyasmit sem, hogy a gépnek „szövegkonstrukciós ügyességet” szükséges használnia, sőt Turing ezt minden bizonnyal definiálatlan és nonszensz kikötésnek tartotta volna, ha abból indulunk ki ahogyan a gondolkdás fogalmát is átkeretezte. A feladat az volt, hogy az univerzális (Turing) gép elveit megközelítőleg impementáló valós számítógép menjen át a teszten – megtörtént.
Z. Karvalics érvelése eddig a pontig klasszikus no true scotsman érvelés (az ő tiltakozása ellenére is), amelyből viszont úgy szabadul ki, hogy újabb limitációkra hívja fel a figyelmet, a „jelentésre való érzéketlenség” miatt, amelyet egy, a számítógép tanítási és működési elvéből fakadó érdemi performancia korlátként mutat be. Innen két lehetőséget látok: az, hogy ezekben az ember továbbra is jobb, és ez mérhető, amit érdemes lenne kidolgozni; vagy pedig a fordításhoz hasonlóan sem az ember, sem a gép nem tudja teljesíteni, így a megvalósuló intelligencia limitációja, de ember és gép között nem tesz különbséget.

És persze sok teljesen gyakorlati kritikánk is lehet, amelyekre nem tértem ki: általánosságban véve az a benyomásom, hogy bár az MI optimalizálások után is nemzetállamokkal összemérhető energiafogyasztása talán most indokoltabbá tenné, mint a ’70-es, ’80-as években, a legtöbb érvelő óvakodik attól, hogy mennyiségi alapon jelezze előre az MI kudarcát valamilyen feladatosztályban „A ChatGPT ezer milliárd paramétert használ, X feladat megoldásához 50 ezer milliárd kellene, de erre már a bolygón nem lesz észszerű lehetőség”. Talán a Moore-törvény által leírt, sokáig töretlen, de mára megakadni látszó kapacitásbővülés élménye az oka ennek az óvatosságnak, minden esetre számomra úgy tűnik, hogy továbbra is működési elvre (pl. „Sztochasztikus papagáj”) és nem mennyiségre hivatkoznak a legtöbben.

7)

Kicsit konkrétabban, Kodaj Dániel (2026) fentebb említett „Mesterséges Unintelligencia” c. kötete ad egy nagyon jó kategória-rendszert az MI elleni érveknek. Kodaj az érveket két nagy kategóriába sorolja, úgy mint „elmeoldali demarkáció” és „számításoldali demarkáció” és ezzel a rendszerető munkával hatalmas szolgálatot tesz a területnek. Az elmeoldali demarkációt 10 alkategóriára osztja, míg a számításoldalit nyolcra és 180 oldalon keresztül tárgyalja a pozíciókat, és úgy gondolom, hogy átfogó áttekintést ad a helyzetről, azonban épp e kötet terjedelme is jelzi, hogy a demarkációs kérdés review szerű tárgyalása egy önálló monográfiát igényel. Az álláspontom szerint a most tárgyalt monográfia felbontási szintjén az általam tárgyalt szerzők is elegendőek a probléma megvilágításához, az MI Etika többi vetületét is figyelembe véve pedig nehéz volna több terjedelmet szentelni a kérdésnek.

Kodaj a Héder: *MI etika új hullámai* valamint a Héder: *Mesterséges Intelligencia – filozófiai kérdések, gyakorlati válaszok köteteket* is elemzi. Egyfelől azért, mert felhasználja a performatív és fenomenológiai siker/kudarc fogalmait, ahogy részben Z. Karvalics is, ami arra utal, hogy ezeket érdemes jobban kidolgoznom a jövőben. Másfelől Kodaj arra jut, hogy Shagrirral (2022) közösen kettőnk számára egy további osztály szükséges, amelynek képviselői nem fogadják el, hogy a tényleges számítógépek teljesítményére a Church-Turing tézis szerinti számításokkal ekvivalens. Ezzel teljesen egyetértek, és hálás vagyok Kodaj-nak, hogy a tisztánlátást nagyban segítő fogalmi térképet hozott létre. Kodaj a *Mesterséges unintelligencia* kötetben részletesen levezeti hogy miért tartja ezt az álláspontot tévesnek. Márton Miklós ugyancsak ezen az álláspontón van a jelen vitában; ellenérveimet lentebb, a 26) ponttól kezdődően fogom megadni.

8)

Z. Karvalicsnak teljesen igaza van abban, hogy érdemes volna kitérnem a leegyszerűsítő számítógép-reprezentációk mellett a túlzottan „hype”-oló reprezentációkra is. A számítógép-reprezentációkat konstruktőri szemmel értem, amelyben állapotok, állapot-átmeneti függvények, szabályozási kör és hasonló elemek kapnak helyet. Mentségemre szóljon, hogy az intelligenciát túl-hypeoló reprezentációk tipikusan nem tesznek állításokat még jobb alkatrészekre, egyszerűen csak csodás fekete dobozként reprezentálják az MI-t, így ezek intellektuális boncolgatása kevésbé érdekes.

Fontos, és megemlítendő kivétel viszont a kvantumszámítógép, amely részletes és pontos leírással reprezentációval rendelkezik, amely eltér a hagyományos számítógéptől, és nem kizárható, hogy a közeljövőben praktikus léptékben (1000+ megbízható qubit) megvalósítható, és valóban hype kapcsolódik hozzá, így kiváló alanya lenne egy ilyen kutatásnak. Általánosságban is elfogadom, hogy nagyon gyümölcsöző kutatás lenne megkeresni azokat a narratívákat, amelyek értelmezhető, jól definiált számítógép-reprezentációból indulnak el, és túlságosan nagy eredményeket extrapolálnak.

9)

Egyetértünk Z. Karvaliccsal abban, hogy a megmagyarázhatatlansági technikák és bizonyos GOFAI megoldások ötvözése az LLM-ekben új performatív MI sikerhullámot indíthat.

10)

Z. Karvalics ezután ismertet egy, a Turing-gép és a ténylegesen megvalósított véges számítógép közötti különbségből fakadó fontos megoldandó problémát, az aritmetikai mélységet. Ezt demarkációs tényezőként mutatja be Neumann nyomán a mesterséges és természetes intelligencia között. Egyfelől ez egy kiváló, a Turing gép és a megtestesült számítógép elválasztására vonatkozó érvelésemet alátámasztó példa, és arra is jól rávilágít, hogy az emberi agy és a számítógép működési elvei nem azonosak. Azonban az intelligencia demarkációja szempontjából ez utóbb csak azzal a ki nem mondott előfeltevéssel együtt érdekes, hogy alapértelmezés szerint az emberi agytól jelentősen eltérő működésű intelligencia, még ha performans is, nem tekinthető igazinak. Ez alátámasztást igényelne.[2] Továbbá, Neumann mindezt az aritmetikai-logikai műveletek kontextusában, az aritmetikai műveletek láncolatára határozza meg, sőt a kiinduló bemenetet leszámítva inputmentes lefutásra, viszont a tényleges számítógépek nem csak aritmetikai-logikai műveleteket hajtanak végre, a közhiedelem ellenére – hanem például szabályozási kört megvalósítva mérnek is és így is tudnak hibát korrigálni.

11)

Egyetértek ugyanakkor azzal, hogy a morális inklúzió kategóriáit szofisztikálni kell. Éppen ezért érvelek a 11.4.3 fejezetben a graduális inklúzió mellett.

12)

Egyetértek azzal, hogy az MI abszolút teljes körű feldolgozásáért az 1950-es évek előttre vissza kellene menni, akár a neolitikus önkioldó csapdákig, mint a kondicionális triggerek előképéig, ám ha választani kell – és sajnos kell – akkor szerintem az 1950-es és még inkább 60-as évektől sokkal fontosabb felvenni a fonalat.

13)

Köszönöm az elismerő szavakat a műszaki tudás természetét taglaló fejezethez, és egyetértek, hogy hasznosan összekapcsolható ez a később tárgyalt junior/medior/senior diskurzussal, amely következő munkák tárgya lehet, mint ahogyan a tervezési idő és futási idő kontextusának további felbontása is. Ezt a dichotómiát a „runtime”, IT-ben bevett fogalmának számomra logikus kiterjesztésével alkottam meg, kicsit jobban kidolgozva valószínűleg önállóan is publikálható gondolat.

14)

Az MI tervezésekor a természet imitációja legélesebben a mesterséges neurális hálókban jelenik meg, ha ezt az interpretációt elfogadjuk, akkor ez a stratégia semmiképp nem nevezhető marginálisnak.

2 Ennek alátámasztása azért is fontos, mert nem kizárható, hogy a természetes intelligencia különféle ágensei is eltérő módon jutnak kívülről hasonlónak tűnő intelligens válaszokig, viselkedésekig, és ilyen alapon kölcsönösen megkérdőjelezhetik egymás intelligenciájának „igaziságát”.

15)

Az átlátszatlanság azon kérdései, amelyek nem a mérnökcsapaton belül értelmezettek természetesen legalább olyan fontosak, ennek a kérdésnek a tárgyalása, ismeretelméleti eszköztárral való kritizálása nagyon időszerű és szükséges lesz legkésőbb akkor, amikor az MI rendeletet kiegészítő átlátszatlanság-szabványok elkészülnek az EU CENELEC testületében.

16)

A 4. fejezetre vonatkozólag megfogalmazott kritika *„Ám itt megint kibújik a lóláb. A még jobb modellezés és a még jobban feltárt („tökéletes modellhez vezető”) út keresése csak akkor volna értelmes kihívás, ha az általános mesterséges intelligencia, az AGI megteremtése lenne a célfüggvény. De AGI-ra nincs szükség, és csak bizonyos típusú komplexitások esetén fontos, hogy képesek legyünk azok kezelésére, más komplexitásoknál nem. Így a „külvilág statisztikai modellezésének igénye” (65) az AGI-várók örök álma marad csupán: nem-ergodikus természetük miatt a „külvilág-alkotó” társadalmi folyamatok alkalmatlanok a statisztikai/sztochasztikus megragadásra.*”

Két válaszom van erre a felvetésre – az egyik, hogy az AGI szükségessége/szükségtelensége nem befolyásolja a tárgyalt kérdéseket érdememben, ahogy például a sakkozó MI szükségessége sem (lehet, hogy sakkozó MI-re sincs semmi szükség, a világ jobb lenne nélküle. Ettől még létezik). Másfelől, bár a megfogalmazásom bevallottan nehezen érthető, a 4.1.4 fejezet úgy értendő, hogy legfeljebb valószínűségek adhatók meg, nem bizonyosságok (az adott példa kontextusában). Elképzelhető, hogy az ergodicitás hiánya releváns bizonyos számítási megközelítéseknél, ám általánosságban véve a számítógépek nem csak analitikus, hanem diszkrét modellezésre is képesek, így nem-ergodikus folyamatok is megragadhatók velük.

17)

Az ötödik fejezetre kapott kritikát, a cselekvőhálózatok és mángorlás fogalmainak mellőzését én elismerésként fogom fel: az előbbit a hálózat tagjaira vonatkozó roppant furcsa metafizikai előfeltevések, az utóbbit az okság és az intenciók vészes fogalmi keveredése miatt kerülöm. Ezeket a 77-ik oldalon említett Dusek által 2006-ban összefoglaló néven puha technológiai determinizmusnak tekintem, de úgy gondolom, hogy a kérdés a két említett elmélet gyanús metafizikai előfeltevései nélkül is kezelhető.

Igaza van Z. Karvalicsnak abban, hogy a technológiába zártság nem abszolút, de ez a lényegi kérdésben, ami itt a konstruktőr által okozott technológiába zártság miatt viselt felelőssége, csak fokozatában változtat.

18)

A 7-9 fejezetekben Z. Karvalics kisebb elnevezési és technikai kivetnivalókat talál, amelyekkel egyetértek, valamint biztat egyes alfejezetek későbbi jobb kidolgozására, amelyek köszönettel megfogadok. A 9-ik, XAI történetének egyes elemeit bemutató fejezet kapcsán felveti, hogy nem lett volna-e jobb csak azt kidolgozni; a helyzet az, hogy innen indulva ugyanoda jutunk el, mint ahová eljutunk az MI kritika-kritika kapcsán: a számítógépek működésének alapvető természetéhez – úgy gondolom, hogy minden nehézsége ellenére a számítógépek természete az a témakör ami nem megkerülhető.

19)

Az Episztemorális tervezés fejezet elején Z. Karvalics a véleményemet kéri abban, hogy *„(…) az „alternatívakeresés” alatt értsük-e megismeréshiányos helyzet tudatosításának nyomán fellépő pót-megismerési igényt, amelynek sikere vezethet új döntési alternatívához?”* – igen, azzal a kitétellel, hogy az „ismerethiány tudatosítása” szükségtelenül antropomorfizáló. A konstruktőr szempontjából az eldöntendő kérdés az, hogy mennyit kell még észlelni mielőtt beavatkozunk, állításom szerint kivételes esetektől eltekintve erre soha nincs megbízható válasz a gép belső modelljének a külvilág általi aluldeterminánltsága miatt.

Egyetértek azzal a felvetéssel, hogy az episztemorálisan tervezett önvezető autós példában nem homológok a manőverek, tehát még az általam javasolt tervezési stratégia is túlságosan leegyszerűsítő. Szerintem ebből a fokozatos bevezetés elve következik, úgy, hogy a legkisebb de tanulságok levonásához elégséges méretű kísérleti övezetet (pl. várost) jelöljük ki, majd iteratívan alakítjuk és bővítjük a kísérletet.

20)

A morális státusz tulajdonítás tévesztési mátrixához: úgy gondolom, hogy a Z. Karvalics által javasolt fogalmi felbontás növelés szinte mindig jó ötlet, bár valószínűleg ezekhez is csak új tudunk eljutni, ismerünk egy átfogó képet, azaz nem tartom megspórolhatónak azt, hogy relatív általánosságokban is tárgyaljuk a kérdést. A másik fontos tényező, amire itt felhívnám a figyelmet az az, hogy sok morális státuszt tagadó álláspont univerzális negatív egzisztenciális állításokkal operál, így téve szükségtelenné a distinkciót. Ha az állítás az, hogy *„semmilyen számítógéppel megvalósított ágens nem rendelkezhet a morális státusz Y fajtájával, a számítógépek X tulajdonsága miatt”,* akkor Z. Karvalics különbségtétele az intelligens exoskeleton és az LLM között hatástalan a vitára.

21)

Jól ismertek számomra a robotok hallgatólagos tudása elleni és a számítógépek emergenciája elleni érvek is, hiszen ez volt a PhD tézisem témája (és ezért is hagytam ki ezt a témát), még ha nem is tartom őket működőnek. Egyetértenék azzal is, hogy a vita kinyitása messzire vezet, de Márton Miklósnak adott válaszomnál mégis kénytelen leszek.

22)
A befejező gondolatkísérlettel kapcsolatban Z. Karvalics kritikája *„A tapasztalatot szerző és tanuló felhasználó nézőpontjához képest az analitikus elemző egészen mást mozgósít a módosított nyelvi kompetenciája segítségével. Erre a különbségre nem lehet nem érzékenynek maradni, semmilyen gondolatkísérletben sem.”* Ezzel egyet tudok érteni, azonban felhívnám a figyelmet arra, hogy még így is egy olyan megismerési szituációról beszélünk, amelyben nem a számítógépek állítólagos természetéről ismert a priori tudásból indulunk ki, amely számomra kulcsfontosságú.

23)
A két konkrét kérdésre adott válaszom:
*Z. Karvalics: Mit lát elveszni a gondolatmenetből, ha az MI-kritikával kapcsolatos korábbi és újabb érvei nem kerülnek bele, s így nem nyitnak felesleges „frontot”. Mi nem mondható el az MI-kritika kritikája nélkül?*

*23.1)*

Az MI-kritika kritikája elsősorban azért fontos számomra, mert bármennyire megosztók is, az MI Etika történetében ezek szervező erőt jelentettek, így a Tudomány- és technikatörténeti bizottság számára is relevánsak. Másodsorban azért, mert ahhoz, hogy a tényleges, megtestesült MI ágens működését mutassuk be szükséges, leleplezni a megtévesztő behelyettesítéseket amelyeket az érvelők használnak. Az MI etikai kérdéseinél ez azért fontos, mert így a valós MI és valós tervezési folyamat limitációira lehetünk tekintettel. De ugyanilyen fontos a transzparencia és a megmagyarázható MI kérdéseinél is mert a magyarázat domináns stratégiája is a visszavezetés – az ismeretlen leképezése valamilyen ismert modellre, így fontos, hogy a modell jó legyen.

*Z. Karvalics: A 2024-ben lezárt dolgozat két friss fejleményre utószó-szerűn kitér. Azóta újabb két év telt el. Ne csak azt említse meg röviden, milyen fejlemények kívánkoznának a kéziratba, ha ma kéne lezárnia, hanem indexelje azokat abból a szempontból, hogy téziseit milyen formában erősítik, árnyalják, finomítják, igazolják vagy akár kérdőjelezik meg.*

23.2)

Az origótól egyre távolodva az alábbi témákat építeném be.

23.2.1)

Az egyik fontos fejlemény Kodaj Dániel *Mesterséges Unintelligencia* c. könyve, amely nemzetközileg is kiemelkedő színvonalú[3], kézikönyv-szerűen használható analitikus filozófiai szintetizálása azoknak a kérdéseknek, amelyek az MI-kritika kritikája kapcsán tárgyal dolgozatom. A 18+1 fő álláspont rekonstrukciója és kortárs vitákig való lekövetése mellett azáltal, hogy az elmeoldali és számításoldali demarkációt megkülönbözteti, óriási előrelépést ér el, amelyet e dolgozat 3. és 4. fejezete is sokkal jobban megírható lenne.

23.2.2)

2025 novemberében a BudPT25[4] konferencián az LAWS-ról (Autonóm Halálos MI) értekeztünk a Kijevi Műegyetem és a Taras Shevchenko egyetem docenseivel valamint hallgatóival. Néhány, fronton már ma is működő megoldás megismerése után úgy gondolom, hogy bármilyen MI Etika monográfia hiányos, ha a LAWS kérdéssel, történetével nem foglalkozik, amelyet a 7-ik fejezet után önálló fejezetként építenék be és az igazságos háború elmélet (JWT) felől közelítenék meg.

23.2.3)

A szoftvermérnöki tudás MI-vel való kiváltásának esélylatolgatásához alkottam meg azt, hogy a munkahelyi szerepkört három különböző tapasztalati és beosztási szinten kezeltem egy helyett. Ez azóta gyümölcsözőnek bizonyult abban, hogy a BME hallgatóinak motivációit fenntartsuk – amikor úgy tűnik, hogy a még kezdetleges szoftverfejlesztői tudásuk felesleges, hiszen az MI sokkal jobb és gyorsabb náluk, akkor lehet biztatni őket azzal, hogy szenior szinten más a helyzet és hogy csak addig kell eljutniuk, az egyetem feladata pedig éppen ez. 2025 nyarán meghívást kaptam az ETH Zürich-re egy workshopra amelynek a témája az MI hatása a mérnökképzésre volt. Erre készülve egy újabb dimenziót fedeztem fel: a munka tétjének hatását a munka megközelítésére. Ezt részben tartalmazza a junior/medior/szenior kategória is, de az elválasztás nem elég világos.

3 … és éppen ezért Angol kiadásra volna érdemes
4 www.budpt.eu

Képzeljünk el egy programozás feladatot, amelynek a specifikációja egy unalmas, de sok részletre való odafigyelést megkövetelő adminisztratív probléma megoldását kéri. Például leírjuk a specifikációban, hogy egy havi könyvelési összesítő algoritmus megírása a feladat, majd megadunk egy alapszabályt és számos adózási kivételt, kivételt a kivétel alól, stb. A hallgatók bármilyen eszközt használhatnak. Az egyik változatban a hallgatók ezt a kurzus pontjainak +10%-ért végzik el, és többször próbálkozhatnak. A másik változatban egyszer adhatják be a kódot, és ha sikerül lakhatást és dupla ösztöndíjat kapnak, ha nem, akkor elveszítik a vízumukat és haza kell költözniük, amit nagyon szeretnének elkerülni (természetesen a valóságban nem lenne etikus ennek a stressznek kitenni a hallgatókat). Az utóbbi, „one-shot" szituációban lévő hallgatóknak mindent nagyon alaposan ellenőrizniük kell, még ha MI-vel is írták a megoldásukat. Persze egy másik MI is tud ellenőrizni, de magát az ellenőrzést is validálni kell, stb.

Egy másik példa: a válasz megírása az MTA doktora tézisemre érkezett kritikákra. Sosem írnám MI-vel a vitapartnerek és az egész intézmény iránt érzett tisztelet miatt. De elképzelhetjük az alteregómat, *Amorális HM*-et. Amorális HM-nek nincsenek gátlásai, de még neki is intellektuális ellenőrzése alatt kell tartania a kezei alól kikerülő mondatokat, mert tudja, hogy Z. Karvalics, Márton és E. Szabó, esetleg a bizottság más tagjai azokat alaposan elolvassa majd és azután meg is kell majd védeni szemtől szemben. Ennek a legegyszerűbb módja, hogy Amorális HM maga fogalmazza meg és írja le a mondatokat, hisz így nem csak a szavakat, hanem azok genezisét is ismeri, ha valamit nem jól írt le, a genezis, mögöttes logika birtokában a vita hevében is van esélye korrigálni. Egyszerűbben szólva a szöveg generálás folyamatát is első személyben kell megtapasztalnia a sikerhez, maga a szöveg nem elegendő.

Ha csak annyi lenne a követelmény, hogy legyen valamiféle válasz, például az archívum kedvéért, és ránézésre kellő terjedelmű és belepillantva érdemi válasznak tűnő legyen, akkor Amorális HM betáplálná a három kritikát a Generatív MI-be és esetleg egy átolvasás után feltöltené a kimenetet.

Mindkét példában ugyanaz a feladat és a hozzá rendelhető szenioritási szint, de ha az intuícióm nem téves, eltérő mértékben és módon használnánk az MI-t. Ez tehát egy újabb dimenzió, ami meghatározza az MI-vel való lecserélhetőségünket. Megfoghatnánk, a számonkérés/ellenőrzés módjánál, vagy használhatnánk a felelősség fogalmát, de úgy vélem, hogy a legegyszerűbb ezt a dimenziót a tét dimenziójának nevezni.[5] Elképzelhető az is, hogy így magyarázható az általam használt másik, ellentétes kimenetelű példa, a kreatív rajzoló problémája is: a legélesebb helyzetben is túl alacsony a tét ahhoz, hogy ne MI-vel próbáljunk nekünk tetsző képet generálni inkább, ezért őt még szenior szinten is kiváltják.

23.2.4)

Úgy látom, hogy az érdemi emberi felügyelet („meaningful human oversight") módjának megválasztása egy olyan etikai kérdés, amelynek tétje igen nagy, és filozófusokat kíván majd a probléma jó keretezése.

23.2.5)

Prágában létrehoztak egy techno-etikai központot CETE-P néven. Ebben a központban az egyik kutatási téma a tudatos és nem-tudatos intelligencia közötti különbség, és ezt az MI-re is alkalmazzák, az MI-t unconscious, de valódi intelligenciaként jellemezve, mindegy intelligens alvajáróként. Lásd

5 „So using AI could become an express ticket to prison" fogalmazta meg frappánsan egy zürichi építőmérnök PhD hallgató, a statikai terveket aláíró mérnök büntetőjogi felelősségére utalva.

Hvorecký és társai (2023), valamint a BudPT 2025 előadást[6]. Úgy gondolom, hogy a zombiérvek és a tudatos MI közötti érdekes fokozatként ez egy nagyon eredeti gondolat.

23.2.6)

A közelmúlt egyik különösen érdekes, témánkra nézve releváns írása a Peter Königs (2025) „The negativity crisis of AI ethics".

**Válaszok Márton Miklósnak**

24)

Márton Miklós kritikája a dolgozat címzettjének, a Tudomány- és Technikatörténeti Bizottságnak a szokványos kérdései tekintetében kevéssé érinti dolgozatomat, a filozófiai tartalmát annál inkább. Úgy gondolom, hogy ebben a bizottságban legitim olyan terjedelemben foglalkozni az MI technikatörténetével, esetenként működési elveivel, ahogyan a dolgozatban teszem és amelyet Márton kritika nélkül hagy, sőt elismer. Természetesen ettől még a dolgozat címe és tartalma filozófiai is: az MI Etika hullámaival foglalkozik, és amellett, hogy Márton szerint „*filozófiai részekben valójában nem bővelkedik a szöveg*" azt a kevés filozófiai állítást amelyet mégis teszek, meg kell védenem, bizottságtól függetlenül.

Az MI lehetőségeivel kapcsolatos filozófiai vitákról:

A szöveg írásakor előfeltevésem volt, hogy még ha a doktori disszertáció valamiféle életmű összefoglalónak is tekinthető, szerencsés azért az originálisabb gondolatokra szorítkozni. Mivel Searle, Penrose, Dreyfus kritikájával másutt hosszan foglalkoztam két monográfiában - Héder (2014, 2020), Collinsszal pedig személyesen és a Polanyiana hasábjain történt cikkváltásban is sikerült eszmét cserélnem (Héder 2012, Collins 2012), az érveket valóban rekonstruáltam itt is nagy terjedelemben, de elismerem, hogy a jelenleginél valamelyest nagyobb terjedelem azért szükséges lett volna.

25)

Örömmel láttam, hogy a szöveg hiányosságai ellenére Márton úgy rekonstruálta az érvet, ahogyan szándékoltam: az általam kritizált érvelési stratégia az, hogy a számítógép feltételezett $A tulajdonságból = *{jellegéből,működési elvéből,modelljéből}*, általam összefoglalóan a számítógép *feltételezett* természetéből valamiféle intelligencia deficit vagy kudarc következik[7]. Az elterjedni látszó fogalmi felosztásom szerint ez lehet perfomatív kudarc $B = {*megverik a sakkban, hallucinál, nem tud fordítani}*, vagy fenomenológiai, amikor az MI perfomancia érdemben nem különbözik a természetes intelligencia performaciájától az adott feladatban, de az $C = *{puszta imitáció, zombi, a gép valójában nem érti mit csinál}*, stb, azaz nem társul az intelligens perfomanszhoz valamilyen mentális élmény vagy állapot, esetleg kválé, amely a természetes intelligenciához társul. A fenti érvelési sémába azért írtam behelyettesíthető értékű változó értékeket, mert ezek értelmezésem szerint ugyanannak a szélesebb kategóriának a szerzőktől függő, de az én szempontomból felcserélhető elemei.

25.1)

6 Juraj Hvorecký Artificial Intelligence and the Unconsciousness – 2025. november 28.

7 Kodaj fogalmával: demarkációs lehetőség jön létre.

Collins esetében $A = „explicit utasításokat követő gép” áll, amelyhez saját, a konvencionálistól eltérő „explicit” fogalmat is alkot, a mimeomorfikus cselekvések segítségével (Collins és Kusch 1998), amelynek bemutatása és szükségszerű, érdemi kritikája nagy terjedelmet igényelne, de a mimeomorfikusság lényege, hogy a cselekedet ugyanúgy néz ki, akárhányszor ismételjük, ezért géppel utánozható; és ha sikerült géppel, tipikusan számítógéppel sikeresen utánozni a cselekvést akkor annak gépnek „a programja” a cselekedet explicit leírása. Collins példája szerint az autóvezetés nem ilyen, a körülményekből és főleg a társas jellegéből fakadóan mindig egy kicsit mást kell csinálni, ezért $B = az autóvezetés nem is automatizálható, a projektek el fognak bukni ($C így fel sem merül).

25.2)

Lucas és Penrose a Héder (2020) 9.1-es fejezetében vagy épp Kodaj Dániel könyvének elmeoldali demarkációt tárgyaló első részének 10-ik alfejezetében „nemteljességi demarkáció”-ként rekonstruált érvében $A = „Turing gép” és $B = „nem lehet képes a Gödel mondatok igazságát felismerni” és $C = az emberi intelligencia szintjét elvileg sem érheti el, mert az emberek képesek a Gödel mondatok igazságát felismerni.

26)

Jól rekonstruálja a kritikámat Márton „*azt állítja, hogy azzal, hogy ezek az érvek a gépi MI-t Turing-gépnek, szintaktikai automatának, vagy explicit szabályvégrehajtónak tekintik, kategóriahibát követnek el.*” Úgy veszem észre, hogy Márton sem vitatja, hogy ha a kritizált szerzők az $A-nál hibás besorolást alkalmaznak, akkor a $B és esetleges $C következtetésről nem kell, hogy meg legyünk győzve. Vitánk abban áll hogy $A-nál van-e kategóriahiba, például akkor, ha az az állítás, hogy „*a számítógép egy Turing-gép*”.

Ezen a ponton szeretném röviden ismertetni a számítást modellezni képes különféle automatákat, mert a jelen vita is bizonyítja, hogy enélkül nem tudunk előrelépést elérni.

27)

A számítástudomány a számításokat „komplexitás” szerint csoportosítja a következő áttételes módon: megadja azt a számítási modellt, amely képes végrehajtani a számítások adott típusát, és ezen modellek komplexitását rendeli magukhoz a számításokhoz.

Konstruktőri szempontból ezek világos megkülönböztetésétől sohasem tekinthetünk el, mert vannak könnyebben, nehezebben megvalósítható számítási modellek és megvalósíthatatlan is.

27.1)

A legkevésbé komplex számítási modell a kombinációs logikai hálózat (Combinational Logic - CL) amely a Boolean aritmetikát tudja megvalósítani.

A komplexitásban következő számítási modell a véges állapotú automata (Finite State Machine - FSM), amelynek lényege egy véges állapothalmaz, és közöttük fennálló állapotátmeneti függvények véges halmaza. A matematikai leírásnak része még néhány szükséges kiegészítés: a bemeneti ABC, a kezdő állapot és a végállapotok halmaza.

27.2)

Ezután következik a veremautomata (Pushdown Automaton - PDA), amely az FSM kiegészítése egy nem korlátos méretű, tehát végtelen nagyra növelhető veremmel[8] és az állapotátmeneti függvények kiegészítésre kerülnek a verem műveletekkel.

27.3)

A Turing gép (Turing Machine - TM) a PDA kiegészítése, ahol a verem helyett végtelen méretű szalag szerepel. Míg a veremautomata kénytelen mindig a verem tetejét olvasni, oda tárolni jeleket, vagy onnan kell elvennie jeleket, azért, hogy a mélyebben tárolt jeleket olvashassa, a szalaggal rendelkező gép előre-hátra léphet a szalagon.

27.4)

A Church-Turing (CT) Tézis szerint „*a function is effectively calculable if its values can be found by some purely mechanical process (...) We may take this literally, understanding that by a purely mechanical process one which could be carried out by a machine.*" (Ez Turing 1937-as Gödel-re és Church-re hivatkozó megfogalmazása). Az erős tézishez szükséges az univerzális gép definíciója, amelyet Turing elkészít és amelyet az utókor Turing Gépnek (TM) nevez.

A tézisnek hatalmas irodalma van. A Church-Turing tézis igazsága filozófiai viták tárgya[9], mivel a megfogalmazása matematikai bizonyítást nem tesz lehetővé. Az idő próbáját viszont kiállta, így a számítástudomány gyakorlatában a bevett nézet, hogy minden ami kiszámítható, TM-el kiszámítható.

Másképp szólva a számítások teljes halmazát a Turing Gép tudja kiszámítani és minden ami számítás Turing Géppel kiszámítható. Az összes számítás egy részhalmaza azokat a számításokat tartalmazza, amelyeket a PDA is ki tud számítani, aminek a részhalmaza az FSM által kiszámíthatók és ennek részhalmaza az CL által kiszámíthatók.

Azonban attól, hogy a számítások egymásba ágyazott halmazokat alkotnak, az automatákra ez nem igaz. ***A Turing Gépek halmazának nem részhalmaza a Véges Állapotú Automata, sem fordítva.*** (ugyanígy a többi kombinációra). Ez könnyen belátható mivel a TM és az FSM definíciói matematikai pontosságúak és egymást kizáró követelményeket támasztanak, tehát diszjunkt halmazokat alkotnak.

Az ezzel a monográfiával, bírálatokkal is szerzett tapasztalataim alapján ma már úgy gondolom, hogy ezt minden filozófiai vita előtt tisztázni kell.

28)

A hardver elemeinek végessége miatt a végtelen verem/szalagméretű TM és a PDA nem valósítható meg, ha a megvalósítás alatt egy olyan bijekciót értünk, amely az absztrakt, ha úgy tetszik platonisztikus állapotgép és a verem/szalag állapotai, valamint a hardver állapotai közötti kölcsönösen egyértelmű leképezés. Ha ezt jelenti a megvalósítás, akkor tehát senkinek a zsebében nincs megvalósított Turing gép. Ha a bijekciótól eltekintenénk, és mégis azt állítanánk, hogy például Turing gépet nagyjából, megközelítőleg megvalósítva az a zsebünkben van, akkor viszont ez a gép nem ugyanazt a függvényt számítja ki, mint a formális Turing gép, ami óriási ár, tekintve, hogy a CT tézis sem igaz erre a gépre.

8 Verem = stack

9 **B. Jack Copeland**, "The Church–Turing Thesis," *Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/church-turing/

Szomorú következménye a helyzetnek, hogy számítógépekkel még az általános iskolában megismert, természetes számokon végzett aritmetikai műveletek sem valósíthatók meg hibátlanul, mivel lehetnek olyan nagy számok, vagy olyan hosszan leírható törtek, amelyek minden számjegyét nem tudjuk eltárolni ezért levágjuk vagy kerekítjük őket. Erre vonatkozik a Z. Karvalics által idézett Neumann-féle fogalom[10], az aritmetikai mélység és a hozzá tartozó tézis, miszerint a hibák az aritmetikai műveletláncokban csak egyre nőnek. Itt Neumann a megvalósított számítógép megszorításairól beszél a TM-hez képest; a TM tökéletesen ki tudja számítani az aritmetikát.

29)

A Kombinációs Hálók és a Véges Állapotú Automaták állapotai, valamint a megvalósult számítógép állapotai között lehet bijekció, szemben a TM-el, persze kvantitatív korlátok itt is lehetnek.

30)

Mire jó akkor a Turing gép a konstruktőr szempontjából? A (szoftver) tervezési időben (lásd 2. fejezet) ugyanis hiába tudjuk, hogy a gép véges lesz, azt csak kivételes esetben ismerjük (ha van megcélzott hardver), hogy mennyire, így nincs egy adott véges állapotú hardver amire tervezünk. Ez a rugalmasság újrahasznosíthatósági, eladhatósági, üzleti célból is előnyös, ráadásul a helyesen végrehajtott szoftvermérnöki módszer optimalizál: a lehető legkisebb memóriába férjen bele, ne pazaroljon; ha a lehető leghatékonyabban valósítjuk meg a számítást akkor a lehető legkisebb hardverigénye lesz; így viszont nem szükséges olyan kérdést feltenni, hogy mekkora a megengedett állapottér méret. Másrészt, az állapottér mérete azért sem tervezhető, mert az az inputtól is függ. A programnyelvek tervezésekor a tervezési és futási idő közötti, monográfiámban felvázolt episztemikus szakadék még nagyobb – évtizedekkel, akár generációkkal korábban kialakított programnyelveket használunk.

Összefoglalva: habár az FSM-el, CL-lel leírt számítások kevésbé komplexek mint a TM számítások, ezt nem jelenti azt, hogy ezekkel tervezni is egyszerűbb. Ezért az általánosan elterjedt nagy programnyelvek Turing-teljesek: azokon bármely számítás leírható, amely Turing géppel is, ami a CT értelmében az összes számítás. Leírható velük például az aritmetika is, szívesen foglalnak újabb és újabb memóriamezőket ahogy jönnek az újabb számjegyek, a végtelenségig. Ennek a hardver vagy az operációs rendszer a valóságban persze véget fog vetni, de a tervezési időben erről esetlegesen lehet információnk, és ha amúgy is a lehető leghatékonyabbra tervezzük a szoftverünket, akkor nincs is ennek az információnak döntést befolyásoló szerepe.

A konstrukciós folyamatban később lesz majd egy fordítási vagy interpretációs lépés, amely a szoftverünkön már adott hardveren futtathatóvá teszi, és innentől az FSM komplexitásnál többet nem fog tudni (cserébe jó esetben működni fog a valóságban).

31)
Visszatérve Márton kritikájához: *„Álláspontom szerint a „Turing-gépnek lenni” – ahogy a többi említett megnevezés is – egy funkcionális predikátum, amely tehát bizonyos funkcionális szerepet tulajdonít annak az individuumnak, amelyről predikáljuk. Az az állítás tehát, hogy a számítógép Turing gép, voltaképpen úgy értendő, hogy a számítógép egy Turing-gép, vagyis betölti a „Turing-gép” predikátum által kifejezett funkciót, vagy másképp: egy Turing-gép megvalósulása (realizációja) – ami egy teljesen értelmes, kategóriahiba-mentes állításnak tűnik.”*

10 Lásd e válasz 10) pontját

Amennyiben a „funkcionális predikátum” illetve a „Turing-gép megvalósulása” bijekciót jelent, akkor ez egy világos, kategóriahiba-mentes, de téves megállapítás.

32)

Itt szeretnék kitérni Kodaj Dániel (2026) ide vonatkozó kritikájára, amelyben a Héder (2020) valamint a jelenlegi doktori művet és arra válaszul külön fejezetet szentel a „Turing gép megvalósítása” kérdésének. Nagy elismerés illeti Kodajt azért, hogy nagyon világosan kezeli ezt a kérdést és további érveléseihez előfeltevés céljára az vizsgálja, hogy a CPU „ekvivalens”-e a TM-el. Az ekvivalencia erősebb viszony, és ha a CPU ekvivalens volna a TM-el akkor valóban el kellene fogadnunk a Gödel nemteljességen alapuló demarkációt is például.

33)

Ismét Márton kritikájához visszatérve, ha a megvalósulást – a fent nevezett funkcionális predikátum teljesülést, vagy ekvivalenciát nem értjük ilyen szigorúan, hanem egyfajta hibatűrő modell-valóság viszonyként képzeljük el, akkor viszont még tarthatatlanabb állásponthoz jutunk.

Tudjuk, hogy ha az aritmetika megvalósítható a TM-en, akkor a TM-re Gödel tételei vonatkoznak. Gödel tételeit ugyanakkor nem tudjuk alkalmazni az FSM-re, mert az aritmetika sem valósítható meg rajta (abban pontos értelemben, ahogyan a matematika ezeket a kérdéseket kezeli, és itt ez a releváns).

Tudjuk azt is, hogy a TM-en vannak megállási problémák, ugyanakkor azt is, hogy ugyanezek nem állnak fent az FSM-ek osztályára, mert azokhoz szerkeszthetünk egy nagyobb, emuláló FSM-et, ami vizsgálja hogy visszatértek-e egy korábbi állapotba, így detektálhatja a végtelen ciklusokat. E helyzet részletesebb jellemzéséért lásd Sóstai (2024) disszertációját, amelyben a megállási problémát a fizikai Church-Turing tézis kontextusában vizsgálja és emlékezett a fizikai realizálhatóság problémájára – ám ez az absztrakt FSM-et nem érinti.

34)

**A tényleges számítógép közelítő modellje tehát a Turing Gép, míg pontos, valós állapot-modell állapot közötti bijekciót megvalósító (tehát nem csak közelítő) modellje a véges állapotú automata. Tudjuk: TM-re működnek Gödel tételei és a megállási kérdés is releváns. Tudjuk azt is: FSM-re nem vonatkoznak Gödel tételei és a megállási probléma sem áll fent.**
Akkor ezzel bizonyítottuk vagy cáfoltuk a *nemteljességi* vagy *eldönthetetlenségi*[11] demarkáció (Kodaj, 10-es és 11-es fejezet)? lehetőségét a tényleges számítógépre?

Márton és Kodaj hasonlóan érvel:

Kodaj: *„A Turing-gépek absztrakt voltából ugyanúgy nem következik, hogy matematikai tulajdonságaik csak a platóni mennyországban relevánsak, ahogy a számok absztrakt mivoltából sem következik, hogy ha van három nyuszink és két répánk, akkor minden nyuszi kaphat egy egész répát.”*

11 A halting problem más megfogalmazása.

Márton: *Érvelése szerint például a „Turing-gép” elnevezés absztrakt platóni ideára utal, és ezért kategóriahibás partikuláris fizikai tárgyakra alkalmazni, mint például egy számítógép (...[itt a funkcionális predikátumra vonatkozó rész HM])*
*Lehet azt gondolni, természetesen, hogy az effajta predikátumok valamiféle absztrakt, platóni entitásokra utalnak – ez jól ismert álláspont a témában –, ám ettől még maguk a predikátumok alkalmazhatók maradnak konkrét fizikai tárgyakra is, ezt még a platonisták maguk sem tagadják.”*

Nem segít azonban ez az érvelés, ha a konkrét tárgyaknak, furcsa módon több platóni entitás is megfeleltethető és ezekből egymásnak ellentmondó következtetés adódik a számunkra érdekes kérdésben (viszont ez a helyzet leleplezheti a vitában uralkodó platonista megközelítés problémásságát).

Ezzel a stratégiával további ellentmondásokhoz jutunk hiszen a számítógép termodinamikai modelleket is megvalósít, vagy például lineáris, másod és harmadfokú végeselemes modellekkel is leírható, de fizikai alakja aligha lehet egyszerre lineáris, másod és harmadfokú is.

35)

Én ugyanakkor azt gondolom, hogy mégiscsak a kategóriahiba a jó jellemzése az általam kritizált érvelésnek. Penrose például (Kodaj fordítása, kiemelés tőlem):

*„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 működtetése. A Gödel-mondatról látjuk, hogy igaz bizonyos számokra [...], noha ezt lehetetlen levezetni az axiómákból. A reflexiós elvekből eredő belátás* ***nem kódolható bele egy formális rendszerbe****, melyben egy algoritmus szerint levezetéseket végzünk.”* (Penrose 1999: 144).

Ez véleményem szerint csak akkor érthető, ha a Penrose a tényleges számítógépet szó szerint formális rendszerrel helyettesíti, tényleg úgy gondolja, hogy az aritmetikát az konzisztensen valósítja meg és ezt kategóriahibának plusz egy tévedésnek tartom.[12] Ha megengedné a valóságnak megfelelően azt, hogy ez egy közelítő modell, akkor lelepleződne a fogalmi elcsúszás.

36)

Minderre rakódik rá az a további tényező, miszerint a tényleges számítógép és az FSM állapotai közötti bijekció sem teljesül minden esetben, továbbá inkább a teljes számítógép bizonyos alrendszereire jellemző mint az egészére. Például ha csak egy egyszerű önvezető járművet vizsgálunk meg közelebbről, hamar rájövünk, hogy az részben analóg: a sebességét nem egy digitális változóban tárolja, hiszen azt a fizikai körülmények – szél, tapadás, stb. - úgy befolyásolják, hogy az a gyakorlatban nem kiszámítható. Ezért inkább minden ciklusban kiolvassák a sebességmérő műszerük értékét, tehát a mozgási sebességére vonatkozó információt a jármű kinetikus energiájában tárolják, részben analóggá téve a működést – az ilyen gépekkel foglalkozik a kiberfizikai rendszerek tudománya. Tiszta FSM modell helyett gépészeti, termodinamikai modellekkel leírható folyamatok érdemben

12 És az sem merül fel benne, hogy a tényleges számítógépet egyébként nem kell szükségszerűen kódolni sem.

befolyásolják a működését (és itt is ugyanúgy fennáll, hogy az absztrakt modell tulajdonságai nem automatikusan vetíthetők le a konkrét számítógépre).

Ennek megmutatását Márton valamiért nehéz feladatnak tartja:
„*Nem lenne persze könnyű feladat, hiszen meggyőzően meg kellene mutatnia, hogy a fizikai paraméterek miért játszanak lényeges szerepet a számítógépek releváns működésében – ilyesfajta érveléssel egyébként a jelölt korábbi könyvében találkozhatunk*” de érveket a szkepszisére nem ad meg.

Valójában semmilyen szabályozási kör, de még egy olyan egyszerű alkatrész, mint a kvarckristályra épülő belső óra sem valósítható meg pusztán digitális modellel. A digitális modellek ciklusszámlálót esetleg ismernek, de a diszkrét állapotok között eltelt időt nem; vagy például eltérő időzítéssel működő, de osztott memóriát vagy osztott adatvezetékeket is használó FSM-ek halmaza (tehát a modern laptop vagy asztaki számítógép) működése sem leírható Márton elképzelése szerint.

37)

Kodaj Mártonnál egy lépéssel tovább viszi a következtetését: hogy ha elfogadjuk ellenvetésemet (vagy a velem egy osztályba sorolt Shagrir ellenvetéseit) annak az a következménye, hogy a számítógép CT értelemben nem számítást hajt végre (a neve ellenére). Kodaj ezt az opciót elveti, holott ez bizonyosan így van, amennyiben egy tényleges rendszert, pl. MI-t egyetlen számításként próbálnánk elképzelni. E helyett CT és nem CT elven modellezhető folyamatok vannak jelen egyszerre, egymásra hatva és a felhasználó ezt tapasztalja meg; ezen felül a kiberfizikai rendszerekhez hasonlóan, de a felhasználóval alkot szabályozási kört.

38)

A céltábla mozgatására, valamint a „Machines cannot do X”-re jó példa a Héder (2020) 2.1-ben bemutatott Gépi fordítás és GO játék példája – mindkettővel kapcsolatban a performatív megvalósítás lehetetlensége volt a vád, ahogy később a kreatív alkotásoknál is. Itt jegyzem meg, hogy önmagában azt a történeti rekonstrukciót is a céltábla mozgatásának tartom, miszerint a korabeli vitákban valójában az sosem volt kérdés, hogy meg lehet-e majd valósítani az adott feladatosztályt, csak az, hogy ez „valóságos” intelligencia lesz vagy valamiféle „imitáció”.[13] Márton ezt még jobban kisarkítja, amikor úgy rekonstruálja a helyzetet, hogy a megvalósíthatósággal foglalkozó filozófiai viták *„az emberi értelemben vett kognitív képességekkel rendelkező MI”* azaz az erős MI kérdés lett volna. A helyzet sokkal jobban leírható úgy, hogy az erős MI kérdés mindig is fontos volt és az erre adott válaszok gyakan a gyenge MI-t egyes eseteit is kizárták. Collins esetében az önvezető járművek lehetetlenek, Penrose esetén a matematikai bizonyítások jó része, ahogy fentebb láttuk, Dreyfusnál pedig szinte minden. Lásd még a Z. Karvalicsnak adott válaszaimat egy hasonló kérdésre fentebb (10), például az ergodicitás vagy az aritmetikai mélység kapcsán.

39)

Bizonyosan igaza van Mártonnak abban, hogy a téma még jobb feldolgozása megkívánna egy elmefilozófia-történetet, amelyre bizonyosan egy önálló doktori mű témája is lehetne. Úgy gondolom, hogy bizonyos, MI megvalósíthatóságát negatívan megválaszoló álláspontok internális kritikával is elutasíthatók, vagy úgy hogy egy, saját leírásukban is szükséges premisszájukat cáfoljuk meg, és épp erre vonatkozik a számítógép természetére vonatkozó levezetésem. Ugyanakkor Márton egy másik

13 Z. Karvalics 5) pontja ide sorolható.

kérdésére válaszolva a pozitivizmus és test-elme identitáselmélet összekapcsolódásának néhány pontját bemutatom alább.

40)

Teljesen jogos Mártonnak észrevétele, miszerint az MI etika kanonizálandó kérdéseinek egyikét sem fejtem ki teljeskörűen, de itt az állítás valóban csak annyi volna, hogy ezek a kérdések látszanak kanonizálódni egy kortárs történeti folyamatban (a kézikönyvekben és oktatásban), persze még ezt is alá lehetett volna támasztani jobban. Nem értek egyet Márton azon állításával, hogy az MI munkaerőpiaci vonatkozásai nem tekinthetők etikai vonatkozásúnak.

41)

A morális ágencia kérdésénél elnézést kell kérnem Mártontól és minden olvasótól, amiért az MI-ben bevett ágens fogalmat egy mondatban használtam az ágencia fogalommal, zavart keltve ezzel. A mesterséges ágens fogalmat MÁ-val helyettesítem ezért, így a kérdéses mondat remélhetőleg érthető lett:

Héder: *„a morális ágencia [az MÁ] a felelősségrevonhatóságra és elszámoltathatóságra vonatkozik; a páciens státusz pedig az MÁ-val szembeni morális kötelességekre, mint például annak az elismerésére, hogy az MÁ szenved. Egy felnőtt, beszámítható ember morális ágens és páciens is egyben, míg egy csecsemő csak morális páciens, mert morális kötelességeink vannak vele szemben, de nem morális ágens, mert nem tarthatjuk cselekedeteiért felelősnek.”*

A másik kérdéses mondatnál kevésbé vagyok biztos a félreértés okában:
Héder *„Egy ijesztő, de az emberi társadalmak történetében nem példátlan logikai lehetőség a morális ágencia, felelősség, megbüntethetőség tulajdonítása a morális pácienssel szembeni kötelességek [elismerése] nélkül: egy háborúban gyakran így látják egymást a felek.*” – az ellenséget ágensnek, azaz tetteikért felelősnek, elszámoltatandónak, kötelességüket megszegőnek tekintik, de páciensnek nem, tehát az ellenségnek nem tulajdonítanak fontosságot a szenvedésének az ellenséggel kapcsolatban nincsenek kötelességeik.

42)

Jogos Márton észrevétele arról, hogy a rendszerben kritikus betöltött szerep és a relációs felelősség a leírt változatban nincs jól megkülönböztetve. Ezt úgy pontosítanám, hogy az előbbi egy rendszer funkcionálása szempontjából kritikus elemre vonatkozik (még akkor is, ha ennek nincs relációja emberekkel vagy más elemekkel), az utóbbit pedig úgy, hogy a társadalom humán tagjaival betöltött reláció függvénye, ezáltal remélhetőleg érthetővé válik a különbség.

43)

Az MI tudatosság és a deontológiai etika szerinti befogadás kapcsolatánál egy elveszett lábjegyzet okozza a problémát. A Jianhua Xie (2021) által felvázolt Kant értelmezés szerint az öntudat teljes morális státuszra jogosít fel.[14]

14 „Kant's deontology provides an important basis for using self-consciousness as the criterion for complete moral status. It is the philosophical source of the equal-status proposition. Kant proposed a far-reaching moral theory. In his view, autonomy is a prerequisite for evaluating the behaviour of moral subjects. Morally permissible behaviours are those that all rational individuals are willing to do under certain circumstances. An explanation of the relationship between

44)
A következő mondatban:
Héder:„*Egy elvi lehetőség a morális megfontolás összekapcsolása az élőlény státusszal. Ez némileg ellentmond a jelenlegi gyakorlatunknak: ha egy ember súlyos és helyrehozhatatlan agysérülést szenved, státusza annyira csökkenhet, hogy mások döntése alapján az illető élete megszakítható.”* Nem a passzív eutanáziára gondolok, hanem, ahogy jelzem is, az agyhalál kritériumra, amely nem azonos a biológiai halál kritériumával. Az agyhalál-kritériumot 70-es évek óta egyre több országban fogadják el a halottá nyilvánítás okaként, akkor is, ha a lélegeztetőgépre kapcsolt beteg és keringése, ebben szívverése működik. Itt nincsen különösebb fogalmi zavar, mert a kontextusban használt „élet” fogalom alkalmasan átalakítható, hogy ne legyen ellentmondás. Pusztán azt az észrevételemet szerettem volna megosztani, hogy ez egy kis tolódásként is értelmezhető abba az irányba, hogy a személy morális státusza a mentális képességek függvénye legyen, valamint, hogy léteznek biológiai élet fogalmak amelyek építenek a sejtek anyagcseréje, a keringésre, egyes kultúrákban szívverésre, amelyek így nem esnek egybe a morális státusz felvetésével.[15]

44)

A 11.5 fejezet végén található egy állítás

*Héder „Nyilvánvaló, hogy a tévesztési mátrix aszimmetrikus. Az utóbbi eset, amikor feleslegesen tulajdonítunk morális státuszt az MI-nek, sokkal kedvezőbb, ezért felvethetjük: ha az MI ágens viselkedése összhangban van egy morális státuszra ma is elismerten érdemes személy vagy entitás viselkedésével (például ember, állat), akkor a bizonyítás terhe azon van, aki a morális státuszt megtagadná az adott ágenstől. (194., kiemelés az eredetiben)”*

Erre Márton kritikája az alábbi:

*Nos, azt hiszem a szóban forgó aszimmetria fennállása korántsem nyilvánvaló. Rövid reflexiót követően szerintem bárki előállíthat valamilyen ellenérvvel, amely arra alapoz, milyen nagy – és várhatóan káros – mértékben változtatná meg világunkat az, ha morális státuszt tulajdonítanánk az MI-nek. Annak kockázata pedig, hogy nem tesszük meg ezt, csak akkor tűnik nagynak, ha az MI morális státuszának legalábbis a lehetőségét valósnak tartjuk. Mindez pusztán azt vonja maga után, hogy a kérdés e sommás megállapításnál és érvelésnél jóval alaposabb kifejtést és argumentációt igényelt volna.*

Ezt nem tartom jó ellenérvnek a felvetésemre négy okból.

44.1)

Egyfelől, az ellenérv csak akkor működik, ha konzekvencialista keretben gondolkodunk. A másfajta etikai keretrendszer mellett elkötelezettek figyelmen kívül hagyhatják. Meggyőződésem szerint viszont a mátrix nem csak ilyen keretrendszerben érdekes: kötelesség-etikában is értelmezhető az az eset, hogy valamiről nem tudjuk, hogy kötelesség-e, de a biztonság kedvéért betartjuk, ha ezzel más kötelességeket nem szegünk meg, így morális viselkedésünket garantálva.

44.2)

artificial intelligence and human beings from the perspective of consciousness”

15 „Az agy halála az egyedi személyiség halála is egyben.” mondja a vérellátó https://www.ovsz.hu/hu/oco/agyhalal

Másfelől, ha spekulatív érvekkel operálhatunk – „az MI nagy és várhatóan káros kockázata” – akkor ezzel ellentétes érvekkel is próbálkozhatunk, amelyek azt jelzik előre, hogy a morális státusz megadása nagyon jó kimenetellel jár. Például, ha a morális páciens státuszt rendeljük sok MI ágenshez, akkor az új világállapotban automatikusan akár milliónyi új érintett is lesz. Ha megengedjük azt, hogy ez nekik átlagban inkább jó, akkor azonnal több millió érintettel tettünk jót, és ezzel kellene szembe állítanunk egy ennél összegezve negatívabb hatást.

44.3)

Harmadrészt, ha ez valóban egy konzekvencialista alapon eldöntött kérdés, akkor a két alternatíva várható jósága közötti különbséget az adott keretben elfogadott kalkulussal lenne kötelességünk kiszámítani, nem pedig csak posztulálni. Nem elegendő, hogy „ *bárki előállíthat valamilyen ellenérvvel, amely arra alapoz, milyen nagy –és várhatóan káros mértékben változtatná meg világunkat”* mert a konzekvencialista keretben nem a negatív következmények feltételezhetősége számít, hanem ezt tényegesen ki is kellene számolni a kalkulus alapján, és továbbra is fenntartom, hogy a bizonyítás terhe azon van aki a javaslattal előáll.

44.4)

A negyedik probléma, hogy ezzel a megközelítéssel a morális kör akár szűkíthető is, például valaki bizonyos feltételezett változó értékek alapján levezetheti, hogy jó volna megvonni az embertől a morális státuszt, majd megtizedelni a fajt, például az ipari állattartás, vagy az ember által okozott tömeges fajkihalás miatt.

Még a 2-ik és 3-ik probléma a konzekvencialista keret nem kielégítő felhasználására utal, az 1. és 4. ellenvetésem magát a keretet kérdőjelezi meg, együttesen úgy gondolom hogy komoly ellenérv Márton javaslatával szemben.

45)

Márton ellenvetést fogalmaz meg azzal a következtetéssel kapcsolatban, hogy a biciklizni képes robotnak hallgatólagos tudása van:
„*Nem világos, miért volna a biciklizni-tudás képessége annak egyértelmű bizonyítéka, hogy a képesség alanya hallgatólagos tudással rendelkezik. Ehhez azt kell feltételeznünk, hogy e képesség gyakorlásához feltétlenül szükség van effajta tudásra, magyarán csak az tudhat biciklizni, aki rendelkezik az ehhez szükséges hallgatólagos tudással. Nos, – bár hangsúlyozottan nem vagyok Polányi szakértő – amennyire tudom, ezt a nézetet Polányi nem képviselte, és józan ésszel is kontraintuitívnek tűnik. Bár a biciklizni-tudás valóban Polányi egyik visszatérő példája a hallgatólagos, képesség jellegű tudásra, tudtommal ő maga sehol nem zárta ki, hogy olyan képességeket, amelyeket a legtöbben hallgatólagos tudásunk alapján vagyunk képesek gyakorolni, egyes – humán vagy nem humán – alanyok nem ilyen tudás, hanem mondjuk explicit (fokális) tudásuk alapján gyakorolják.*”

Márton Miklós itt téved, Polányi egészen pontosan azt az álláspontot képviselte amit Márton megkérdőjelez.

45.1)

Polányi szerint minden tudás részben vagy egészben hallgatólagos, így az itt leírt eset is: pusztán explicit tudással semmilyen személy sem képes semmire.

45.2)

Márton Miklós megközelítése ugyanakkor nem példátlan: nagyon hasonlóan érvel Harry Collins a Tacit and Explicit Knowledge (2019) c. könyvében, így ennek recepciójában a Polányi kutatók köreiben találhatók meg az ellenérvek. A legjobb ezek között Gulick (2023) rekonstrukciója, de lásd még cikkváltásomat Harry Collins professzorral a Polanyiana hasábjain (Héder 2012, Collins 2012).

45.3)

A gépek hallgatólagos tudása a PhD tézisem volt, az originális állításokra törekedve – a doktori pályázat követelményei szerint - az MTA doktora téziseimben ezt a 14 éve indult vitát nem szerettem volna rekonstruálni, de a kérdésre válaszolva a legtömörebb formában összefoglalom a főbb pontokat.

A kerékpározó személy (akár természetes, akár mesterséges) esetére a legtömörebb magyarázat talán így adható elő: nem kizárható, hogy valaki explicit utasításokat hajt végre és ezáltal jön létre a teljesítmény, például valaki kalkulációk segítségével kerékpározik esetlenül, de sikeresen. Ám ekkor is használja az explicit szimbólumok manipulációjának hallgatólagos képességét. Az erre adható esetleges ellen-ellenérv nem működik rekurzívan: ez úgy nézne ki, hogy a személy a kerékpározáshoz használt szimbólum manipulációt, kalkulációkat sem hallgatólagosan végzi el, hanem azokat is még alapvetőbb explicit szimbólum manipulációk segítségével végzi el – ez a lánc egyszer biztosan megszakad, hiszen az performanszhoz szükséges agyi idegek vagy izmok működése egy ponton már biztosan nem lesz valami explicit manipulációként jellemezhető.

46)

Fontos, hogy míg a hallgatólagos tudás az állatvilág sztenderdje, addig az artikuláció és ezzel párhuzamosan az explikált tudás használatának képessége erre rárakódva csak az embernél alakul ki, és megsokszorozza annak képességeit (Lásd a tankönyvünket (Héder, Paksi 2020)). Az artikuláció képességével szimbólumokat hozhatok létre saját magam vagy hozzám hasonlók (nyelvhasználat vagy kultúra szempontjából) számára, amelyeket a fokális figyelmembe helyezve manipulálhatok, vagy mások ezeket megértve részben vagy egészben hallgatólagos tudásra tesznek szert, például tényeket ismernek meg, vagy megtanulnak zongorázni.

47)

Egy alternatív levezetés a figyelem szerkezetére alapul: ahogy fentebb bemutattam, még az így elsajátított, nagyon explicitnek tűnő tudás, például képletek ismerete és alkalmazása is a jelentéstársítás és kompetens szimbólum manipuláció hallgatólagos képességeinek felhasználásával történik, és ezek a képességek a figyelem szempontjából is csak járulékosak maradnak a teljesítmény közben.

48)

A Polányi tankönyvünk írása során szerzett saját tapasztalataim alapján egyetértek Márton Miklóssal abban, hogy sokaknak mindez roppant kontraintuitívnek hat, talán azért, mert Polányi téziseivel szemben sokan a hallgatólagos tudást a magasabb rendű, kicsit titokzatos, ritka, valahogyan magasabb

minőségű; és főleg az emberre jellemző tudásnak tartják; nem pedig Polányival összhangban a nyelvhasználat és artikuláció nélkül is működő, ősi, az állatvilágból örökölt, egyszerű tudásnak.

49)

Polányi ismeretelméletén belül viszont épp Márton felfogása kontraintuitív és radikális: a programnyelven megfogalmazott állítás a humán programozó által artikulált explicit szimbólumsorozat, amelyet egy másik humán programozó megérthet. Azt feltételezni, hogy a számítógép fokális figyelmébe helyezi és annak segítségével végrehajtja a programot: erős antropomofrizmusnak tűnik.

50)

Fontos az is, hogy az explicit tudás relációban áll a személlyel akinek a személyes tudásának része. Először azé, aki explikálja, majd azoké akik elsajátítják. Ha minden személy eltűnik, de a könyvtárakban ott maradnak a könyvek, akkor ott nincs tudás. Ha később egy régész megtalálja és megfejti, akkor ismét egy személlyel kerül relációba és ismét lesz tudás. Elsőre talán szokatlan, de meglehetősen konzisztens felfogása ez a tudásnak.

A program a programozó explicit tudása – ha azt gondolnánk, hogy a számítógépnek van explicit tudása az nem eleve lehetetlen, de mindenképp ez a nehezebb állítás. A könnyű állítás az, hogy hallgatólagos tudása van, annak ellenére, hogy az ember programozza. A program itt olyan szerepbe kerülhet, mint a gének, amelyek az állat DNS-ében kódolva vannak, de nem mondanánk, hogy az pl. a rovar a viselkedését a génjeinek végrehajtásával valósítja meg.

51)

Ahogy Walter Gulick összefoglalja a robotok hallgatólagos tudására vonatkozó téziseinket: „*We typically learn to ride a bicycle without knowing the physical rules that enable us to stay upright and make progress. But the rules for balance and forward motion can be explicated and employed by a robot [Héder és Paksi 2012]. Polanyi would agree.*” Azaz az egyensúlyozás szabályai explikálhatók (a mérnökök által) és azokkal robot készíthető, de ez nem jelenti azt, hogy a robot fokális figyelmébe a program futtatását helyezi, ahogyan a hallgatólagos tudással rendelkező rovar sem a DNS-ét, mint valamiféle utasítássort hajtja végre.

52)

A legerősebb ellenérv a tézisemmel szemben az volna, hogy a személyes tudás fogalmát az élővilág számára tartjuk fent. Habár Polányi maga az élőlényeket is gépek tágabb kategóriájába sorolta (Polányi 1968).

53)

Valóban sok munka született az emergencia, redukció fogalmáról (abban, hogy a rétegzett valóság fogalmáról könyvtárnyi irodalom született volna az elmúlt időkben, Mártonnál kevésbé vagyok biztos). Azonban ezen elméletek nagyobb része, a gyenge vagy ismeretelméleti emergencia egyszerűen nem tűnik relevánsnak az ebben a fejezetben vizsgált kérdésben, mivel nem teszi lehetővé, hogy a fizikalista leírásokra nem redukálható működési elvekről, valamint rétegzett ontológiáról beszéljünk. Hogy a gyenge emergenciát miért zárom ki ebből a levezetésből, arról másutt írok (Héder (2013), kibővítve

Héder és Paksi (2019)), ahogy arról is, hogy mi a jelentősége a működési elveknek Héder (2017) és a doktori értekezésemben is (Héder 2014).

53.1)

Megjegyzem, hogy a gyenge emergencia elméleteit az ezeket képviselő filozófusok kiválóan alkalmazhatnák ugyanakkor az MI megmagyarázhatóság kérdésének újrakeretezésére, különösen a gépi tanulásra.

Érdemes szót ejtenem az emergencia kategóriáinak egy ortogonális megkülönböztetéséről, a szinkron vagy diakrón emergencia distinkcióról (Héder 2017, p.7.). A diakrón emergencia egy entitás jelenlegi és megelőző állapota közötti viszony jellemzője (természetesen az idő egyirányú fogalma vagy az állapottér állapotainak egymásutánisága szükséges a fogalomhoz). Z. Karvalics László például megkérdőjelezi, hogy az MI létrehozása evolúciós vagy emergencia folyamat-e részben a dolgozatomban, részben Szathmári Eörs különféle kinyilatkoztatásaiban, és ekkor a diakrón emergenciára utal. Ezzel szemben a szinkron emergencia a rendszer egy adott pillanatban fennálló tulajdonságai vagy jellegzetességei közötti viszony, erős emergencia esetén például a valóság ontológiai rétegzettségére utal.

A fejezetem az erős, szinkron emergencia kategóriáján belül működik, és itt már egyáltalán nem olyan sűrű a szakirodalom. Ez a lehatárolást más írásaimban elvégeztem, de minden bizonnyal itt is segítette volna az olvasót.

54)

Ebben a szűkített kérdésben fontos, a dolgozat keretrendszeréhez releváns előrelépés történt mióta a fejezetet lezártam: J. M. Fritzman *Collapsing strong emergence's collapse problem* 2024 júniusában megjelent cikkében hatékonyan kerüli el a lefelé okozás roppant problémás kérdését, és az általam is előszeretettel alkalmazott „Landauer elv"-re emlékezteti az olvasót. Landauer, az IBM chiptervezője az 1960-as években pontosan ismerte a Turing gép és a megtestesült számítógép lehetőségei közötti eltérést, és az entrópia fogalmára építve így fejezte ki azt: „*az információ bármilyen logikailag irreverzibilis manipulációját — például egy bit törlését vagy két számítási útvonal összeolvasztását — az információfeldolgozó berendezésben vagy annak környezetében található, információt nem hordozó szabadsági fokok entrópiájának megfelelő növekedése kell, hogy kísérje.*" (Bennett 2003 rekonstrukciója Landauer 1961 levezetésének, ford. tőlem). Leegyszerűsítve e tétel azt magyarázza el, hogy a számítás energetikai ára a chipben keletkező hő, és hogy ez teljesen nem eliminálható. Fritzmann az emergencia vitájában, az érvelőtől függően megengedett vagy épp kizárt redukciós lépést információfeldolgozási számítási műveletként keretezi – igen meggyőzően –, ezáltal a redukciónak fizikai korlátokat szabva. Fritzmann érve nem kikezdhetetlen, de jól példázza azt a folyamatot amelyet „új hullám"-ként jellemzek dolgozatomban: egy kifinomultabb, pontosabb képe a számításnak más megvilágításba helyez korábbi vitákat, ez esetben „az erős emergencia új hulláma"-ként is jellemezhetném Fritzmann érvelési stratégiáját, még ha annak igazságáról egyelőre nem is akarok állást foglalni.

55)

Márton: *Szintén nehezen értelmezhető Polányi elgondolásainak alkalmazásai közül az, hogy a szerző úgy véli, a MI-vel aló interakciók tapasztalata alapján olyasfajta hallgatólagos, személyes tudásra tehetünk szert e rendszerekről, amelyekben kritikai szemüvegünket félretéve megbízhatunk.*

Polányi bizalmi programjában, amennyiben bármivel kapcsolatban tudásra teszünk szert, azt így tesszük. Itt nem sorrendben haladok Márton érvelésében, egy mondatot átugorva: „*Számára azon képesség jellegű tudások érdemlik ki e bizalmi, kritikától mentes státuszt, amelyeket egy-egy tudományos iskola generációk során kitermelt magából”* nem kizárólag: a személyes tudás szerkezetéből következően mindenfajta tudást részben kritikamentesen használunk, például beszéd közben nem tudjuk a szavak jelentését állandóan kritika alá venni, amikor egy lépést teszünk, izmok tucatjai stabilizálnak és az egész összeköttetésben van a füljáratainkkal, amikor elménkben matematikai szimbólumokat manipulálunk, akkor is sem tudunk reflektálni a háttérben működő kognitív apparátus esetleges hibáira. Polányira jellemző, hogy ugyanaz a gondolatot különböző léptékű jelenségekre alkalmazza: való igaz, hogy tudósok közösségében kitermelt hagyományra is kijelent hasonlókat.

Nagyon jónak tartom Márton ellenérvét, „*Nos, amennyire tudom, Polányi posztkritikai tudományfilozófiájának e központi elképzelése nagyban támaszkodik a hagyomány fogalmára. (…) Itt [az MI esetén -HM] egyelőre nem beszélhetünk valamiféle fennálló, relatíve stabil hagyományról, amelybe a személy – akár mint MI kutató, akár mint egyszerű használó, interaktáló személy – alapos képzés során beavatódna.”* Az MI-vel kapcsolatban, ha az általam javasolt szűkebb definíciót alkalmazzuk, akkor megismerésre rendelkezésre álló idő eddig kb. 70 év volt, ami több generációt felölel. Azonban a közelmúlt hatékonyságnövekedése után előállt generatív MI különböző, új célja a megismerési folyamatnak, akkor inkább 4-5 évünk volt csak. Ugyanakkor Márton ellenvetése csak azt jelenti, hogy a vizsgálat nem aktuális, mert nem telt el kellő idő – nem pedig azt, hogy elvileg lehetetlen.

Ennek ellenére úgy vélem, hogy Márton ezen érve az MI megismerésének egy nagyon is valódi akadályát mutatja be, amelyet ki lehetne egészíteni azzal is, hogy az élőlények megismerésével szemben nem áll rendelkezésre a potenciálisan hasonló biológiai alap, amely segít egy másik biológiai személy megértésében. Ha mindezt elfogadjuk, akkor 12-ik fejezet védhető maradványa inkább csak egy attitűd. A megismerési folyamatokból sohasem tudjuk eliminálni a tévedés logikai lehetőségét: a lehetőségek korlátlanok a kritikára. A bizalmi program koncepciója arra hívja fel a figyelmet, hogy amennyiben bármilyen megismerési folyamat eredményét elfogadjuk tudásként, azt olyan elemekbe vetett bizalom által tesszük, amelyekben logikailag akár kételkedhetnénk is. Egyszerűbben, az információ elfogadása tudásként bizalmi aktust igényel. Ebben az episztemológiában a ténylegesen megvalósult MI esetében sem lehet ez másképp, fejezetem pedig arra hívja fel a figyelmet, hogy ha a bizalmi aktust eleve kizárjuk teljesen spekulatív, logikai kényszerítő erő nélküli előfeltevésekre hivatkozva (pl. Gödel tételei, stb.) akkor az MI, mint mesterséges személy elvileg sem megismerhető számunkra; valójában nem megismerési, hanem tagadási programot működtetünk.

56.1)

Márton Miklós hiányolja a „know how” és „know what”, másutt, például Ryle (1945) által „knowing how”/”knowing that”-ként bevezetett megkülönböztetés alkalmazását vagy tárgyalását ebben a fejezetben, mivel azt kulcsfontosságúnak tartja a hallgatólagos tudás tárgyalása szempontjából. Nem értek ezzel egyet. Egyfelől a hallgatólagos tudás koncepciója mindkettő „know” üzemmódban kulcsfontosságú, a számítási folyamatok fogalomkészlete pedig egyáltalán nem különbözteti meg ezeket – a számítás tervezésekor ez a különbségtétel korlátozottan tűnik gyümölcsözőnek, hiszen az enciklopédia szoftver és a kerékpározó robot is hasonló architektúrán és FSM-el működnek, eltéréseik pedig nem így írhatók le.[16]

16 Az informatikus képzésben olykor érvelnek valami olyasmi mellet, hogy: míg a robot vezérléshez tipikusan imperatív nyelvek kézenfekvőek, az enciklopédikusnak nevezhető információ esetén a logikai programozás, például az általam a dolgozatban röviden bemutatott Prolog és újabb iterációi praktikusak. Ugyanakkor Turing teljes nyelvekről lévén szó

56.2)

A Polányi ismeretelméletével ismerkedők egyik szokásos felvetése, hogy az explicit tudás a „know that”, míg a tacit tudás a „know how”-val hozható párhuzamba, azonban a személyes tudás egy részletesebb leírásával, amit a dolgozatomban is bemutattam, ez már nem összeegyeztethető – a szimbólumok manipulációja és az artikulációval létrehozott explicit kijelentések („know that”) a legfejlettebb hallgatólagos képességeinket igénylik, másfelől zongorázás („know how”) elsajátítása is kezdődhet explicit szimbólumokra való fókuszálással.

57)

Héder: *„Polányi azt állítja, hogy a személy nem redukálható a fizikában vagy más tudományágakban ismert fogalmakra (ahogyan azt az ő idejében meglehetősen pozitivista szemlélettel felfogták).” (201.)*
Márton: *Azt hiszem a zárójeles történeti megjegyzés nem stimmel. A Személyes tudás keletkezése – az 1958-as megjelenés előtti évek – idején még éppen csak megszületőben voltak – és általános megrökönyödést keltettek – azok az elgondolások, amelyek az emberi elmét (amelyről most fölteszem, hogy a személy legfontosabb konstituense) fizikai állapotokkal és folyamatokkal igyekeztek azonosítani. Amennyire tudom, a logikai pozitivisták korábban nem köteleződtek el azon nézet mellett, hogy az emberi személyek fizikai vagy más természettudományos „fogalmakra” volnának redukálhatók (ez egyébként itt egy valódi kategóriahiba a szövegben, hiszen egy entitással – mint például egy személlyel – kapcsolatban értelmetlenség fölvetni, hogy redukálható-e vagy sem bizonyos fogalmakra).*

Mártonnak igaza van a zárójeles megjegyzésében, ez hibás megfogalmazás, természetesen nem fogalmakra redukálhatók hanem természettudományos fogalmak segítségével.

Márton történeti megfigyelését nem tartom megalapozottnak. Természetesen az elme redukálhatósága mindig is vita tárgya volt és bármilyen korban kiválthatott megrökönyödést, de a Személyes Tudás keletkezésekor ez a gondolat már sem ritka, sem újszerű nem volt; Polányi pedig kiváló rálátással bírt minderre.

Lássunk egy egy példát: Alan Turing és Polányi egyszerre dolgoztak Manchesterben egy rövid ideig. A Turing teszt történetét kutatóktól tudjuk (Hodges 2007), hogy ekkoriban a 34 éves Turingot leginkább a számítógépek gyakorlati megvalósíthatóságának kérdései érdekelték és csak az 58 éves, már filozófussá átlényegült Polányi unszolására kezdett maga is filozófiainak nevezhető kérdésekkel foglalkozni; ennek az interakciónak az eredménye lett az imitációs játék megfogalmazása (a Turing teszt) is.

1949 október 17-én Manchesterben panelbeszélgetést rendeztek „Discussion on The Mind and the Computing Machine.”[17] címmel (Steiger 2019, Blum 2010) Tudjuk, hogy a szeminárium-sorozatot Dorothy M. Emmet szervezte, de csak sejtjük, hogy a konkrét téma Polányitól érkezhetett: „*Can thinking be mechanical?”*[18]

Jelen volt Turing-on és Polányin kívül Geoffrey Jefferson, aki a British Medical Journal-ban „*The mind of mechanical man*” címmel publikált cikket pár hónappal korábban, 1949 nyarán. Jelen volt a neuro-fiziológus J.Z. Young, aki változatos, élő polipok és békák agyába történő sebészeti beavatkozáskor

ezek felcserélhetők, és a valóságban fel is cserélik őket.

17 https://www.turing.org.uk/sources/wmays1.html

18 Főleg abból, hogy a jegyzetek azzal kezdődnek, hogy Polányi felvetésére válaszolnak a résztvevők.

megfigyelt viselkedésváltozásra alapozta olyan álláspontot vett fel, amely szerint a viselkedés egyes komponense az agy egyes komponenseinek feleltethetők meg. A matematikus Max Newman a fentebb részletesen tárgyalt Gödel érvelés korai változatával próbálkozott („*a jó kérdés az, hogy a gép tud-e Gödel-cikkeket írni*”).

Ezek a természettudósok tehát az elmét fizikai állapotokkal és folyamatokkal látszottak azonosítani vagy legalábbis ennek a lehetőségeit vizsgálták, persze talán filozófiai reflexió nélkül. Jogos felvetés lehetne ezen a ponton, hogy a megnevezett tudósok nem pozitivista filozófusok, pusztán a pozitivizmustól független feltehetőleg redukcionista tudósok. Ám Polányi a pozitivizmust lazábban (is) értelmezte: trendnek, korszellemnek tekintette amely a természettudósok köreiben terjedt.

Ugyanakkor Polányi a direkt összefüggést is jól ismerte a pozitivizmus és a test-elme identitáselmélet között. Úgy tűnhetett akkoriban, hogy bizonyos empirikus pszichológiai eredmények, valamint plauzibilis, megfelelő eszközökkel empirikusan tesztelhető pszichológiai feltevések és a logikai pozitivizmus egymást erősítő viszonyban álltak.

Nyolc évvel a Személyes Tudás megjelenése előtt, 1950-ben jelent meg először Herber Feigl „The mind-body problem in the development of logical empiricism” c. szövege, beszámol többek között a bécsi kör elmúlt 25 évének három monista megközelítéséről és említi Edwin Boring 1933 munkáját is, amelyben amaz identitáselméletként hivatkozik az egyik álláspontra. Feigl fontos alakja volt a logikai pozitivizmusnak, egyben az iskola emlékezete is, amennyiben szintetizáló, összefoglaló, történeti szövegek, szerkesztett kötetek, visszaemlékezések is megjelentek tőle, egészen az 1980-as évekig. Feigl 1958-ban tovább mélyítette elköteleződését a The ‘Mental’ and the ‘Physical’”(1958) c. esszében. (Később ezt részben finomította, részben visszavonta, de ez már a Személyes Tudás megjelenése után történt. Egyébként Polányival levelezett is a ’60-as években).

A híres amerikai pszichológus, Boring hatása nem meglepő a bécsi kör filozófusaira: a kapcsolat a pszichológusok és a filozófusok között élő volt – például Bühler vagy Popper kapcsán – de még pontosabb volna úgy fogalmazni, hogy a diszciplínák még nem váltak szét, például maga Boring is csak évekkel később érte el, hogy kiválhasson a filozófia tanszékből.

Feigl 1981-es visszaemlékezéséből tudjuk, hogy ő Schlick-et és Waissmannt „fenomenológiai pozitivista” álláspontban látta már az 1920-as években.
„*Under the influence of Carnap and the early Wittgenstein, Schlick and Waismann were converted to a sort of phenomenalistic positivism during the middle twenties. Their brilliant and powerful arguments overwhelmed me temporarily (...)”*(Feigl 1981, pp. 9–10)

Mivel Polányi bátyja, Károly a Galilei kör első elnöke volt, Polányinak jó hozzáférése volt a budapesti majd később a bécsi pozitivistákhoz is. A Polányi archívumban található feljegyzések és a megjelent műveiben található hivatkozások okán is megalapozottan feltételezhetjük, hogy követte Feiglt és Schlick-et is; azt is láttuk, hogy a manchesteri agytudósok megközelítése milyen volt, és hogy ennek is voltak előzményei, pl. Boring[19] munkája amely a pszichológián belül nagyon ismert volt. Nyugodtan mondhatjuk tehát, hogy az 1950-es évek végére Polányi már évtizedes távlatban több fronton találkozik azzal amit Márton így jellemez *„az emberi elmét fizikai állapotokkal és folyamatokkal igyekeztek azonosítani”*.

19 Egyébként Boring maga még korábbi előképekre mutat vissza, de valahol meg kell állnunk a történetben...

Érdekességként megjegyezhető, hogy Polányi kivételesen jó beágyazottsága korának filozófiai életében, valamint személyes kapcsolata Turing-gal lehetővé tette számára, hogy már a Személyes Tudásban megfontolja és kritizálja a Gödel-demarkációt a számítógépek ellen:

*„Automatizálás általában*

*Az axiómák megsokszorozása, amit Gödel fedezett fel, nyilvánvalóan bizonyítja, hogy egy logikai következtető gépet működtető személy olyan tudásmennyiséghez juthat informálisan, melyet az ilyen gép műveletei nem tudnak bizonyítani, jóllehet ezek a műveletek sugallják, hogy ehhez a tudáshoz könnyű hozzáférni. Ez azt bizonyítja, hogy az értelem képességei meghaladják egy logikai következtető gép képességeit. De még szembe kell néznünk azzal az átfogóbb problémával, amit az automatikus célzóberendezések, a robotpilóták stb. vetnek fel, vagyis azok a gépek, amelyeknek a teljesítménye jóval meghaladja a logikai következtetésekét." (*Személyes Tudás, 3-ik rész 9. fejezet).
Polányi tehát már nagyjából 70 évvel ezelőtt észrevette, hogy sem az elmét, sem a – mai szóval – autonóm robotokat nem lehet a logikai következtető géppel azonosítani, ha a limitációk megállapításáról van szó.

58)

Márton: *„nem egészen világos sem az, mit ért itt [206.oldal a mérnöki tudományok teleologikus jellegéről] a szerző a tudományok természetén, sem az, mifajta teleológiai elemekről volna itt szó. Amennyire tudom, manapság nagyjából konszenzus övezi azt az álláspontot, hogy a biológia nem tartalmaz valódi, kiiktathatatlanul teleologikus magyarázatokat, mint ahogy azt is, hogy a biokémián és a sejtbiológián keresztül egységesíthető a kémiával és a fizikával, a szó egy bizonyos értelmében redukálható ezekre."* Ezt az álláspontot reduktív fizikalizmusnak nevezik és valóban jól ismert, de azért az erős kijelentés, hogy konszenzus övezné. A tudomány egységesítésére történt utalás miatt Polányi valószínűleg pozitivistának is nevezné. Ugyanakkor kétségtelen: ha valaki a reduktív fizikalizmust tartja helyes álláspontnak, akkor a kérdéses szakaszt a dolgozatomban, amely az ontológiailag emergens számításokról elméletét mutatja be, tévesnek kell tartania.

59)

Jogos Márton kritikája a rossz megfogalmazású mondatommal kapcsolatban. „*A gépek nem képesek X-re*" helyett annak egy rokon érvelését, a „*A gépek nem reldelkeznek X tulajdonsággal*" kellett volna használnom, amelyet azonban be sem vezettem. Itt nem képesség, hanem tulajdonság alapú demarkációról van szó. A behelyettesítés eredménye így „*A gépek nem rendelkeznek biológiai alapokkal*". Ez roppant plauzibilis[20], ezért ha a biológiai alapokkal rendelkezés relevanciája igazolt lenne, akkor a felvázolt elmélet valóban léket kapna.

**Válaszok Márton „további apró részletek" alatt felsorolt kritikáira**

60)

Az aluldeterminánltság fogalma egy pontosabb definíciót valóban megérdemelt volna, de úgy vélem, hogy az 50-ik oldalon leírt állapot azért érthető „*De az, hogy mely tulajdonságokat hagyjuk el, és melyeket reprezentáljuk a modellben, döntés kérdése, és ez a döntés általában aluldeterminált: hacsak nem egy matematikai modellt képezünk le egy másikba, általánosságban nem tudhatjuk, hogy mely*

20 Mivel itt a gépek felépítéséről van szó, a gépeket megalkotó ember, mint diakrón biológiai ok, nem számít.

*tulajdonságokat kell meghagynunk és melyeket szabad eldobnunk ahhoz, hogy a modell még épp megfeleljen a céljainknak, például, hogy előrejelzéseket olvashassunk le róla. A fentiek világossá teszik azt is, hogy ugyanannak a modellezettnek beláthatatlanul sok modellje létezhet: ez a szám a modellező által meghozható döntések kombinációinak száma."* A dolgozatban végig ebben az értelemben használom: a döntéshozás terében minden dimenzióban ki kell választanunk egy értéket, de bizonyos dimenziók vagy azon belül tartományok esetében semmilyen külső szempont nem választ ki vagy zár ki egy értéket, azaz nincs determináló körülmény – de mégis választunk. Az episzemikus távolság növekedésével (1.2 fejezet) az aluldeterminált ság egyre nő. Ezt a tézis járom körül a dolgozat különböző pontjain (néha esetlen megfogalmazásban). Azt is fenntartom, hogy a terv és a végtermék között is kölcsönösen aluldeterminált viszony áll fenn: ha a tervet implementáljuk, annak során adott pontokon úgy kell döntéseket hoznunk, hogy nincs megalapozottságra mód, és ha a végtermékből szeretnénk visszafejteni azt, hogy mi volt a végtermék terve vagy modellje, akkor ugyanez a helyzet. Úgy vélem, hogy ez konzisztens a tudományfilozófia általi aluldeterminált ság fogalommal; és szándékosan nem a specifikáció fogalmáról írok. Mindez egy madártávlati leírása annak a helyzetnek, amit a 34) pontban vázoltam fel a számítás és a modellje között.

61)

A kontraktárius etikára vonatkozó kritikákra: köszönöm szépen a jogos észrevételt a meta-etikai szó hibás használatára vonatkozóan. Az általam az alku érdekében megadott, a javasolt szabály praktikusságát hangsúlyozó érv valóban a hasznosságra apellál, de úgy gondolom, hogy ez nem ellentmondás: bármilyen alkufolyamatban találhatunk a javaslat hasznosságára hivatkozó érveket. Ám ahhoz, hogy az érvet valódi szabály-utilitarianistaként érvként fogadhassuk el, nem elegendő posztulálni a hasznosságot; meg is kellene indokolni valamely kalkulus segítségével, hogy valóban ez a legjobb szabály-alternatíva, amit nem teszek meg. Még fontosabb, hogy is az input elégtelensége és a nagy fokú bizonytalanság miatt nem is tartom kivitelezhetőnek e számítást, ezért javaslom e kevesebb igazolt feltevéssel is működő modellt.

62)

A modularitás fogalmát az 2.3.1-ben mutatom be az enkapszuláció fogalmára visszavezetve. Ezzel a modularitás fogalommal operálva a közvetlenül az idegsejtekre ható, vérkeringéssel utazó kémiai transzmitter, pl. hormonok működése sem összeegyeztethető. De úgy gondolom, hogy még egy kevésbé precíz modularitás fogalomnak sem felel meg ha, „(Márton:) *léteznek különböző funkciókért felelős, többé-kevésbé jól beazonosítható agyterületek*", mert még a köznyelvi modularitás fogalomnak – amit a bútoroktól kezdve az egyetemi tanterveken át az házgyárakig bezárólag mindenfélére használnak - része a modulok világos lehatárolhatósága.

63)

Egyáltalán nem tartom meggyőzőnek Márton, a sajátommal ellentétes jellemzését a klasszikus és neoklasszikus közgazdaságtani iskolákról. „*E koncepciókban szereplő egyén valóban ideálisan racionális, az instrumentális racionalitás értelmében, azaz mindig a célját a számára legvalószínűbbnek tűnő módon elősegítő cselekvés mellett fog dönteni. Azonban, amennyire tudom, ezen egyéni cél mibenlétére vonatkozólag nem tételeznek föl semmit e közgazdaságtani elképzelések, az tehát bármi lehet. Vagyis az egyén éppúgy szakadatlanul igyekezhet saját hasznát maximalizálni, mint mondjuk altruista céljait megvalósítani – viselkedése e megközelítés szerint leírható az általuk preferált közgazdasági modellekkel."* Tény, hogy önmagában az instrumentális racionalitás koncepciója neutrális arra nézve, hogy mi az mikroökonómiai ágens milyen célt választ ki. De ha őszintén nem lenne ezen

felül is előfeltevése az említett közgazdaságtanoknak a tipikus mikroökonómiái ágens céljaira nézve, akkor sem egy alapkamat emelést, sem egy adócsökkentést, vagy bármilyen más szakpolitikai döntést nem lehetne indokolni: hiszen lehet, hogy az ágensek ki nem állhatják, ha a megtakarításaik túl könnyen nőnek, vagy magasabb adókat preferálnak és nem pedig alacsonyabbakat.

64)

Márton: „*Nem világos, milyen értelemben használja itt [a 2. fejezetben – HM] a szerző a normativitás fogalmát. A kontextus alapján úgy tűnik, leginkább az instrumentális normativitás értelmében, amit általában nem tekinthetünk valódi normativitásnak, hiszen pusztán egy cél-eszköz viszonyt ragad meg. Ezt fontos volna jelezni, mert enélkül könnyen az a benyomása támadhat az olvasónak, hogy itt hasonló értelemben van szó normativitásról, mint mondjuk a morálfilozófiában.*
Mivel a deskriptív tudományos kimenetekhez akarom hasonlítani a mérnöki tudomány kimeneteit, egyik értelmezésben sincs ellentmondás, így én itt nem látok problémát. Hacsak nem az Márton állítása, hogy az instrumentális-normatív leírás valójában deskriptív leírás. Akárhogy is, nem pusztán cél-eszköz leírásról van szó, ez tévedés: elképzelhető, hogy a mérnöki tudományt körüllengő mítoszok miatt valaki úgy értelmez egy tervet, mint egy cél-eszköz leírást: „amennyiben a cél X funkció megvalósítása, úgy Y célra vezető megvalósítás”. Ez túl optimista értelmezés, a műszaki tervek lehetnek célszerűtlenek, de megvalósíthatók, így működésképtelen műterméket kapunk; valamint irracionálisak és megvalósíthatatlanok is. Továbbá az építészettől az ipari termékeken át a szoftverrel bezárólag igen fontos szerepe van a mérnöki tervekben az esztétikának, amelyet a terv normatíve előír, de nem cél-eszköz viszonyban kezel.

65)

Köszönöm szépen az észrevételt, valóban hibás a mondat: nem a számok ontológiai kategóriájára akartam utalni, hanem a számokra.

66)

A 6. fejezettel kapcsolatban az MI megmagyarázhatóság és a humán döntéshozó megmagyarázhatóság közötti kettős mérce abban áll, hogy a 9.3-as fejezetben bemutatott legkevésbé pontos lokális magyarázat modell is megadja azokat az input token/pixel/adatpont + mesterséges neuron aktivitás párokat vagy input token/pixel/adatpont + modell tulajdonságokat (a gépi tanulás *feature* fogalmára utalok), amelyek kritikusak a predikcióhoz. A szociálpszichológia által meggyőzően bemutatott, a körülmények általi, a döntéshozó által olykor fel sem ismert meghatározottság meg sem közelíti ezt a szintet. A másik nagy különbség, hogy az említett szociál/morálpszichológiai kísérletek a jelenség létét tudják alátámasztani, a XAI elvárása pedig a folytonos, minden meghozott döntéshez mellékelhető magyarázat.

**Válaszok E. Szabó Lászlónak**

67)

E Szabó László címre vonatkozó észrevétele jogos, az valóban szűkebb, mint a dolgozat tartalma, tehát az MI filozófia új hullámai indokoltabb lenne.

68)

A mesterséges entitásokra vonatkozó episztemikus státuszunk: úgy vélem, hogy a lényegben egyetértünk. Az intuíció az, hogy amennyiben a PVC molekula mérnöki tervezés eredménye, a tervek, amelyek alapján megvalósult, tudástöbbletet képviselhetnek a PVC molekulát megismerni vágyók számára. Természetesen ez nem igaz azokra, akik a tervet nem ismerik. Ebből tehát az következik, hogy az emberi fajnak, csoportszinten tudástöbblete lehet a mesterséges entitásokkal kapcsolatban, amennyiben feltehetjük, hogy a tudást megosztják.

69)

A technológiába zártság kapcsán jogos E. Szabó észrevétele arról, hogy a fogalmat érdemes tovább cizellálni. Egy megközelítés, amelyet még csak vázlatosan dolgoztam ki, az idődimenzió világos kiemelését, valamint a kontroll fogalmának oksági láncokra való lefordítását javasolja, de úgy, hogy nem feltételez monokauzalitást.

Így világossá válhat, hogy a technológia megalkotásánál a fontos döntést meghozók nem kontrollálják szó szerint a későbbi generációkat – például azért, mert már nem is élnek. E helyett, például az MI-vel előállított technológiába zártságot egy olyan oksági láncként képzelhetjük el, amely viszonylag lokalizált, akár egyszemélyi döntésekből indul, és az idő előrehaladtával százmilliónyi ember életére van hatással. Minden felhasználó saját maga is befolyásolja ezt a hatást, mivel ko-kauzális viszonyba kerül a mesterséges ágenssel: a viselkedését részben maga, részben az MI okozza. Az egyének elvileg képesek úgy alakítani a helyzetet, hogy felülírják, kiküszöböljék az mesterséges ágens korábban eldöntött működési módjából adódó következményeket, de a tervező csapattól induló oksági lánc ezen a ponton annyi részre ágazott el, hogy valódi többségi társadalmi viselkedésváltozás szükséges ehhez a felülíró művelethez. Olyan előre haladott állapot is elképzelhető, mint például a globális felmelegedésnél, hogy a technológiahasználó saját hozzájárulása már nem számít, mert a mozgásba lendült természeti folyamatokra nagyon kis ráhatással tud bírni.

70)

Köszönöm szépen E Szabó László észrevételét arról, hogy az elmefilozófiai fizikalizmust és a funkcionalizmust felcserélhetően használom, ez valóban hiba amelynek kiküszöbölése a két releváns fejezetet jobbá tenné.

Héder Mihály
Budapest, 2026. május 21.

**Felhasznált Irodalom**


Bennett, C. H. (2003). Notes on Landauer's principle, reversible computation, and Maxwell's Demon. Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics, 34(3), 501-510.

Blum, P. R. (2010). The Immortality of the Intellect Revived. Michael Polanyi and his Debate with Alan Turing. *Knowing and Being: Perspectives on the Philosophy of Michael Polanyi, Newcastle upon Tyne*, 173-184.

Boring, E. G. (1933). *The physical dimensions of consciousness*. The Century Co. New York.

Collins, H. (2012). Symbols, strings and social cartesianism: response to Mihaly Heder. *Polanyiana 21*, 59-70.

Collins, H. (2019). *Tacit and explicit knowledge*. University of Chicago press.

Collins, H. M., & Kusch, M. (1998). *The shape of actions: What humans and machines can do*. MIT press.

Donath, J. (2020). Ethical issues in our relationship with artificial entities. *The oxford handbook of ethics of AI*, 53-73.

Gulick, W. B. (2023). Machine and person: reconstructing Harry Collins's categories. AI & Soc 38, 1847–1858 . https://doi.org/10.1007/s00146-020-01046-3

Feigl, H. (1950). The mind-body problem in the development of logical empiricism. *Revue internationale de Philosophie*, 64-83.

Feigl, H. (1958). The 'mental'and the 'physical'. *Minnesota studies in the philosophy of science*, *2*(2), 370-497.

Feigl, H. (1981). *Inquiries and Provocations: Selected Writings, 1929-1974* (Vienna Circle Collection 14), R. S. Cohen (ed.), Dordrecht/Hingham, MA: D. Reidel/Kluwer.

Fritzman, J. M. (2024). Collapsing strong emergence's collapse problem. *European Journal for Philosophy of Science*, *14*(2), 24.

Héder, M. (2012). Explicit Knowledge in the Philosophies of Harry Collins and Michael Polanyi. *Polanyiana 21, 45-58.*

Héder, M. (2013). A gyenge emergencia kritikája Polanyiana 22 : 1-2 pp. 18-25. , 8 p.

Héder, M. (2014). Emergencia és hallgatólagos tudás a gépekben. PhD disszertáció, BME.

Héder, M. (2017). Emergent Computing and the Embodied Nature of Computation. *Polanyiana* 26.

Héder, M. (2020). *Mesterséges intelligencia: filozófiai kérdések, gyakorlati válaszok*. Gondolat.

Héder, M., & Paksi, D. (2012). Autonomous robots and tacit knowledge. *Appraisal*, *9*(2), 8-14.

Héder, M., & Paksi, D. (2019). A Criticism of Weak Emergence. *Polanyiana*, *28*, 1-2.

Héder, M. & Paksi, D. (2020). *A személyes tudásról*. Akadémiai Kiadó.

Hodges, A. (2007). Alan Turing and the Turing test. In *Parsing the Turing test: philosophical and methodological issues in the quest for the thinking computer* (pp. 13-22). Dordrecht: Springer Netherlands.

Hvorecký, J., Marvan, T., & Polák, M. (Eds.). (2023). *Conscious and unconscious mentality: examining their nature, similarities, and differences*. Taylor & Francis.

Jefferson, G. (1949). The mind of mechanical man. *British Medical Journal*, *1*(4616), 1105.

Kodaj, D. (várhatóan 2026). *Mesterséges unintelligencia.* Eötvös Kiadó.

Königs, P. (2025) The negativity crisis of AI ethics. *Synthese* **206**, 277. https://doi.org/10.1007/s11229-025-05378-9

Lenat, D. B., Prakash, M., & Shepherd, M. (1985). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI magazine*, *6*(4), 65-65.

Lighthill, J. (1973). 'Artificial intelligence: a general survey', *Science Research Council, Artificial Intelligence: A Paper Symposium*, London: SRC, 1973, pp. 1–21.

Penrose, R. (1999) The Emperor's New Mind: Concerning Computers, Minds, and The Laws of Physics. Oxford University Press.

Polányi. M. (1958) *Personal knowledge: Towards a post-critical philosophy.* Routledge & Kegan Paul.

Polanyi, M. (1968). Life's Irreducible Structure: Live mechanisms and information in DNA are boundary conditions with a sequence of boundaries above them. *Science*, *160*(3834), 1308-1312.

Ryle, G. (1945, January). Knowing how and knowing that: The presidential address. In *Proceedings of the Aristotelian society* (Vol. 46, pp. 1-16). Aristotelian Society, Wiley.

Shagrir, O. (2022). *The nature of physical computation*. Oxford University Press.

Sóstai, Z. (2024). Empirical Constraints and the Computational Unpredictability of Physical Systems: Exploring the Physical Church─Turing Thesis and the Halting Problem 131 p.
Eötvös Loránd Tudományegyetem (ELTE), Filozófiatudományi Doktori Iskola, konz.: E. Szabó László
Disszertáció benyújtásának éve: 2024, Védés éve: 2025 Megjelenés/Fokozatszerzés éve: 2025

Steiger, A. (2019) The Imitation Game: Polanyi vs. Turing and Why it Matters to Human Dignity. The Polanyi Society website, draft.

Turing, A. M. (1937). Computability and λ-definability. *The Journal of Symbolic Logic*, *2*(4), 153-163.

Turing, A. M. (1987). Computing machinery and intelligence (1950). *Mind*, *59* (236), 33-60.
Xie, J. (2021). An explanation of the relationship between artificial intelligence and human beings from the perspective of consciousness. *Cultures of Science*, *4*(3), 124-134.