Opponensi vélemény
Héder Mihály: **Az MI etika új hullámai**
című MTA Doktori értekezéséről[1]

Z. Karvalics László, 2026 január

1.

Kezdjük a cím kulcsszavával, az *MI etikával* kapcsolatos néhány megfontolással – különös tekintettel arra, hogy a terminus maga is egy, a diszciplináris térben zajló kettős fogalmi és narratív helyfoglalás lenyomata. Ha az egyik oldalról nézünk rá, akkor a „klasszikus" etikai hagyomány és apparátus kiterjesztése a társadalmi valóság egy új, technológiai, gazdasági és kulturális elemeket ötvöző jelenségvilágára, ha a másik oldalról, akkor a mesterséges intelligencia valóságátalakító hatására – és annak újként felmerülő dilemmáira - adott társadalomelméleti és bölcseleti reflexió egyik segédcsapata (amely leginkább a szabályozási és standardizációs igényeket elégíti ki, gyakorlati szempontok sorát „termelve" meg számukra).

Szögezzük rögtön le: a szerző e második közelítéstípus képviselője. S noha elmélyülten tárgyalja a szabályozás és a kanonizáció kérdéseit, valójában az etikai tudás gyakorlati „lefordíthatósága", alkalmazhatósága, a tervezési folyamatokba való beolvasztása a „küldetése": legizgalmasabb javaslata, az *episztemorális tervezés* elsősorban ezt a célt szolgálja. S nagyon is hasznosnak és hatékonynak bizonyul abban is, hogy meghaladja a hagyományos követelést, hogy *„etikus viselkedést implementáljunk számítógépes programokba"* (Málik, 2025), (amelyek, ugye ab ovo nem viselkednek, hanem programkódot futtatnak). És bizony itt is jelentősége van a „megvalósíthatósági résnek" („implementation gap"), amely arról a nehézségről szól, hogy az MI alkotók köreiben megértett és konszenzusosan el is fogadott normák implementálása is nagyon nehéz programkóddal (13).

Tízből kilenc kutató bizonyára egyetértene azzal, hogy az „MI etika" (és rokon kifejezései, valamint előzmény-fogalmai[2]) *metonimikusan szerkesztett* szakszavak: valójában nem a mesterséges intelligenciának van etikája, amit a birtokos szerkezet sugallna, hanem az „MI etika" rövid változata egy hosszabb definíciónak: *„a mesterséges intelligenciával kapcsolatos valamennyi etikai kérdés vizsgálata, elsősorban a felelősségteljes tervezés, fejlesztés és felhasználás problémakörére érzékenyen"* - vagy a szerző még frappánsabb verzióival: *„az MI fejlődésével párhuzamosan felmerülő etikai kérdések"* (9.o.), *„az MI megbízhatósága és etikus alkalmazása"* (162.o.) *„az MI etikus tervezése"* (214.o.). Hasonló

[1] Az értekezésből vett idézeteket és illusztrációs részeket minden esetben zárójelbe tett oldalszám alapján adom meg.
[2] AI Ethics vagy Ethics of AI, Ethics of Autonomous and Intelligent Systems, Robot ethics (Roboethics). Information ethics, Computer ethics, Machine ethics, Cyberethics

megoldásot választ Ansari (2023): *„az MI etikai összefüggései”* és Fási (2025): *„etikai kihívások és megoldások az MI alkalmazásában”*, majd *„az MI etikai integrációja”*.[3]

Amikor az MI etikából „etikus MI” lesz, egy lépéssel közelebb kerülünk ahhoz, hogy a rövidség a leegyszerűsítés felé sodorja a fogalomalkotást. Nem kell hozzátenni, hogy itt is *„az etikai megfontolások érvényesítéséről van szó olyan cselekvőhálózatokban, amelyekben MI elemek is vannak”*, vagy röviden: „*etikusan tervezett MI”* a voltaképpeni tárgy. Mégis minden alkalommal, amikor az „etikus MI” szakszöveg, tudományos értekezés címében jelenik meg, az első kihívás megbizonyosodni arról, hogy a szerző komolyan gondolja-e, hogy az MI az etikus, vagy metonimikusan használja a kifejezést.[4]

Már csak az ilyen helyzetek miatt is szükséges némiképp inkriminálni az MI etika és az etikus MI kifejezést. De van rá két további, jó okunk is.

Nyilvánvaló, hogy minél ingoványosabb a definíciós altalaj, annál könnyebb az MI-fejlesztésben és felhasználásban serénykedő szereplőknek a folyamatokba való lényegi beépítés helyett imitáltan felmutatni az „etikus viselkedést”, amelyet a greenwashing mintájára megalkotott „AI ethicwashing” (MI-etikai mosdatás) fogalma igyekszik lefedni.[5] S minél elterjedtebb az etikai mosdatás, az annál jobban megváltoztatja az AI-etikával kapcsolatos érzékelésünket, és szakadatlan nyomást gyakorol a trivializálásra, ami viszont könnyen vezet az AI-etikai alapelvekkel való szembefordulóshoz (ethicbashing) (Schultz, 2024).

Másfelől e fogalmak látványosan kiszabadultak már a jelentések e szabatos és többé-kevésbé konszenzuális ketrecéből, és paradiskurzusok sora született meg azóta, hogy *„Putnam sürgette a robotok morális státuszának tisztázását”* (18). Ma már egyre több helyen olvasunk arról, hogy mik a csapdái a gépi döntéshozatalnak, hogyan tehetők felelőssé

---

[3] A terület egyik, 2021-ben indult szakfolyóirata ezt a dilemmát úgy oldja fel, hogy finom distinkcióval az „AI and Ethics” címet választotta az AI Ethics helyett, amelynek tartalma a küldetésnyilatkozatuk szerint „az MI fejlesztése során következményként felmerülő etikai, szabályozási és politikai kérdések” vizsgálatát fedi le (ethical, regulatory, and policy implications that arise from the development of AI). https://link.springer.com/journal/43681/aims-and-scope

[4] Prisznyák Alexandra tanulmánya az Etikus AI-ról például azzal a mondattal kezdődik, hogy a *„GPT–3 2020-ban gondolkodó robotként definiálta magát”* (Prisznyák, 2023:169). Az olvasó ilyenkor megretten, mert tart attól, hogy a gépnek (robotnak, szoftvernek) episztémikus ágenciát tulajdonítva az etikus MI-t majd nem metonimikusan érti a kifejtés. A konkrét esetben gyorsan megnyugszunk, hogy szerencsére nem erről van szó, a tanulmány alapos, izgalmas és időszerű. Mégis, ott motoszkál a gyanú, hogy miért tett engedményt a PhD-jelölt szerző a bulvárnak, miközben bizonyára tisztában van vele, hogy nemcsak etikai, hanem minden további szakmai kontextusban teljesen irreleváns, hogy egy program a neki feltett kérdésre válaszolva hogyan határozza meg önmagát. Azt, hogy egy nagy nyelvi program „micsoda, mit gondoljunk róla”, azt nem tőle kérdezzük meg, hanem magunk alkotunk rá definíciót, modelleket, „verbális stratagémát”.

[5] A kifejezés 2019-ben került a köztudatba, amikor az EU szakértői testületet állított fel az MI világához szükséges etikai alapelvek összerendezése érdekében.A testület egyik tagja, Thomas Metzinger bírálta szenvedélyesen az álszent gyakorlatot, s hamarosan csatlakozott hozzá az MI-iparág „élő lelkiismerete”, Karen Hao is (Hao, 2019). A korábbiak értelmében itt is azt látjuk, hogy az „AI ethicwashing”-hoz képest mennyivel pontosságra törekvőbb az *„ethicwashing in AI”* formula használata (https://aiethicslab.rutgers.edu/glossary/ethics-washing/.) A „digital ethicwashing” változat az MI-felhasználás mellé a mindennemű, MI nélküli adatkezelési szabályokat rendeli – és mostanra már kiterjedt irodalma nőtt.

döntéseikért a gépek, s ha van felelősségük, jogaik is vannak-e. S itt már nem elég elintézni azzal a kérdést, hogy természetesen a döntést is metonimikusan értelmezzük, vagy a döntést idézőjelbe tesszük, hogy érzékeltessük, hogy ez azért másfajta dolog, nem elsődleges értelmében kell használni.[6] Minden konkrét esetben pontosan meg kell tudni mondani, hogy az episztemológiai autoritás átruházásának (Golden, 2025) melyik formája, szintje valósult meg, s *pontosan mit értünk döntés alatt*.

Emiatt érzem problematikusnak, hogy Héder érvelésében a morális státusz kérdése azért vetődik fel olyan élesen (182), mert *„az MI az ember helyébe lép társas interakciókban, döntéshozásban”*. Metonimikusan, ugye. A szabatos megfogalmazás az volna, hogy „egyre több olyan helyzetet ismerünk, ahol a rutinszerű döntési helyzetekben *MI segíti jól definiált döntési feladatok végrehajtását*, és ahol korábban face to face humán partnerrel folytatott párbeszéd vagy interakció kvázi-alanyaként jelenik meg, teljesítménye magasrendűsége okán bizonyos szituációkban elfeledtetve azt, hogy nem emberi cselekvő).

Talán épp ezért lehet szerencsés az etika kontextusában a MI helyett is egy tágabb kategóriát használni, ahogy Donath (2020) teszi, aki *a „mesterséges létezőkkel”* (artificial entities) kapcsolatos *etikai kérdéseket* taglalja.[7]

Egyáltalán nem felesleges kitérő ez a fogalmi bakafántoskodás. Teljes mértékben egyetértek a szerzővel, hogy *„az MI etika legnagyobb kihívása a nyelvi, módszertani, ontológiai és egyéb rejtett előfeltevések mentén kialakuló számtalan hajszálrepedés és az így létrejött, egymáshoz nem magától értetődően illeszkedő, de egyenként roppant ígéretes részeredmények egységes kezelése”* (13.o.), s hogy *„a jelenleg folyó kanonizációs törekvések egyik fontos eleme a terminológiai tisztázó munka”* (14.o.) Emiatt végtelenül fontos, hogy a metonimikus fogalomhasználat letisztuljon. Erre egyébként sok oldalról indultak meg erőfeszítések. Az MI etika helyett az *„MI rendszerek etikus dizájnját”* (ethically designed AI systems) javasolja használni Beale (2022). A villamosmérnökök emblematikus szakmai szervezete, az IEEE a szabványosítási törekvéseiben is már *„etikusan vezérelt”* (ethically driven) (Prestes, 2021) és *„etikai szempontokhoz igazított”* (ethically aligned) rendszerekről beszél (Houghtaling, 2024).

A szerző – Lem nyomán izgalmasan mozgósított – „konstruktőri nézőpontjához” a fenti meghatározások már alkalmasabbnak, pontosabbnak tűnnek, mint az MI etika. Más kérdés, hogy Héder Mihály a *limitációk* felől értelmezi a konstruktőr megismerési helyzetét (14-15.o.), és nem érinti azt a kérdést, hogy miképpen volnának *normatívan* beilleszthetőek a tervezési és irányítási folyamatokba például egy progresszív társadalom-vagy jövőkép elemei[8]. Pedig ebből az irányból is lecserélődik az „MI etika”-kifejezés: Floridi (2018) már azt keresi sok-sok szerzőtársával, hogy *„a mesterséges intelligenciát okosan használó jó*

---

[6] Szerzőnk is így tesz: *„triviálisan értelmezzük, hogy akár egy termosztát stb. is képes „döntést” hozni* (103).

[7] S eközben (Donath, 2020:53-54), a korábban mondottak szellemében, kettéválasztja az általános, kapcsolat-alapú etikai kérdéseket (ethical issues in our relationships with artificial entities) a tervezési folyamat és a működtetés során felvetődőktől (ethical issues in the design and deployment of artificial entities).

[8] Héder másfajta normativitást tárgyal alaposan és felkészülten (23-24.o.): azt, ami a tervezéskor a még nem létező technológiai megvalósítástól elvár valamit.

társadalomnak milyen lenne az etikai keretrendszere” (Ethical Framework for a Good AI Society).[9]

A rigorózus fogalmi tisztázásra való igény nem önmagában áll. Egy friss tanulmány úgy látja, hogy az MI-etikának a mindennapi életben egyre több ponton csődöt mondó hagyományos és domináns megközelítéseivel és keretrendszereivel szemben is szaporodnak a fenntartások és a kifogások, és erős az igény ezek megújítására. És ha a gyakorlati-alkalmazási meg a tipikus regulációs kihívások körmére nézünk, sok esetben jutunk vissza a fogalomhasználati zavarok kiigazításának szükségességéhez (Afroogh, 2026). Másképp: a fogalmi tisztázás segíthet a ráépülő gondolati felépítmények stabilizálásában is.

Emiatt gondolom azt, hogy az értekezés fokozatos felépítésében meghatározó bevezető szakaszok, amelyek az „MI szkepszist” bírálják vagy az „MI kritikájának” kritikáját adják, véleményem szerint elmarasztalhatóak a nem feltárt metonimicitás, s a diskurzus ezzel összefüggő leegyszerűsítése miatt.

2.

Az, ahogyan Héder Mihály az MI-kritikát (máshol: *MI-szkepszist*, megint máshol: az *MI felett mondott negatív ítéletet*) kezeli, gyakorlatilag *egyetlen, leegyszerűsítő narratívára* érzékeny[10]:

---

[9] Ez és az ehhez hasonló, a Humanistic AI, AI for Social Good irányhoz sorolt megközelítések természetesen túlmutatnak az értekezés választott tárgyán, de filozófiai szempontból (is) nagyon sok relevanciát hordoznak.

[10] A 2020-as könyvben még három kritikatípus szerepelt, érdemesnek látom érinteni az értekezésbe át nem emelt kettőt. Az MI teljesítménye „puszta imitáció” (Héder, 2020:13) állítást *nem tartom kritikai narratívának*, hanem a *„mi az MI-nek az emberi kognícióhoz sorolt feladattípusok kimeneti paramétereivel megegyező teljesítménye mögött álló műveleti képesség forrása”* kérdésre adott válasznak. Az állítás semlegessége, leíró volta akkor illan el, amikor a „több, mint puszta imitáció” tézissel, vagy a „lám, az intelligens entitásoknak új fajtája született” állítással szemben kell megerősíteni, hogy *nem*. (S bár Héder Mihály nem foglal határozottan állást a kérdésben, úgy érzem, ő nem lenne ennyire eltökélt a „nem” kimondásában). Önmagában tehát a „puszta imitáció” nem kritikai narratíva, hanem ellen-narratíva. Nem az MI kritikája, hanem az MI-vel kapcsolatos, helytelen(nek vélt) állítások vitatásakor kerül, kényszerűen, kritikai szerepbe. S lám, abban a pillanatban mindennek újra jelentősége lesz, hogy az MI jelenlegi hullámának teraszain a jövőbeni fejlesztésekről esik szó, az általános célú MI (AGI) és a szuperintelligens MI (ASI) narratíváival. Csakhogy ha a legtökéletesebb imitáció is csak olyan, mintha intelligencia állna mögötte, de nem áll, akkor az AGI és ASI logikai lehetőségként tárgyalható ugyan, de valóságos fejlesztésüknek se értelme, se esélye nem látszik most. Dagdelen Demet szavaival „*A ChatGPT természetesen nem mutat valódi intelligenciát, de olyan hihető utánzata az emberi gondolkodásnak, amilyet korábban elképzelhetetlennek tartottunk* (Demet, 2025). Illúziógépezet (Šekrst (2025). Héder szerint azonban az MI valódi története ott kezdődik, hogy megjelentek az első, „az emberi intelligencia sebességéhez mérhető, *potenciálisan általános célú,* intelligens viselkedést mutató” megoldások (10.o.). Vagyis nála már a puszta imitáció is az AGI lehetőségét tételezi – miközben kellőképp óvatos, hogy ne nevezze intelligensnek a megoldást, csak „intelligens viselkedést mutatónak” (vagyis imitációnak). Most akkor a tudóstársadalomnak az a része, amely az egész AGI-diskurzust (s ennek révén implicite a közveszélyes ASI-t) – egy irányba mutató, de változatos érveléssel – elutasítja, az mivel szemben is szkeptikus-kritikus? Nem lehet megkerülni, hogy állást foglaljunk „az MI intelligenciaforma-e vagy nem” kérdésében. Héder (2020:18) felvetés-szerűen képviselt álláspontja („a mesterséges neurális hálók egyre komolyabb megvalósításaival kapcsolatban merülhet fel, hogy a hasonlóság az emberi aggyal nagyon is megalapozott”) azonban kicsusszan a dichotómiából, hisz mindebből az következne: *talán, nem tudjuk*. De ezt is csak azért állíthatja, mert nem elég pontos a *hasonlóság*-meghatározás. Mi hasonlít? Milyen részfeladat esetén?

„*az MI nem fog működni bizonyos intelligens viselkedést igénylő feladatok esetén”.* De lám, mégiscsak képessé lett rá, és ezzel a kritika „szavatossága” már le is járt volna azzal, hogy *„az MI, mint kutatási terület eredeti célkitűzései megvalósultak”* (9.o.).[11]

Az „MI nem fogja tudni” tézisek az MI-vel kapcsolatos egykori vitavilágoknak már akkor is csak az egyik, nem is a legjelentősebb részét tették ki.[12] Ez a vita részben ugyan a

---

Milyen mély a rendszerszintű analógia? Az iszappenész erőforrás-elérésre optimalizált útvonal-alakítása hasomló ahhoz, ahogyan bizonyos kihívások esetén az emberi agy útvonalat tervez. A sok közül egyetlen aspektusban érvényesül hasonlóság, a kimenetben, s bár Aono Masashi erősen keresi az agy és a nyálkagomba közti hasonlóságot, valójában épp az a csodálatos, hogy az iszappenész központi idegrendszer nélkül képes erre a teljesítményre. Sok példát lehetne hozni még az élő rendszerek (a gombák mellett a növények és alacsonyabb rendű állatok világából is), de „hasonlóságokat” az élettelen természet erői és az agy működése között is találunk – belátható, hogy sokkal erősebben alátámasztott állításegyüttessel lehet csak a dolog végére járni.

S mindezeken felül az „eredeti célkitűzések megvalósulásának” tétele mögött még ott tülekedik a sor elejére az a kiegészítő állítás is, hogy majd a másodlagosak is meg fognak valósulni. Ha az „eredeti” célkitűzés az volt, hogy az algoritmizálható agymunkát váltsuk ki azokon a területeken, ahol ennek komoly értelme és hatása lehet, szinte automatikusnak tarthatnánk, hogy e néhány kognitív műveletet követően nemcsak további, hanem minden művelet algoritmizálhatóvá válik, ez csak számítási kapacitás és még okosabb modellek kérdése.

A „harmadik irányt” (Héder, 2020:14), a „katasztrofális következményeket hozó MI-siker” – látomásokat tudományos értelemben szintén nem kritikai, hanem egyszerűen olyan disztópia-narratíváknak tartom, ahol egy bonyolult, sokváltozós állapotteret jellemzően egy-két forgatókönyvre egyszerűsítenek le, azokat is sok esetben parodisztikusan megválasztott „trigger-eseményekhez” kötik. Jellegzetes „gondolatkísérleteik” nevetséges szcenáriók, amelyekhez hasonlókat a gőzgépre vagy az elektromosságra is legyárthattak volna kreatív középiskolások. Enigmatikus képviselőjük, Nick Bostrom például nem korszakos gondolkodó, aki a fel nem tárt veszélyek bányáiba vezet expedíciót, hanem imposztor, aki a nonszensz határát súroló argumentációval foglalt el helyet a figyelemgazdaságban.

[11] Ezt az állítást azért további elaborációnak lenne érdemes kitenni, annál is inkább, mert a 208. oldalon erős érvként visszatér, immár a gépek hallgatólagos tudásának „megkérdőjelezése” formájában. A tézist véleményem szerint mindig ki kell egészíteni azzal a szemponttal, hogy a gépi tudás *„milyen mértékben”* valósult meg, *melyik felhasználási területen*. Amikor ezeket a sorokat írom, alig 24 órája történt, hogy az angolul egy kukkot sem tudó arab sofőrrel jóízű beszélgetést sikerült folytatni egy alig másfél órás úton, a Vörös-tenger mellett. Arabul mondta okos telefonjára a közlendőjét, majd ellenőrizte a hangfelismerés helyességét a szöveges átiraton, és elküldte angol fordításra. Elolvastam, angolul válaszoltam, ezt ő már arabul olvasta. Eltársalogtunk. Az eszköz és a mögötte álló két különböző MI-technológia remekül teljesített - általában. Olyasmit tett lehetővé, amit évtizedekkel korábban lehetetlennek gondoltunk volna, s ma elérhető árú szolgáltatásként bizonyít és keresletet épít. De a megvalósulás korántsem tökéletes: néha sokszor kellett ismételni egy-egy mondatot, hogy a kívánt szöveg induljon fordításra. Kérdéses, hogy mekkora zajtűrésig működőképes. A fordítás hibátlannak bizonyult, amikor banális logisztikai részletekről volt szó, és erősen elnagyolttá, olykor szinte érthetetlenné vált, ha komplexebb kulturális tartalom került elő. S illesszük most emellé mindazt, amit a nagy nyelvi modellek hibáiról, hallucinációiról, az értekezlet-összefoglalókat készítő ágensek megmosolyogtató badarságairól tudunk: valóban *mindent* elértünk már az eredeti célkitűzések közül?

[12] Tele vagyunk valódi MI-kritikai helyzetekkel, amelyekből egyébként felépíthető lenne egy átfogó kritikai tipológia, kezdve a *technológiára magára* vonatkozó kritikákkal (mint például a 2024-ben és 2025-ben megjelent MI-fejlesztési moratórium-követelések (abszurd) világával), folytatva a valamely *technológiai részmegoldással* kapcsolatos kritikával (elég, ha a gyilkos robotok kérdéskörére, ill. azok betiltásának követelésére gondolunk), vagy *a megvalósult technológiai részmegoldás bevezetésével, használatával kapcsolatos fals helyzetekkel* (pl. a gépi döntéshozatal emberekre vonatkozó ügyekben, a technológia energiafelhasználási igényének, gazdaságosságának figyelmen kívül hagyása, stb.)

tudományos szövegtermésbe is beszivárgott, de valójában azon túl, elsősorban az üzlet és a bulvár világában pusztított. A szakmai dimenzióban pedig a „valamire nem lesz képes *valahogy”* vélemény érvényességét nem bontotta le utólag, ha „*máshogy”* később mégis képes lett rá. A „nem lesz valamire képes” (mint alternatív jövőkép) az MI minden létező, sikeres, működő megoldásának és érdemének pozitív kontextusú kezelése és tényként való elfogadása, vagy akár az MI további fejlesztéseinek szükségességével való egyetértés mellett is megfogalmazódhatott. Nem minden, az MI majdani, hipotetikus képességeinek elérésére vonatkozó *interpretációt és előrejelzést* érintett, csak azok *néhány fajtáját*. Kifejezhetett fenntartást egy túlzó és/vagy alátámasztatlan előrejelzéssel szemben. Megkérdőjelezhette a megjelenített fejlesztési irány szükségességét, értelmét, racionalitását, gazdaságosságát (ld. az „ami kimaradt” fejezetet (217.): „az MI túlzott energiafelhasználása miatt érdemes újragondolni néhány felhasználási terület gazdaságosságát és értelmét. A szerzőnél: „nagy környezetterhelésű jószág fogyasztása” (218). Érezhette nem meggyőzőnek azokat a módokat, ahogyan a fejlesztő képessé akarta tenni valamire az MI-t. És mai napig nem minden esetben az a forgatókönyv, hogy csak idő kérdése, és lám, „már tudja is az MI”. Számtalan olyan előrejelzés él, ahol az MI „még (mindig) nem tudja, és belátható időn belül sem fogja tudni”.

A „no true scotsman”[13], Héder kedvvel citált „érvelési hibája” valójában nem az MI-kritikusok ellenőrzőjét rondítja csak el, hanem *mindkét vitázó félét*, mert a gond nem abból fakad, hogy valakik szűklátókörűen lehetetlennek vélnek valamit, mások viszont, érvényesebb átlátás birtokában, jobban ítélik meg a kifutás esélyeit. *A baj a teljes problématér elégtelen felépítettségéből fakad – és ez közös konceptuális vétek.* Melyik MI melyik részképességének melyik modalitására vonatkozik az előrejelzés és a kritika? Ha valaki a hatvanas évek elején a „gépi fordítás sohasem fog működni” állítást tette, nyilvánvalóan maga is túlzó fenntartást fogalmazott meg annak fényében, ahogy aztán az első megbízhatóan működő fordítóeszközök megjelentek a piacon. Csakhogy egyáltalán nem mindegy, hogy milyen nyelvek, milyen szövegosztályok és milyen összetettségű fordítási kihívások esetén tartja valaki érvényesnek a fordításteljesítmény kiterjeszthetőségét. Az *„előbb-utóbb* ***minden*** *fordítási feladatra lesz* ***tökéletesen/teljesértékűen*** *alkalmas MI”* állítással kapcsolatos azon megfontolás, hogy némi

---

[13] A „no true scotsman” csak elfedi egy jóval komolyabb kérdés megoldatlanságát, kitárgyalatlanságát. Éppen a Héder által lábjegyzetben (13.o.) bemutatott példa árulkodik róla. *„Míg a gép csak korábban látott zenékből összegyúr egy statisztikailag megfelelő új változatot, addig az igazi zeneszerző valami többet csinál, például kifejezi a lelkét, az inspiráltság állapotában alkot stb.”* Nem az a kérdés, hogy az MI vagy a „gép” inspiráltan alkot-e, hanem az, hogy alkot-e? Hát *nem alkot*. Már a komponál ige használata is félreviszi a diskurzust, a mi az „igazi” komponálásról szóló válasz megérkezése előtt. Nem jelent ugyanis meg egy alternatív komponista (festő, verselő, stb.) aki immár a humán zeneszerzők mellé sorakozik fel. A zenekomponáló és a többi MI egy emberi utasításokat végrehajtó szoftver, amelynek kimenete az instrukciós bemenet függvénye. Az egész logika rossz, amibe belepasszírozzák a problémát, az „igaziság” keresése csak kétségbeesett menekülési kísérlet a rosszul megkonstruált induló állítás kalodájából. Az a tapasztalat pedig, hogy a legerősebb sakkprogramok megverik a legerősebb sakkozókat és a gyengébb programokat, az erős sakkozók megverik a kevésbé erőseket, ám *már közepes sakkozók közepes sakkprogramokkal is megverik a legerősebb sakkprogramokat*, ugyancsak feladja a leckét: ugye hogy nem az „ember versus MI” szerkezetben kell feltenni az alapkérdéseket, hanem az „ember vs. MI-vel megtámogatott emberi tevékenység” szerkezetben! De hiszen nem éppen ez volt az „eredeti célkitűzés” azzal szemben, hogy „fejlesszünk minél többre képes intelligenciaalternatívát”?

óvatosság esetleg indokolt egy-egy szövegtípus (például irodalmi alkotások) esetén, és a teljesértékűség helyett a *kielégítőség vagy az interkulturális megértést nem zavaró mivolt* is elégséges lehet, mértéktartónak tekinthető (ld. a hosszabban bemutatott egyiptomi sofőr esetét). Annál is inkább, mert épp a fordításnak ezekben a speciális tartományaiban még a humán intelligencia teljesítményével kapcsolatban is sok jogos fenntartás és kívánalom létezik. És mindez nem üres elvi okoskodás: néhány orvosi képalkotó berendezés által készített felvételek MI-támogatta diagnosztikájában épp azt tapasztalja az orvostársadalom, hogy van olyan detektálandó betegség/elváltozástípus és a hozzá tartozó képtípus, amelyek esetében alacsony a gépi hatékonyság, és van olyan, ahol magasabb – de a bizonytalanságok forrása sok esetben éppen az, hogy a vizuális tartalom értékelésében az az emberi elme is ellentmondásosan és pontatlanul viselkedik, ahonnan a megerősítéses tanulás visszacsatolásainak származnia kéne.[14]

Ugyanez a kérdés visszatér a 3. fejezetben. Héder inkriminálja azokat, akik „ítéletet mondtak az MI felett" (49), nem is akárhogy: „szalmabáb érvekkel: az érvelők egy része egy addig meg nem valósult entitás tulajdonságai alapján".

Csak ismételni tudom magam: egyrészt nem az MI-vel magával, hanem az MI teljesítményét értelmező állítások túlzásaival szemben jelenik meg a kritikai pozíció. Másrészt mi van az „érvelők másik részével" – akkor, és miért nem látjuk a mai érvelőket, akik a megvalósult MI-entitásokkal (pontosabban: az azok teljesítményét értelmező túlzásokkal) kapcsolatban kritikusak? Ha mai kritikusokat tűznénk tollhegyre, már el is illanna a ryle-i kategóriahiba? Hol van az utolsó 4-5 év elképesztő szövegtermése, miért álltunk meg Penrose-nál, Dreyfusnál? Ha pl. kizárólag az anti-Hinton kánont vennénk szemügyre, jól látható lenne, hogy milyen fontos azokat a fals kiindulópontokat feltárni, amelyek alapján pánikdiskurzusként pertraktálható az MI veszélyvilága, d felismernénk, hogy a legnagyobb AI-boostereknek (a csodavároknak) és az apokaliptikus jövőképek megfogalmazóinak pontosan ugyanaz a hibás kiindulópontja. A hibás kiindulópontok kritikája pedig nem az MI kritikája. Héder nagyon óvatos az MI-fejlesztések leállítását követelő 2023-as moratórium nagynevű aláíróival szemben (118) – a felhívás annyi tarthatatlan kiindulópontot, ostobaságot, félmítoszt tartalmaz, hogy az ember lecserélésének valóban elcsépelt toposzának használata a legkevesebb, amit fel kell róni nekik. (A kézirat elkészítése óta ráadásul a moratórium kritikájának magának nőtt hatalmas irodalma. S bizony fontos részlet, hogy az aláírók közül Bengio és sokan mások különböző nemzetközi szervezetek tanácsadó testületeiben tűnnek fel aztán, hogy etikai állásfoglalásaikhoz hozzájáruljanak. Meghívóik mit sem törődnek azzal, hogy a tudományos diskurzusban az alarmista nézeteik meghaladottnak számítanak, és nem vehetőek komolyan.)

Mindig csak az egyik szélsőség jelenik meg Hédernél: „a leegyszerűsítő számítógép-reprezentációk (mint például az egyszerű statisztikai gép) egyre kevéssé lesznek alkalmasak az MI kapcsán felmerülő problémák, filozófiai kihívások elemzésére" (52.). A gépnek intelligenciát túlhypeoló számítógép-reprezentációk ugyanúgy nem lesznek alkalmasak a

[14] Azért éltem ezzel a hasonlattal, mert a szerző hosszan ismerteti Polányi példáját a tüdőröntgennel (201).

kihívások elemzésére. S hasonlóképp: a nem leegyszerűsítő számítógépreprezentációkkal (amelyek hasonló konklúzióra jutnak, mint a leegyszerűsítők) mi a helyzet?

Ha felsorolja a „rossz érveket az MI ellen", önálló alfejezetben, miért nem kapnak helyet a „jó érvek" is? Ráadásul itt is a problématér leegyszerűsítésével állunk szemben. A determinisztikus versus valószínűségi automatizálás két nagy világára osztott térben még a leegyszerűsítőnek nevezett állítások is igazak a determinisztikus automatizálást igénylő, egyszerűbb környezetekben – a komplex és szabad kimeneteket megengedő valószínűségi automatizációs helyzetekben pedig világosan kirajzolódik a valódi kérdés: világos, hogy van „képességtöbblet". Izgalmas boncolgatni, „hogyan lehetséges ez", de a frontvonal ott húzódik, hogy a képességtöbbletből ki milyen ontológiai következtetést von le: ez már a gépi tudatosság felé tett első lépés volna? Az MI-ügynök döntési szabadsági terének megjelenése még nagyobb döntési szabadságok, és így növekvő gépi autonómia felé vezet-e? [15] A kérdés nem a levegőben lóg: amikor az MI-inklúzió fordított érveléssel kerül tárgyalásra (191-192), pontosan a teljesítményhez vezető különböző 'hogyanok'-ra épülnek a megfontolások.

3.

Hasznosnak, eredetinek és jól eltaláltnak tartom a szerző előző könyvéből megismert, itt csak megemlített modalitás-párost, az MI-kritika narratíváit remekül elrendező *performatív* ill. *fenomenológiai* siker (illetve kudarc) fogalmat (Héder, 2020:13)[16] – úgy tűnik azonban, hogy mindez enyhe és implicit *performációfetisizmussal* párosul (az MI performatív sikerei a 2020-as könyvben önálló fejezetet is kaptak).

Héder nagyon komolyan veszi például a Turing-próbát, mintha annak óriási jelentősége lenne a performatív siker szempontjából[17]. *„Amikor 2024 májusában a ChatGPT4 minden eddigi próbálkozásnál meggyőzőbben teljesítette a Turing-tesztet, az már egy mínuszos hírt is alig ért meg"* (9.o.) Nem véletlenül: a Turing-tesztnek való nekifutásoknak egyre kisebb a jelentősége. Hiszen egyrészt nem szöveges tartalommal már a múlt század ötvenes (!) éveinek végén megkülönböztethetetlenné vált a gépi az emberi performanciától,

---

[15] 2026 januárjának felparázsló Moltbot/Moltbook diskurzusa tökéletesen illusztrálja ezt az elvi alaphelyzetet, mert gyakorlatilag azonnal két táborra osztotta az interpretációk világát: az egyik tábor az AI-ügynökök közösségi médiájában egy új gépi civilizáció felé tett lépést lát, a másik tábor ezerféleképp mutatja be, hogy a gépnek tulajdonított teljesítmény mögött hogyan áll a képletből kifelejtett ember (Holtz, 2026). S véleményem szerint elég alaposan megismerni, hogy pontosan mi történik a Moltbook-platformon, hogy belátható legyen: azzal, hogy kimutatjuk a gépi oldal túlinterpretálásának a hibáit, nem váltunk MI-kritikusokká. Ld. különösen: https://o-mega.ai/articles/how-moltbook-works-the-ultimate-guide-2026

[16] Itt a szerző saját 2020-as könyvének a 13-18. oldalához irányít, az ottani állításokat is beleszőttem a véleményembe.

[17] Ez már a korábbi könyvében is tapintható volt: amikor úgy látja, hogy „*a Turing-teszten való egyre jobb teljesitéssel úgy tűnik, hogy egyre közelebb kerülünk e korlátok – t.i. a szocializációs kompetencia nélküli teljesítményelérés leküzdéséhez"* (Héder, 2020: 43). Ennek a mondatnak pedig nem lett volna szabad bekerülnie az értekezésbe: *„A saját értelmezésünk kérdése, hogy a jelenlegi MI technológiát, amely átmegy a Turing teszten, és könnyen rávehető, hogy jogokat követeljen, ezen előrejelzések bekövetkezésének tartjuk-e"* (189). Mi vegyük rá az MI-t, hogy jogokat követeljen, sőt, ha ügyesek vagyunk, követel is, hiszen rávehető? Ezeket a fordulatokat tartsuk meg a bulvárnak.

amikor a checker game-ben az MI legyőzte az embert.[18] 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. Ezért immár limitjei is áthelyeződnek, abba a tartományba, ahol a jelentésre való érzéketlensége a korlát: a speciális, a szövegkorpuszban nem vagy nem elég pontosan reprezentált helyzetek, szituációk, poliszémiák, áthallások, metaforák, egyedi képalkotások, bizarr szöveghelyzetek, szórend-függő árnyalatok világába. Ez pedig újra csak *nem* a true scotsman probléma friss mutációja. Ha a jelentésértés-képtelenség alapvető gát, akkor a megmagyarázhatóságban tett előrelépés sem ígéri a „tökéletes szimuláció" elérését.

> … a gépi tanulással, az emberiség írásos örökségének nagy részét felhasználó módszerekkel előállított, Turing-teszten átmenni képes MI-be hogyan integrálhatók bele a megmagyarázhatóság hőskorának, a 1980-as éveknek a megoldásai. Várakozásom szerint ez az integráció meg fog történni, mert elvi akadálya nincs, és közben pontosan azokra a problémákra adna megoldást, amelyet ma a legfenyegetőbbnek éreznek: a hallucinálóra, valamint a félrevezető, és emiatt megbízhatatlan MI kérdésére. Ha a teljes értékű intelligenciaként működő nagy nyelvi modellt tartalmazó ágensek következő generációja úgy tud majd számot adni a következtetéseiről, mint a '80-as évek szakértői rendszerei, akkor az egy újabb MI hullámot indíthat be. (163.o.)

A megmagyarázhatóság jelenlegi legerősebb próbálkozása, a *neurális régészet* (amelyet 2027-re várnak kifejlett módszerré válni) sokat ígér, de valójában éppen azon a területen, ami eddig a performativista kánonon kívül maradt: a *„hogyan produkálja a teljesítményt"* kérdésekre adott válaszokkal. Tegyük fel, hogy számos ponton eljutunk a révén megmagyarázhatóságig. Ez kétségkívül előrelépést fog majd eredményezni azokon a területeken, ahol a hibás működés a megmagyarázhatóság hiányára volt visszavezethető. Akár

---

[18] Ennek a szempontnak az erős párjára figyelmeztet egy friss interjúban Yann LeCun, az AGI-előrejelzések bajnokaival szemben. Hiába a szövegekre alapított képesség, ha egyébként az emberi intelligencia maga alapvetően nem a szövegeken keresztül épül fel és nyilvánul meg: sem az érvelés, sem a tervezés, sem a hosszú távú emlékezet, sem a fizikai világ megértése: becslése szerint egy gyerek, mire betölti a negyedik életévét, addigra már ötvenszer több információval találkozott, mint a világ legnagyobb nagy nyelvi modellje https://thenextweb.com/news/meta-yann-lecun-ai-behind-human-intelligence . Itt az ideje, hogy dekonstruáljuk a szöveget, mint intelligencianyersanyagot. Az elmeműveletek és a jellé tárgyiasított, externalizált elmetartalmak viszonyában soha nem a pillanatnyi jel-állapot (a szöveg), hanem az arra vonatkozó és jelentést ismét az elmében felépítő mozzanatok a jelentősek. A szöveg épp azért alacsonyan lógó gyümölcs, mert maga is kognitív redukció eredménye, és a szavakon keresztül csak korlátozottan vezet vissza a világban való komplex szenzoros jelenlét szakadatlan áramlásban létező információs univerzumához. Ezért is teljesen értelmetlen kreativitásversenyre hívni az emberi elmét az LLM-ekkel, ahogy egy friss, szenzációként tálalt, de teljesen értelmetlen tanulmány teszi (Bellemare-Pepin, 2026). Hiszen hiába ír szabályos haikut az LLM, jobbat is, mint az egyébként haikuval nem bíbelődő „humán célcsoport", ha azt haikuként, jelentéssel értékelni csak egy szöveget olvasó ember képes. Versecske algoritmikus előállítása nem különösebben komoly teljesítmény (csak a performációfetisisztáknak), már LLM nélkül is több évtizede képes rá a kódoló ember és a számítógép, vagy még korábban a predigitális szótárcsa) – a haikut értelmes közlésként felfogni, az igen. Ez pedig (csakis) az emberi agyban játszódik le, és nem máshol.

még újabb MI-hullámot is gerjeszthet (ráadásul közben végre eltüntethetné a tárgylemezről azt az alarmista paradiskurzust, amely a gépi oldal valamiféle „különösségét, rejtélyes többletét” látva a megmagyarázhatóság hiányában, ezt is argumentumként veti be a gépi intelligencia léte, fenyegető mivolta, vagy emberit meghaladó képessége igazolásaként). Csakhogy az új MI-hullám, ha lesz, a jobb szimulációból merít erőt, de továbbra is szimuláció marad, mindössze csökkenti majd azoknak a speciális szöveghelyzeteknek a számát, ahol nem nyújt megoldást (s ilyen értelemben sem AGI-előkészítő). A limit-tisztázás itt nagyon lényeges, hiszen épp a hibakezelés-béli eltérésnek van óriási jelentősége Neumann János briliáns felismerése szerint a számítógép és agy által vezérelt idegrendszer működésében (Neumann, 1959/1972:108-109). Ezt korábban így foglaltam össze:

> Egy egymásra következő számolásműveletek millióit elvégző gépi feldolgozó rendszer akár egyetlen számítási pontatlansága egyre több hibához vezet a ráépülő műveletek során: „felnagyítódik”, sokszorosodik. Ezt … Neumann *aritmetikai mélységnek* nevezi, amely a feldolgozás rendjét meghatározó *logikai mélységhez* is igazodik. Az emberi agy egészen másképp működik: nem törekszik aritmetikai pontosságra. A jelölői helyzetek (amelyek értelmezik és irányítják a feldolgozási műveleteket és az egyes elemekhez rendelt jelentéseket) nem szabatosak, nem szigorúak és nem kód-szerűek: az aktuálisan egymásra vonatkozó jelentések együttállásából születnek. Statisztikus természetűek, mert nagy sokaságra alkalmaznak jelentést, emiatt a tévedések (akár szaporodó) száma sem érinti a viselkedést meghatározó helyzetértékelést. *Az aritmetikai „romlásért” cserébe magasabb logikai megbízhatóságot kapunk. (Z. Karvalics, 2022:11-12)*

Másképp: a megmagyarázhatósággal sem leszünk képesek áthidalni az aritmetikai mélység és a logikai megbízhatóság közt tátongó űrt – és ez szinte pontosan rímel Turing felemelt mutatóujjára, aki visszatérően utal *„a sikeres működés és a gondolkodás közti különbségre”* - amire más kontextusban Héder is hivatkozik (Héder, 2020: 48).[19] Tehát nem ismeretelméleti fontoskodás minden pillanatban emlékeztetni az azonos kimeneti teljesítmény mögött álló mechanizmusok különbözőségére – a predikcióérzékeny és gazdag kontextusokkal építkező vitaterekben ennek óriási és sarkalatos jelentősége van.[20] Héder konstrukciójában is: amikor azt írja, hogy *„ami egy elme-funkcióját képes betölteni (márpedig ez empirikusan tesztelhető) az elme is”* (183), azt elégtelen formulának érzem[21]: oly sokféle

---

[19] Ezt újabban szemantikus pareidoliának hívja Floridi (2026)

[20] Ez különösen ott jelent erős nyomást, ahol a történetiség jelen van, legyen az a homonímák esete a nyelvtörténetben, a karcinizáció az evolúcióban, vagy a régész dilemmája, hogy a barlangi tűznyom a Homo tűzkeltő tevékenységének eredménye, vagy a magas foszfortartalmú denevérürülék öngyulladása miatt alakult-e ki. Presner (2024) könyvében ezt úgy fogalmazza meg, hogy a számítógépek más módon „olvasnak”, mint az emberek, teljesítményük nem jobb és nem rosszabb, mint a miénk, hanem más. Emiatt lehetnek a segítségével olyasmikre képesek a történészek, amire nélkülük nem.

[21] S természetesen ugyanez igaz a társ-állításra (*az MI intelligens viselkedésre való képessége, mint az MI megkülönböztető sajátossága*) (185). Az intelligens viselkedés kontinuumának melyik darabjára gondolunk? Már valamilyen viselkedéskomponens teljesítése is elég (ld. blob), vagy szuperkomplex viselkedést kell a gépnek produkálni – miközben kifejtésünkben következetesen „az intelligens viselkedésre emlékeztető” formulával kellene élni – s minden következtetés ugyanolyan érvényes marad!

élőlénynek van elméje, s oly sok elmefunkció aggregálódik, hogy a potenciális elmefunkciókból például a különböző közösségi és civilizatórikus szintek emberei egészen más kognitív konfigurációkat építenek, egészen más funkcionális környezetekhez, Legalábbis szofisztikálni, elaborálni kéne ezt a tézist – mert abból, hogy a kör már bővült az állatok morális befogadásával, még nem következik semmi, ha univerzális kvantorként gond van azzal az állítással, hogy valami hézagmentesen betölti az elme funkcióját. (Emlékeztetek a blob-példára: az elme útvonalválasztás-optimalizációs funkcióját remekül teljesíti, senki nem tulajdonítana azonban morális státuszt a nyálkagombának.) Ettől persze bizonyos MI rendszerek inkludálhatóak morálisan, de nem elmeanalógiás természetük, hanem más megfontolások okán. A megbízható családi robottárs morális inklúziója egészen más természetű lenne, mint egy katasztrófaelhárításkor túlélőt kereső serpentoid roboté: vagyis különböző morális inklúziótípusokról kell beszélnünk, ha nem „episztémikus ideológia", hanem kizárólag pragmatikus szempontok alapján soroljuk osztályokba.[22]

4.

Erősen egyetértek a szerzővel *a történeti kontextus hasznosságát* illetően.[23] Amit MI-nek nevezünk, az valóban az 1950-es évek terméke, s a korai víziók sem igénylik a visszalépést az időben. S a megmagyarázhatóságnak az értekezés szempontjából oly lényeges problémaköréhez sem lehetett volna közel kerülni a hetvenes évek elejéig visszanyúló történeti rekonstrukció nélkül (142-151).

Ám a programozás és az algoritmikus gondolkodás legkorábbi formáinak keresése igencsak fontos és a „klasszikus" etikai dilemmák szempontjából is relevánsnak mondható. Gondoljunk az Endrei (1992) által a programozás eredetének tekintett neolitikus önkioldó csapdák esetére: már itt is megjelennek az emberi felügyelet nélküli használat/műveletvégzés dilemmái és kockázatai, az azonnali korrekciós/módosítási igény lehetetlensége miatti fenntartások, a nem kívánt végeredmény kialakulásának befolyásolhatatlansága, stb. [24]

Másrészt az MI „előtörténete" is számos olyan szempontot vet fel, tanulságot hordoz, amelyek befolyásolhatnak kortárs diskurzusokat. S lám, a műszaki tudás anatómiájának, a második fejezetnek *is a kontrollképesség történeteként felfogott históriai kontextus* ad keretezést – a szívemnek erősen kedves módon.

---

[22] Emiatt is érdekes, hogy Héder csak ismerteti a humanoid robotokkal kapcsolatos morális befogadás-párti érveket, ám se pro, se kontra nem foglal állást (184). A Moravec-gondolatkísérletet szerintem ebben a kontextusban felesleges extremitás (pedig még vissza is tér a 187. oldalon), mert valójában a másutt a tárgyalásból kizárt digitális elmetranszfer unokatestvére – miközben praktikusan egy kiborg-lét, az emberi test gépi elemekkel való megerősítésének kérdése valódi kihívás, különösen, ha a pacemakerekkel kezdődő, majd neurális interfészekkel tranzakcióképessé tett bénultakkal folytatódó korrekciós univerzumok mellé belép a beültethető elemekkel operáló interfészvilág, a beültetett chipekkel, stb.). Felesleges a sci-fi, amikor a szemünk előtt ott a valódi kihívás.

[23] Apró pontatlanság, de nem a hatvanas évek végén, hanem az elején (1961) születik meg az információs társadalom kifejezés Japánban, és a 6.o. 2. lábjegyzetben hivatkozott mű is ezt így taglalja.

[24] Azóta ráadásul már tudjuk, hogy a mérgezett nyilak használata is sokszoros „programozás" eredménye, és legalább 70 ezer éves algoritmikus technológia. Borzongató elképzelni, hogy egy paleolit jogvédő mennyi biztonsági kockázatot találhatott volna az akkori gyakorlatban…

5.

Hasznosnak és találónak tartom, hogy Héder Mihály a műszaki tudás természetét egy önálló fejezetben tette nagyító alá. A tárgykör irodalmában nem vagyok jártas, de a gondolatmenetet és az irodalomjegyzék alapján úgy vélem, hogy nem gazdag nyersanyagból kellett a szerzőnek szűrnie, hanem szegényes nyersanyagból kellett érvényes gondolati szerkezetet létrehoznia. Ugyanakkor felrovom a szerzőnek, hogy a dolgozat későbbi pontján (127) – nagyon helyesen – elkülönített junior, medior és szenior tudáshordozókkal nem kapcsolja ezt a gondolatmenetet össze.

Messzemenően egyet lehet érteni a fő tézisekkel (a műszaki tudás *normatív*, *célszerű/ funkcionális*, inherens *morális* dimenziója van és *prediktív* is). A technológia tervezési és létrehozási idejének szétválasztása termékeny szempont, úgy tűnik, a szerző nem indexelte, hogy ez az ő leleménye. Én szívesen látnám még a kísérleti üzemi és a rendszerbe állítási idő megkülönböztetését is, mert az a nyelv, ahogyan a technológia teljesítményéről, tudásáról, megoldó erejéről beszélünk, más-más modalitásokkal jelentkezik az egyes fázisoknál (és ez korántsem mindegy az MI aláaknázott fogalmi világában, ahol a Tervező/Konstruktőr szerep olyan távol kerül már a „Megbízhatóan Dohogó Masina" szerepétől, hogy nagyon könnyű a gépi ágencia felé elbillenni, mert már alig látjuk a folyamatok mögött a konstruktőrt.

Nincs semmi meglepő abban, hogy a szerző a dekompozíciós mozzanatot ragadja meg, és elemzi nagy felbontásban. Annyi megjegyzésem van csak, hogy ez nem pusztán a konstruktőri helyzet sajátja. Ahogyan az egyes technológiák összeérnek, ahogyan végfelhasználókat megatechnológiák szolgálnak ki, ezek „összegabalyodott" (entangled) természete szükségszerűvé teszi a sokféle kompetenciához való igazítást, amit a dekompozíciós elv tud megvalósítani/lefordítani. (A kézműves/céhes világban még nem volt szükség dekompozícióra, a manufaktúrákban és a gyárüzemben fokozatosan egyre többre.)

Szellemes és élvezetes az egyes dekompozíciós rétegek bemutatása, láttató erejű a jármű-példa, szakmai oktató anyagokra emlékeztetően pontos az egyes módszertanok bemutatása.

Az MI-dekompozíciós erőtér megrajzolásakor számomra zavaró azzal kezdeni a behatolást a problémamezőbe, hogy *„ha a mesterséges intelligencia tervezésekor a természet imitációja a stratégiánk …"* (45.o.) – ha egyszer ilyen stratégia csak elvileg vagy periférikusan létezik (s az idézett IBM-Watson példa sem erre az imitációra épül). Ahogy a repülőgép sem a madarakat/a természetet imitálja (csak csapkodó szárnyú, röpképtelen formájában), az MI-fejlesztések számára is csak az a fontos, hogy a végén a hiányzó kognitív műveletvégzési kapacitás rendelkezésre állhasson. Az a lényeg, hogy a biológiailag/társadalmilag/kulturálisan adott korlátokat meg lehessen haladni, s ehhez a természet imitációja csak egy a sok lehetőség közül.

6.

Az átlátszatlansággal kapcsolatos felvetések fontossága nem kérdés. De csak a mérnökcsapatokon belül értelmezhető? A fejlesztések mögött álló informatikai cégek

menedzsmentjére nem kéne kiterjeszteni (ha már a stack, a szeparációs lánc velük kezdődik)? És azokra a politikusokra, akik MI-fejlesztésekkel kapcsolatos stratégiai és forrásallokációs döntéseket hoznak?

A 4. fejezet remek példákat hoz arra, hogy az élethelyzetek sokfélesége milyen tervezői limitációkat eredményez, s hogy mennyire reménytelennek tűnik varratmentes modellt építeni pl. az önvezető járművek számára.

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.

Némi opponensi fejfájást okoz az 5. fejezet is. Nagyon is szabatos ugyan a technológiai determinizmus kritikája, de nem világos, hogy miért csak ez az alacsonyan lógó gyümölcs van kéznél. A technológiai determinizmus (és annak disztópiára fejnehéz) elméleteit már rég túlhaladtuk, változatos módon megcáfoltuk – hol vannak a nem technológiai determinista elméletek, a mángorlás (mangle) vagy a cselekvőhálózatok világa, amelyben nemcsak ember és technológia, hanem a környezet is a modell része?

A technológiai zártságnak pedig szintén egy leszűkítő karakterisztikáját kapjuk (82.):

> annak a lehetősége, hogy a technológia kontrolljára rendelkezésre álló időablak bezáródik, és a függés a technológiától olyan erős, hogy a technológiát nélkülözhetetlenné, a társadalom technológiai szintjét és állapotát pedig visszafordíthatatlanná teszi.

Az időablak a cselekvők adott köre számára záródik be, és nem a visszafordítás az egyetlen forgatókönyv a társadalmat függésbe hozó technológiával szemben, hanem a technológia cseréje vagy lényeges jellemzőinek átalakítása is. Héder meggyőző módon sorolja azokat a szempontokat, amelyek mentén *„a mesterséges intelligenciát több tényező is különösen alkalmassá teszi a technológiai bezártság előidézésére, ezáltal visszafordíthatatlan változások okozására"* (84) és helyesen állapítja meg, hogy MI-feladatosztályonként a zártság ereje azért különböző – és valóban, a szabályozási düh mögött valóban megérezzük ennek a nyomását.

Csak megjegyzem, hogy a szabályozási düh mögött azért azok a paradiskurzusok is ott vannak és hatnak, amelyeket a metonimikus fogalomhasználatnak, az AGI-és ASI-allúzióknak és a disztópikus forgatókönyveknek köszönhetünk. Azért fontos a Héder által a későbbiekben javasolt megközelítés és logika, mert „ridegen konzekvens" szerkezete éles ellentétben áll és hasznos és tárgyilagos keretrendszert kínál a hisztérikus szabályozási törekvések legtöbbször tarthatatlan kiindulópontjaitól (pl. az emberi faj kiirtásával, a ránk növő új intelligens fajjal kapcsolatos badarságoktól) való eltávolodáshoz.

És ismét csak helyesen tér vissza a szövegben a szabályozással kapcsolatban az MI különböző formáival kapcsolatos megkülönböztetési igény (96-97), amelynek fontosságára a korábbi kontextusok esetében is utaltam (másképp: ezt az elvet az MI-kritika kritikájánál is érdemes pl. figyelembe venni).

Teljesen egyetértek azzal, hogy az *eudaimonia* vezéralapelvvé tétele sok kérdést vet fel, önmagában az is egy értekezés tárgya lehetne, hogy mi mindennek kellene mellélépnie egy komplex alapelv-hálóban[25] - amelyek valóban egy „technológiai felvilágosodás" (110) erkölcsi sormértékéül szolgálhatnak.

7.

Az automatizálási torzítás jó arányérzékkel kibontott tárgykör. Azt hiányolom belőle, aminek épp a nagy leíró erejű tárgyalás a bizonyítéka: hogy t.i. egy konstruktőri nézőpont könnyűszerrel be tudhatja építeni a tervezési folyamatba azokat az elemeket, modulokat, effektusokat, amelyek tudatosítják a fogyasztóban a kockázatokat, illetve egyértelművé teszik, hogy pontosan milyen típusú segítséget kapnak, mit nem szabad szem elől veszteniük.[26]

Nem tartom szerencsésnek viszont a „selective adherence" szelektív MI-kritikának fordítani – sokkal inkább direkt negligálásról van szó, és az MI-kritikát magát erősen elhasználta másra Héder.

Nagyon szórakoztató és tanulságos az MI-képgenerátor alaposan kidolgozott alfejezetecskéje, a parciális egyensúlyt illusztrálandó. Ha a disszertáció többi része is ebben a stílusban, ilyen könnyedséggel, elegáns módon öltött volna testet, még nemzetközi sikert is lehetne jósolni egy fordításnak (mint azonban később kitérek rá, a mű szerintem nem alkalmas ebben a formájában a fordításra és nemzetközi megjelenésre.)

Nem tartom viszont pontosnak az interpretáció végét: „az MI csak egy helyettesítő terméke az embernek" és „az MI minden fontos paraméterben imitálni és meghaladni képes az embert egy adott feladatosztályban" megközelítések szerintem pontatlanok. Két lényeges mozzanat van: a képalkotó teljesítmény demokratizálásának eszköze lehet az MI (amit eddig a mesterek tudtak csak, arra most az MI segítségével más is képes lehet), s amire a mesterek eddig nem voltak képesek, az elgondolható képek megvalósításához kaphatnak segítséget (amit a filmipar a CGI-dömpinggel fényesen bizonyít évtizedes távlatban).

Ám a szoftvermérnöki szakma átalakulására vonatkozó előrejelzést olyannyira fontosnak találom, hogy azt sérelmezem, hogy ezt a tárgykört nem bontja ki még alaposabban, még operatívabban. Ez az egyik olyan terület, ahol a szerző diskurzusteremtőként tudna belépni a nemzetközi térbe, mindenképp érdemes ezt az egyetlen témakört tovább alakítani friss publikációkban.

Egy ilyen közleményben bizonyosan erős pont lehet az algoritmizált méltányossággal kapcsolatos gondolatokat tartalmazó 8.fejezet talán két legfontosabb tézise (139): annak illusztrálása és bizonyítása, hogy

- *„a meta-etika MI ipar általi felfedezése zajlik, annak ismert dilemmáival és divergens előfeltételeivel egyetemben"*(vagyis a filozófusok által bejárt gondolati ösvények

[25] Ehhez csak tipp-ként: az Aquinói Szent Tamástól eredő hármas megkülönböztetése a „jónak" jól árnyalhatja az MI különböző „arcait": a megfelelésekhez könnyű utat építő *bonum utile* és *bonum delectabile* mellett a talányosabb és komplexebb arcot mutató *bonum honestum* kibontásával lehet érdemes bíbelődni.

[26] Ez nem működik pl. a lelki tanácsadóvá, baráttá, házastárssá emelt GPT-k esetében – a sok korláttal, tudáshiánnyal, traumákkal küzdő individuumok nem kapnak meg minden segítséget, hogy egyértelmű legyen, mivel állnak szemben.

egyre relevánsabbak a technológiai fejlesztésben, s hogy a fontos diskurzusok hamarabb indulnak el az egyetemek tanári szobáiban, mint a tervezőlaboratóriumokban)

- *… „mindez holisztikusabb, a méltányosság-metrikák egyszerűségével szembeni kritikusabb világképhez vezet”* (vagyis a mérnöki és statisztikai irály mellett a társadalomtudományi narratívák felértékelése és beemelése felé vezető nyomás erősödését konstatálhatjuk).

A transzparencia múltjával és jövőjével foglalkozó fejezetet érzem a dolgozat legkompaktabb, szakirodalmat leggazdagabban mozgósító, a szempontokat innovatívan egymás mellé rendező, gördülékenyen megírt részének. Olykor az az érzésem támadt, hogy a szerző ebben a megmagyarázhatóság-probléma köré épített kisvilágban mozog a legszívesebben és legmagabiztosabban – s talán az is szerencsés lehetett volna, ha a teljes dolgozat ebből nő ki, megspórolva az MI-kritikával kapcsolatos „felvezetéseket”.

8.

Kíváncsi vagyok a szerző véleményére, hogy az „etikai problémát döntésként értelmező” kiindulópontokkal szembeni kritikájában megjelenik-e implicit módon az a szempont, hogy a döntés, mint a cselekvésválasztáshoz vezető hosszú út utolsó mozzanata, épp amiatt, hogy a pillanatnyilag adott ill. mozgósított, már előzetesen jelen lévő sémák alapján születik meg[27], szükségszerűen hordozza magában a hiba lehetőségét (is), a tévedés sokfajta forgatókönyve okán (ezt érti-e bizonytalanságkezelési probléma alatt?), s emiatt sokkal inkább azoknak a sémáknak a működésére érdemes figyelnünk (ezt értsük-e episztémikus kapacitás alatt?), amelyek döntéshelyzetben aktualizálódnak (a Boyd-féle OODA-loop 2. „O”-ja, az orientáció fontosságát kiemelve).

Amikor a szerző azt írja, hogy

> döntést igénylő valós szituációk, talán néhány speciális eset kivételével, mindig az alábbi struktúrát követik: *válasszunk egyet a nyilvánvaló alternatívák közül, vagy próbáljunk újakat találni* (170)

az „alternatívakeresés” alatt értsük-e megismeréshiányos helyzet tudatosításának nyomán fellépő pót-megismerési igényt, amelynek sikere vezethet új döntési alternatívához[28]?

A fékezéses-kikerüléses-sávváltásos (ütközés-elkerülős) példafeladathoz egyetlen kiegészítő kérdésem van: jól gondolom-e, hogy a *„más szociotechnikai rendszer”* fogalmában (MI + szabályozás + közösségi ismeret + konszenzus) (180) nem foglaltatik

---

[27] S ezeknek természetesen része mindaz, amit a Polányi-féle hallgatólagos tudásként a szerző tárgyal a 12- fejezetben. Ha az emberi viselkedésben és döntésben egyfajta hiba/tévesztés/szuboptimalizáció valósul meg, akkor kell-e ezt szimulálni gépi rendszerben – illetve a gépi rendszernek az emberi megoldási utat kell-e követnie? Az LLM-hallucináció – amelynek a mechanizmusát nemrég sikerült, talán, megfejteni, programozói hiba, és nem viselkedési karakterisztika (Kalai, 20215).

[28] Amit jól körülír pl. ez a helyzet: „az autó időnként lelassít, mert „önbizalma” egy adott küszöbérték alá esett, így *további megfigyelésekre van szüksége*” (175).

benne az a megfontolás, hogy az episztemorálisan tervezett önvezető jármű esetében nem szükségszerűen homológok a manőverek, ha figyelembe vesszük, hogy a közlekedés átállításakor számos hagyományos művelet-és paramétertípus is átalakul: az önvezető járműökoszisztémában más helyzetekre (is) készülni kell, és ezek a helyzetek is alakítottak (pl. a szenzorokkal felszerelt intelligens út, a folyamatos adatkommunikációs a járművek között) – nem véletlen, hogy a létező flottarendszerekben néhány egyszerű elv szinte varratmentes megoldást ígér. Ha nem, milyen nevet adjunk ennek a sajátosságnak, s hogyan érdemes kiegészíteni ezt a tételt (amellyel egyébként messzemenően egyetértek, és fontos üzenetnek tartom a szakmai közösség felé):

> Az önvezető jármű esetén ezt a metszetet a közlekedés minden résztvevője számára érvényes, elfogadható jármű-viselkedésként értelmezhetjük, már ha létezik egyáltalán ilyen metszéspont. Ez azt jelenti, hogy az efféle viselkedés definiálásáért felelős iparág sikeresebb lehetne, ha ajánlásokat fogalmazna meg és az ezekkel kapcsolatos kompromisszumot keresné, ahelyett, hogy erkölcsi igazságokat próbál kergetni. (177.)

S hasonlóképpen, nagyon erősen rezonál bennem az egyik tanulság, amelyben a magam MI-tudáskormányzási (knowledge governance) narratívájának fontos komponensét látom viszont, használni és idézni is fogom:

> „a tudás-generálás nagymértékben segíti a megvalósítást, de csak addig a pontig, amíg a ráfordított erőforrások meg nem haladják a bennük rejlő tudás értékét a cél eléréséhez … szükségünk lenne egy olyan módszerre, mellyel megbecsülhetőek a tudás megszerzéséhez szükséges ráfordítások és az ezzel járó előnyök”. (174)

Virtuóz érvnek tartom a morális státusz tulajdonításával kapcsolatban, hogy *„a tévesztési mátrix aszimmetrikus … amikor feleslegesen tulajdonítunk morális státuszt az MI-nek, sokkal kedvezőbb”* (194). Mégis, kicsit kötözködök: ha pl. a tervező módszertanilag tulajdonít feleslegesen morális státuszt egy MI-rendszernek a csendes laboratóriumában, az egészen más megítélés alá esik, mint amikor egy MI-teoretikus nagy nyilvánosságot kapó, milliós példányszámú bestsellerben ostoba példák sorával követeli a morális státuszt (emlékezzünk csak a látványos, de veszélyes érvre: azért vagyok udvarias az LLM-mel, mert nem akarom, hogy haragudjon rám, amikor majd átveszi a hatalmat a gép).

És persze ez megint odavezet, hogy szerintem nem jó fogalmi keret a nagyon fontos „MI tapasztalása” c. 12. fejezetben „az MI morális státusza”, mert azt az illúziót kelti, hogy létezik egy ontológiai entitás, „ami az MI”. Héder maga emeli ki, hogy az episztémikus feladatokat ellátó szoftver némiképp más, mint mondjuk a felfekvéses beteg megfordításában segédkező exoskeletonba épített intelligencia. Ahány karakterisztikusan különböző tipikus és már használatban lévő MI-létező, *annyifajta morális státusz.*

Tehát szerintem éppenhogy nem átfogó entitásként kell kezelni, hanem különböző, egymással leginkább össze nem kapcsolt technológiai megoldásokra szelektív teóriákat kell kidolgozni. Az átfogó entitás-képzet egyetlen lépéssel az egyik nagyon kevéssé valószínű Gép-világagy vagy Terminátor-forgatókönyv előszobája, aminek éppen a javasolt distinkciót nélkülöző volta a csalásforrás.

Nagyszerű a posztkritikaiként bemutatott Polányi-gondolatmenet, ezer finom szálat sző a témakör szövetébe, mégis, szerintem hallgatólagosnak mondható gépi tudás (a bicikliző robot esete) nem létezik (ld. megint: blob), csak interpretációs bravúr, hogy annak tekintsük egy gondolatmenetben. De ami még lényegesebb, sem a Polányi által használt értelemben, sem az univerzális an használt értelmében nem tartom emergenciajelenségnek, ami az MI különböző rendszereiben és rendszereivel történik[29]. Az MI létrehozása nem evolúciós folyamat, ahogy Héder állítja (208): *az MI-megoldásokkal élő társadalom saját makroevolúciójában azonban nagy szerepet kapó és még nagyobbra hivatott technológiai komponens*. Ez a különbség posztkritikai szempontból is nagyon jelentős – a mérlegelési tér szinte ugyanolyan marad, ha nem magát az MI-t tekintjük evolúciós alanynak (és szükségtelenné teszi az MI-kritika olyan megközelítését is, ahogyan az a disszertációban történik). (Ezt a vitát azonban külön folytassuk le, mert a nézetkülönbség nem érinti a dolgozat értékepével kapcsolatos véleményemet.)

Az izgalmas gondolatkísérletbe is kedvvel kötök bele. Kétségtelen, hogy saját személyes tudásunkat növeli minden interakció (nem az MI-vel, hanem annak különböző részrendszereivel – az MI-univerzum nagy részével nem vagyunk kapcsolatban, mindenki más MI-kisvilágokat tapasztal meg). Természetesen, hogy eközben tanulunk is. Hirtelen belesodródunk a popperi három világ őserdejébe. A tapasztalati térrel az első világban járunk, ahonnan rögtön be is hatolunk a második világba, amikor Polányi arra vezeti vissza kiemelkedő intellektuális fölényünket, hogy képesek vagyunk legalább részben kifejezni magunkat *a nyelv segítségével* (mind evolúciós, mind egyedfejlődési kontextusban *artikulációnak* nevezve azt).[30] Ám amikor metaszintre ugrunk, állításokat megfogalmazva az MI ontológiai természetéről, azzal a popperi harmadik világba emelkedünk, ami eltávolít a közvetlen tapasztalattól, és részben intézményessé teszi a megismerést a közvetített tapasztalat mozgósításával (szakirodalom, kísérletek, diskurzusok ismerete), részben magasrendű módszert ültet a nyelv fölé, a tudományos beszédmóddal, annak fejlett módszereivel, állítás-és érveléslogikával, amelynek révén magasabb szinten kívánunk számot adni a vizsgált jelenségekről. A tapasztalatot szerző és tanuló felhasználó nézőpontjáűhoz képest az analitikus elemző egészen mást mozgósít a módosított nyelvi kompetenciája segítségével. Erre a különbségre nem lehet nem érzékenynek maradni, semmilyen gondolatkísérletben sem. Annál is inkább, mert a „kiegészítésben" előkerülő ember-gép

---

[29] Ez természetesen azt is jelenti, hogy Szathmáry Eörsnek az evolúcióképes mesterséges intelligenciával kapcsolatos téziseivel sem értek egyet, amelyet előadásokban, interjúkban és publikációkban fejt ki, rendkívül magas színvonalon. Minden MI-művelet tervezett és irányított, még akkor is, amikor evolúciós vagy emergenciafolyamatok szimulálására utasítjuk. Fontos viszont megjegyezni, hogy az élő anyagból épített (Gánti Tibor „lágy automatájára" emlékeztető elképzelhető architektúrák) mesterséges intelligencia rendszerek esetében már sokkal több nyitott kérdés merül fel – de ez még nagyon távoli technológiai jövő. Mi veszik el, ha nem úgy fogalmazunk, mint Héder (212): *„az MI-nek nagy, kifinomult, összetett és nyughatatlan belső struktúrája van, amelyen keresztül a funkcionalitás emergál"* – hanem így: *„minden, különböző képességekkel felruházott MI-rendszernek nagy, kifinomult, összetett és nyughatatlan belső struktúrája van, amelyen keresztül egyre több és egyre összetettebb funkciók ellátására teszik őket képessé alkotóik"*?

[30] S egyáltalán nem banális itt arra utalni, hogy a gépi kód is egyfajta mesterséges nyelv, ami utat nyit a műveletvégzési tér kitágításának.

hibrid problémának[31] épp a szuperperszonalizált világa felé (is) haladunk, vagyis minden egyes ember megtapasztalás-és interakcióélménye valóban része az MI-vel kapcsolatos megismerési térnek.

9.

A disszertációval kapcsolatban egyetlen, összegző állítást teszek.

Szövegműként egyenetlen minőségűnek, a részekből szerkesztettség jeleit erősen magán viselőnek tartom, ami hol magas rendű értekező próza, hol az oda-és visszautalások, mi következik-típusú fordulatok nagy száma miatt nehézkes, és nem teljes igényességgel megformált olvasmány.[32] Fogalmi felépítettségét is számos ponton inkriminálta m.

Gondolkodói tettként, tudományos teljesítményként azonban üdítő színfolt. Nincs egyetlen, önmagáért való rekapituláció, nekrofil tudásfelmondás benne. Minden összefügg mindennel. A szerzőnek egymásba soha nem olvasztott diskurzusterekben kell navigálnia, nagy-nagy önállósággal, eredetiséggel, innovatívan. S még ha ezt néhol sikerületlenül, nehézkesen is teszi, a diszciplináris leckefelmondók unalmas, steril és monoton kórusa helyett folyamatosan kihívó és nyugtalanító, eredeti motívumokat, dallamokat hallunk megszólalni. Nagyon sok mély, a szerző által reflektíven ki nem emelt szempontot, fogalmi hozzájárulást, az irodalomfolyamban nem szereplő eredeti gondolatot találunk, a mű igazi „treasury of insights". Az opponencia műfaji jellege miatt fókuszáltam a kritikára:

> **a disszertációt komoly szakmai teljesítménynek tartom, Héder Mihály izgalmas és felkészült gondolkodó, eredményeit elegendőnek tartom az MTA doktora cím odaítéléséhez, az értekezés nyilvános vitára bocsátását javaslom.**

Kérem, hogy a bírálatokra adott válaszában az itt-ott megfogalmazott kérdéseken túl térjen ki az alábbi két kiegészítő kérdésre is:

---

[31] Jóleső érzés, amikor egy megjelenés előtt álló tanulmányomban a népszerű ember-gép szimbiózis fogalom nehézkességével és pontatlanságával szemben megformált érveimet látom visszaköszönni a tőlem teljesen függetlenül megformált szövegben.

[32] Ezt a szöveg szerkesztettsége is tükrözi. A kiérlelt, önállóan megformált részek gördülékenyek, hosszabb bekezdésekkel operálnak, a nem elég alaposan megmunkált részek rövidebbek, lakonikusak, nélkülözik az enyhet adó példákat, töredékesek, néhány soros fragmentumokból építkeznek. Nem segít, hogy Akadémiánk befogad térközzel és behúzás nélkül tördelt kéziratokat-. Ez az üzlet világából átszivárgó helytelen gyakorlat szembe megy az értekező próza, a tudományos szövegformálás évszázados hagyományával. Ezzel a gyakorlattal szemben nem győzöm visszatérően hangoztatni a fenntartásaimat, itt is ezt teszem. Jól látszik, hogy még a pongyolaságok is jobban kiütköznek egy ilyen oldalképen. Csak néhányat gyűjtöttem össze, mert nem lényegtelen, de elhanyagolható probléma, de azért tükrözi a megformálás alaposságával kapcsolatos némely aggályt. Pl. (1) a 115.o. 7.7.-es fejezetre hivatkozik – az valójában a 7.6. (2) 119.o. ChatGPT *adopciós* (sic) görbéje (3) 211.o – bántó egybeírás (háziállatbőrében).

1. 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?
2. A 2004-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.

Budapest-Kőszeg-Vonyarcvashegy. 2025 október – 2026 január

Z. Karvalics László

Hivatkozások:

Afroogh, Saleh et al. (2026) Against AI ethics: challenging the conventional narratives. *AI and Ethics* 6, 125. https://doi.org/10.1007/s43681-025-00978-0

Ansari, Waisuddin (2023): The Ethical Implications of Artificial Intelligence on the Human Society. *SSRN Electronic Journal* January DOI: 10.2139/ssrn.4512301

Beale, M. (2022): The Need for Ethically Designed AI Systems *Cloudfactory*, June 9. https://www.cloudfactory.com/blog/ethics-of-ai

Bellemare-Pepin, Antoine et al. (2026): Divergent creativity in humans and large language models. *Scientific Reports* (16) 1279 https://doi.org/10.1038/s41598-025-25157-3

Demet, Dagdelen (2025): A hájp velejárói (Csabai Máté interjúja) *Magyar Narancs* január 14. https://magyarnarancs.hu/tudomany/a-hajp-velejaroi-278279

Donath, Judith (2020): Ethical issues in our relationship with artificial entities In: Dubber, M.D. et al. (Eds): *Oxford Handbook of Ethics in AI* Oxford University Press, 53-76.

Endrei, Walter (1992): *A programozás eredete* Akadémiai Kiadó

Fási, Csaba (2025): Etikai kihívások és megoldások a mesterséges intelligencia alkalmazásában. *GLOSSA IURIDICA*, 12 (3-4). pp. 121-140. ISSN 2064-6887 https://doi.org/10.55194/GI.2025.3-4.6

Floridi, L., Cowls, J., Beltrametti, M. *et al.* (2018): AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines* 28, 689–707 https://doi.org/10.1007/s11023-018-9482-5

Floridi, Luciano (2026): AI and Semantic Pareidolia: When We See Consciousness Where There Is None (revised version) (January 12) *SSRN*: https://ssrn.com/abstract=5309682 or http://dx.doi.org/10.2139/ssrn.5309682

Golden, Dániel (2025): Mesterséges intelligencia és társas tudás *Magyar Tudomány* (186) 8:1596–1603 DOI: 10.1556/2065.186.2025.8.14

Hao, Karen (2019): In 2020, let's stop AI ethics-washing and actually do something MIT Technology Review Dec.27. https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/

Holtz, David (2026): The Anatomy of the Moltbook Social Graph January 31 at https://github.com/daveholtz/moltbook_scraper

Houghtaling, M. A. *et al*. (2024): Standardizing an Ontology for Ethically Aligned Robotic and Autonomous Systems *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 54, no. 3, pp. 1791-1804 doi: 10.1109/TSMC.2023.3330981

Kalai, Adam Kauman et al. (2025): Why Language Models Hallucinate Sep.4. https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf

Málik, József Zoltán (2016): Ethical Considerations in Digital World. *Jogelméleti Szemle* 2016/2 80-90.

Neumann, János (1972): *A számológép és az agy* Gondolat, 1972 (2.kiad.) Eredetije: The Computer and the Brain Yale University Press, 1959

Presner, Todd (2024): *Ethics of the Algorithm: Digital Humanities and Holocaust Memory* Princeton University Press

Prestes, E. *et al*., (2021): The First Global Ontological Standard for Ethically Driven Robotics and Automation Systems *IEEE Robotics & Automation Magazine*, vol. 28, no. 4, pp. 120-124 doi: 10.1109/MRA.2021.3117414

Prisznyák, Alexandra (2023): Etikus AI: Javaslat az európai uniós megbízható AI-szabályozás hiányosságainak áthidalására és a gyakorlati implementáció támogatására *Gazdaság és Pénzügy* (10) 2: 169-195. DOI: 10.33926/GP.2023.2.4

Schultz, Mario D. et al. (2025): Digital ethicswashing: a systematic review and a process-perception-outcome framework. *AI and Ethics* (5) 805–818 https://doi.org/10.1007/s43681-024-00430-9

Šekrst, Kristina (2025): *The Illusion Engine: The Quest for Machine Consciousness* Springer

Z. Karvalics, László (2022): *Az infodémiától a tudáskormányzásig. Világjárvány és világrendszer információkutatói szemmel* L'Harmattan, Budapest