

Hivatalos bírálói vélemény Héder Mihály „Az MI etika új hullámai” c. MTA doktori értekezéséről

Héder Mihály „Az MI etika új hullámai” címmel MTA doktori értekezésént benyújtott szövege meglátásom szerint számos izgalmas, a releváns hazai és időnként még a nemzetközi szakirodalomban is újdonságnak tekinthető gondolatot, meglátást tartalmaz. Átfogó koncepciót, összefüggő gondolatmenetet ugyanakkor nem fedeztem föl a szövegben, bár az igaz, hogy – ahogy azt a szerző maga is többször hangsúlyozza – egyfajta megközelítésmód, perspektíva, nevezetesen a konstruktóri szemszög érvényesítése koherensen végigvonul az értekezésen. A szöveg sokkal inkább lazábban összefüggő elemekből építkező mozaik- vagy kaleidoszkópszerű struktúrát mutat. Jómagam, filozófus lévén, érdeklődéssel és örömmel olvastam a szakterületemtől távolabb eső, a mesterséges intelligenciával (a továbbiakban: MI) kapcsolatos számtalan különböző tervezési, működtetési, betanítási, sőt gazdasági és társadalmi kérdésekről.

Az értekezés informatív betekintést nyújt a tervezőmérnöki dekompozíció lényegébe, illetve legalapvetőbb tulajdonságaiba és módszereibe; a technológiába zárttság fogalmába és megvalósulásának okaiba; az MI szabályozás különböző formáiba, nehézségeivel, illetve a szabályozások előtt álló feladatokkal együtt; az MI-re jellemző transzparenciahány technikai és architektikai okaiba; az MI konstruktőrök azon erőfeszítéseinek történetébe, amely arra irányul, hogy az MI képes legyen saját működésének megmagyarázására; illetve az MI emberi személyeket is érintő működése során a méltányosságot biztosítani igyekvő formális, algoritmikus megoldások főbb jellemzőibe és problémáiba. A sort valószínűleg még tovább lehetne folytatni. A szöveg e részei számomra valószínűvé teszik, hogy a szerző alaposan ismeri és érti e témákat, magabiztosan és gondosan kezeli annak elsődleges és másodlagos szakirodalmát, egyes kérdések kapcsán önálló koncepcióval, megoldási javaslattal is képes előállni. Ugyanakkor, nem lévén e területek kompetens szakértője, mindebben természetesen inkább csak a benyomásaimra hagyatkozom. Az értekezés szövege továbbá többnyire transzparens és jól követhető, a szerző általában világosan, jól fókuszálva, érthetően fogalmaz – ahol mégsem, azt külön jelzem az alábbiakban.

Az értekezés fenti erényeinek elismerése mellett a továbbiakban, kompetenciámnak megfelelően a szöveg filozófiai részeivel, filozófiai aspektusaival foglalkozom, tekintettel arra, hogy címe és vállalt fő témája szerint alapvetően morálfilozófiai értekezésről van szó. E szempontból sajnos már egyáltalán nem túl pozitív az összkép. Meglátásom szerint filozófiai részekben valójában nem bővelkedik a szöveg, méghozzá két értelemben sem. Egyrészt ugyanis bár a cím ígérete szerint az értekezés központi témája az MI etika, az etikai vagy ezt megalapozó filozófiai tartalmi részek kisebbségben vannak az egész művet tekintve. Természetesen tisztában vagyok vele, hogy nem igazán lehet apriori elhatárolni a filozófiai és a nem-filozófiai témákat egymástól, mégis határozottan az a benyomásom, hogy a szerző kevés helyen foglalkozik olyan témákkal, amelyeket a morálfilozófusok, és más filozófiai diszciplínával foglalkozó szerzők az MI-vel kapcsolatos publikációikban tárgyalnak. Ebben a bevett értelemben voltaképpen

az értekezés (a Konklúzió előtti) két utolsó (11-12.) fejezete, illetve a 3., a 7. és a 10. fejezet egyes, hosszabb-rövidebb részei tekinthetők filozófiai fókuszúnak. A jelölt továbbá arról sem győzi meg az olvasót, hogy olyan témákat tárgyal, amelyek bár az MI filozófiai szakirodalmában nem kapnak helyet, de helyet kéne kapniuk, mivel olyan problémafelvetések, fogalmi újítások vagy újfajta megközelítésmódok szerepelnek részletesen kifejtve és alátámasztva benne, amelyek filozófiailag relevánsak.

Nem találhatunk sok filozófiát a szövegben ugyanakkor a szó egy másik értelmében sem. Nemcsak tartalmilag, módszertanilag sem bővelkedik az értekezés a filozófiai részekben. Nem találkozzunk egyes problémák részletes, alapos – fogalmi vagy történeti – bemutatásával, elemzésével; a problémák megoldására irányuló átfogó koncepciók mélyreható prezentációjával és kritikájával (ez alól az utolsó fejezet első részében szereplő Polányi-elemzés jelent némi üdítő kivételt); és végül egyetlen tárgyalt kérdés kapcsán sem mutatja be a szerző részletesen a szembenálló álláspontok melletti és elleni érveket. Ezek helyett sajnos legtöbbször igencsak felszínes, rövid utalás történik az MI-vel kapcsolatos filozófiai, morálfilozófiai problémákra, álláspontokra, az érvek rekonstrukciója inkább csak odavetett, a fogalmak használata sokszor pontatlan és nem igazán következetes; mindez különösen visszás annak fényében, hogy a szerző egyes helyeken határozott, sőt nagyívű állásfoglalásokat tesz átfogó filozófiai kérdésekről. Mindezek miatt benyomásom szerint a szöveg olvasója nem kap alapos és kellően részletezett képet sem arról, milyen filozófiai jellegű problémákat vet föl az MI fejlődése, és főleg nem arról, mit jelent filozófiailag reflektálni e problémákra. Az alábbiakban a fentiek alátámasztása és illusztrációjaként néhány különösen neuralgikus helyet fogok bemutatni, elemezni a szövegből.

Az MI lehetőségével kapcsolatos filozófiai probléma kezelése

Az egyik, filozófiai szempontból legproblematiszabb rész az értekezés 3. fejezete, annak is főleg utolsó, „Rossz érvek az MI ellen” címet viselő szakasza. Bár a cím, és a szakasz tartalma – a fejezet korábbi megállapításaival együtt – nyilvánvalóvá teszi, hogy a szerző itt egyes, valamilyen értelemben MI-ellenes érvekkel vitatkozik, valójában a szövegből semmi pontosabb nem derül ki. Először is, nem tudjuk meg azt, milyen értelemben irányulnak ezek az érvek az MI ellen – az MI lehetőségét kérdőjelezi meg vajon, a lehetőség valamelyik, közelebbről nem meghatározott (logikai, metafizikai vagy fizikai) értelmében, vagy az MI valós létezése, esetleg az MI fejlesztése ellen szólnak? Másodszor, egyáltalán nem derül ki a szövegből, mik is volnának a szóban forgó érvek; a szerző nemhogy részletes logikai szerkezetük felvázolására nem vállalkozik, de még informális módon sem ismerteti a szóban forgó érvelések lényegét. Végezetül, az is eléggé homályosan és nehezen érthetően derül csak ki, miért volnának e közelebből nem azonosítható érvek a szerző szerint rosszak.

A szöveg egésze alapján a leginkább az az értelmezés tűnik helyesnek, amely szerint a szerző e szakaszban – és az azt megelőző releváns részekben – az MI lehetőségével szembeni filozófiai érvekkel foglalkozik. Sajnos ezek közül egyiket sem mutatja be vagy elemzi részletesen, a

nevezetes érvek közül Searle, Penrose, Dreyfus és Collins érvét pusztán megemlíti a szerző. Egy érvet kicsit hosszabban – nagyjából egy oldalon (52-53.) keresztül – tárgyal a szöveg, nevezetesen azt, amely szerint míg az MI pusztán explicit szabályokat követ működése során, addig az emberek ezen túlmenő, hallgatólagos tudással – és, gondolom, ebből fakadó intelligens viselkedéssel (de ez már az én hozzátételem) – rendelkeznek. Bár nem mondja ki, nagy valószínűséggel úgy értelmezi a szerző ezt az érvet, hogy annak konklúziója szerint e különbségből arra következtethetünk, hogy a gépi MI nem tekinthető intelligensnek (valamilyen, releváns értelemben). Mint láthatjuk, az érv szerkezete itt is igencsak hiányosan kerül megfogalmazásra a szövegben, de legalább valamelyest láthatók a valószínűsíthető kiegészítés körvonalai. Azonban az érv még e kiegészítésekkel sem igazán érthető, pláne nem értékelhető, hiszen a szerző nem definiálja mit ért explicit (és esetleg implicit) szabálykövetésen, és miért állna ez szemben a hallgatólagos tudással – ehhez például hasznos lett volna a szabálykövetés terjedelmes filozófiai irodalmát valamelyest megidézni. Éppígy nem egészen világos, miért vonná maga után e különbség (ha tényleg fennáll) az MI lehetetlenségét. Ugyanakkor még egyszer hangsúlyoznám, hogy a többi kritizált érv még ennyire sincs kibontva.

Arra a kérdésre, hogy miért volnának ezek a csak „hallgatólagosan” bemutatott és elemzett érvek rosszak, a szerző úgy válaszol, hogy azért, mivel ezen érvek kategóriahibán nyugszanak. (51., 54.) Miután nagy vonalakban ismerteti a kategóriahiba fogalmát, azt állítja, hogy azzal, hogy ezek az érvek a gépi MI-t Turing-gépnek, szintaktikai automatának, vagy explicit szabályvégrehajtónak tekintik, kategóriahibát követnek el. Érvelése szerint például a „Turing-gép” elnevezés absztrakt platóni ideára utal, és ezért kategóriahibás partikuláris fizikai tárgyakra alkalmazni, mint például egy számítógép. Nos, ez az érvelés vagy nagyon hiányos, vagy nyilvánvalóan téves. 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. Hogy a predikátumok jelölnek-e valamiféle entitásokat, és ha igen, azok ontológiailag hogyan jellemezhetőek, milyen ontológiai kategóriába tartoznak, az egy további kérdés, amely széleskörű metafizikai vizsgálódások tárgya. 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. Azt állítani, hogy ez kategóriahiba, a kategóriahiba fogalmának bizonyos mértékű félreértéséről tanúskodik, hiszen pontosan olyan, mintha valaki azt állítaná, hogy kategóriahiba az íróasztalomat bútornak nevezni, mivel a bútor egy absztrakt funkcionális predikátum, az íróasztalom pedig nagyon is konkrét fizikai tárgy, például van hőmérséklete. Éppígy, amikor a jelölt azt állítja, hogy „... bár sokan a fejükben gondolkodnak Turing-gépekről, senkinek sincs egy Turing-gép a zsebében.” (54.) az épp olyan, mintha azt állítaná: „Bár sokan a fejükben gondolkodnak bútorokról, senkinek nincs egy bútor a szobájában.” Hiszen csak székek, asztalok és szekrények vannak ott (bár persze ez is kérdéses lehet).

Egy szó, mint száz: absztrakt funkcionális predikátumokkal jellemezni konkrét partiklárís fizikai tárgyakat nem kategóriahiba, ezt lépten-nyomon megteesszük hétköznapi nyelvhasználatunk során.

Valószínűsítem ugyanakkor, hogy a jelölt igazából nem akarta valóban kategóriahibából fakadóan értelmetlennek nyilvánítani az általa kritizált érvek bizonyos (közelebbről meg nem nevezett) premisszáit. Az értekezés végén (205.) például már arról beszél, hogy elutasítja a számítógép Turing-gépre történő redukcióját – ami előfeltételezi, hogy az általa hamisnak tartott elgondolás mégiscsak értelmes állítás. Valószínűsítem, hogy sokkal inkább arról van szó, a szerző úgy gondolja az említett funkcionális predikátumok általi jellemzések félrevezetőek, vagy nem ragadják meg jól a gépi MI lényegét. Ahogy azt a 3. fejezet egy részében (50-51.) megállapítja, az ilyesfajta jellemzések tekinthetők a számítógép modelljeinek, és – teszem hozzá – mint minden modell esetében, kérdéses lehet, hogy egy adott kontextusban valóban adekvát-e a kérdéses dolgot az adott modellel reprezentálni. Ha tehát úgy érvelt volna a jelölt, hogy mivel a számítógépek konkrét fizikai tárgyak, amelyek fizikai paraméterekkel jellemezhetők, és e jellemzés döntő, lényeges abban a kérdésben, képesek lehetnek-e MI-ként funkcionálni, éppen ezért inadekvát e fizikai jellemzőktől elvonatkoztatva a 'Turing-gép' vagy az 'explicit szabálykövető' modelljét alkalmazni rá – nos, ez valóban izgalmas, érdekes argumentáció lehetne, de sajnos ilyesmivel nem áll elő a szerző az értekezésben. 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, lásd (Héder 2020: 95-105.) –, vagy hogy miért nem adekvát, ha a számítógép működését explicit szabálykövetésként írjuk le. Szerintem e kérdések alapos és kimerítő tárgyalása már önmagában egy fél nagydoktori értekezést igényelne. Mindenesetre még egyszer kiemelném: mindez pusztá spekuláció a részemről, az értekezés szövegében ugyanis sem a kritizált érvek rekonstrukciója, sem az ellenük történő érvelés pontos mibenléte nem azonosítható.

Ehhez a témakörhöz tartozik még értelmezésem szerint az a még a Bevezetésben szereplő megállapítás, miszerint a MI intelligencia megvalósulásának tagadása sokszor a „cél tábla mozgásával” történik, azaz az ellenérvek az MI fejlődésének megfelelően mindig újabb és újabb aspektusból, újabb és újabb tulajdonságok feltételezett hiánya alapján tartják lehetetlennek az MI megvalósulását (13.). Hogy itt pontosan milyen érvekre gondol a szerző, az sajnos megintcsak nem derül ki, így nem is nagyon tudjuk eldönteni, mennyire hihető ez az állítása. A kapcsolódó lábjegyzetben érvelési hibaként mutatja be ugyanezt, azzal, hogy végső soron az „emberi” fogalmának újra- és újradefiniálásával mindig tagadni lehet az MI megvalósíthatóságát.

Nos, nem tudom eldönteni a szerző pontosan mifajta érvekre gondol e kritika során, mivel nem említ egyet sem, pláne nem elemzi a kritizált szempontból azokat – magyarul meg sem kísérli alátámasztani állítását –, a legutóbb említett megjegyzés alapján valószínűsítem, hogy – részben legalábbis – itt is az MI lehetőségével szembeni filozófiai érvekre utal. Abban igaza van szerintem a jelöltnek, hogy – legalábbis amennyire ezt meg tudom ítélni – az MI

megvalósíthatóságával kapcsolatos filozófiai viták központi kérdése mindig is az volt, hogy vajon mesterséges létezők, konkrétan digitális számítógépek rendelkezhetnek-e kognitív képességekkel és állapotokkal abban az értelemben, amelyben hétköznapiak úgy véljük, mi emberek rendelkezünk ilyenekkel. Persze, ha ez a megállapítás nagyjából igaz, akkor az ezt tagadó álláspontok melletti érvekre nem áll a szerző célpontmozgatás-vádja, hiszen ezen érvek célja mindig ugyanaz: az emberi értelemben vett kognitív képességekkel rendelkező MI. Persze az a kérdés, hogy pontosan miben is állnak az emberihez hasonló kognitív képességek, magyarul mit értünk az ember kognitív képességein, állapotain, mi adja ezek azonosságfeltételeit, nos e kérdések az utóbbi 50-60 év elmefilozófiájának központi kérdései, amelyekben természetesen nincs egyetértés a téma kutatói közt. E terület problémáiról, eredményeiről, vitáiról sajnos szinte egy szó sem esik az értekezésben, ami lehetetlenné teszi, hogy akár csak vázlatos képet kapjunk az MI megvalósíthatósága körüli filozófiai vitákról. A szöveg vége felé a szerző ugyan megállapítja (álláspontom szerint helyesen), hogy az MI morális inklúzióját valamely kemény (intrinzikus) feltételhez kötő álláspont jó eséllyel támaszkodik az elme funkcionalista felfogására (183., 186., 189-90.). De hogy pontosan miért, az nem igazán derül ki a szövegből, a jelölt ugyanis nem fejt ki részletesen – de még vázlatosan sem igazán – a funkcionalista álláspontot, sem az MI-állapotok és az emberi kognitív állapotok közti hasonlóság vagy azonosság megalapozását a funkcionalista keretben megteremtő Reprezentációs Elme elméletét és a hozzá kapcsolódó komputacionalista elképzelést. Pláne nincs szó ezek továbbgondolásairól, kritikáiról, mint például a reprezentációk intencionalitáson keresztüli értelmezéséről, az intencionalitás naturalizálhatóságának problémáiról, vagy az intencionalitás és a fenomenális tudatosság viszonyának kérdéséről. Összességében tehát úgy tűnik, a jelölt nem foglalkozik érdemben az MI megvalósíthatósága körüli filozófiai kérdésekkel, ám mégis határozottan elutasítja az e kérdést negatívan megválaszoló álláspontokat.

Az MI etika főbb témáinak tárgyalása

Az értekezés 7. fejezete hivatott a szerző meghatározása szerint (111.) szerint az MI etika filozófiai szakirodalomban kanonizálódó kérdéseivel foglalkozni. Sajnos azonban az olvasó e fejezetben sem kap ízelítőt az e témakörhöz kapcsolódó filozófiai problémafelvetésekről, kidolgozott álláspontokról, érvekről, vitákról. Ezek helyett lényegében egy felsorolásszerű katalógusát kapjuk az MI-vel kapcsolatban fölmerülő etikai aggályoknak, röviden – nagyjából egy-egy oldal terjedelemben – meghatározva azok lényegét. Ezen etikai aggályok filozófiai megalapozása, különböző nézőpontú megközelítései és megoldási javaslatok nem kapnak sajnos helyet. Sem azt a kérdést nem tárgyalja a szöveg, milyen normatív etikai, morálfilozófiai alapelvekből vezethetők le e problémák, és jelent-e különbséget konceptualizálásukra nézve, ha különböző morálfilozófiai keretelméletekből indulunk ki, sem pedig azt, a különböző normatív etikai megközelítésmódok milyen különböző megoldási javaslatokat implicálják e problémának. Nem is beszélve az e konceptualizációk és megoldási javaslatok melletti és elleni érvek rekonstruálásáról. Messze a legnagyobb terjedelmet ezek helyett e fejezetben az MI elterjedése által okozott

munkaerőpiaci változások mikro- és makroökonómiai elemzése teszi ki, ami kétségkívül izgalmas téma, de semmiképpen nem tekinthető szorosan a filozófiai etika területéhez tartozónak.

Illusztrációként külön foglalkoznék a felelősségrés problémájának tárgyalásával, amely nagyjából másfél oldalnyi terjedelmet kapott az értekezésben (113-115.) (habár korábban egyszer már megemlíti a szerző az önvezető autók kapcsán, 63.). A jelölt itt helyesen állapítja meg, hogy a problémát elsősorban az MI viszonylag nagyfokú autonómiája, és ezzel összefüggésben az a tény váltja ki, hogy emberi ágensek nem képesek a működését teljes mértékben kontrollálni. Ez utóbbival kapcsolatban a probléma elméleti előfeltevéseként fogalmazza meg a szerző a felelősség kontroll-feltételét. Mindeközben ugyanakkor sajnos egyáltalán nem foglalkozik azzal a kérdéssel, a felelősség melyik értelmét, melyik felelősségfogalmat alkalmazzuk, amikor a felelősségrés fogalmáról beszélünk. Vajon itt morális, jogi, szerep- vagy oksági értelemben vett felelősségről van szó? És az elszámoltathatóság (hibáztatás és dicséret) értelmében vett felelősségről, vagy a kárviselési kötelezettségről beszél a szerző? Ezek egyáltalán nem mellékes kérdések, mert például a jogi felelősség kérdése egészen más problémákat vet föl, mint a morális – míg az utóbbi elsősorban valóban a kontroll és az MI morális státuszának kérdéséhez vezethet el, addig az előbbi inkább a kárviselés kockázatának allokációjával kapcsolatos kérdések felé terel minket (amely egyébként több jogtudós szerint nem vet föl különösebb elméleti nehézséget, a polgári jog jelenlegi eszközeivel, jogintézményeivel megoldható ugyanis). De még ha fel is tesszük, hogy itt a morális felelősségről van szó, akkor sem evidens a kontrollfeltétel relevanciája, léteznek ugyanis olyan – például konzekvencialista alapú – felelősségelméletek, amelyek szerint nincs szükség ennek figyelembevételére a felelősségtulajdonítási döntések megalapozott meghozatalához. Minderről sajnos szintén nem esik szó a témát tárgyaló rövid részben, mindössze egy bekezdésben, minden elméleti, morálfilozófiai kontextus nélkül megemlíti egy olyan –Champagne és Tonkens nevéhez fűződő – megoldási javaslatot, amely nem a közvetlen oksági befolyás képességét tekintené a felelősségtulajdonítás kritériumának.

Az MI morális státusza körüli vita bemutatása

Filozófiai szempontból a legkidolgozottabbnak álláspontom szerint az MI morális inklúziójának kérdéskörét tárgyaló 11. fejezet tekinthető. Sajnos azonban még ez is csak áttekintő, tág értelemben ismeretterjesztő jellegű szöveg, nem szerepelnek benne ugyanis részletes kifejtések, elemzések, érvelések, még a szerző saját megoldási javaslatai, az ártalom-aszimmetriára hivatkozó inklúzió melletti érvelés, és az MI sajátos, esetleg graduális morális státuszáról szóló koncepció esetében sem. Ehelyett megint inkább csak a probléma lényegének kissé felszínes, első megközelítés-szerű bemutatását, az inklúzió melletti álláspontok rövid taxonómiáját, a mellette szóló érvek vázlatos ismertetését kapjuk, az ellenérvek pedig épphogy csak megemlítődnek, mindössze két bekezdésben. Mindezt ráadásul mindenféle egységes alapállás vagy perspektíva nélkül kínálja föl az olvasó számára a szerző, ami néhány esetben ugyanannak a témának (pl. a szimpátia táguló köréről szóló elképzelésnek, a tudatosság mint kemény kritériumnak) a többszöri, redundáns és esetleges előkerülését eredményezik.

Jónéhány fogalmi következetlenség, érthetetlen vagy alátámasztatlan kijelentés, gondolatmenet szerepel továbbá még ebben a fejezetben. Például rögtön a probléma ismertetésénél így ír a szerző:

... a morális ágencia a felelősségrevonhatóságra és elszámoltathatóságra vonatkozik; a páciens státusz pedig az ágenssel szembeni morális kötelességekre, mint például annak az elismerésére, hogy az ágens 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. (182.)

Egyrészt nyilvánvaló az ellentmondás a két mondat között – míg az első szerint a páciensnek vannak kötelezettségei az ágens felé, addig a második szerint épp fordítva. Véletlen elütésre is gyanakodhatnánk, ám akkor sem volna érthető, hogyan szándékozta a szerző meghatározni a morális ágens és páciens fogalmait, az első mondatban ugyanis mindkét meghatározás valóban az ágensre vonatkozik: épp az indokolja, hogy a morális ágensek felelősségrevonhatóak és elszámoltathatóak legyenek, hogy kötelességeik vannak a morális páciensek iránt (akik persze lehetnek egyben ágensek is). Mellesleg a szerző e sikerületlen meghatározások után explicit módon kijelenti, hogy az MI morális inklúzióját érintő kérdés szempontjából nem választja külön az ágensi és a páciensi státuszt, ami szerintem nem a legszerencsésebb teoretikus döntés, lévén a két státusz tulajdonítása melletti és elleni érvek elég nagy mértékben különböznek egymástól.

Három sorral lejjebb ez áll: „Egy ijesztő, de az emberi társadalmak történetében nem példátlan logikai lehetőség a morális ágencia, felelősség, megbüntethetőség tulajdonítása a morális pácienssel szembeni kötelességek nélkül: egy háborúban gyakran így látják egymást a felek.” (uo.) Nem érthető, mire gondol itt a szerző – a háborút viselő embercsoport tagjai tényleg úgy látnák ellenfeleiket, mint akik felelősséget viselnek anélkül, hogy valamiféle kötelezettségeik lennének (például az ártatlan civilek biztonságának biztosítására)? Aligha. Szerintem legtöbbször sokkal inkább úgy vélik, a másik fél nagyon is rendelkezik ilyen kötelességekkel, ám megszegi azokat.

Szintén ugyanitt a jelölt ismerteti a morális inklúzió melletti érvelés három lehetséges módját, a kemény feltételre, a nagyobb rendszerben betöltött szerepre hivatkozó érveléseket, és az ún. relációs megközelítést. Különösebben nem részletezi ezek mibenlétét, amely így azt eredményezi, hogy nem lesz világos, mi különböztetné meg az utóbbi két koncepciót. Hiszen a relációs megközelítést is úgy határozza meg, mint „amely a morális befogadást a társadalmi kapcsolatok hálózatában betöltött pozíció alapján ítéli meg.” (183.) Mindkettő tehát egy nagyobb rendszerben betöltött szerepre, és így az MI nem intrinzikus, hanem relációs tulajdonságaira hivatkozik. Nem ártott volna tehát itt sem a bővebb kifejtés.

Ugyanezen az oldalon ezt olvashatjuk: „Egy másik kemény feltétel, a tudatosság elegendő alapot nyújthat a deontológiai etika szerinti befogadáshoz.” (183.) A tudatosságra mint kemény feltételre való hivatkozás ebben a kontextusban általában azzal a céllal történik – ahogy azt a szerző is említi máshol –, hogy megalapozza a szenvedésre és az élvezetre való képességet, és így a morális páciensi státuszt. Az ilyesfajta érvelések pedig egyértelműen nem deontologikus,

hanem konzekvencialista keretben futnak le, lévén a szenvedés és az élvezet növekedése vagy csökkenése bizonyos cselekedetek vagy folyamatok következményeiként valósulnak meg. Persze nem lehetetlen, hogy deontológiai keretben is adható hasonló konklúziójú érvelés, azonban ez közel sem bevett vagy evidens, így nem ártott volna, ha a jelölt valamelyest kifejtette volna, mire gondol.

Néhány oldallal később az alábbi megfogalmazással találkozhatunk: „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.” (192.) Nem egészen érthető, mire gondol a szerző itt. Egyrészt talán túlzás a passzív eutanáziát a jelenlegi gyakorlatnak tekinteni, másrészt a mellette szóló érvek éppen a szóban forgó személy morális státuszára, morális méltóságára vagy a szenvedés csökkentésének követelményére alapoznak – nem világos tehát, miért volna ellenpélda az eutanázia az élőlények mint élőlények morális státuszára.

A fejezet vége felé a morális inklúzió melletti saját érvének lényegét így foglalja össze a szerző:

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)

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 nagy-nak, 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.

A fent kiemelték csak a legszembetűnőbb problémák a fejezettel kapcsolatban, abban sajnos még további következtetések, pontatlanságok, kifejtetlenségek is találhatók. Mindez főleg azért őszintén sajnálatos, mivel épp ez a fejezet az, amely az értekezés címének és megadott témájának megfelelően az MI-vel kapcsolatos egyik legfontosabb kurrens filozófiai etikai témáról hivatott informálni az olvasót.

Emergencia, redukció és hallgatólagos tudás az MI-vel kapcsolatban

Ami az utolsó, 12. fejezetet illeti, itt először kapunk egy alapvetően informatív és korrektnek tűnő rövid, nagyjából 8-9 oldal terjedelmű összefoglalást Polányinak a személyes tudáson, a hallgatólagos tudáson és egy sajátos emergencia-fogalmon alapuló filozófiai koncepciójáról. Ezek után pedig, némileg még rövidebb terjedelemben e koncepciónak az MI-vel kapcsolatos

etikai, filozófiai problémákra való alkalmazhatóságát igyekszik megmutatni a szerző. Megíté-
lésem szerint ez utóbbi rész több szempontból is erősen problematikus.

Először is – számomra legalábbis – kevésbé válik világossá, pontosan hogyan kéne az MI-re és az azzal kapcsolatos problémákra alkalmazni Polányi koncepcióját. Ha jól értem valami olyas-
miről volna szó, hogy egyrészt az MI valamiféle emergens, sem fizikai sem tisztán szintaktikai
tulajdonságokra nem redukálható létező volna, másrészt pedig valamiféle hallgatólagos, sze-
mélyes tudással véteződünk föl a gépi MI működéséről szóló tapasztalataink során, amelyben
– Polányi elgondolásának megfelelően – kritika nélkül megbízhatunk. Ezen analógiák azonban
egyrészt, ismétlem, csak nagy vonalakban, kissé impresszionisztikusan bomlanak ki ahogy a
fejezet gondolatmenete halad, hiszen a szerző nem áll elő a kulcsfogalmak – hallgatólagos tu-
dás, személyes tudás, emergencia – pontos definícióival, és nem mutatja meg részletesen,
elemről-elemre, hogyan kellene ezeket az MI-re és az azzal kapcsolatos problémákra alkal-
mazni. Másrészt, éppen a pontosabb definíciók hiánya és az elnagyoltság miatt nehéz is érté-
kelni az alkalmazás és az analógiák adekvát mivoltát. Egy helyen például a szerző amellet ér-
vel, hogy nemcsak mi rendelkezhetünk hallgatólagos tudással az MI-ről, hanem bizonyosfajta
MI rendszerek maguk is rendelkezhetnek ilyen tudással. Erről így ír:

... a robotok képesek olyan feladatok végrehajtására, mint például a biciklizés, valamint a viselkedésük más
módon történő koordinálására, kérdések megválaszolására, intelligens és sikeres cselekvésre. Ez szemben állt
azzal az állásponttal, amely szerint csak az emberek rendelkeznek hallgatólagos tudással, ezáltal csak mi le-
szünk képesek olyan feladatokat megoldani, amelyekhez az szükséges. (207.)

Nem világos, miért volna a biciklizni-tudás képessége annak egyértelmű bizonyítéka, hogy a
képesség alanya hallgatólagos tudással rendelkezik. Ehhez azt kell feltételeznünk, hogy e ké-
pesség gyakorlásához feltétlenül szükség van effajta tudásra, magyarul csak az tudhat bicik-
lizni, 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, képességjellegű 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 exp-
licit (fokális) tudásuk alapján gyakorolják. Önmagában tehát a sikeres feladatteljesítésből, a
performatív sikerből nem vonhatunk le egyértelmű következtetést a kompetencia mibenlé-
tére. Érdeemes lehet ebből a szempontból összevetni az emberi sakkozók (nagy mesterek) sak-
kozni-tudását az olyan régivágású MI rendszerek (GOFAI) sakkozni tudásával, amelyek még
nem a gépi tanulás módszerével, hanem algoritmusok betáplálása révén tettek szert e képes-
ségekre, mint pl. a Deep Blue. Azt hiszem, továbbá, hogy Polányitól nem állt távol az az elkép-
zelés, hogy hallgatólagos tudáson alapuló készségeink explicit tudásként kezdték pályafutásu-
kat, csak aztán a begyakorlás révén váltak hallgatólagos, a feladatok végrehajtása során jár-
ulékosan jelenlévő tudássá.

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. Nos, amennyire tudom, Polányi posztkritikai tudományfilozófiájának e központi elképzelése nagyban támaszkodik a hagyomány fogalmára. Számára azon képességjellegű tudások érdemlik ki e bizalmi, kritikától mentes státuszt, amelyeket egy-egy tudományos iskola generációk során kitermelt magából, amelyeket ezért a tudósjelölt képzése során mintegy magába szív, és amelyek módszertani útmutatóként megszabják, hogyan kell bizonyos feladatokkal megbirkózni, egyben képessé teszik az e hagyományba beavatódott személyt erre. Nos, szerintem az MI-vel való interakció esetében ezek a feltételek nem állnak fenn. Itt 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. Így megítélésem szerint ez az analógia kissé sántít, ám ennek alaposabb megítélését persze, mint oly sokszor az értekezésben, a pontos definíciók, a pontos fogalomhasználat és a részletes kifejtések hiánya megakadályozza. Az a fajta leírás és érvelés, amellyel a fejezet utolsó szakaszában találkozunk egyébként inkább tűnik egyfajta fenomenológiai megközelítésnek, mintsem a személyes-tudás fogalom alkalmazásának, vagy a szerző explicit utasítása szerinti gondolkísérletnek. Gondolkísérletként nem nagyon azonosítható, mi volna a pontos kísérleti helyzet vagy forgatókönyv. És az sem válik világossá, igazából mi mellett kíván érvelni az MI-vel való interakció tapasztalatának fenomenológiája révén a jelölt.

A legfontosabb probléma e fejezettel azonban igazából az, hogy egyáltalán nem tükrözi a téma kiterjedt szakirodalmának ismeretét. Ameddig pusztán Polányi nézeteinek filológiai igényű feltárásáról, exegéziséről van szó, addig teljes mértékben elfogadható, hogy csak az ő írásaira, és az ezekre reflektáló szakirodalomra támaszkodjék az elemző. Onnantól fogva azonban, hogy – akár Polányi művei, akár más elképzelések által inspirálódva – a szerző az MI jellegzetességeinek feltárását és az MI-vel kapcsolatos problémák megoldását gondolja véghezvinni az emergencia, a személyes tudás, a hallgatólagos tudás stb. fogalmai révén, álláspontom szerint muszáj szembesülnie az e témákban született kurrens elképzelésekkel. Márpedig az a helyzet, hogy mind a propozicionális (know what) és képességszerű (know how) tudás – a személyes és hallgatólagos tudás szempontjából kulcsfontosságú – viszonyáról, de még inkább az emergencia, a redukció, a rétegzett valóság fogalmáról könyvtárnyi szakirodalom született az utóbbi évtizedekben, amelyet legalább főbb vonalaiban, a legfontosabb megkülönböztetéseket, érveket és álláspontokat ismertetve, érdemes lett volna a szerzőnek az elemzésbe bevonnia. Ennek hiányában egyszerűen nem igazán érthetők és értékelhetők azon állásfoglalásai, amelyek az MI és/vagy a számítógépek emergens jellegéről, redukálhatatlanságáról vagy a már említett hallgatólagos tudásukról szólnak. Nem derül ki ugyanis, pontosan mit ért itt emergencián, redukción, avagy nem explicit képességen a szerző, milyen érvek alapján tulajdonítja ezeket az MI-nek, vagy nekünk. Számtalan finom megkülönböztetésen alapuló különböző változatai vannak ugyanis forgalomban e fogalmaknak és érveknek a kortárs szakirodalomban. E hiányosság sajnos tovább erősíti az értekezés elnagyolt, kissé felszínes jellegéről szóló olvasói benyomást.

Ez a fejezet sem mentes továbbá a pontatlan megfogalmazásoktól, fogalmi inkonzisztenciáktól, alátámasztatlan állításoktól. Íme néhány. A szerző egy helyen így fogalmaz: „Polányi azt állítja, hogy a személy nem redukálható a fizikában vagy más tudományágakban ismert fogalmakra (ahogyan azt az ő idejében meglehetősen pozitivistá szemlélettel felfogták).” (201.) 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). Már csak azért sem, mert az ilyesfajta metafizikai redukciónak már az ötletét is értelmetlennek, vagy legalábbis tudománytalannak tartották volna.

Kicsit alább a következő megfogalmazással találkozhatunk:

A tudományok ismeretének struktúrája visszatükrözi saját tárgyát, ezért a *mérnöki tudományok* és a biológia bizonyos teleológiai elemeket és a működési elvek tanulmányozását is magában foglalja. Minden arra irányuló kísérlet, hogy ezen tudományágak természetét a fizikáéhoz és a kémiáéhoz közelítsük – különösen a teleológiai összetevő eltávolítása révén – károsnak bizonyult (Polányi erre a biológia példáját említi, de mindez a mérnöki tudományokra is igaz). Ha pedig mindez rendszeressé válik, az a valóságérzékelés elvesztéséhez, gyengüléséhez vezet. (206., kiemelés az eredetiben)

Nos, nem egészen világos sem az, mit ért itt 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. Persze lehet, hogy itt nem ezekről a szempontokról van szó, ám hogy akkor milyen egyéb megfontolásokról, az sajnos nem derül ki a szövegből.

A fejezet vége felé pedig ezzel a magyarázattal találkozhatunk:

Egy menedék a robotok emergenciájával és hallgatólagos tudásával kapcsolatos következtetés elől az lenne, ha azt követelnénk, hogy csak biológiai lények, azaz állatok sorolhatók ezekbe a kategóriákba. Ekkor „de a gépek nem képesek X-re” sémában az X helyére a biológiai alapokat helyettesítenénk be.

Ha elvégezzük a javasolt behelyettesítést, ezt kapjuk: A gépek nem képesek a biológiai alapokra. Ez így nyilvánvalóan értelmetlen állítást eredményez, így talán valamiféle konkrét képességre mint biológiai alapra gondol a szerző. Vagy valamely képesség biológiai alapjára (de persze ekkor is nehezen érthető lesz a mondat, hiszen a „nem képesek X-re” helyébe valamiféle képeséget vagy annak gyakorlását lehetne csak behelyettesíteni, nem annak alapját). Nem jöttem rá, valójában mit akart közölni itt a szerző.

További apróbb észrevételek

Jónéhány további fogalmi pontatlanság, inkonzisztencia, elnagyolt megfogalmazás, kevésbé alátámasztott kijelentés található még az értekezés szövegében. Bírálatomat ezek közül néhánynak a bemutatásával szeretném befejezni.

- A szerző előszeretettel használja az aluldetermináltság fogalmát, ugyanakkor sajnos nem mindig pontosan ugyanabban az értelemben. Első előfordulásakor lábjegyzetben említést tesz a fogalom tudományfilozófiai eredetéről, ugyanakkor nem magyarázza el azt – ezt később, a 73. oldalon valamelyest pótolja –, ellenben hozzáteszi, hogy a fogalmat bevezető szerzőkhöz hasonlóan „tágabb” (nem tudjuk meg, minél tágabb) értelemben fogja használni azt. A fogalom eredetileg arra a gondolatra utalt, hogy a tapasztalati, kísérleti evidenciák alapján sokszor nem lehet dönteni arról, a rivális és egymással inkompatibilis magyarázó elméletek közül melyiket válasszuk. Az evidenciák tehát aluldeterminálják az elméletválasztást. Ehhez képest az első előfordulás eképpen hangzik a szövegben: „A konstruktóri csapat helyzete az elképzelhető problémák extrém aluldetermináltsága miatt teljesen más morális keretekben érthető meg, mint a sofőré.” (17-18.) Nem világos ugyanakkor, mi által lennének a problémák aluldeterminálva. A kontextus alapján úgy tűnik, itt a kifejezés nagyjából a „bizonytalan”, „homályos körvonalú” értelmében szerepel. Máshol az önvezető autók tervezési problémáival kapcsolatban így ír: „Minden terv aluldeterminált az általa elkészült, megfogható végtermékkel összehasonlítva és viszont.” (63.) Itt elsőre úgy tűnik, arról van szó, hogy a tervek szükségképp kevésbé specifikáltak, mint a végtermék, a tervezők nem tudnak ugyanis minden későbbi körülményt előre látni. Itt talán „alulspecifikált”-nak adhatjuk vissza leginkább a kifejezés értelmét. Ez az értelmezés azonban megdől az utolsó, kissé odavetett „és viszont” kitételen, a többé vagy kevésbé specifikáltság ugyanis nem lehet szimmetrikus viszony. Így nem igazán derül ki, milyen értelemben használja a fogalmat itt a szerző. Hasonlóan nehezen értelmezhető a fogalom abban a mondatban, amely „a terv és az implementáció kiküszöbölhetetlenül aluldeterminált” viszonyáról beszél. Egyes helyeken aztán az eredeti tudományfilozófiai jelentésben, vagyis egy döntés, választás valami általi aluldetermináltsága értelmében használja a szerző a kifejezést. (50., 52.) Az esetleges félreértések elkerülése, és az értelmezés megkönnyítése szempontjából érdemes lett volna a különböző értelmekre különböző kifejezéseket használni, vagy az eredetitől eltérő használatokat jelezni, elmagyarázni.

- A 10. fejezetben fontos szerepet kap a kontraktárius (vagy kontraktariánus – az előbbi a bevettebb magyar, az utóbbi a szerző által használt alak) megközelítésmód. Ez a szerző meghatározása szerint azt az álláspontot testesíti meg, mely szerint „a jóként vagy megfelelőként elfogadott normák csupán egy olyan társadalmi kompromisszum eredményei, melyet valójában az érintett felek önös érdekei motiváltak. Ez az álláspont szemben áll más metaetikai álláspontokkal, mint például a deontológia.” (177.) Tekintsünk most el attól, hogy a deontologikus etika nem metaetikai, hanem normatív etikai álláspont, és fogadjuk el ezt a meghatározást. Ekkor viszont érthetetlen lesz a fogalom további alkalmazása, amelynek egy központi

példája az önvezető autók előtt álló morális probléma megoldása lesz a szövegben, amellyel kapcsolatban a jelölt így ír:

ilyen helyzetekben az önvezető autó intenzív fékezéssel reagálhat, de soha nem térhet ki. Ez a javaslat magában hordozza a jó, szabály-alapú rendszerek ismertetőjegyeit: könnyen megvalósítható és érthető, emellett kiszámítható viselkedést eredményez. (178.)

Nem világos, ez a javaslat, miért volna a fenti értelemben kontraktárius, hiszen nem tudjuk, az érintett felek mely önös érdekei közti kompromisszumként születik meg. A javaslat indokolása fényében inkább tűnik egyfajta szabályutilitarista megoldásnak, hiszen a szerző érvelése szerint leginkább az szól mellette, hogy elterjedt használata és ismerete esetén a legjobb következményekhez vezet. Ettől még tekinthetnénk akár kontraktárius elgondolásnak is, de ezt nem alapozza meg a szöveg.

- A 2. fejezet egy helyén az alábbi gondolatokkal találkozunk:

... megpróbálkozhatunk annak a kérdésnek a feltevésével, hogy a természetes intelligenciát milyen modulok alkotják, és ezek között milyen interfészeket tudunk azonosítani? Ami azt illeti, a kognitív tudomány korai szakaszában, amikor spekulatív számítógépmodellekkel próbálták leképezni az intelligenciát, előállítottak néhány hipotézist. Csakhogy az elme szigorú modularitása, mint tudományos modell, nem volt sikeres. Helyette egy szövevényes, idegi és hormonális utakon is önmaga részeivel összeköttetésben álló neurológiai leírását kaptuk az emlősök és az ember agyának. (45.)

Bár jómagam nem vagyok sem kognitív tudós, sem agykutató, amennyire meg tudom ítélni, ezek a megállapítások túlságosan sommásak, és ebben a formában semmiképp sem érvényesek. Egyrészt az elme moduláris elmélete legjobb tudomásom szerint a mai napig valid – persze részleteiben vitatott – elképzelés a releváns diszciplínákban, ahogy az agy moduláris elrendeződése is, vagyis az az álláspont, hogy léteznek különböző funkciókért felelős, többé-kevésbé jól beazonosítható agyterületek. Természetesen ezen tág értelemben vett moduláris keretelméleteken belül folyamatosan változó, pontosabbá váló konkrét elméleti elképzelések léteznek. Amennyire látom továbbá, az hogy az agy neurális hálózatának leírása szövevényes rendszert mutat, nem cáfolja a fentieket, hiszen evidensen kompatibilis az elme modularitásával – a moduláris elme fizikai alapjaként miért ne szolgálhatna egy szövevényes hálózatként leírható neurális rendszer? –, de még azt hiszem az agy modularitásával is, amennyiben azt kiegészítjük a plaszticitás tézisével.

- Szintén túl sommásnak tűnik ugyanebben a fejezetben az a megjegyzés, mely a „... szakadatlanul a saját hasznát kereső egyén koncepciójára épülő közgazdaságtanok”-ról beszél. (22.) Nem tudom, pontosan milyen közgazdaságtanokra gondol itt a szerző, de én személy szerint nem ismerek ilyet. Talán egy félreértésről van szó, és valójában a saját *célját* szakadatlanul kereső egyén koncepcióját alapul vevő klasszikus és neoklasszikus közgazdaságtani iskolákra gondol. 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.

- Szintén a 2. fejezetben:

Észszerűbb ezért a műszaki tudást minden szinten átjáró normativitást tekintenünk a megkülönböztetés alapjának. A deskriptív tudományok nem tudnak számot adni arról, hogy hogyan kell azokat a normatív terveket előállítani, amelyek a célszerű és hatékony technológiát eredményezik. (25.)

Nem világos, milyen értelemben használja itt 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.

- A 3. fejezetnek a kategóriahibákkal foglalkozó részében olvashatjuk, hogy „... ezt a számot kéknek nevezni kategóriahiba, mert a számok ontológiai kategóriája nem rendelkezik színnel, így kék sem lehet.” (49.) Ez egészen biztos, ugyanis egy kategória tényleg ne rendelkezhet színnel, ezt fölvetni valóban kategóriahiba volna. Ám ehhez fölösleges megemlíteni a számokat, hiszen nemcsak a számok kategóriájára, de az összes kategóriára igaz ez. A szerző azt hiszem valójában nem a számok kategóriájáról akarta állítani, hogy nem rendelkezhet színnel, hanem e kategória alá tartozó feltételezett entitásokról, vagyis a számokról.

- A 6. fejezetben az MI szabályozás kapcsán olvashatunk arról, hogy

Végül, igen meggyőző érvek támasztják alá azt az állítást, hogy az emberi döntéshozatallal összehasonlítva az MI magas szintű átláthatóságára vonatkozó követelmény nem más, mint kettős mérce. Ugyanis, az embertársaink által kivitelezett döntéshozási folyamatok átláthatósága is igen alacsony. Nyilvánvaló, hogy nem áll módunkban az agyat bármilyen érdemleges szinten vizsgálni egy döntés meghozatala közben, sőt, maga a döntéshozó sem képes – még a legjobb szándékkal sem – mindenről beszámolni, hiszen a kapcsolódó folyamatok egy része még magát a döntést meghozó személy előtt is homályos.

Ez nagyon izgalmas meglátás, de nem világos pontosan mire gondol a szerző, és mivel támasztaná alá állítását. Érdeemes lett volna az utóbbi évtizedek nagyhatású kísérleti szociálpszichológiai és morálpszichológiai eredményeit ezen a ponton bevonni az elemzésbe, hiszen azok éppen azt a szituacionista elgondolást látszanak támogatni, hogy az ágensek cselekvéseit, döntéseit nagymértékben befolyásolják az adott cselekvési szituáció olyan vonásai, amelyekről maga az alany nem szerez tudomást, azok nem tudatosulnak benne.

Összesített értékelés

Az értekezés szövegének a bírálat elején számbavett, filozófiaiak tekinthető részei a fent kifejtett hiányosságok miatt álláspontom szerint nem érik el az akadémiai doktori minősítéshez szükséges színvonalat. Sem a tartalom filozófiatudományi megalapozottsága, sem a kifejtés

alapossága nem mutat olyan kvalitásokat, amelyek elvárhatók volnának egy sikeres akadémiai doktori értekezéstől.

Ugyanakkor javasolom az értekezés nyilvános vitára bocsátását, amire két indokom van. Egyrészt előfordulhat, hogy az általam jelzett kritikai észrevételek kapcsán egyes kérdésekben tévedek, vagy rosszul értelmeztem az értekezés valamely részét. Ez egyaránt kiderülhet a többi bírálóból, valamint a jelölt érdemi válaszából. Másrészt, ahogy azt a bíráló elején jeleztem, az értekezés számos, filozófiaiak nem igazán nevezhető részt tartalmaz, amelyek kritikai megítélésére kevésbé tartom magam kompetensnek. Előfordulhat tehát, hogy bírálótársaim bírálata alapján a Bíráló Bizottság úgy ítéli meg, hogy az értekezés egyéb érdemei miatt, filozófiatudományi hiányosságai ellenére alkalmas a doktori fokozat odaítélésének megalapozására. Az említett kérdésekben álláspontom szerint a Bíráló Bizottságnak kell a védés során és után döntenie, ez indokolja tehát javaslatomat.



Budapest, 2026. február 17.

Márton Miklós