Az MI etika új hullámai

Héder Mihály

2024

MTA Doktori értekezés
Filozófiai és Történettudományok Osztálya
Tudomány- és Technikatörténeti Osztályközi Tudományos Bizottság

Tartalomjegyzék

Előszó

1981. október 19-én Tokióban Moto-Oka Tohru a mesterséges intelligencia (MI) történetének egyik legfontosabb beszédét tartotta.[1] Az általa vezetett „ötödik generációs számítógépek” projekt első nemzetközi konferenciáján foglalta össze víziója lényegét arról, hogy az MI milyen szerepet játszhat a ’90-es évek társadalmában, ha megfelelően valósítják meg.

A japán kormány által busásan támogatott tíz éves program sokkolta a világot és felrázta az MI télben szendergő versenytársakat. A támogatás nagy mértéke és legalább egy évtizedes ígérete komoly állami elköteleződést mutatott. A technológiai terv káprázatosan ambiciózus volt, de köszönhetően annak, hogy Moto-Oka és egy szűkebb csapat már 1978-ban elkezdett dolgozni rajta, nagyon részletes és jól alátámasztott, ezáltal hihető is volt. Ráadásul mindez a „japán minőségi forradalomként” is leírható technikatörténeti fejezet érett szakaszában történt, amikor az olcsó kvarcórával, zsebszámológéppel, valamint a megbízható és megfizethető kisautóval szerzett tapasztalatok miatt a világon szinte senki nem kételkedett a japán ipar képességeiben.

Az ötödik generációs számítógép program kimondott célja Japán társadalmi kihívásainak megoldása volt a annak jelentős átalakításával. Az azonosított problémák: a fenyegető elöregedés, a gazdasági fejlettség megrekedése a gyáriparon kívül és ezáltal a társadalmi olló kinyílása, és Japán még mindig fennálló elzártsága a nemzetközi élettől.

A felvázolt megoldás pedig a felhasználóval természetes nyelven beszélni képes, ezáltal mindenki számára használható, nyelvek között fordítani tudó mesterséges intelligencia volt, amely mindemellett az akkor uralkodó szakértői rendszerek képességeivel is rendelkezett. Funkciói között megtalálható információ keresése adatbázisokban, logikai következtetés, és az olyan formálisan kifejezhető problémák megoldása, mint az útvonaloptimalizáció vagy a számítógéppel segített mérnöki

[1] Moto-Oka Tohru (1982).

tervezés. Ez, a köz minden tagja számára elérhető eszköz az 1990-es évek „információs társadalmát”[2] volt hivatott létrehozni.

Moto-Oka tökéletesen tisztában volt a program potenciáljával, és ebből fakadóan a projektcsapat felelősségével. Ezért embereit három hasonló méretű egységre osztotta: az elméleti munkatársakra, a számítógép mérnöki kérdéseivel foglalkozó szakemberekre és a társadalmi hatásokat elemző csapatra. Ez utóbbi vezetője Karatsu Hajime lett.

Karatsu fontolóra vette a jövőkutatás eszközeinek használatát, de annak elvi korlátai miatt elvetette azokat. Ehelyett felvázolt egy Japán számára kívánatos jövőbeli társadalmat, és azt a kérdést tette fel, hogy a jelenből milyen úton lehet eljutni oda.[3] Ezzel az ötödik generációs számítógép program az akkori informatikai projektekhez képes szokatlanul nyíltan tartalmazott egy normatív elemet, amely már a tervezés legelején hangsúlyosan megjelent.

Az MI etikájával a 2020-as években foglalkozó kutató számára lenyűgöző, hogy a ma vitatott témák közül mennyi megjelent már több mint 40 évvel ezelőtt is: a magánélet védelme, a méltányos MI működés, a munkaerőpiac megváltoztatása és az MI hatása az egyénre mind vizsgált témák voltak.

Könyvemet az ehhez hasonló élmények motiválják: nem az egykoron divatos koncepció, a „jövő-sokk”, hanem a „múlt-sokk”. A számítástechnika szakirodalmára mindig is jellemző volt, hogy néhány állócsillag megemlítésén kívül a közelmúlton túl, 2-3 évnél távolabbra már nem tekint vissza, mert az annál régebbi tudásra az elavultság gyanúja vetül a Moore törvénnyel leírt folyamatos fejlődés mellett. A mesterséges intelligencia ezt csak tovább élezte, a jelenleg népszerű nagy nyelvi modell technológiák, és az azokról írt cikkek naprakészségét hónapokban, sőt, hetekben mérik.

Talán ez az attitűd az oka annak, hogy bizonyos témák periodikusan, reflektálatlanul ismétlődnek az MI etikájának irodalmában is. Például Moto-Oka és csapata több olyan kérdést is újként tett fel, amivel kapcsolatban Norbert Wiener és a vele vitatkozók

[2] Aiso (1985). Maga az „Információs Társadalom” kifejezés is Japánban született meg az 1960-as évek végén, lásd Z. Karvalics (2007).

[3] Karatsu (1982).

releváns hipotézisekre jutottak az 1950-es években, és amelyeket sokan ma ismét új kérdésként vetnek fel. Az MI etika területére az elmúlt évtizedben beözönlött bölcsészek sokkal hosszabb időtávokon gondolkodnak, és megnyugtatják a műszakiakat abban, hogy valójában több ezer éves etikai problémákkal hadakoznak, így nem szabad elkeseredniük, ha azok rövid távon nem oldódnak meg. Ugyanakkor a 20. századi MI története e bölcsészek számára néhány nevezetes fejleményt leszámítva alig hozzáférhető, mert sok MI megoldás népszerűvé, elterjedtté sosem vált műszaki terminológiával megalkotott tervekben és ma már használhatatlan forráskódokban lelhető csak fel.

Azonban, amikor sikerül beazonosítani egy, a kortárs MI etikai vita számára is releváns, nagyrészt elfeledett forráshalmazt, az szerencsés aranyásó hangulatába hozhat minket.

Könyvemben az MI etika történetének teljes katalogizálására nem vállalkozom, inkább bizonyos kiválasztott kérdésekben villantom fel a történeti kontextus hasznosságát, miközben néhány saját MI-etikai tézist mutatok be.

Azért, hogy a saját téziseimet világosan elválasszam azok mondandójától, akiket elemzek, egyes szám első személyben fogalmazok. Amellett, hogy sokan stiláris okokból, a blogszerűség veszélyét elkerülendő sem javasolnák ezt egy tudományos monográfia esetén, egy további probléma, hogy az „*én így gondolom, én úgy veszem észre, én arra jutottam*" és hasonló megfogalmazások nem engedik felszívódni a szerzőt. Márpedig a szerző háttérbe vonulás hasznos lehet van azért, hogy az érvek dominálhassanak. Ez a stiláris választás tehát a témába való bevonódásomnak és az érvrendszerem befejezetlenségének a transzparens beismerése.

I. rész

1. Bevezetés

A mesterséges intelligencia (a továbbiakban MI) történetét ismerők köreiben általánosan elfogadott, hogy e technológia megítélését a kezdetektől napjainkig nagy amplitúdójú hullámzások jellemezték. Az aktuális technológiai áttöréseket felfokozott várokozások követték az alkalmazások és üzleti következmények terén. Ezek az elvárások olyan rövid időtávokkal operáltak, amelyek eleve kudarcra predesztinálták őket.

A 2020-as évek újdonságot hoztak abban, hogy az MI-vel kapcsolatos fő üzenet nem egy ígéret a közeljövőre, hanem az, hogy *az MI megérkezett*. A különféle állítólagos, hatalmas akadályokat – úgymint a kreativitásra való képesség, összetett, teljes körű intellektusra épülő feladatok teljesítése, változatos írásos tesztek teljesítése érettségitől jogászvizsgáig–, az MI iparág az elmúlt években leküzdötte. Az MI-k új hullámával kapcsolatos hírek özöne annyira eltompította a köz figyelmét, hogy amikor 2024 májusában a ChatGPT4 minden eddigi próbálkozásnál meggyőzőbben teljesítette a Turing-tesztet, az már egy mínuszos hírt is alig ért meg.[4]

Könyvem megértéséhez fontos kiinduló feltételezés, hogy az MI, mint kutatási terület eredeti célkitűzései megvalósultak. Ezt a feltételezést nem fogom tételesen, teljeskörűen alátámasztani, bár egyes alkalmazási osztályokat viszonylagos részletességgel bemutatok. Az MI jelenlegi technológiai érettségi szintjén az értekezés szempontjából már nem számít, ha esetleg akadnak is még olyan makacs, intelligenciát igénylő feladatosztályok, amelyek további fejlesztéseket követelnek meg.

Ennél sokkal fontosabbnak tűnik ugyanis az MI fejlődésével párhuzamosan felmerülő etikai kérdések története és várható jövője.

Könyvemben bemutatom a jelenleg kanonizálódó MI etika területét, és kritikai elemzésnek vetem alá a legfőbb irányokat.

Az elemzés és a kritika nem egy külső, kellő történeti távolságban elhelyezkedő szemlélő szemszögéből készül, hiszen kortársai vagyunk az MI napjainkban

[4] https://arxiv.org/html/2405.08007v1

tapasztalható fejlődésének, ami fontos limitációt jelent kutatásaim számára. Az időbeli távolság és az utólagos bölcsesség előnyeiről azért mondok le, mert a vizsgált folyamatok, az MI körül formálódó normatív rendszerek megértésére most is hatalmas az igény. Hozzáteszem, hogy az MI előző hullámai már lezártnak tekinthetők és ezek történeti elemzése sokat segíthet a jelen megértésében – ki is fogok térni néhány történetszálra. Összességében arra invitálom tehát az olvasót, hogy merüljünk alá együtt ebbe a most még igen képlékeny világba és bizonyos alaptézisekre támaszkodva próbáljuk megtalálni azt, ami időtállónak bizonyul.

E könyvben az MI kezdeteit az 1950-es évekre helyezem, ellenállva a tendenciának, amely ennél egyre korábbi dátumokat javasolna.[5]

[5] A legfőbb indokom erre az, hogy elektronikus számítógéppel megvalósított, ezáltal az emberi intelligencia sebességéhez mérhető, potenciálisan általános célú, intelligens viselkedést mutató megoldással ekkortól találkozott az emberiség, így az MI ekkortól empirikusan vizsgálható. Így Babbage és Lovelace munkáját vagy Čapek és Lang vízióját a robotokról (R.U.R ill. Metropolis) pre-MI munkáknak tekintem.

1955-1965: szimbólumok vs. számok
1955 - : szimbólumok vs. folytonos rendszerek
1955-1965: problémamegoldás vs. felismerés
1955-1965: pszichológia vs. neuropszichológia
1955-1965: teljesítmény vs. tanulás
1955-1965: soros vs. párhuzamos
1955-1965: heurisztika vs. algoritmusok
1955-1985: interpretált vs. fordított programnyelvek
1955-: szimuláció vs. mérnöki analízis

1960-: az emberek lecserélése vs. segítése
1960-: episztemológia vs. heurisztika
1965-1980: keresés vs. tudás
1965-1975: erő[*deduktív - HM*] vs. általánosság
1965-: kompetencia vs. teljesítmény
1965-1975: memória vs. feldolgozás
1965-1975: problémamegoldás vs. felismerés #2
1965-: tételbizonyítás vs. problémamegoldás
1965-: mérnöki vs. tudományos megközelítés

1970-1980: nyelv vs. specifikált feladatmegoldás [task]
1970-1980: procedurális vs. deklaratív reprezentáció
1970-1980: keretek [*Minsky keretelmélete - HM*] vs. atomok
1970-: észszerűség vs. érzelmek

1975-: játék vs. valós feladat
1975-: soros vs. párhuzamos #2
1975-: teljesítmény vs. tanulás
1975-: pszichológia vs. idegtudomány #2

1980-: soros vs. párhuzamos #3
1980-: problémamegoldás vs. felismerés #3
1980-: procedurális vs. deklaratív reprezentáció

1. ábra: Newell tematikus áttekintése az MI történetéről 1955-1980 között

Hogy mennyire nehéz is a vállalkozás, amibe belekezdek, azt jól illusztrálja Alan Newell 1982-es összefoglalója[6] az MI addigi történetével. Newell a mostanihoz hasonló léptékű növekedéssel akart számot vetni, hiszen az 1950-es évek és a 1980-as évek között a terület a sokszorosára nőtt. Az alábbi intellektuális térképen Newell dichotómiákat vázol fel, amelyek nagyjából évtizedes felbontásban ábrázolják az MI akkori dilemmáit.

[6] Newell „Intellektuális kérdések a mesterséges intelligencia történetében” c. 1982-es esszéje 1983-ban jelent meg egy kiadványkötetben, lásd (Newell 1983)

Allen Newell összegzésének jelentősége abban áll, hogy emlékeztethet minket arra, hogy mennyire esendő a kortárs technikatörténet. Az általa felsorolt „intellektuális kérdések” legalább fele jelentéktelen vagy a jelenlegi kérdésekre hatástalan, ilyen például a teljesítmény és a tanulás közötti megkülönböztetés[7], vagy az episztemológia és a heurisztikák közötti ellentét.[8] Egy másik részüket, mint például a lista elején (lásd az 1. ábrát) a „szimbólumok vs. számok” el sem tudnánk magyarázni egy mai egyetemi hallgatónak, mert az uralkodó keretrendszerben a számok szimbólumok. Hasonlóan, az interpretált nyelvekhez is adott a röptében fordítás, szigorúan soros MI architektúrát pedig csak szándékos és jól célzott erőfeszítéssel lehetne létrehozni és persze semmi értelme nem lenne.

Ennek az oka az, hogy az MI alkotás módszerei, de különösen a módszerek mögött álló, az intelligencia működésére vonatkozó előfeltevések jelentősen megváltoztak.

A tény, hogy e dichotómiák közül ma igen keveset gondolnánk relevánsnak intő példa arra, hogy a most tapasztalt sarkalatos kérdések ugyancsak meghaladottak lehetnek néhány év múlva.

Az ’50-es évektől napjainkig értelmezett, leszűkített intervallum is számos izgalmat tartogatott, gyakorlatilag minden évtizedre jutott jó néhány olyan kordokumentum, amely negatív ítéletet mondott az MI felett és egyben minden további támogatás megvonását javasolta, a fellángoló lelkesedést naivitásnak bélyegezte.

A '60-as évek relatív[9] kudarcot hoztak a gépi fordítás terén, a '70-es évek meghatározó dokumentuma a Lighthill jelentés[10], a '80-as években a „tudásmérnökök” és a logikai

[7] Az igen erős számítási teljesítménykorlátok miatt akkoriban a válaszidő órákat, (vagy akár) napokat jelentett; erre vonatkozik a dichotómia: időben szeretnénk választ kapni vagy azt szeretnénk, hogy a rendszer tanuljon? Természetesen nem állítom, hogy a válaszidő kérdése teljesen lényegtelen, csak azt, hogy a mi vizsgálati szintünkön nem fog szerepet játszani.

[8] Newell számára ebben a kontextusban az episztemológia a tudomány módszerei által történő tudásszerzést jelentette, míg a heurisztika olyan módszer általi tudásszerzést, amelyről tudjuk, hogy nem tudományos. Ma mindkettőt az ismeretelmélet alá vennénk, de ami még fontosabb, hogy az MI ágensek episztemikus képességeiben mindkettő megjelenik, de egyikre sem ezekkel a címkékkel referálnak.

[9] A kudarc az elvárásokhoz képest értelmezendő: „orosz-angol fordítás az évtized végére” (Héder 2020, 21).

[10] Ez a jelentés annyira lesújtó értékelést adott az MI-ről, ahogy az Egyesült Királyságban később egy teljesen más néven éledt újjá a terület: „Intelligent Knowledge-Based Systems (IKBS)”. Lásd (Lighthill 1973), (Héder 2020, 79).

programozás[11] projektjei futottak zsákutcába, míg a ’90-es években a nagy szemantikai projektek[12] és a szakértői rendszerek ütköztek falakba.

A 2020-as évekre azonban a számos MI tél alkalmával hiányolt MI képességek és eredmények elkészültek és széles körben használhatók. Ennek a tagadása kizárólag a céltábla mozgatásával érhető fel. Ez általában a feladat újradefiniálásával áll elő. Ilyen például az, hogy nem elégszünk meg a funkcionáló megoldással, azt is elvárjuk, hogy az MI közben „tudja”, hogy mit csinál; nem elég hogy a befogadó kreatívnak tekinti a produktumot, kapcsolódjon hozzá az alkotás élménye, stb. Előző könyvemben ezt a problémát részletesen kifejtem és a *performatív* ill. *fenomenológiai* siker közötti megkülönböztetéssel kezelem[13]. Máskor az MI-szkepszis egyszerűen a „*no true scotsman”*[14] érvelési hibára épül.

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. A második legnagyobb kihívás a „megvalósíthatósági rés” („implementation gap”), amely arról a nehézségről szól, hogy még az MI alkotók köreiben megértett és konszenzusosan el is fogadott normák implementálása programkóddal nagyon nehéz.

---

11 Egy fontos példa erre a fentebb bemutatott „ötödik generációs számítógép”, amely a Prolog logikai programozási nyelv köré épült.

12 Lásd a FrameNet és a Cyc projekteket (Héder 2020, 76), valamit ide sorolom a szemantikus web 2000-es évekre datálható kifulladását is. Mindegyik újraéledése terítéken van, különösen a szemantikus webé, lásd (Héder 2013).

13 Ennek a kifejtését lásd (Héder 2020, 13-18)

14 Ezt az érvelési hibát Anthony Flew (1975) nevezte el így. A következőről van szó:
„-Egy skót sohasem tenne ilyet!
-XY, aki történetesen skót, épp ezt tette.
-Egy *igazi* skót sosem tenne ilyet!”
Az MI esetében az érvelési hibát ebben a formában látjuk viszont: „Az X feladatra, amely összetett intelligenciát/kreativitást/bölcsességet igényel, a gépek soha nem lesznek képesek” jelenti ki valaki. Néhány hónappal, évvel később az MI képessé válik megoldani a feladatot úgy, hogy az minden konvencionálisan elvárt paraméternek megfelel. „Az *igazi* X-re azonban most sem képesek a gépek”. Egy példa kedvéért az X helyére behelyettesíthetjük a rajzolást vagy a zenék komponálását; ekkor megtudjuk a fenti érvelőtől, hogy míg a gép csak korábban látott zenékből összegyúr egy statisztikailag megfelelő új változatot, addig az *igazi* zeneszerző valami többet csinál, például kifejezi a lelkét, az inspiráltság állapotában alkot stb. Ez a stratégia a végtelenségig alkalmazható, hiszen eljuthatunk oda, hogy az *igazi* egyenlő lesz az *ember*ivel, az *ember*inek viszont mindig lehet olyan definíciója, ami az MI-t kizárja.

Mindkét probléma közismert a terület kutatói számára, és éppen ezért a jelenleg folyó kanonizációs törekvések egyik fontos eleme a terminológiai tisztázó munka. Könyvem amellett, hogy ennek a munkának egy részét is bemutatja, megpróbál nagyobb távolságba tekinteni – és természetesen ezáltal még esendőbbé válik – anticipálni az MI etika új kihívásait és az ezekre adható válaszokat, azaz felvázolni egy megalapozott sejtést az MI etika új hullámairól.

Minden releváns MI etikai eredményt lehetetlen bemutatni vagy akár csak átlátni. Ugyan az általános célú generatív MI (pl. ChatGPT, Claude, stb) nem a legelterjedtebb MI alkalmazás – ugyanis az az internetes keresőmotorok kifinomult gépi tanulásra is épülő világa, amit szó szerint mindenki használ, akinek van internet-hozzáférése – ez az alkalmazás az, aminek a köz percepciójában a domináns jellemzője , hogy mesterséges intelligencia dolgozik benne. Mivel ezt az alkalmazás típust is százmilliók használják, és a felhasználók között felülreprezentáltak az akadémikusok, nem csoda, hogy a jelenség hatalmas, minden bizonnyal havi több száz, akár ezer cikkben mérhető MI etikával kapcsolatos vélemény-, és észrevételhullámot váltott ki. A munkák jelentősen eltérnek a problémakeretezés, az előfeltevések, a módszertan szintjén is – mondhatni, tudományos forradalom zajlik az MI etika világában.

Éppen ezért nagyon értékes számunkra valamiféle rendezési elv, amellyel megküzdhetünk az áradattal. E könyvben az alábbi **kulcsfontosságú kihívásokhoz** képest fogom elemezni és kritizálni a különféle álláspontokat és forrásokat, azonnal elismerve, hogy ez minden bizonnyal nem az egyetlen mód és teljesen biztosan tökéletlen is. Allen Newell 1982-es elemzéséhez hasonlóan én is visszatérő problémákat próbálok azonosítani (ennek minden kockázatával), azonban a lehetséges válaszokat nem dichotómiaként képzelem el.

1.1 Az MI konstruktőr szemszöge

Az egyik kulcsfontosságú kihívás, ami köré könyvem szerveződik a sikeres MI-t létrehozó folyamattal kapcsolatos. A kihívás abban áll, hogy a tendenciózusan fekete dobozzá váló gépi tanulás, valamint a rendszerek méretéből és komplexitásából fakadó munkamennyiség hatalmas, több tucatnyi közreműködőből álló csapatot feltételeznek. Ezen csapatok egyik tagja sem képes intellektuális ellenőrzése alatt tartani a teljes rendszert, és a fejlesztési módszertan megválasztása önmagában is értékválasztás. A

fenti tények észben tartását *konstruktőri*[15] megközelítésnek nevezem és igyekszem szisztematikusan alkalmazni.

Ez a kérdés explicit formában jelenleg alig van az MI etikával foglalkozó filozófusok látóterében[16], ahol a fejlesztési folyamat leegyszerűsített képével, és gyakran a fejlesztést végző vállalat korlátlan döntési potenciáljával (és az ehhez a hatalomhoz tartozó kritikával) számolnak. Eközben a konstruktőri helyzetből fakadó limitációkra való reflexió hiánya miatt átsiklunk afelett, hogy más ágenst hoz létre az evolúció és egy esendő emberekből álló csapat; és mindkettő különbözhet az ideális vagy kívánatos MI ágenstől.

Keserű pirulaként kell lenyelnünk a tényt, hogy az MI számunkra fontos alkalmazásaiban, a képességei maximumán szükségszerűen moduláris felépítésű – ugyanis az episztemikus korlátok miatt egyetlen tervező nem lenne képes átlátni – és a mikroökonómia mérnökökre is vonatkozó törvényei miatt számottevő modul-újrahasznosítással jár, ez pedig jelentős kötöttségeket vezet be a fejlesztésekbe.

A dekompozíció, azaz a modulokra való felbontás kényszerét az evolúció nem hordozza magán, és sok gondolkodó sem számol vele, amikor az MI etikához akar hozzájárulni. Azonban a modularizálás kényszere nagyon is valóságos, és egyáltalán nem elhanyagolható.[17]

A szoftverfejlesztést és ennek részeként az MI fejlesztést is jellemző dekompozíciós kényszert és az általánosan használt agilis fejlesztési módszertanok létezését, valamint annak limitációit a jövőben nem lesz lehetséges figyelmen kívül hagyni. Az MI tervezés etikai téttel rendelkező filozófiai kérdései közé tartozik az is, hogy a rendszerek episztemikus és normakövető képességeit egyszerre kell létrehoznia a fejlesztő csapatoknak, mert működésük egymás jelenlétét feltételezi – ezáltal *episztemorális* tervezési kérdések jönnek létre.[18]

Az MI funkcionális dekompozíciója és a választott fejlesztési módszertan olyan döntések eredménye, amelyeket gazdasági és kulturális preferenciák legalább annyira

[15] A kifejezést Stanisław Lem (1972) *Summa Technologiae* c. kiváló művéből kölcsönöztem.
[16] Megkockáztatom, hogy az elmúlt évtizedek MI-filozófiai vitáiban ez fontosabb kérdés volt, mint ma, lásd például Dennett gondolatait a témában (Héder 2020, 53-59).
[17] Ezt 2. fejezetben vezetem majd le.
[18] Lásd a 10. fejezetet.

vezérelnek, mint a konkrét feladatra alkalmazott racionális megfontolások. Ezáltal a tényleges MI ágensek belső felépítésének részleteire többnyire szociálkonstrukcionista magyarázatokat tudunk csak találni.

1.2 A tervezés, mint előrejelzés

A konstruktőri szemszögből érthetjük meg azt is, hogy a tervezés fontos episztemikus korlátja az, hogy a gép viselkedésére vonatkozó morális dimenziójú döntéseket a tervezés során kell meghozni, míg az ember a saját viselkedésére vonatkozó morális döntéseket képes lehet akár „élőben” produkálni. Természetesen az MI is „élőben” hoz döntéseket, azonban a morális dimenziót a tervezőhöz társítjuk.

Ez a kihívás, bár a konstruktőri alápállásból ered, külön figyelmet érdemel. A probléma illusztrálására képzeljünk egy teljes, 5-ös szintű[19] önvezetésre képes járművet. Magától az autótól semmilyen mértékű morális tevékenység nem várható.[20] Ehelyett, ezeknek a döntéseknek a tervezési idő[21] során kellene megszületnie. Itt válik fontossá az episztemikus képességeink korlátozott volta.

Ember által vezetett autók esetén, úgy tűnik, a felelősségvállalás olyan munkamegosztás alapján történik, melyet minden érintett – legalábbis nagyrészt – elfogad. A járművek tervezéséért és a minőségbiztosításért az autó gyártója vállalja a felelősséget. Ezek garantálják, a jármű szabványoknak való megfelelését.

Ehhez társul a jármű tulajdonosának felelősségi köre, pl. a gyártó által meghatározott időközönként a fékfolyadék cseréje, vagy a jármű használatát illető bizonyos (pl. hőmérsékleti) korlátozások betartása.

Továbbá a közlekedés minden résztvevője rendelkezik előre definiált felelősséggel, beleértve az útkezelő hatóságot is, akik pl. megfelelő állapotú közlekedési jelzésekért felelősek.

És végül, természetesen ott a sofőr.

[19] A Society of Automotive Engineers felosztásában, lásd: https://www.sae.org/standards/content/j3016_201806/

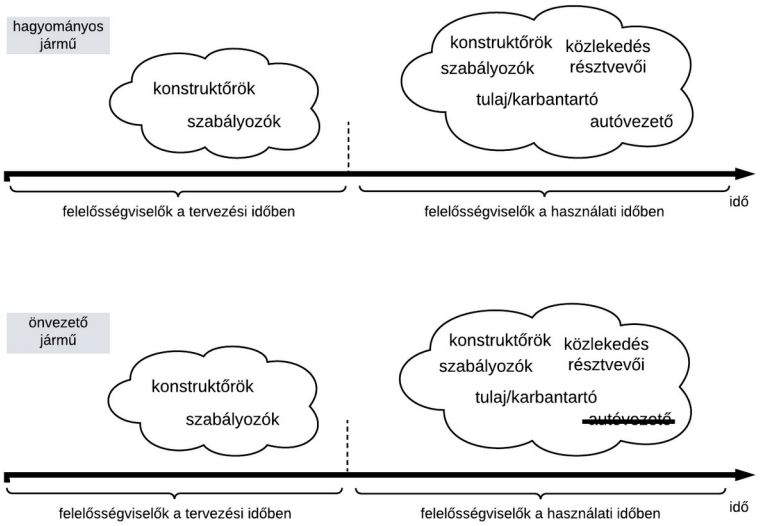
[20] Az általam feltételezett új MI etika hullámokban ez nem feltétlenül lesz így, mert felmerül, hogy morális ágensekké válhatnak az MI-k (lásd a 12. fejezetet), de ez egy meglehetősen újszerű gondolat.

[21] A „tervezési idő” magyarázatához lásd a 2. fejezetet.

Baleset esetén ez a „munkamegosztás”, pontosabban felelősség-megosztás kulcsfontosságúvá válik. A sofőr, valamint a közlekedés egyéb résztvevői, sőt még a közútkezelő hatóság tevékenysége is vizsgálat tárgyát képezi. A balesetért és az okozott kárért felelős személy a körülmények függvényében, és a résztvevők cselekedetei alapján lesz megnevezve.

Az önvezető autók megjelenésével azonban úgy látszik, a sofőr eltűnése egyfajta vákuumot teremtett. Világos, hogy a társadalom szemszögéből maga az önvezető autó nem bír olyan morális ágenciával, hogy egy baleset esetében a felelősség rá hárulhatna. Egyfajta „felelősségrés” jön tehát létre.[22]

Így tehát a felelősséget újra kell osztani a rendszert megalkotó mérnöki tervezőcsoport, a forgalom többi résztvevője, valamint a közútkezelő között.



2. ábra: felelősségrés kialakulása az önvezető járművek esetén

Csakhogy a tervező csoport sokkal nagyobb episztemikus távolságra találja magát a problémától, mint a sofőr: időben, térben, szituációs tudásban is messze van. A konstruktőri csapat helyzete az elképzelhető problémák extrém aluldetermináltsága[23]

[22] Lásd a 7.2. fejezetet.

[23] Aluldetermináltság (underdetermination) a tudományfilozófiában használt kifejezésre utal. Bár bizonyosan vannak elődjei is ennek a gondolatnak, a tudományos elméletek síkján mégis általában Pierre Duhem-nek (Duhem 1914), egyébként pedig Quine-nek (1951) tulajdonítják, aki minden tudásterületre kiterjesztette azt. E fejezetben a kifejezést tágabb jelentésben értelmezzük, pontosan úgy,

miatt teljesen más morális keretekben érthető meg, mint a sofőré. Ezt a különbségtételt olykor nem teszik meg, lásd a „kit üssön el az önvezető autó” típusú kísérleteket,[24] amely valójában a sofőr episztemikus hozzáférését tételezik fel, ezért teljesen félrevezetőek. Ahogy később látni fogjuk, a felelősségrés jelensége nem ismeretlen az MI etika számára, ám a kevés vita, amely erről a problémáról szól, nagyrészt az elszámoltathatóságot vizsgálja, nem pedig az episztemikus okokat. Ezt a nagy kihívást jelentő szempontot, a *tervezési idő kihívás-t* is igyekszem szisztematikusan érvényesíteni, amelynek eredményeképp újszerű következtetésekre jutok.[25] Úgy tűnik, ennek a problémának egy változatát ismeri fel Zódi[26] is a jogi szabályozás kontextusában, tehát nem csak az MI etika hanem az MI jog is foglalkozik a kérdéssel.

1.3 A számítógépek metafizikája

Az harmadik nagy kihívás metafizikai természetű – a terminológia és a kérdések tisztázása után elkerülhetetlen lesz a számítógép ontológiai státuszának vizsgálata, mivel az egyes területeken felmerülő etikai kérdések, mint például az MI ágenciája vagy kreativitása, függ a számítógép-képünktől. Önmagában, a filozófián belül ez egy időszakosan, újra és újra felmerülő kérdés, amelyet részben a fikciós művek is reprezentálnak a szélesebb közvélemény számára. Az újdonság itt abban rejlik, hogy míg például a ’60-as években, amikor Putnam[27] sürgette a robotok morális státuszának tisztázását, még csak elképzelhető sem volt az a rendszerkonfiguráció, amely egy emberi módon kommunikáló géphez ma tartozik. Ehelyett a kognitív tudomány korai eszköztárával, a szimbólumfeldolgozás és az alkalmazott logika, esetleg programozható heurisztikák lehetőségeinek fejlettebb verzióit vizionálták.

Ma azonban a sikeres MI ágens megérinthető, vizsgálható gép, amely egyfelől akuttá teszi a kérdést, másfelől teoretikus feltevésből egyszerű megfigyelési problémává teszi azt, hogy hogyan is működnek ezek a gépek.[28]

ahogy ezek a tudományfilozófusok tették.

24 Lásd a 4. fejezet bevezetőjét!

25 Lásd a 4., 8., 9. és a 10.1-es fejezeteket.

26 Zódi (2024) az EU MI rendelet körüli narratívákat vizsgálja. A szabályozás körüli viták keretezésére könnyű megoldás az, hogy a nagyobb szabadságot kívánó cégek és a nagyobb szigort kívánó képviselők és jogászok küzdelméről van szó. Zódi azonban ezt – nagyon gyümölcsözően – elveti, és helyette az *ex-ante* és *ex-post* megközelítések közötti kötélhúzásként mutatja be a folyamatot.

27 Lásd a 11. fejezetben.

28 A 3-as és a 4-es fejezet ebben a szemléletben készült, és a 12-ik fejezet is épít erre a megközelítésre.

Végül, az MI társadalomformáló hatásával kapcsolatban az MI etika előtt álló kihívás a közgazdaságtan bizonyos téziseinek beépítése. Az MI etika szempontjából a legjelentősebb *az MI agresszív tendenciája a technológiába zártság előidézésére.* A tény, hogy a szabályozók soha nem látott erővel próbálnak hatással lenni nem csak a termékre, hanem a *tervezés* folyamatára,[29] implicit elismerése ennek a potenciálnak.[30]

Az MI inherens tulajdonságaiból fakadó technológiába zárási potenciálja bármely más technológiánál erősebb lehet; míg az MI-vel elérhető majdnem teljes automatizálás különös és azonnali figyelmet érdemel.[31]

[29] Lásd a 6. fejezetet.
[30] A szabályozók és a rájuk hatással lévő szavazók (és közvélemény) a jelek alapján érti, hogy *időben* kell szabályozni, mert azután késő lesz.
[31] Lásd a 7.7. fejezetben, illetve kisebb részben az 5.5. fejezetben.

2. Az MI és a műszaki tudás

E fejezetet egy egyszerű tézissel szeretném kezdeni: az emberiség, sőt, az állatvilág története felfogható úgy, mint a személy által birtokolt, egyre hatékonyabb kontrolláló képesség a története.

Az élőlények világának, az embert is beleértve *egyedei* vannak, amelyek önálló, a környezetüktől elkülönülő létezőknek tekintünk. Az, hogy hol és miért húzódik határ az egyed és a környezete között, egy konszenzusos választ nélkülöző metafizikai kérdés, ami a filozófusokon kívül nem sokakat zavar.

Ugyanilyen módon tisztázatlan, ám mégis nagyrészt elfogadott az is, hogy ezek az egyedek saját célokkal és preferenciákkal rendelkeznek, Polányi Mihály terminusával élve önérdek-centrumoknak tekinthetők.[32] Az önérdek érvényesítése pedig az egyed környezetének, a reprodukciójának, sőt, saját lényének kontrollálását igényli – ez utóbb némi körkörösséget is létrehoz.

Az élőlények közötti fejlettségi különbségek kifejezésének egyik módja a kontroll képességek szerinti sorrendezés. Ebben kétségkívül az ember jár az élen, hiszen ő képes a környezete, a reprodukciója és saját maga legnagyobb mértékű kontrollálására.

Ezt a magas szintű kontrollt a tudása teszi lehetővé. E tudás magas szintje az egyik alapja az ember megkülönböztetésének az állatvilág többi részétől: az eszközhasználat és a nyelvhasználat – a megkülönböztetés szokásos alapjai – mind-mind az emberi élet különféle aspektusainak kontrollálására valók.

A hatékonyan kontrollálni képes civilizációra jellemző feladatmegosztás és specializáció a tudás előállítását sem kerülte el. A gyakorlatias, túlélést, boldogulást és reprodukciót biztosítani képes tudás egyre több árnyalatot vett fel, és létrejöttek a tudományok, amelyek nagyrészt a valóság megismerésére törekednek. A természettudomány különösen sikeres lett a kontrollhoz való hozzájárulása miatt. A kontroll eléréséhez nagyon hasznos a világ előrejelzésének képessége, a természettudomány pedig a saját területén ebben verhetetlen: milliméterre, mikroszekundumra, grammra pontosan tud előre jelezni, ha megfelelően használjuk.

[32] Polányi ontológiájáról bővebben a 12-es fejezeten írok.

Ez a tudás abban is újdonságot hozott, hogy leíró jellegű, és nem előíró vagy alkotó, mint az ókori és középkori emberiség tudásának a döntő része. Azonban a leíró tudás sikere és kiváló módszerei nem jelentik azt, hogy nincs szükség többé alkotó tudásra – a munkamegosztás és az intézményesülés folyamatának a másik ágán így jött létre a műszaki tudás koncepciója.

E fejezet a műszaki tudás jellegzetességeiről szól: szerepét a könyv céljának elérésében az a (nyilvánvaló) tény adja, hogy az MI is műszaki alkotási folyamat, konstruktőri munka eredménye.

A műszaki tudás egyik fajtája az, ami egy jó tervben, hatékony eljárásban, alkalmas programsorokban megtestesül és általában precizen dokumentált. Egyszerűbben szólva, mérnöki alkotás eredménye, egy mérnöki terv vagy eljárás leírása jeleníti meg.

A másik az a tudás, amely ezek előállításához, tervezéséhez szükséges és sokkal kevésbé precízen dokumentált. Ez tehát az, amit a mérnök az alkotási folyamat során felhasznál a terv előállításához. A műszaki tudománynak erős aspirációi vannak arra nézve, hogy ezt a tudást is olyan pontosan dokumentálja, mint magukat a terveket és egyéni műszaki alkotásokat, azonban ezek az aspirációk csak kis mértékben teljesülnek. Hiába a törekvés a tervezés és alkotás megbízható és szubjektív elemektől mentes módszerei leírásának, nem létezik olyan mérnöki terület, ahol ne lenne nagy jelentősége a szavakba nehezen önthető megérzéseknek, és az esztétikai érzék műszaki alkalmazásának, annak minden szeszélyével együtt. A hallgatólagos tudás[33] elismert eleme a mérnöki tudásnak.

Mint minden tudományban, a műszaki tudományban is jelen van tehát a hallgatólagos összetevő, a tudásnak egy olyan formája, amely expliciten nem kifejezhető vagy átadható, így élő és folytonosan megújuló tudáskultúrában, esetünkben a műszaki kultúrában hagyományozódik át és fejlődik. Ezek az elemek azonban legalább részben explikálhatók, ezáltal megnyílik a reflexió és a felülvizsgálat, továbbfejlesztés lehetősége. A reflexiós folyamat az első lépése a műszaki tudás természetének megértése.

[33] Lásd a 12. fejezetben.

Ennek a könyvek az egyik alapvető feltevése, hogy a műszaki tudás természete a technológia természetéből adódik. Az állítás önmagában nem túl izgalmas, a legtöbb olvasó elméjében e pillanatban valószínűleg nem okoz mást, mint egy szünetjelet. Igazsága viszonylag könnyen belátható, hiszen mindenfajta tudás tükrözi a tárgyának természetét – miért épp ez lenne kivétel?

Például a fizika által vizsgált világ uniform – bármely két elektront vesszük az univerzumban, azokra egyforma törvények vonatkoznak, mindegy, hogy milyen messze vannak egymástól térben vagy időben. Ezért azután a fizikában és a rokon természettudományokban univerzális törvényeket és modelleket keresnek, vizsgálnak és alkalmaznak.

A társadalomtudomány alanyai, az emberek pedig épp az ellenkező helyzetet okozzák: mivel egyáltalán nincs közöttük két egyforma, a társadalomtudományban vagy a jelenségeket a maguk esetlegességében leíró ideografikus tudást, vagy a statisztikus, erősen absztrakt állításokat fedezzük fel, de univerzális törvényeket semmiképp. Ráadásul mivel az ember információt dolgoz fel – tűnődik, hogy mi lehet a kísérlet célja, amiben rész vesz, esetleg elolvassa, amit róla írtak és ezzel megváltoztatja viselkedését – a tudás megszerzése az alany visszacsatolásai miatt olyan nehézségekkel járhat, ami a természettudományban fel sem merül.

Az persze nincs bizonyítva, hogy a természet valóban uniform, azonban ezzel a fizika nem foglalkozik egy pillanatra sem – mivel eddig semmi nem mond ellent a természet uniformitásának, másfelől pedig az uniformitás elvetése az egész episztemikus vállalkozás alól húzná ki a szőnyeget, ezért a kérdés sosem aktuális. Hasonlóképp, az sincs tisztázva, hogy az embereknél hol, milyen tulajdonságoknál húzódik a határ a kvázi-univerzális antropológiai jellegzetességek – amelyek mégiscsak lehetővé tesznek valamilyen törvényszerűségeket[34] – és a valóban soha meg nem ismétlődő attribútumok között. Tehát mindkét esetben találhatunk metatudományos kérdéseket, amelyeket az adott diszciplína nem képes vizsgálni pusztán a saját eszköztárára hagyatkozva.

A műszaki tudás jellegzetességeinek és a technológiák jellegzetességeinek összekapcsoltsága könyvem szempontjából az MI konstruktőri tudás és az MI

[34] Például a szakadatlanul a saját hasznát kereső egyén koncepciójára épülő közgazdaságtanokat.

természete közötti kapcsolatra szűkül. A műszaki tudomány metatudománya kevéssé kutatott a társadalom és természettudományokhoz képest. Ez utóbbi területek professzionális művelői ismerik és lépten-nyomon elismerik a kritika felett nem álló, esendő előfeltevéseiket, és azt, hogy ezek hogyan korlátozzák az adott tudomány jellemző módszertanainak sikerét. A mérnökökre ez kevéssé jellemző, ezért e ponton indokolt egy történeti megközelítésű bevezető.

Fentebb a „technológia” kifejezést használtam, hogy ne a „gépre” szűkítsem le a mondandóm tárgyát. A technológia fogalma is erősen összetett és semmiképp sem kanonizált; azonban a mi céljainkra egy tág definíció/meghatározás a legalkalmasabb: mindenféle fizikai és virtuális gépet, járművet, eljárást, mesterséges rendszert értsünk alatta.

A technológia mesterséges, ami azt is jelenti, hogy létrehozzák azt. Másképp fogalmazva, a technológia a *tervezésekor* – ezt a szakaszt mindvégig „tervezési időnek” fogom nevezni – még *nem létezik*, ezért empirikusan nem is vizsgálható, cserébe könnyen alakítható[35], látszólag az alkotó szabadságát adva a tervezőmérnöknek.

Ráadásul, ahogyan az 1.1. fejezetben hangsúlyoztam, szinte minden technológia eléggé összetett ahhoz, hogy csapatmunkára legyen szükség a tervezéséhez majd a megalkotásához.

E két tényező okozza azt, hogy a sikeres mérnöki munka eredménye – egy tervrajz, egy előírás, vagy egy utasításokat tartalmazó kódsor - mindig *normatív*,– még ha soha *nem* is *teljes mélységben specifikált*. Ezzel rögtön egy óriási különbségre találtunk rá a deskriptív természet- és társadalomtudományokhoz képest, melyek hagyományosan tartózkodnak a normatív állításoktól, helyette a világot a maga esetlegességében szeretnék leírni.

Természetesen az, hogy a világ milyen, önmagában is vita tárgya. Azonban az, hogy a világ *milyen legyen* – hiszen még egy hétköznapi ingatlan-tervrajz is átalakítja egy kicsit a jövőbeli világot– polémia egészen új szintjét hozza létre. A normatív termékeket (terveket, utasítássorokat) előállító műszaki tudás kötelezően tartalmaz egy

[35] Ebből fakad az úgynevezett Collingridge dilemma – lásd a 6. fejezetben.

prediktív elemet: azt jósolja, hogy a terv követése működő technológiát eredményez. Ez merőben eltér más tudományoktól, ahol egy jelenség puszta magyarázata akár előrejelzés nélkül is önálló eredménynek tekinthető.

Azáltal, hogy bármely mérnöki tevékenység során, még ha nagyon kis léptékben is, de a világ jövőjének valamilyen kisebb-nagyobb részletét tervezzük, a mérnöki alkotómunka a döntéshozók felelősségét is viseli. Emiatt a műszaki tudomány a tudományok között *szokatlanul erős morális dimenzióval* is bír.

Végül, a technológia kimondottan *célszerű*, az ember természet, más emberek, vagy akár saját maga feletti kontrollját szolgálja. A célszerűség teszi lehetővé azt, hogy hatékonyságról beszéljünk, ami alatt a céljaink minél jobb megközelítését értjük. A természeti jelenségeknek nincs hatékonyság dimenziója, társadalmak esetén pedig csak akkor beszélhetünk hatékonyságról, ha valaki eldönti, hogy mi is a cél, és ezzel instrumentálja, azaz technológiává teszi az embert vagy a csoportot a cél elérése érdekében. A célhoz szorosan kapcsolódó fogalom a *funkció*, amely ugyancsak bőséges és nem konvergáló irodalommal rendelkezik, ám számunkra itt elegendő annyi, hogy a funkció a cél fogalmának részletesebben, szabatosabban kifejtett változata.

Tehát a technológia természete azt okozza, hogy a műszaki tudás *normatív, célszerű/funkcionális*, inherens *morális* dimenziója van és *prediktív* is.

2.1 Alap- és alkalmazott kutatások

A tervezésben foglalt predikciók előállításához a műszaki tudomány egy része – az, amelyik gépeket, épületeket vagy egyéb olyan alkotásokat hoz létre, amelyek sikere erősen fizikai természetű – kiválóan ki tudja használni a világ szabályosságait, így műszaki tudomány nagy hasznát veszi a természettudománynak. Ezért adná magát az ötlet, hogy nevezhetjük akár *alkalmazott természetttudományoknak* is a műszaki tudományokat. Ám egy ugyanilyen meggyőző erejű gondolatmenettel akár azt is állíthatnánk, hogy a műszaki tudományok *alkalmazott etikának* minősülnek, hiszen ez legalább ennyire igaz.

Az *alkalmazott* jelzővel egy további probléma az, hogy nem megkülönböztető, minden tudomány alkalmazott valamilyen megismerési módszer vagy elv tekintetében. Arról se feledkezzünk meg, hogy minden tudomány felhasznál technológiákat a megismerési folyamathoz, tehát alkalmazott technológiának is tekinthető.

Az alkalmazott jelző a műszaki tudomány kontextusában talán akkor nyer értelmet, ha arra utalunk, hogy a tevékenység (ember által kitűzött) célok elérésére *alkalmazott*, másképp szólva *célszerű*, például egy MI ágenst próbál létrehozni.

Ezzel a Vannevar Bush által a második világháború végén kifejtett[36] megkülönböztetéséhez jutunk el, amely főképp a finanszírozásra vonatkozik: az alaptudományokat a létező világ vizsgálatáért, míg az alkalmazott tudományokat a céljaink szolgálatáért érdemes finanszírozni. Ez a megkülönböztetés azonban hamar értelmét veszti: természetesen a létező világ vizsgálata is szolgálhatja a gyakorlati céljainkat, ahogyan a lehetséges technológiák és az általuk elérhető jövő-állapotok is lehetnek haszonvágytól mentes kíváncsiságunk tárgyai.

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. A munkának ezen szintetizáló fázisa önálló szabályszerűségekkel bír, amelyek a technológia jellegzetességéből adódnak.

Amikor Bush megkülönbözteti az alap- és az alkalmazott tudományokat, az alaptudományokra is úgy gondol, mint a társadalom sikerességéhez, jólétéhez hozzájáruló ismeretgyártási tevékenységre, amely hozzájárul egy általánosan magas tudáskultúrájú társadalom fenntartásához, amely társadalom, ha akar (vagy a körülmények kényszerítik) nagyon konkrét célokat nagyon hamar el tud érni.[37] Közismert, hogy Bush 1945-ös tudományképét alapjaiban határozta meg az tapasztalat, amit a Manhattan-terv és a közelségi bomba projektek felügyelete során szerzett: ennek lényege, hogy a háború megnyeréséhez a technológiai áttörések, azokhoz pedig a kellően erős tudományos miliő vezetett. A különbséget tehát nem is

[36] Bush (1945)

[37] Ebből a szempontból roppant érdekes, hogy a műszaki tudás elemei, különösen az MI létrehozásához szükséges informatikai tudás nem bír ugyanazzal a kulturális státusszal, mint pl. a történeti tudás. Az informatikai ismeretek műveltségi önértéke kisebb, és ennek az MI etikájára is hatása van.

elsősorban a célszerűségben, hanem a célszerűség időhorizontjában látta: az alapkutatásról lehetetlen megmondani, hogy mi lesz belőle (ha lesz egyáltalán bármi) akkor, amikor a finanszírozásáról kell dönteni. Ki sejthette volna, hogy mi lesz az atom kutatásának haszna, ha lesz egyáltalán? Ki hitte volna, hogy az ionoszféra földfelszín feletti távolságának mérésére – tehát a meteorológia egyik, nem túlságosan központi kérdésének vizsgálatára – kiötölt műszerből lesz a radar? Bush ezt nem képzelhette el. Időben ugorva, ki gondolta volna, hogy a Wikipedia információs dobozainak szerkesztése a mesterséges intelligencia által ismert tudáshorgony forrása lesz?[38]

Ezek az időhorizontok és a kapcsolódó bizonytalanság arra kényszeríti a társadalmat, hogy az alapkutatást ne a gazdasági megtérülésért, hanem más elveken támogassa. Ezek az elvek pedig a tudósok belügyének tekinthetők, ezért a tudományos döntési folyamat alapköve az egyenrangú érdeklődésűek (peer-ek) általi megmérettetés és bírálat

2.1.1 Műszaki alapkutatás

Ha belátjuk, hogy az alapkutatás főleg a sikerkritériumokban tér el az alkalmazott kutatásoktól, – amennyiben a célja az általánosan magas tudáskultúra művelése, de nem valamilyen ennél konkrétabb eredmény–, akkor hamar rájöhetünk arra is, hogy az alapkutatás koncepciója egyáltalán nem kizárólagosan a természet- vagy társadalomtudományokra értendő.

A magas szintű tudáskultúrának tagadhatatlanul része a magas szintű műszaki tudás is, amely előfeltétele az MI etikájának. Az MI területén is megfigyelhető az a fajta bizonytalanság, ami a tudáselőállítás minden más területén is jellemző: marginális projekteknek előre nem látható transzformatív következményei lesznek, máskor pedig a transzformatívnak gondolt technológiai áttörések kiábrándítóan érdektelennek bizonyulnak. Mégis, könnyen belátható, hogy a kifinomult műszaki kultúra fenntartása összességében hasznos.

Műszaki alapkutatás például a tervezés és az esztétika viszonyának a vizsgálata, valamint a különféle tervezéselméletek kutatása; sőt, a probléma reprezentáció és a

[38] Itt a Linked Open Data-ra és a DBpedia projektre utalok (Héder 2013), amely tényalapot biztosít számos MI rendszer számára.

műszaki termékek reprezentációjának kutatása, a minőségbiztosítás és a hatáselemzés módjai is.

Azonban a műszaki tudomány nem tudja megérteni az összes, saját területén belül felmerülő kérdését a maga eszközeivel: például, ha a mesterséges entitások ontológiai természetéről van szó, ebből a nézőpontból metatudományos kérdéssé válik a kutatás.

Azt az állítást tehát, hogy a műszaki tudás természete a technológia természetéből adódik, úgy is átfordíthatjuk, hogy a műszaki tudás jellegzetességeinek megértéséhez a technológia körüli metafizikai kérdéseket kell megválaszolnunk, és máris érdekesebb állítást kapunk.

2.1.2 Tervezéselméletek és az MI

A műszaki tudás egyik fontos ága a tervezéselmélet. Ennek egyik fajtája deskriptív: ekkor egy külső szemlélő pozícióját felvéve, egyfajta mérnökszociológiai keretben megvizsgáljuk azt, hogy mit tesznek a tervezők a tervezési folyamat során anélkül, hogy megpróbálnánk eldönteni, hogy mit kellene *tenniük.* Ám mivel az ilyen elméletek nem elégítik ki a konstruktőri munka tudásigényét – hiszen útmutatást szeretnénk kapni a tervezéshez – a preskritpív tervezéselméletek sokkal inkább tartoznak a műszaki tudomány központi kompetenciáihoz.

Ezeket a tervezéselméleteket – iskolától függően – gyakran látják el olyan jelzőkkel, mint hogy „szisztematikus” tervezés vagy „racionális” tervezés, annak hangsúlyozására, hogy professzionális, esetlegességet kerülő, elvi alapokon nyugvó tudásról van szó.

A különféle posztulált elvi alapok mentén a tervezéselméleteket csoportosíthatjuk is.

Mivel a tervezés egy folyamat, minden tervezéselméletnek strukturálnia kell azt valamilyen végrehatható lépésekre. Ez a különbségtétel egyik módja: az, hogy szekvenciális, mérföldkövek mentén elválasztható *szakaszokból* álló folyamatot gondolunk el, vagy *iteratív,* a korábbi fázisokba vissza-visszatérő, kérdéseket újra megnyitó eljárást.

Az iteratív folyamat különösen népszerű a szoftverfejlesztésben, így az MI konstrukcióban is. Pártfogói[39] általában megdőltnek és elbukottnak tartják a szakaszos, vízesés-szerű tervezési módszert, azonban ez a helyzet túlzott leegyszerűsítése.

Az iteratív folyamat tagadhatatlan előnye, hogy jobban kezeli a folyamat előrehaladott szakaszában felmerülő változtatási igényeket. Ezáltal a megrendelőnek, értékgazdának vagy bárkinek, aki a követelmények meghatározásának erejét birtokolja, nagyobb a szabadsága: nem kell előre megsejtenie minden egyes majdan felmerülő követelményét, ami azért roppant hasznos, mert így előrehaladott tervezési részeredmények birtokában, vagy akár kézzel fogható prototípusok segítségével, sokkal jobb episztemikus szituációban hozhat döntéseket.

Azonban ez a folyamat drágább a szakaszok és fázisok ismételt megnyitása miatt, és a hozzáadott érték nem mindig garantált, sőt az sem, hogy a megcélzott felhasználó bevonható a folyamatba.

A másik óriási probléma az iteratív megközelítéssel, hogy a késői, alapokat érintő változtatási igények miatt kevéssé kompatibilis bizonyos mérnöki területekkel (pl. az űrbe kilőtt eszközt már nehéz és drága módosítani).[40]

Mivel (például) a szoftver tervek relatíve kis költséggel módosíthatók, sőt, választott megoldásoktól függően maga a készülő szoftver is viszonylag könnyen alakítható, ezen a területen a kivitelezés szakaszából is vissza tudnak ugrani a tervezéshez, ami az iteratív megközelítést különösen vonzóvá teszi. Későbbi szakaszokban, amikor a szoftver elterjedt, megfordul a helyzet, és a szoftverek de facto leválthatatlanná, és olykor nehezen módosíthatóvá is válnak.[41]

Ugyanakkor egy épület létrehozásakor a kivételezés szakaszából visszaugrani a tervezési szakaszba roppant költséges vagy akár kivitelezhetetlen is lehet. Így egy ilyen projektnél az iteratív jelleg a tervezés folyamatára szorítkozhat, amely még mindig lehet előnyös, de jóval kisebb előrelépés egy szakaszolt eljáráshoz képest.

[39] Például az agilis módszertan követői (Fowler és Highsmith 2001).
[40] A James Webb teleszkóp utólagos, űrsétával elvégzett javítása kivételként erősíti a szabályt és demonstrálja a magas költségeket.
[41] Lásd az 5.5-ös fejezetet.

A folyamat strukturálása azonban csak egy kisebb része a preskriptív tervezéselméletek témakörének. A nagyobb rész az, hogy pontosan minek is kellene történnie az egyes szakaszokban?

2.2 A dekompozíció

A konstruktőri alaphelyzet sajátja, hogy az egészet - ami a gép víziójában, specifikációjában van kifejezve-, részekre, kisebb csoportok által megérthető és előállítható elemekre kell bontani. Ennek a felbontásnak a pontos módja a tervezéselméletek csoportosításának a másik fontos megkülönböztetési elve.

Szinte minden tervezési elv top-down, azaz az egésztől induló és a részleteket folyamatosan kibontó metódust követ. Ennek az az oka, hogy maguk a követelmények, a rendszer főbb funkciói is így (magas szintű leírásként) állnak rendelkezésre. A feladat az, hogy a magas absztrakciós szinten körülírt egész produktum részleteit kibontsuk, a lehető legalaposabb specifikációs szinten, ezzel a kivitelezőknek egyértelmű utasítást hagyva (az interpretációs szabadságuk elvétele mellett). Egy egészet bontunk tehát részekre, innen származik a *dekompozíció* kifejezés.

A folyamat átfogó strukturálása mellett (iteratív-e vagy sem) tehát a lépcsőfokok meghatározása és a dekompozíció vezérelve az, ami megkülönbözteti a különféle tervezéselméleteket. A lépcsőfokoknál gyakran bevezetnek egy „koncepció” szintet, ami itt nem fogalmiságot, hanem megoldási elvet követ (lásd „koncepcióautó”), azzal az igénnyel, hogy itt még a konkrét megoldástól független leírását kapjuk a létrehozandó terméknek. A pontos határ azonban vitatott, ahogy az is, hogy célszerű-e, sőt, akár csak lehetséges-e koncepcionális szinten alkotni úgy, hogy a végső megvalósítás kötöttségeit és adottságait nem vesszük figyelembe.

2.2.1 A dekompozíció alapfogalmai

A mérnöki tervezési folyamat dekompozícióként való felfogása mellett fontos gazdaságossági érvek szólnak, amelyeket a mérnökök nagyon erősen szem előtt tartanak.

A kérdés kulcsa az alkatrészek, komponensek, részegységek előállításának mennyiséggel csökkenő költsége. Ez igaz mindenféle gép, épület, elektronikai eszköz, de még az immateriális szoftverek tervezése során is.

A mennyiséggel csökkenő költség azt jelenti, hogy minél többet állítunk elő egy adott alkatrészből, általában annál alacsonyabb a következő egységre jutó költség. Ez költség-átlag értelemben is igaz: bizonyos egyszeri költségek, mint a tervezés, a gyártósor beruházása stb. több egységen oszlanak el, így a költség minden egységre egyre kisebb. Az ilyenkor szokásos érvelés szerint a gyártósorok méretgazdaságossága, valamint az újrakonfigurálásuk magas költsége együttesen arra szorítanak minden tervezőt, hogy lehetőleg olyan elemekből építkezzen, amelyek nagy mennyiségben hozzáférhetők. De sokkal érdekesebb számunkra a marginalista megközelítés[42], az 1870-1890-es évek közgazdaságtani fordulata. Ebben a felfogásban nem átlagot nézünk, hanem adottnak veszünk egy már megtermelt mennyiséget és ebben a pontban döntünk a következő egységről. Az egységek számunkra alkatrészeket jelentenek és a döntés az, hogy egy meglévő alkatrészt hasznosítsunk-e újra a tervünkben, vagy újat alkossunk: azt a határt keressük tehát, amikor az új alkatrész konstrukciója jobban megéri.

Ennek jelentőségét az MI-re nézve akkor érthetjük meg, ha például a szoftverek tervezését vizsgáljuk: itt igazából nincs szükség gyártósorokra, bármely komponens újrahasznosításának költsége egységesen gyakorlatilag nulla, így a határköltség függvény teljesen lapos. Ám egy alkatrész, építőelem használatának nem a beszerzés az egyetlen költsége. A megismerésébe, megértésébe vagy ha új komponensről van szó, akkor az előállításába fektetett intellektuális munka legalább annyira jelentős. Ez az egyik oka annak, hogy a re-invenció („újra feltalálni a kereket”) az egyik legnagyobb tabu minden mérnöktudományban. Ráadásul a komponensekbe fektetett intellektuális munka nem csak a tervezés során merül fel, hanem a minőségbiztosítás, a karbantartás és a termék nyugdíjazása, esetleges újrahasznosítása során is minden érv az ismert komponensek használata felé mutat, nem is beszélve a szabványokról és egyéb előírásokról.

Összességében tehát a mérnöki tervezési folyamat során az ismert, bevált alkatrészek és összetevők akárcsak egy erős gravitációs vonzás, maguk felé húzzák az új terméket.

[42] Steedman (1997)

Ez alól olyan esetekben találunk ritka kivételeket, amikor műszaki paradigmaváltásra van szükség, mint például az újrahasznosítható űrrakéták vagy az elektromos járművek esetében. A meglévő alkatrészek ugyanis, amellett, hogy felhasználásuk roppant költséghatékony, mind a termék előállítása és karbantartása szempontjából, mind a mérnök kognitív munkája tekintetében, kötöttséget is jelentenek.

2.2.1.1 A funkcionális dekompozíció

A technológia egyik jellegzetessége a teleologikus jellegéből következően az, hogy funkcióval vagy funkciókkal bír. Egy technológia által betölteni kívánt funkció végső soron az alkotó vagy felhasználó céljaira vezethetők vissza, de gyakran nem közvetlenül: az alkotó vagy felhasználó egy nagyobb egységgel, rendszerrel szemben rendelkezik funkcionális elvárásokkal. Ahogy láttuk, minden technológia rész-egész viszonyokból épül fel, ennek megfelelően míg a felhasználó számára a rendszer a funkciók teljesítésének a forrása, addig a rendszer a maga részeivel, alkatrészeivel szemben támaszt funkcionális követelményeket.

A rendszer kibontása alkatrészekre a tervezés során döntések sorozata, amely magas tudást, és világos vezérlő elvet igényel.

Az egyik jelölt az elképzelt egész részekre való bontására az elvárt funkciók kibontása alfunkciókra, majd azokhoz megvalósítás rendelése. Ezt nevezik *funkcionális* dekompozíciónak[43], a funkciók elemzését általánosságban pedig *funkcionális következtetésnek*. Az alkatrészek összefüggései így az alfunkciók összefüggéseit fogják követni. Mivel az alkatrészek is alkatrészekből állnak, ez egy rekurzív módszer.

Természetesen a funkcionális dekompozíció egyik nagy kihívása az, hogy ne pusztán a dekompozíciós probléma új elnevezése legyen, hanem adjon valamilyen konkrétabb fogódzót.

A műszaki tudomány története során többféle módszert létrehoztak.

2.2.1.2 Algoritmikus dekompozíció

Az algoritmikus dekompozíció a rendszer elvárt viselkedését megvalósító folyamat megtervezésére való. Folyamatközpontúsága miatt gyakoribb a gyártástechnológiai

[43] van Eck és társai (2007)

eljárások és a szoftverek tervezésekor. Habár akár egy épület vagy egy jármű életciklusát is számos folyamat összességeként foghatjuk fel, ezeken a területeken nem jellemző az, hogy a tervezés alapja a folyamat lenne.

Az algoritmikus tervezés során az adott – még mindig részben funkcionális – rendszerleíráshoz próbálunk rendelni egy, a feladatot megoldó folyamatot, majd ennek a lépéseit fentről indulva olyan mélységig dolgozzuk ki, hogy az egyes lépésekhez konkrét gépek vagy komponensek rendelhetők, valamint összeállítható a teljes rendszer is.

Az algoritmikus dekompozíciót úgy is felfoghatjuk, mint egy recept vagy útmutató készítését (a majdan működő gép számára), ahol az egyes lépéseket egyre nagyobb mélységben írjuk le. A módszertan a ’60-as és ’70-es évek programtervezéséhez, valamint az üzleti és gyártási folyamatok tervezéséhez lett kifejlesztve, és a mai napig használatos, például a Business Process Reengeneering területén. Azonban egy kellően komplex rendszer túl sok párhuzamos, és részben egymásra ható folyamatot tartalmaz ahhoz, hogy ezzel a módszerrel könnyen kezelhető legyen, ezért például a szoftvertervezés során az algoritmikus dekompozíciót meghaladott módszernek tartják.

2.2.1.3 Egy példa a dekompozíció jelentőségére

Az MI etika szempontjából kulcsfontosságú dekompozíciós kérdés megértésére talán a legmegfelelőbb egy, a szoftvernél megfoghatóbb példa: az autók dekompozíciója.

Vessünk össze egy erősen munkamegosztás alapú (pl. német) autóipari szektort és az új fejezetet nyitni kívánó elektromos autóipari céget (pl. Tesla).

Ez előbbinél erősen specializált cégek óriási mennyiségben, így meredeken csökkenő határköltséggel képesek előállítani sztenderd alkatrészeket. Mivel minden modern autóban van gyújtáselosztó vagy klíma, ezek gyártására cégek specializálódnak (pl. Bosch). Néhány cég ezáltal gyakorlatilag az iparban szereplő összes gyártó termékébe beszállít, így *horizontálisan*, az adott műszaki feladatban az ipar teljes szélességében megjelennek. Ugyanennek az állapotnak az egyik kifejeződése a jármű-platform, amely arra való, hogy többféle típus ugyanazokra az alapokra épüljön. A kevéssé

beavatottak is észreveszik, hogy az Audi és a Volkswagen vagy a Toyota és a Lexus egyes típusainak műszaki tartalma között óriási az átfedés.

Mindennek az az előfeltétele, hogy a tipikus autó-dekompozíció állandó. Ugyanakkor ahhoz, hogy az azért mégiscsak eltérő autók dekompozíciója állandó lehessen, az elemek sokoldalú felhasználhatóságára van szükség.

Az elektromos autó akkumulátorának – a fizikai törvényei miatt – a hatékonyságát nagyban növeli, ha temperált, azaz adott hőmérséklet-tartományban tartózkodik, és bizonyos esetekben (pl. a töltés előtt) előfűtött vagy előhűtött az adott tartományra. A kezdeti német elektromos autók (pl. e-Golf) a belső égésű motoros autók dekompozícióját követték, tehát az autó nagy részét változatlanul hagyva a hajtásláncot lecserélték. Csakhogy ezekben a járművekben a klíma az utasteret hűti/fűti, tehát vagy egy második megoldás kell az akkumulátor számára, vagy a temperálásról teljesen lemondanak. A szokásos dekompozíció gyors megvalósítást tett lehetővé, azonban be is zárta a típusokat a szuboptimális hatékonyságba.

Az alapoktól elektromos hajtásra tervezett típusoknál hamar rájöttek, hogy az integrált fűtő-hűtő rendszer sokkal hatékonyabb, azonban ez újfajta dekompozíciót és ezáltal teljesen új alkatrészeket igényel. Ennek folyományaként lett a Tesla *vertikálisan integrált* vállalat, amely házon belül állítja elő a lehető legtöbb felhasznált alkatrészt. Mivel a mérnöki termékek, ahogyan az autók is rész-egész viszonyokból, hierarchikusan épülnek fel, a tervezés kezdőpontjához képest (ami az általános leírást jelenti), a kidolgozás egyre mélyebb rétegeibe kell eljutni. A „rétegekben” és mélységben való gondolkodás – amely persze nem szó szerint, a talaj irányába mutató mélységként értendő – vezet a technológián belüli vertikális dimenzió érzetéhez: ezért a vertikális integráció kifejezés használata.

A vertikális integráció tehát nem üzleti döntés, hanem a dekompozíciónál megengedett szabadság – amely teljesen újfajta, hatékonyabb lebontást eredményez – ára, amelyet a kivitelezéskor és a cég felépítésében kell megfizetni. Hogy az ár milyen nagy lehet, azt a Tesla esetében jól mutatja a házon belül megkísérelt akkumulátor-gyártás, amely

sokáig nem volt sikeres és a Panasonic beszállítói pozícióit erősítette.[44] Az üzleti döntés pusztán abban áll, hogy az új alkatrész kezdeti költségét megfizetik.

A fenti példával szemben a hétköznapi mérnöki tervezési folyamat során a dekompozíció nem „naiv” abban a tekintetben, hogy várhatóan nem az elképzelhető legcélszerűbb komponenseket és alkomponenseket fogja eredményezni, hanem nagyon nagy mértékben olyanokat, amelyek egyébként rendelkezésre állnak. Így a top-down tervezési megközelítésbe bottom-up elemek kerülnek – nyílt (keretrendszerek, platformok, tervezési minták) vagy kevésbé nyílt (rutin, copy-paste) módokon. A top-down és az ezzel járó legcélszerűbb tervezés, valamint a bottom-up és az attól elválaszthatatlan kompromisszumok közötti egyensúly a tervezési folyamat egyik legfontosabb döntése.

A dizájnelmélet kutatása egyelőre nem tart ott, hogy empirikus adatokkal rendelkezhessünk arról, hogy a dekompozíció során milyen mértékű az ismert, bevált megoldások újrahasznosítása egy átlagos projektben egy adott mérnöki területen. Anekdotikus beszámolókból, tanácsokból sejthetjük azt, hogy az újrahasznosítás aránya nagyon magas és az általános ajánlás annak a maximalizálása.

Ez azt jelenti, hogy új komponenst ott alkalmazunk, ahol a komparatív előnyhöz feltétlenül szükséges (például a piacon újszerű funkció megvalósítása) vagy pedig feltétlen piaci elvárás (a terméknek önálló formatervvel kell rendelkeznie, nem lehet másolata egy másik terméknek).

2.3 Szisztematikus dekompozíciós módszerek

Fentebb bemutattam a legfőbb motivációkat amellett, hogy a mérnöki tervezési folyamatot dekompozícióként fogjuk fel, valamint két alapvető megközelítést, a funkcionális és az algoritmikus dekompozíciót. Emlékeztetőül: az 1.2 fejezetben

[44] Nem a filozófiai okfejtéshez tartozik, de a technológia, esetünkben az elektromos autó tipikus dekompozíciójának változásából megkísérelhetünk üzleti jóslatokat is megfogalmazni. Az itt tárgyalt esetben az új szereplő új dekompozíciót hozott létre, amelynek kivitelezése vertikálisan integrált vállalatot igényelt. Az ezzel járó extra költségeket sikeresen finanszírozta meg a piacról és sikeres terméket hozott létre, így az új dekompozícióba fektetett költség gyümölcseit élvezi. Ha azonban a tipikus dekompozíció evolúciója lelassul – úgy is kifejezhetnénk ezt, hogy megszilárdul a műszaki paradigma – akkor a horizontális iparszervezés előnyei elkezdenek majd nőni. A tipikus komponensekre szakosodott cégek jönnek majd létre, amelyek a csökkenő határköltségnek köszönhetően olcsóbban tudják majd a megfelelő minőséget előállítani az ipar számára, így a vertikálisan integrált vállalat fokozatosan hátrányba kerül majd.

bemutatott érvek mentén el kell fogadnunk, hogy a *tervezési idő* az MI-vel kapcsolatos morális deliberáció elsődleges helye, így az MI etika hatóköre, amelynek ezért nem elegendő csak rendszerek viselkedésével, kimenetével foglalkoznia.

A tervezés tevékenysége kiemelkedően releváns az MI szempontjából, emiatt érdemes áttekintenünk a kapcsolódó, legfontosabb, szisztematikus dekompozíciós módszereket. E módszerek előterjesztői mind a „céhes" tudást, mind a pontosabban nem leírható mérnöki szemléletet és hallgatólagos tudást szeretnék valami konkrétabbra, a tudomány fejlődésének megfelelően explicit és reprodukálható lépésekre bontani.

2.3.1 Objektum alapú dekompozíció

A tervezéselméletek között a teoretikusan legjobban megalapozott és minden bizonnyal egyben a legelterjedtebb az objektum alapú dekompozíció. Elterjedtsége és kidolgozottsága összefügg annak az iparnak a méretével és jelentőségével, amiben megszületett: a szoftverfejlesztéssel.

Ahogyan hamarosan látni fogjuk, a szigorúan vett objektum alapú dekompozíció lépései nem egyformán hasznosak mindegyik mérnöki területre, hanem elsősorban a szoftvertervezésre, adatmodellezésre, bürokratikus folyamatok definiálására, kiberfizikai rendszerek és mesterséges intelligencia tervezésére alkalmas. Ugyanakkor az objektumalapú dekompozíció hatalmas irodalmában és gyakorlatában kidolgoztak számos kapcsolódó elvet, amely bármilyen mérnöki tervezési folyamatot nagyban segíthet.

Ezek egy része az ideális modulok/alkatrészek jellemzőivel kapcsolatos, a másik részük pedig a tervezési folyamat strukturálásával.

Az objektum alapú dekompozíciós módszer hivatalos neve az Object-Oriented Analysis and Design (OOAD)[45]. Az OOAD kezdeteit 1967-re datálják, amikor az első kimondottan objektum-orientált programnyelv, a Simula megjelent. Ezután az 1980-as évektől egyre népszerűbb lett, az 1990-es években valósággal berobbant és mára az egyeduralkodó dekompozíciós elvnek tekinthető a szoftverek területén. Ugyanakkor épp a mesterséges intelligencia az a szoftverfejlesztési alterület, amely kihívás jelent

[45] Booch (1990).

számára, mégpedig azért, mert itt a tanulóadat felépítésének jelentősége olyan nagy lett, hogy elkezdte felülírni az itt bemutatott elveket.

Az OOAD ugyanazt a dekompozíciós feladatot oldja meg, amit a 2.3 fejezet ismertetett: adott egy leírás egy elképzelt rendszerről, amely főleg funkcionális elemeket tartalmaz, azaz elmondja, hogy mit kellene tudnia a rendszernek. Ezt fel kell bontani addig, amíg implementálható elemeket kapunk – nagyjából ebből áll a kapott terv. Itt is igaz az, hogy elvileg minden egyes komponenst készíthetnénk újonnan, az adott tervre szabottan; azonban a valóságban inkább minél több olyan elemet szeretnénk kapni a dekompozíció végén, amely már kipróbált, bevált és hozzáférhető.

Az OOAD sajátosságainak bemutatásához a nem szoftvermérnök olvasó számára hasznos lesz két fontos körülmény említése. Mindkét körülmény a szoftverek, mint gépek virtuális jellegével kapcsolatos. Az első, a fent már említett rugalmasság más mérnöki termékekhez képest: egy épület esetében (valósznűleg) sosem jutna eszébe a megrendelőnek utólag még egy mélygarázs szintet kérnie, mert belátná, hogy ez nem, vagy csak a projekt újrakezdésével kivitelezhető, azonban a szoftverek területén a hasonló kérések mindennaposak. Filozófiai értelemben persze ez pontatlan, de a konstruktőr számára a szoftver természete olyan, hogy a fizikai világ törvényeivel gyakorlatilag nem kell foglalkoznia a tervezés során[46], ami meglehetősen nagy szabadságot ad. Ezért a szoftvertervezők számára a legnagyobb kihívás a változások menedzsmentje.

A másik egyedi tényező a szoftverek kapcsán az, hogy mivel a feladatuk általánosságban véve információfeldolgozás, ezért legtöbbször a külvilág releváns szerepének a reprezentációját tartalmazzák, amiről az információ szól.

Egy mainstream, OOAD-t alkalmazó szoftvertervező a szoftver belső felépítését objektumokból és az azonos típusú objektumokat leíró osztályokból alakítja ki. Ha egy klasszikus műveltségű filozófust beíratnánk egy OOAD kurzusra, az illető arról számolna be, hogy az OOAD egy mereológiai gyakorlat, bizonyos token-type[47] megkülönböztetésekkel.

[46] Természetesen az áramfogyasztás sosem elhanyagolható, különösen a kriptobányászat és az MI tervezés során. Azonban az általános szoftverfejlesztési gyakorlatban a legutóbbi időkig ez a megfontolás nem játszott szerepet.

[47] Wetzel (2018).

Az OOAD tervező ugyanis úgy védekezik a változtatási kérelmekből fakadó áttervezések ellen, hogy a valóság minél permanensebb osztályaiból építi fel a szoftver belső reprezentációját, hiszen azokat az ügyfél nem lesz képes felülírni. Egy példa segítségével hamar érthető lesz az elv lényege: a „személy” osztály, amely tehát az individuális személyeket reprezentáló személy-objektumokat tartalmazza, egységesen jól használható, sőt, projektek között újrahasznosítható, hiszen a személyek mindenütt ugyanazokkal az alaptulajdonságokkal rendelkeznek. Ezért a „személy” osztály beépítése a szoftverbe OOAD alapon professzionális. A jó modularizáció fentebb bemutatott ismérvei alapján az ehhez az osztályhoz tartozó programkód, konfigurációs állományok és dokumentáció a létrehozása tehát elvileg jó befektetés, mert megtérül majd a jövőben – a gyakorlatban pedig a tervező biztos lehet benne, hogy létre sem kell hozni, mert valaki úgyis megtette már.

Nem professzionális viszont a „részleg”, „tanszék”, „osztály”, „főosztály”, „csoport” és hasonló osztályok bevezetése, mert ezek valóságos jelöletei sem hosszú életűek, helyette javasolt a „szervezeti egység” osztály használata. Tekintve, hogy részleg, főosztály stb. elnevezések nem túl permanensek, hajlamosak összevonódni, szétválni stb., a szervezeti egység osztályban ezek csak puszta címkék lesznek, amik tetszőleges pillanatban átírhatók. Azonban az OOAD-t alkalmazó tervező azt is megfigyeli, hogy vezető mindig van, így azt a szervezeti egység lényegi részeként reprezentálja, azzal, hogy a titulus sem kőbe vésett, hasonlóan az egység nevéhez.

Ugyanez a helyzet a relációkkal: a szervezeti egység személyekből áll, de ez a viszony nem ugyanolyan, mint ahogyan a koktél az összetevőiből áll. Ezért az OOAD kereteiben megkülönböztetik az aggregációt (pl. a szervezeti egység és tagjai között) és a kompozíciót (pl. a koktél és összetevői között). A különbség abban fedezhető fel, hogy amikor a befoglaló egység megsemmisül, akkor a befoglaltak megsemmisülnek-e vagy sem: a meggyszirup *kompozíció* viszonyban szerepel a koktélban, ezért ha a koktél megsemmisült, akkor a meggyszirup is. Ugyanakkor a személyek *aggregáció* viszonyban vannak a szervezeti egységgel, ezért ha a szervezeti egységet bezárják, akkor a személyek nem törölhetők az adatbázisból. Láthatjuk, hogy a párhuzam a filozófia mereológia ágával[48] nem túlzás, a különbség az, hogy az OOAD-nek magnitúdókkal több művelője van.

[48] Varzi (2019).

Az OOAD tervező tehát anticipálja a jövőbeli változtatási kéréseket és ehhez a világ kellően permanens reprezentációját próbálja kialakítani, amely paraméterezéssel formálható a kívánt állapotra, nem pedig átprogramozással. Ha sikeresen teszi, akkor a következő megrendelői kérést mosolyogva, a konfiguráció átírásával teljesíti, ha kudarcot vall, annak az a tünete, hogy egyszerű változtatási kérések hatalmas áttervezést és sok hónapot vesznek igénybe.

Az OOAD-nek a megfelelő reprezentáció, azaz az objektumokra való dekompozíció létrehozásához komplett, kvázi-filozófiai eszköztára van. Ezek között több olyan elemet is találhatunk, amely a szoftvertervezésen túl is jól használható.

Először is, az *abszrakció* szintje olyan legyen, hogy az a lehető legszükségesebb elköteleződéssel járjon és ne függjön az implementációtól. A „szervezeti egység” például egy vállalati ügyviteli rendszer kontextusában azért lehet jó absztrakció, mert nem szűkít le feleslegesen például „főoszály” specifikusságra, hiszen a tervező észreveszi, hogy az felesleges (természetesen ez esetfüggő, lehet a főosztály speciális és saját objektum-osztályt kívánó, itt csak egy szokásos példát hozok), és nyilvánvalóan bármely programnyelven és adatbázisban kifejezhető, tehát nem implementáció-specifikus. Habár ez az OOAD elv egy az egyben nem fordítható át másfajta műszaki termékek tervezésére, azt fogjuk látni, hogy létezik a megfelelője a funkcionális dekompozícióban, nagyjából a koncepcionális szint körül. Tehát az absztrakció kérdése expliciten megjelenik, azonban konkrét válasz nincs rá, pusztán az, hogy kellően absztrakt legyen a változások elviseléséhez, de ne annyira, hogy értelmetlenné váljon.

Az *enkapszuláció* elve abból az elvárásból következik, hogy egy nagy rendszer semelyik komponense ne függjön egy másik komponens valamilyen belső részletétől. Az enkapszuláció azt várja el, hogy a komponenseknek jól definiált interfésze legyen, más komponensekhez való csatlakozáshoz, azonban erről az interfészről rejtsük el a belső állapotot. Másképp fogalmazva a komponens enkapuszlálja, hordozza magában mindazt, ami a működéséhez kell (és ne legyen rejtett összefüggésben a belső működése más komponensekkel), és ezeket határolja is le. Az OOAD-t követő tervező a modulok könnyű és mellékhatás-mentes cserélhetőségét várja ettől az elvtől, amely nagyon releváns az MI tervezésnél is.

Az enkapszulációhoz szorosan kapcsolódó követelmény a *modularitás*. Ha az enkapszuláció és az absztrakció elveit betartjuk, akkor esélyünk van arra, hogy a megoldásunk erősen moduláris legyen. Az enkapszuláció biztosítja, hogy a komponensek között laza csatolás álljon fent, a belsejükben pedig erős kohézió, mindez azonban még nem jelenti azt, hogy a komponens könnyen hozzáférhető és cserélhető, mert például a rendszer teljes leállítása szükséges hozzá. A modularitás ebben a dimenzióban segít, illetve abban, hogy a komponenseket nagyobb egységekbe – modulokba – csoportosítsuk, és az implementáló technológiától függ, hogy mi ennek a megfelelő kivitelezési módja.

A negyedik OOAD alapelv pedig a *hierarchikusság*, amely az eddigi három követelményen felül azt is megkívánja, hogy világos és szükség szerint bővíthető alá-fölérendeltségi viszonyokat kapjunk a dekompozíció végére.

A fenti leírásból kiderülhetett, hogy az objektum alapú dekompozíció sokkal nagyobb hangsúlyt fektet a nem-funkcionális követelményekre – változásállóság, javíthatóság, komponensek cserélhetőségekre – mint a funkcionális követelményekre. Ez azt a felismerést tükrözi, hogy legalábbis a szoftverfejlesztés esetében a sikerhez – azaz a jó szoftvertervhez – ezek a tényezők erősen hozzájárulnak. Valószínű, hogy a funkcionális és nem-funkcionális követelmények közötti egyensúlyozás eltérő mérnökdiszciplinákban eltérő megközelítéseket eredményez, és ezek a különbségek a tervezett mérnöki termék természetére vezethetők vissza. Tehát a szoftvertervező a szoftver képlékenysége miatt a változásra koncentrál, míg például az építész az utólagos módosítás óriási költsége miatt a funkciók minél teljesebb teljesülésére.

Mindez csak egy újabb érv amellett, hogy a tudás, esetünkben a dekompozíciós módszer jellege a tudás tárgyának jellegét tükrözi. Ez egy újabb ok arra, hogy a gépek, azon belül különös figyelemmel az MI természetével foglalkozzunk majd a harmadik fejezetben.

2.3.2 A funkcionális alap

A *funkcionális alap*[49] (functional basis) megközelítésében a műtermék egy input-output kapcsolatot valósít meg, ez maga a funkció. Ez a leginkább gépészek számára

[49] van Eck és társai (2007).

kifejlesztett módszer az MI-vel vezérelt robotok kialakításában kaphat[50] szerepet. Az input-output közötti összefüggést tárgyi vonzattal rendelkező igés mondatokkal fejezi ki (valamit csinál valamivel), maga a termék pedig kezdetben egy fekete doboz és ezáltal is releváns az MI területére. A tervezett terméket leíró mondatokra vonatkozó megkötés meglehetősen elterjedt, a géptervezéstől kezdődően az agilis szoftvertervezésig találkozunk sablonokkal és elvárásokkal arra nézve, hogy a funkcionális követelményeket hogyan érdemes kifejezni.

A funkcionális alap módszere úgy horgonyozza le a fekete doboz felnyitásának a problémáját, hogy kijelenti: a három lehetséges folyam az input és az output között az *anyagi*-, az *energia*-és a *jelfolyam*. Ezt voltaképpen egy ontológiai elköteleződésnek is tekinthetjük.

Ha ezt elfogadjuk, akkor a következő lépés olyan alfunkciók meghatározása, amely a három folyam legalább egyikén transzformációt hajt végre, vagy a folyamok közötti konverziót végez. Az anyag, az energia és a jel típusai, valamint az ezeken végezhető összes funkció halmazát végesnek tekinti ez a megközelítés.

Több próbálkozás is történt ezek teljes számba vételére, a legnagyobb hatású ezek közül Hirtz és társaié[51]. Ez az összeállítás például a jelek két fajtáját különbözteti meg (állapot is kontrol), míg az anyagoknál számba veszi a három alapvető halmazállapotot, a plazmát, illetve a halmazállapotok keverékeit (pl. köd, jeges folyadék). Az energia formáival is hasonlóan jár el.

Ezután felsorolja a három folyamon végrehajtható műveletek nyolc fajtáját (elágazás, vezetés, összekötés, mennyiség manipuláció, konvertálás, ellátás, jelzés, támogatás). Mindegyiknek van 3-5 alesete és al-alesete, de itt véget ér a hierarchia rendszer. Például a mennyiség manipulálásnak az indítás, a szabályozás, a módosítás és a megállítás az alesetei. Ezen belül a módosításnak a növelés, a csökkentés, a formázás és a kondicionálás. Természetesen ezeket mind azonosítókkal látták el, az előbbi négy művelet például a 4/c/i,ii,iii,iv.

[50] A feltételes mód oka, hogy a módszer jóval kevésbé elterjedt az OOAD-hez képest. Nehezen elképzelhető olyan informatikus, aki nem ismeri az objektum orientált fejlesztést, de nagyon is elképzelhető olyan mechatronikai mérnök, aki még nem hallott az itt bemutatott funkcionális alap módszerről.

[51] Hirtz és társai (2002).

A módszerben foglalt erős és kritizálható előfeltevés az, hogy minden művelet lefedhető az azonosított alapkategóriákkal.

Hogy el tudjuk képzelni a rendszer működését, nézzük néhány példát a publikációból, amely a rendszert javasolja:

- 4/c/i (növelés) – „a nagyító a vizuális *jel*et növeli, ami a papírról érkezik”
- 4/c/ii (csökkentés) – „a meghajtott csavarhúzó áttétele a forgatási energia folyamá[nak sebességét] csökkenti”
- 4/c/iii (formázás) – „az autóipar nagy présnyomói az anyagot alakítják át”
- 4/c/iv (kondicionálás) – „az elektromos eszközök védelme érdekében a biztosíték kondicionálja az elektromos energiafolyamot úgy, hogy abban tüskék és zaj ne legyenek.”

Egy konverzió pedig úgy néz ki, hogy pl. a villanymotor az elektromos energiafolyamot materiális forgó mozgássá alakítja.

Amint a példákból látható, az alapállapot az, hogy mindegyik manipuláció mindegyik folyamon értelmezhető, azonban ez alól vannak kivételek.

A *funkcionális alap* megközelítésben az első feladat az, hogy véglegesítsük a teljes rendszer leírását, másképp fogalmazva a megrendelővel vagy értékgazdákkal folytassuk le a szükséges kommunikációt addig, amíg eljutunk egy olyan fekete doboz rendszer tárgyas ige alapú leírásig, ami kielégítően lefedi az igényeket. Ez egy vízesés típusú tervezési eljárást vetít előre, bár ez nincs kőbe vésve: a funkcionális alap elvű dekompozíciót elvileg újra és újra elvégezhetjük egy iteratív projektben.

A tárgyas-igés alakú mondatok utalnak a funkciókra és a folyamatokra is: előbbi az ige, utóbbi a tárgyból kikövetkeztethető. A következő lépés az, hogy megállapítsuk, a három folyammal (anyagi, energia, jel) minek kell történnie sikeres esetben egy idősíkon ábrázolva. Ebből felépíthetjük egy dekompozíciót, amely szintén tárgyas ige formátumban van kifejezve, de ezúttal már alfunkciókra bontva. Ezt a folyamatot addig folytatjuk, míg mindenütt eljutunk a funkcionális alap által adott primitívekig. Végül, ezeket az időbeliség alapján rendezett folyammódosító alfunkciókat

alkalmasan össze kell kapcsolnunk, és meg is kapjuk a funkcionális modellt. Ez a modell ezen a ponton az elvárt inputból az elvárt outputot állítja elő, és kizárólag a *funkcionális alap* könyvtárban megtalálható alapvető alfunkciókat és műveleteket tartalmazza. Ennek előnye, hogy a funkcionális alapkönyvtár tartalmaz alkalmas építőelemeket, gépelemeket, szoftveres függvénykönyvtárakat stb. így a dekompozíció eredménye olyan, ami könnyen kivitelezhető az ismert elemek felhasználásával.

2.3.3 A funkció-viselkedés megközelítés

A funkcionális alapra való visszavezetés azonban a funkcionális dekompozíciónak csak az egyik változata. A funkció-viselkedés[52] megközelítés nem alapfolyamokkal dolgozik, hanem csak arra koncentrál, hogy a tervezett rendszernek hogyan kell „viselkednie”. Ez a módszer is tárgyas-igés leírásból indul ki, ám minden ilyen mondat mellé („tegye X-el Y-t a rendszer”) egy rendszer behaviorális leírást társít, ami azt próbálja megválaszolni, hogy milyen viselkedéssel valósítsa meg a rendszer az adott mondatot. A viselkedés leírása és konkrét megvalósítása között szakadék áthidalása egy további lépést igényel. A módszernek része az is, hogy a viselkedési leírásban állapotokat használ, a működést pedig állapotok közötti átmenetnek tekinti az időbeliségi sorrend megadásával, tehát egy funkciókat kielégítő, idő-reprezentációt is tartalmazó állapotgépet javasol, mint a tervezés eszközét.

Ehhez hasonló megközelítés köszön vissza az MI szabályozásában akkor, amikor a szabályalkotó általánosságban és implementációs részletek nélkül fogalmazza meg a rendszerrel szemben elvárt viselkedést: ilyen például az EU MI rendelet.

Ez a dekompozíciós elv is előre megismerhető elemekre kívánja visszavezetni a tervezési folyamatot. Ezek könyvtára azonban az nem anyagon, energián, jeleken működtethető alfunkciókat tartalmazza, hanem tervezési mintákat és funkció-prototípusokat. Az FBS funkcionális dekompozíciós eljárás számítógéppel támogatott.

Az első lépésben a tervező választ egy funkcionális sablont a sablonok repozitóriumából. A funkció sablonokhoz előre megadott dekompozíciós felbontás tartozik, amelyekhez a második lépésben csak a paramétereket kell megadnia a tervezőnek. Ezt nevezték el „task” vagyis feladat dekompozíciónak. A harmadik

[52] Umeda és társai (1996).

lépésben a fizikai dimenziók megadására kerül sor, amelyek a konkrét megvalósulás alkatrészeit fogják meghatározni.

A negyedik lépésben annak az ellenőrzése történik, hogy a rendszer egyáltalán megtervezhető-e fizikai ellentmondás nélkül, illetve meghatározzák, hogy milyen egyéb rendszerekre lesz szükség a működéséhez – más szóval ez a függőségek felfedezésére. Ezek a függőségek ok-okozati elven vannak kifejtve. Ha mindez megvan, akkor az FBS modellező (szoftver) egy Qualitative Process Abduction System nevű eljárás segítségével elvégzi az alfunkciók fizikai megtestesítését. Ez természetesen nem minden esetben lehetséges, illetve gyakran többféle alternatíva is levezethető a funkcionális leírásból.

Az ötödik lépésben a tervező összeköti a részletekre vonatkozó megvalósítási javaslatokat, ezáltal egy „behaviorális hálózatot” hozva létre. A hatodik lépésben – számítógépes módszerről lévén szó – a kialakított megoldás szimulációja történik. Itt tehát két lépésben történik dekompozíció: egyrészt a funkcionális sablonok választásakor a második, feladat-dekompozíciós lépésben, majd oksági szempontból egy újabb dekompozíciós lépéssel él a negyedik pontban.

Az FBS módszer tehát sokkal erősebben épít meglévő építőelemekre, mint a funkcionális alap módszer. Az utóbbi esetében is nagy hangsúlyt kapott a megoldások újrahasznosítása, azonban az FBS-nél – mivel számítógépes tervezési rendszerről van szó – nagyon meghatározó az, hogy a funkció sablonokat hogyan táplálták be a rendszerbe.

2.3.4 A funkcionális szintézis módszere

A funkcionális szintézis Chakrabarti és Bligh[53] által kidolgozott módszere szintén funkcionális elvű dekompozíció, azonban az alapja az eddig ismertetett két módszerhez képest is erősebben támaszkodik egy „ismert dizájn megoldások” halmazra vagy könyvtárra, amelyeket strukturális és fizikai síkon ír le.

Ezáltal a funkcionális szintézis módszere tartalmazza a legerősebb bottom-up elemet, amennyiben a teljes tervezett rendszer funkcionális leírásához azonnal adott méretű és megfelelő fizikai tulajdonságokkal bíró komponenseket keres. Ezáltal a funkcionális

[53] Chakrabarti és Bligh (1996)

alap és az FBS módszerekhez képest, amelyek elsősorban géptervezéshez készültek, ám megfogalmazásuk (a funkcionális alap esetén a jelfolyam-átalakítás, az FBS esetén a viselkedési leírás) lehetővé tette, hogy területek között hordozhatók legyenek valamelyest. A funkcionális szintézis módszere kimondottan gépek területére alkalmas, azon belül is erőátviteli rendszerek tervezésére.

Ebben a módszerben hasonlít a „funkció” leginkább arra, amit a matematikai fogalom takar: két változó közötti összefüggésre. A módszer során matematikai input-output leírásokat rendelnek az alkatrész-könyvtár elemeihez, majd ezen keresnek a magasabb szintű, hasonlóan megfogalmazott követelményeket kielégítő kombinációkat, akár számítógéppel. A módszer lehetővé teszi a rekurzív keresést, hiszen értelemszerűen bármilyen matematikailag kifejezett input-output összerendelés felbontható lépésekre, amelyek újabb lépésekre bonthatók és így tovább.

A módszer nyilvánvaló hátránya a nagyon erős beszállási követelménye: a feladatnak kifejezhetőnek kell lennie a módszer által megkövetelt matematikai nyelven, a megoldás-könyvtárnak pedig elég nagynak kell lennie ahhoz, hogy legalább esély legyen arra, hogy használható eredményt kapjunk.

Ugyanakkor ennek a módszernek a teoretikai érdekessége, hogy elveti a feltételezést miszerint lehetséges megoldás- és körülményfüggetlenül tervezni. Az MI etika esetén is felismerhetjük ezt a kényszerű kontextualitást. Míg a korábban bemutatott funkcionális dekompozíciós módszerek implicit ígérete az volt, hogy mindenfajta probléma kifejezhető és kezelhető általuk, ott erről szó sincs. Az általánosság és a gyakorlatban való bevethetőség között ellentmondást találunk: minél általánosabb a módszer, annál több kiegészítés és kreativitás kell az alkalmazásához, ám minél kézzelfoghatóbb, azonnalibb eredményt remélhetünk tőle, várhatóan annál kevésbé általánosan használható.

2.4 A mesterséges intelligencia dekompozíciója

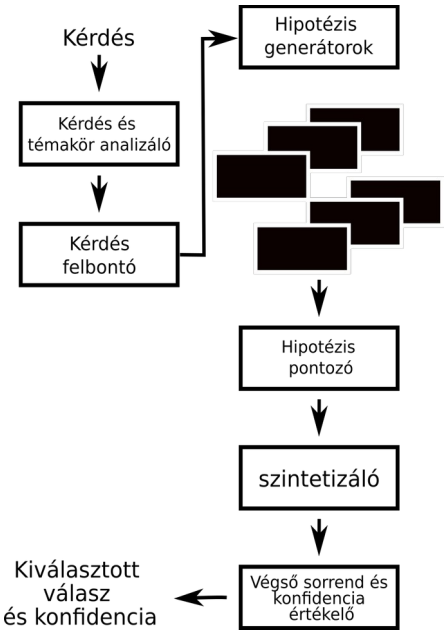
Ezen a ponton érdemes erősebben visszacsatolni a dekompozíció kérdését a könyv fő témájához, az MI-hez. Természetesen, mint minden gép, a szoftverek, azon belül az MI tervezése is dekompozíciós lépések soraként érthető meg.

Ha egy általános mesterséges intelligenciára gondolunk, akkor a következő kérdés adódik: milyen komponensei, moduljai vannak egy mesterséges ágensnek?

Ha a mesterséges intelligencia tervezésekor a természet imitációja a stratégiánk, akkor 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.

A következő kérdés az, hogy a mesterséges intelligenciát megvalósító hardver esetében lehetséges-e és kívánatos-e a szigorú modularizáció (erős enkapszuláció és elvágólagos interfészek) elengedése? A válasz valószínűleg *nem* és *nem*. Hiszen a gép dekompozíciója a tervező csapat munkamegosztását is szabályozza; másfelől a szigorú dekompozíció óriási mérnöki, és egyben üzleti előny.

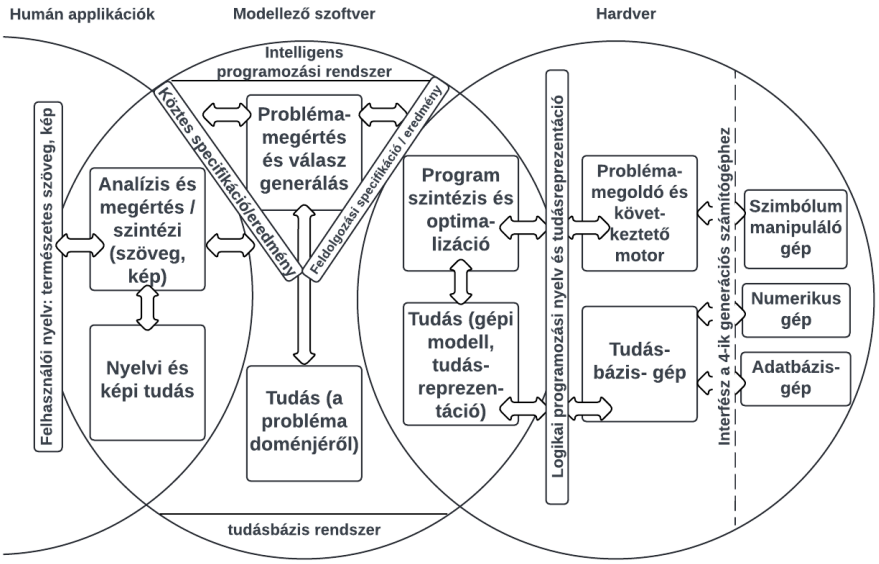
Az alábbi ábra az IBM Watson dekompozícióját mutatja. Láthatjuk, hogy ez nem a természet imitációjára épül, hanem egy erősen párhuzamos, egymástól független modulok által végrehajtott munkafolyamatot valósít meg. Az ábra nem fedi fel, de ezek között olyan modulok vannak, mint a Wikipédia teljes adatbázisa, vagy filmadatbázis, a hozzá tartozó keresőmotorokkal együtt.



3. ábra: Az IBM Watson dekompozíciója, amely sikeresen megoldotta a Jeopardy! nevű műveltségi vetélkedő intellektuális kihívását.

A következő ábra az ötödik generációs számítógép felépítését mutatja be.[54] Látható, hogy az 1981-ben készült ábra elég nagy összhangot mutat Newell 1982-es MI-történeti elemzésével, amely ettől minden jel szerint függetlenül készült. Itt is megjelenik a szimbólumfeldolgozó gép, a numerikus gép és az adatbázis-gép, mint külön kategóriák, amelyek ma már nem értelmezhetőek. A logikai programozási nyelv és a tudáskifejező nyelv – az ábra jobb oldalán – a prolog nyelvet takarja, annak „tudás” és „adat” dobozaival. Az ábra bal oldalán látható természetes nyelvfeldolgozó és szintetizáló, a gép nyelvére fordítani képes modulok akkoriban még nem voltak megvalósíthatók, viszont az a tény-adatbázissal lehorgonyzott nagy nyelvi modell, amelyet a Watson demonstrált először és a 2020-as években vált a széles közönség számára elérhetővé, pontosan ezt a funkciót valósítja meg.

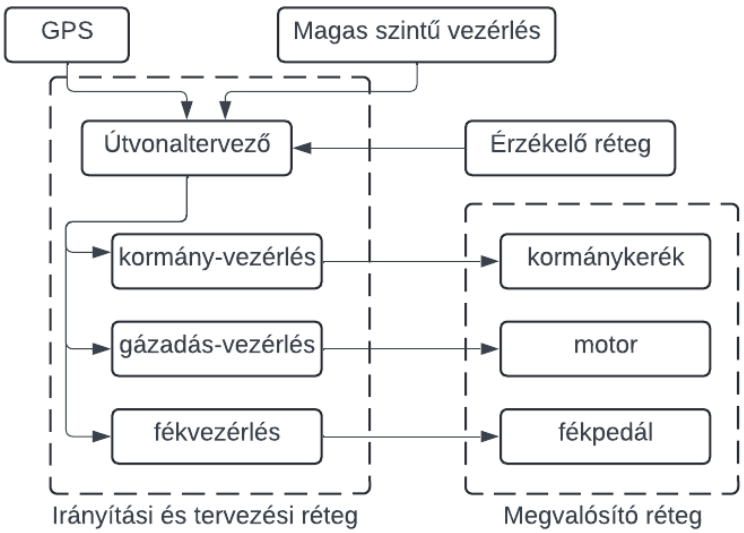
[54] Moto-Oka és társai (1982).



4. ábra: Az 5-ik generációs számítógép- Moto-Oka ábrájának magyar fordítása (HM).

Végül, az alábbi ábra egy önvezető jármű architektúráját mutatja be.[55] Látható, hogy ebben a dekompozícióban a „hagyományos” autó komponensei is benne vannak: a kormánykerék, a gáz és a fékpedál, amin keresztül az irányítás történik. A jelenleg elérhető, önvezető funkcióval rendelkező járművek természetesen valóban rendelkeznek ezekkel a modulokkal, amelyek később egy alternatív dekompozícióban eltűnhetnek. Az architektúra ábra bal oldala az irányításelméletben ismert Érzékelj-Modellezz-Tervezz-Cselekedj[56] dekompozíciót valósítja meg.

[55] Broggi (2012), lásd bővebben Héder (2020, 118).
[56] Sense-Model-Plan-Act: SMPA



5 ábra: Egy kortárs önvezető jármű dekompozíciója

A dekompozíciós döntések általánosságban véve megelőzik az implementációs döntéseket, és utóbbiak számára előfeltételként szolgálnak.

Éppen ezért az MI etika egy következő hullámában várhatóan előtérbe fog kerülni az a kérdés, hogy milyen *rendszer-dekompozíció etikus*, és némileg háttérbe szorul majd az, hogy melyik komponens hogyan *viselkedjen.* Egy egyszerű példa kedvéért, az önvezető járművek kapcsán kétféle tervezési iskola létezik: az, amelyik használna lidart (lézeres távolságmérő radart) és az, amelyik kizárólag videokamerás érzékelésre szorítkozik, mint ahogyan az ember is. Az implementálhatósági rés miatt gyakorlatilag nem tudjuk előírni, hogy *mit csináljon az önvezető autó* – helyette arról rendelkezhetünk, hogy milyen felépítésű legyen az önvezető autó, vagy bármilyen MI. Ez persze meghatározza az MI episztemikus képességeit is, amelynek etikai vonzatairól a 10. fejezet szól.

3. A gépek természete

Ahogyan azt a Bevezetőben is hangsúlyoztam, a mesterséges intelligencia körüli filozófiai viták az elmúlt évtizedekben nagyrészt egy olyan technológiáról szóltak, amelyet csak elképzelni lehetett, mégpedig az adott korban éppen rendelkezésre álló kezdeti megoldások extrapolációjával, illetve az azokat létrehozó mérnökök víziói alapján.

Talán ez a bizonytalanság vezetett oda, hogy gyakoriak lettek a szalmabáb érvek: az érvelők egy része egy addig meg nem valósult entitás tulajdonságai alapján mondtak ítéletet a mesterséges intelligencia felett.

Ennek a problémának a tipikus megjelenési formája a számítógép modelljének (pl. Turing-gép, véges automata) és a tényleges, megtestesült számítógépnek az összekeverése[57] – és ezáltal annak a súlyos érvelési hibának a megvalósítása, amelyet Gilbert Ryle kategóriahibának nevezett el.[58]

A kategóriahiba lényege, hogy egy entitást egy hibás érvelés erejéig olyan ontológiai kategóriába sorolunk, amibe nem tartozik bele, azután az entitással nem kompatibilis tulajdonságot társítunk hozzá vagy következtetéseket vonunk le. A szokásos példák „a 7-es szám kék”, vagy, ha szó szerint vesszük „a szombat nyugovóra tért”.

Természetesen ahhoz, hogy kategóriahibával vádolhassunk valakit, szükségünk van egy saját metafizikai világképre (amihez képest a vizsgált érvelés hibás), és fontos, hogy reflektáljunk a világképünk esendőségére. Ebben a fejezetben néhány MI-filozófiatörténeti példát mutatok be, hogy azután a kérdést a 12. fejezetben részletesen tárgyaljam.

Visszatérve a kategóriahibák bevezető példáira: ha elfogadjuk, hogy a hetes szám nem materiális létező (pl. emberi fogalmi konstrukciónak, emergens entitásnak vagy a matematikai platonizmust elfogadva, tőlünk független, de nem materiális ideának tekintjük azt) akkor mondhatjuk, hogy ezt a számot kéknek nevezni kategóriahiba, mert a számok ontológiai kategóriája nem rendelkezik színnel, így kék sem lehet.

[57] Lásd: Héder (2014) kisebb részben Héder (2020).
[58] Magidor (2024).

3.1 A számítógép modelljei

Az asztalon most előttem lévő számítógép anyagi tulajdonságokkal rendelkezik. Tömege, színe van, termodinamikai folyamatok zajlanak benne, kapacitása véges, sebessége korlátozott és az anyagfáradás, illetve végső soron az entrópia miatt az élettartama is véges: a statisztika alapján előbb-utóbb egy kondenzátor vagy kritikus memóriacella hibája miatt használhatatlanná válik.

Ennek a számítógépnek számos, potenciálisan végtelenül sok modellje lehet. Ne feledjük, a modellezés egyik fontos lépése az egyszerűsítés, bizonyos kevéssé fontos részletek elhagyása. Ez szükségszerű, hiszen a modellt azért készítjük el, mert könnyebb bánni vele, következtetéseket levonni belőle, mint a modellezett objektummal.

Például, ha a modellezett objektum a naprendszer maga, akkor annak egy virtuális modelljével sokkal könnyebben boldogulunk, szimulálhatjuk az égitestek mozgását, előre-hátra lépdelhetünk az időben, amit a modellezett objektummal nem tehetünk meg.

De az, hogy mely tulajdonságokat hagyjuk el, és melyeket reprezentáljuk a modellben, döntés kérdése, és ez a döntés általában aluldeterminált: hacsak nem egy matematikai modellt képezünk le egy másikba, általánosságban nem tudhatjuk, hogy mely tulajdonságokat kell meghagynunk és melyeket szabad eldobnunk ahhoz, hogy a modell még épp megfeleljen a céljainknak, például, hogy előrejelzéseket olvashassunk le róla. A modell alapú gazdasági predikciók és a meteorológiai előrejelző rendszerek megalkotói szakadatlanul ezzel az aluldeterminálttsággal birkóznak.

A fentiek világossá teszik azt is, hogy ugyanannak a modellezettnek beláthatatlanul sok modellje létezhet: ez a szám a modellező által meghozható döntések kombinációinak száma.

A számítógéphez is rengeteg modellt alkotnak, mindig a célnak megfelelően: a hőelvezetéshez termodinamikai modell készül, az ergonómiához fából vagy egyéb alkalmas anyagból kifaragott modell, az érintésvédelemhez a műanyagok vastagságát veszik figyelembe, és így tovább. A modelleket úgy is csoportosíthatjuk, hogy a

számítógép megvalósulásának mely fázisában hasznosak, például a 2. fejezetben bemutatott dekompozíciós szoftver-modellek a még megvalósítandó szoftverről szólnak, míg más modellek, például a 9.fejezetben bemutatott, fekete doboz MI-t magyarázó post-hoc modellek a már létező rendszert közelítik.

A filozófiai vitákban kiemelkedő azonban a Turing-gép, vagy olyan egyéb matematikai automaták, amelyek nem Turing-teljesek (pl. véges automaták), és leggyakrabban a szabatosan nem definiált, de a jóindulat elve alapján a matematikai modellek közé sorolható „szabályvégrehajtó gép", „statisztikai számításokat végző gép" és társaik. mi is az állítás ebben a bekezdésben? a kiemelkedő?

A Turing gép primátusa nem véletlen: a híres Alonzo Church-Alan Turing tézis[59] (évszám) szerint ugyanis minden kiszámítható matematikai függvény működtethető a Turing-gépen.

A Turing gép, csakúgy, mint a hetes szám, nem-fizikai entitás, számos olyan tulajdonsággal bír/rendelkezik, ami fizikai entitásokra nem is értelmezhető, de az ideákra annál inkább: nincs kiterjedése, tömege; nem kopik, az entrópia nem hat rá; tökéletes és örök; végtelen tárhellyel rendelkezik.

Másutt számos példát részletesen bemutattam[60] az MI filozófia 1950-2010 közötti történetéből, amelyek az akkor csak elképzelt, de plauzibilisen megvalósítható MI-ről próbáltak előre ítéletet mondani. Érvelésem szerint ez a történet a legjobban úgy ragadható meg, ha az érvelőket a számítógép-modelljeik alapján helyezzük el egy térképen.

Így érthetjük meg, hogy Searle[61] kínai szoba érve – amely szintaxis-gépként modellezi a számítógépet–, miért tűnik kategóriahibának az ellenérvelők számára. Penrose és Mermin *A császár új elméje*[62] narratívája a számítógép (egy modellje) és az emberi agy közötti különbséget használja. Hubert Dreyfus számos munkája[63] és Harry Collins

[59] Copeland BJ (1997)
[60] Héder (2014, 2020)
[61] Searle (1980)
[62] Penrose és Mermin (1990).
[63] Dreyfus (1992), lásd Héder (2020, 21-23).

elemzése a robotikáról[64] is az érvelő céljainak megfelelő modellválasztásra épülnek, amely választás azonban aluldeterminált.

3.2 Dekompozíció és a gépek természete

Az előző fejezetben a mérnöki tervezés lényegeként bemutatott dekompozíciós probléma sem kezelhető teljeskörűen ontológiai feltevések nélkül: láttuk, hogy például az OOAD[65] működtetése is felfogható filozófiai, a világ ontológiai kategóriáiról elmélkedő tevékenységként, a szoftvertervezők pedig nem deklarált mereologistákként.

Másfelől, ha felfedezzük, hogy a gépek önmaguk speciális ontológiai státusszal bírnak, úgy az örökké problémás funkció és működés fogalmak könnyebben kezelhetővé válnak. Továbbá, az MI morális státuszának[66] valamint az MI ágensek értékelésének[67] is a gépek metafizikájának vizsgálata az alapja.

A könyv későbbi részében a gépeket emergens entitásokként mutatom be, kiterjesztve, de egyben tömörítve is az ezzel kapcsolatos korábbi munkáimat[68]. Ennek a gép-ontológiának az elfogadása nem szükséges a könyv további érvei többségének az elfogadásához. Az MI etika általam várt új hullámainak megértéséhez elegendő annyit elfogadni, hogy az MI-metafizika sokkal nagyobb reflektorfényt fog kapni, mint eddig, és hogy a leegyszerűsítő számítógép-reprezentációk (mint például az egyszerű statisztikai gép) egyre kevéssé lesznek alkalmasak az MI kapcsán felmerülő problémák, filozófiai kihívások elemzésére.

3.3 Rossz érvek az MI ellen

Egy gyakran felhozott kategóriahiba az MI megértése kapcsán annak „explicit” természetére hivatkozik. Az érv néha odáig megy, hogy azt állítja, a robotoknak vagy számítógépeknek csak explicit programjuk van[69], míg az emberek másfajta, speciális, például „hallgatólagos” tudással rendelkeznek – ez utóbbit a 12. fejezetben fejtem ki.

64 Collins (2012).
65 Lásd a 2.3.1. fejezetben.
66 Lásd a 11. fejezetet.
67 Lásd a 12. fejezetet.
68 Héder (2019).
69 Collins (2010)

Ha a számítógépet mikroszkóp alá helyezzük, akkor azt látjuk, hogy az elektronikus töltéseloszlás, amely a program betöltésekor a gépben kialakul, számunkra számításként interpretálható eredményt hoz bizonyos permanens fizikai tulajdonságainak és a töltéseloszlás változó fizikai tulajdonságainak együttműködése által.

Az az állítás, hogy egy számítógép vagy robot (amely elkerülhetetlenül tartalmaz egy számítógépet) „explicit” utasításokat követ, összekeveri a megfigyelő és a megfigyelt viszonyát – a megfigyelő egyik modelljében explicit a programkód, a számítógépnek azonban nincs ilyen perspektívája.

A helyzet olyan, mintha azt mondanánk, hogy egy rovar (vagy bármely más, lényegében tanulásra képtelen állat) DNS-e egy olyan explicit utasításkészlet, amelyet a rovar követ. Az adott DNS az emberi tudós explicit tudásának részévé válhat a kérdéses rovarról, de ez természetesen nem jelenti azt, hogy a DNS által lehetővé tett képesség explicit tudásként jelenik meg a rovar számára.

Természetesen, mielőtt léteztek volna emberek, akik képesek lettek volna explicit tudást artikulálni, a DNS ugyanúgy működött. Hasonlóképpen, a számítógépeket olyan emberek hozzák létre, akik a programkód (részben) explicit nyelvét használják a kommunikációra egymás között. Nagy valószínűséggel a számítógépet soha nem készítették volna el, ha nem lenne ez a részben explicit nyelv, amelyet az emberek használhatnak. Azonban az explicit tudás a számítógépről nem ugyanaz, mint az a fajta explicit tudás, amelyet a számítógép birtokolna.

A gépi tanulás egy olyan példa, amely egy újabb módon mutatja be, miért nem tartható az, hogy a gépek explicit utasításkészleteket követnek. A gépi tanulás esetében azok, akik azt állítják, hogy a számítógépek csupán azt csinálják, amire programozták őket, dilemma elé kerülnek. Ha egy gépet úgy programoznak, hogy tanuljon, majd az alapján hajtson végre műveleteket, amit megtanult, akkor még mindig mondhatjuk-e, hogy ez csupán programvégrehajtás? Azok a viselkedési formák, amelyeket a tanuló gépek mutatnak, különösen az előre nem látható, néha problémás viselkedések, nagyon megkérdőjelezik az „ez csak egy program végrehajtása” érvet. Ez az érv leginkább akkor érvényesül, amikor mind a program, mind a bemenet előre meghatározott; ilyenkor nincs valódi véletlenszerűség, ezért az algoritmus kimenete

teljes mértékben a bemenet és a kód alapján meghatározott. Azonban alig léteznek ilyen rendszerek, mivel az érzékelők és a külvilág reprezentációja az MI alapvető jellemzői közé tartoznak, akár egy robottestben, akár egy hagyományos számítógépben működnek.

Kategóriahiba, ha összekeverjük a gépek készítőinek vagy felhasználóinak modelljeit a gépek saját attribútumaival. Ez egy téves redukció – csakhogy a redukció szokásos irányához képest itt nem az anyag alacsonyabb szintű alkotórészeivel azonosítjuk a komplex jelenségeket, hanem a számítógépet a róla alkotott absztrakt modellel tévesztjük össze. Ezt akár „felfelé irányuló redukciónak” is nevezhetjük (az ontológiai viták elismerhetően önkényes figuratív nyelvén az absztrakciók általában „fölötte” helyezkednek el a konkrét dolgoknak). Néha egy gépet véges automataként tudunk modellezni, néha Turing-gépként (amely nem véges), néha pedig egy megtestesült matematikai függvényként. De modellezhetjük őket termodinamikai rendszerekként, valamint nem-determinisztikus, és így feltörhetetlen, véletlenszám-generátorokként is[70]. Emellett léteznek olyan modellek is, amelyekben a gép fő jellemzői az érzékelők és a végrehajtók, és a cél az egész rendszer irányításának megteremtése; ezt hívják kiber-fizikai rendszerek megközelítésének, aminek egyes elemeit a 4. fejezet mutatja be.

Egy világos és sokat kritizált példája ennek a kategóriahibának az az állítás, hogy mivel a formális rendszerek, beleértve a Turing-gépet is, Gödel nemteljességi tétele alá esnek, a számítógépek valamilyen módon eredendően korlátozottak az emberi elmével szemben[71]. Ez az érv azt feltételezi, hogy elfelejtjük, hogy a számítógépek valójában nem Turing-gépek vagy formális rendszerek, amelyek matematikai absztrakciók – hanem eközben fizikai rendszerek is, amelyeket a természetben is megtapasztalhatunk. Miközben igyekszem elkerülni a matematikai platonizmus problémáját, megkockáztatom a következő leírást: bár sokan a fejükben gondolkodnak Turing-gépekről, senkinek sincs egy Turing-gép a zsebében.

Míg a Gödel nemteljességi tételei és az MI közötti vita talán a legtisztább és leginkább megcáfolt példa, ez csupán a jéghegy csúcsa. Például olyan érveket is megfogalmaztak, amelyek azt követelik, hogy a számítógépeket egy olyan modellel

[70] Héder (2020, 105)
[71] Lásd például Penrose (1994)

azonosítsuk, amely „szintaxist” ismer, de „szemantikát” nem[72], vagy amely „szabályokat követ”[73].

A helyzetet tovább bonyolítja az olyan nyelvezet használata, amely például azt állítja, hogy néhány jelenleg igen népszerű MI szolgáltatás „nagy nyelvi modell”. A veszély abban rejlik, hogy néhányan ezt szó szerint veszik és a rendszert a modelljével azonosítják, ahelyett, hogy megértenék: valójában olyan gépekről van szó, amelyeket olyan rendszerekre vonatkozó tudással és módszertannal hoztak létre, amelyek könnyen alkalmazkodnak a nagy nyelvi modell reprezentációjához.

Ezek a megtévesztő behelyettesítések[74] ahhoz vezethetnek, hogy veszélyesen félreértjük az MI természetét, és ezáltal elhomályosítjuk az MI megmagyarázhatóságának lehetőségeit és az MI iránti esetleges erkölcsi kötelezettségek lehetőségét. Valóban, bármilyen tényleges MI rendszerrel kapcsolatos vitában érdemes tisztázni, hogy hol és mikor helyezkedik el az MI az idő-tér kontinuumban, valamint megközelítőleg mekkora tömeget képvisel és mennyi energiát fogyaszt; más szóval, emlékeztetnünk kell magunkat arra, hogy sok tekintetben osztozunk ugyanazon természet jellemzőin.

[72] Searle (1980)
[73] Larson (2021)
[74] A „megtévesztő behelyettesítés” kifejezés hétköznapi nyelven is érthető, de a 12. fejezetben bemutatott poszt-kritikai megközelítés terminus technicusa is egyben.

4. Az MI episztemikus átlátszatlansága

A gépek előző fejezetben bemutatott, működési elvek mentén megragadható természetének az az egyik következménye a tervezés filozófiájára nézve, hogy a helyes működés lényege nem ragadható meg a természettudomány fogalmaival. Az így előállt helyzet sajátos episztemikus karakterrel rendelkezik, amely a mesterséges intelligencia etikájára különösen nagy hatással bír.[75]

Az MI, vagy az autonóm rendszerek etikai kérdéseit taglalva a vita tárgya gyakran az, hogy mit tegyen, vagy épp mit ne tegyen egy rendszer egy adott szituációban. A lehetséges kimenetelek egyértelműen definiáltak, habár csak kezelhető számban, amely gyakran nem több kettőnél. Az ilyen becslések jellege precíz és határozott, nem pedig valószínűsíthető – probabilisztikus. Más szóval, tökéletes információkból dolgozunk. Ez különösképpen igaz a nagyközönség részére szánt problémafeltárások esetén.

Kiváló példa erre a MIT (Massachusettsi Műszaki Egyetem) híres Moral Machine kísérlete (2016 –). Lényegét tekintve a Moral Machine egy – a MIT Media Lab által összeállított – igen kreatívan kivitelezett közvélemény-kutatás azzal kapcsolatban, hogy a közvélemény szerint két lehetséges választás közül melyiket kellene egy önvezető autónak meghoznia. Választani persze csak kettő meglehetősen szerencsétlen végkimenetel közül lehet, a felmérés alanyainak mégis döntést kell hozniuk, azzal kapcsolatban, vajon öregeket, vagy fiatalokat, férfiakat vagy nőket, kutyákat vagy embereket, bűnözőket vagy orvosokat stb. mentenének meg inkább egy autóbaleset esetén[76] – feltárva a feltárva preferenciáikat. Ez az a probléma-reprezentáció, amelyet az 1.2 fejezetben leírt episztemikus távolság miatt korlátozott hasznosságúként jellemeztem.

A fenti felmérés során 2016-2018 között 40 millió egyéni válasz gyűlt össze és szolgáltat hasznos információkat. A kísérlet megítélése nem egyértelműen pozitív. A kritikusok között Saxena[77] megkérdőjelezi azt, hogy vajon megfelelően tükrözi-e mindez a közvélemény – kultúránként akár eltérő – értékítéletét, ennélfogva alkalmas-

[75] E fejezet (Héder 2023) tovább gondolásával készült.
[76] Awad és társai (2018)
[77] Saxena és társai (2009).

e arra, hogy algoritmusok tervezési folyamatában használják. Mások[78] pusztán feltáró jellegűnek tekintik a kísérlet során kapott eredményeket amelyeket egyéb módszerekkel igazolni kell. Természetesen, a MIT Moral Machine csak egy a számos, hasonló jellegű erkölcsi dilemma-felvetések közül.[79] Az efféle problémakeretezések haszna megkérdőjelezhetetlen a közvéleményben uralkodó morális értékeket felmérő eszközeként. Több mint 400.000 válaszadója révén e tanulmány képes volt rávilágítani az országok és régiók közötti, kulturális alapokon nyugvó, erkölcsi megközelítésben és értékítéletben uralkodó különbségekre.

Az episztemikus távolságból adódó problémát a 1.3-as fejezetben vezettem be. Ez – röviden – abban áll, hogy a fentihez hasonló modellek azt a látszatot igyekeznek kelteni, mintha az autonóm rendszer tervezéséért felelős csoport képes lenne meghatározni azt, mit tesz majd a rendszer egy adott, jól meghatározható helyzet során. Ez azonban nem fedi a valóságot, mivel a tervezőcsoport nem láthatja teljes bizonyossággal előre, mi lesz az általuk tervezett rendszer reakciója egy bizonyos helyzetben, valamint az sem garantált, hogy az adott rendszer pontosan, megfelelően értelmezi majd a felmerülő szituációt. Ezen felül, a különféle alternatív kimenetelek száma meglehetősen magas, ráadásul sosem pontosan meghatározható.

Más szóval, a tervezőcsoport episztemikus értelemben korlátozott, legalábbis a legtöbb kísérlet által lemodellezett, tökéletes információkkal rendelkező helyzetekkel összehasonlítva. Minden, kellően összetett rendszer esetén (és az összes, az MI-t illető viták során szóba kerülő rendszer igen összetett), a tervezők ugyan igen nagy hatalommal rendelkeznek teremtményeik fölött, még sincs a kezükben olyan magas szintű irányítási lehetőség, ami lehetővé tenné, hogy egyszerűen lekódoljanak egy kívánt utilitarista, deontológista, vagy bármely más erkölcsön alapuló viselkedést.

Az MIT Moral Machine és az ehhez hasonló probléma-modellek egyike sem tervezési időben megválaszolható dilemma; ezek *in-situ (helyi)* döntési problémák. Abban a pillanatban, amikor egy sofőr a kormányhoz ül, megjelenik az in-situ erkölcsi összetevő. Az autonóm rendszerek elvetik ezt, így teljesen kizárják az in-situ *(csak az adott helyzetre vonatkozó)* döntéshozás lehetőségét. A gépi autonómia megvalósulásához minden morális döntést már a tervezés során meg kellene hozni; ám

[78] Kaplan és Haenlein (2020).
[79] Goodall (2014a és 2014b), Lin (2014), Lin (2015).

akkor a jövőbeni helyzeteknek csak egy része előrelátható, vagy szimulálható. A várt, vagy előre látható körülmények az összes lehetséges dolognak csak kis részét képezik. Ennek a helyzetnek a felismerésével jött létre az MI etikában a „felelősségrés” fogalma (lásd 7.1.2 fejezetet).

Az ismeretbeli hiány két jelentős okra vezethető vissza: a rendszert körülvevő környezet átlátszatlanságára, és magának a rendszernek az átlátszatlanságára. Ez a fejezet az utóbbival foglalkozik: az átlátszatlanságnak és ellenállásnak a tervező akaratával szemben magán az autonóm rendszeren belül jelenlevő forrásaival. Ezután áttérek arra, hogy mindez csak a rendszer hallgatólagos ismeretéhez köthető, összetett probléma-e.

Példáimat ebben a fejezetben az önvezető autók köréből merítem; ugyanakkor úgy gondolom, hogy a feltárt igazságok minden autonóm rendszerre érvényesek.

Számos, az algoritmusok átláthatóságát gátló akadályt sorolok fel, azt állítva, hogy ezek közül sok elválaszthatatlanul kapcsolódik az autonóm rendszerek ember általi tervezését magában foglaló helyzetekhez, ezért eredendően leküzdhetetlenek.

A fejezetben központi jelentősége lesz a *rétegek* és az *emergencia* kifejezéseknek. E meghatározásokkal kapcsolatban fontos, hogy ebben a fejezetben mindig ismeretalapú (episztemikus) kategóriákra vonatkoznak, azaz nem szükséges a később, a 12. fejezetben bemutatott hierarchikus ontológia melletti elköteleződés az itt bemutatott tézisek elfogadásához. Szerencsére úgy tűnik, hogy az episztemikus emergencia lehetősége, valamint a mérnöki tervezői tudás rétegeinek episztemikus differenciálódása a gépi architektúrák összefüggésében gyakorlatilag soha nem képezte vita tárgyát.

Ahhoz, hogy a következő alfejezetek érthetőek legyenek, be kell vezetnem néhány fogalmat.

Az *átlátszatlanság*[80] a mérnök és alkotása között fennálló episztemikus korlátra utal. A létrehozott rendszerek annyira összetetté és ön-módosítóvá válnak – gépi tanulás útján –, a mérnöki csoport pedig olyan nagy, hogy egyetlen személy nem képes mindezt megérteni, átlátni. Ebből fakad az átláthatóság hiánya, más szóval az átlátszatlanság.

[80] Az angol nyelvű irodalomban „epistemic opacity”.

Mindez annak ellenére áll fenn, hogy szoftverek esetén elvileg az utolsó részletig minden visszakövethető.

A 2. fejezetben már említett *tervezési idő* egy mérnöki fogalom: a rendszer tervezését magában foglaló időablakra vonatkozik. Az informatikai mérnöki tudományban *futási idő* a szoftver futására vonatkozó időablak. E kettő megkülönböztetése kulcsfontosságú számunkra a két helyzet között tátongó hatalmas ismeretelméleti szakadék miatt.

Technológia stack[81] a mérnöki megoldások réteges architektúrájára utal. A kifejezés valószínűleg a programozás területéről ered, ahol az mérnökinformatikusok programcsomagokban és halmazokban gondolkodnak, melyek különböző szoftver-rétegek munkáját egyesítik: például, az operációs rendszerben olyan program-keretrendszer futhat, amely cserébe valamilyen „üzleti logikát” futtat. A lényeg, hogy az effajta szeparáció több értelemben is lehetővé teszi a munkamegosztást: a magasabb szintű programok az alacsonyabb szintűek felé delegálják a feladatokat; humán szinten pedig a különböző rétegeket más-más emberek fejleszthetik (ennek szükségszerűségét szintén a 2. fejezet tárja fel részletesen). Így, a legáltalánosabb feladatok az alsóbb rétegek szintjén zajlanak (ahogy például az operációs rendszer kezeli az állományokat és a hálózatot), hogy a felsőbb rétegek minden részletre kiterjedően, egy konkrét feladatra koncentrálhassanak. Ez a rétegződés tovább folytatódik lefelé, a hardveres szinten, tehát a *technológia stack* kifejezés a gépezet teljes egészére vonatkozik.

4.1 Az autonóm rendszerek átlátszatlanságának forrásai

E fejezet központi tézise szerint az autonóm rendszer viselkedése nem tervezhető előre – legalábbis olyan mélyreható részletességgel nem, amelyet egy érték- vagy etikai rendszer létrehozása megkövetelne. Ennek egy elégséges, és mindig jelenlévő oka a rendszer átlátszatlansága, amelynek több forrása van.

Az egyik ilyen forrás a gépi tanulási rendszerek – különösen az neurális hálózatok – híresen magas bonyolultsága. Ez számos vizsgálat tárgyát képezi.[82]

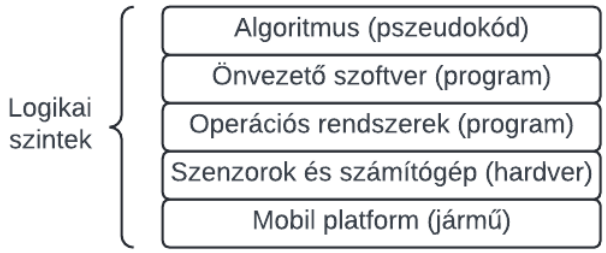
[81] A magyar informatikai szakfordításokban a stack-et „veremként” fordítják, ám ezt a kifejezést sem kellően ismertnek, sem megfelelően kifejezőnek nem tartom, ezért hagyatkozok az angol szóra.

[82] Pasquale (2005), Diakopoulos (2014), Burrell (2016).

Innen eredeztethető az MI etikában a *transzparencia* és *megmagyarázhatóság* kulcsszavakkal leírt, nagyon széles körben kutatott/vizsgált altéma, melynek történetét a 9. fejezet mutatja be.

A bonyolultság önmagában is igen nagymértékben alátámasztja az átláthatatlanságra vonatkozó aggodalmakat, de vannak egyéb, kevéssé ismert indokok is, amelyek az MI etika új hullámaiban játszhatnak szerepet.

4.1.1 Rétegek közötti interakciók vs. enkapszuláció



6. ábra: Egy önvezető autó technológia stack-je, amely az önvezetés sikeres megvalósulásáért felelős.

Az átláthatatlanság forrásainak megértéséhez érdemes megvizsgálni az autonóm rendszerek réteges architektúráját, és azt a legalább episztemikusan emergens viselkedést, amely a rétegek között jön létre. Egy felsőbb rétegben, az architektúra bármely pontján, olyan események zajlanak le, melyeket az alsóbb szintektől elérő szakmai nyelven és vizsgálati eszközökkel elemeznek.

A felső szintek mindig szolgáltatásként veszik igénybe az alsóbb szintek tulajdonságait, *elvágólagos interfészek* mentén, úgy, hogy az *enkapszuláció*[83] miatt a működés részleteihez nem férnek hozzá.

A fenti állítás általánosítható, miszerint az autonóm rendszerek nem rendelkeznek olyan megbízható önellenőrző rendszerrel, amely képes lenne megállapítani, hogy egy adott szinten jelen vannak-e a megfelelő működést biztosító feltételek – azaz megfelel-e az alsóbb szint állapota[84]. Léteznek állapotfelmérések, például egy számítógép képes

[83] A fogalmak magyarázatát lásd: 2.3.1 fejezetet.

[84] Természetesen számos önellenőrző funkcióval rendelkeznek, azonban a magasabb szintek ezeket ugyanúgy szolgáltatásként veszik igénybe az alacsonyabb szintektől, e viszony minden hibalehetőségét vállalva.

a saját fájlrendszerének vizsgálatára, és felismeri azokat a jeleket, melyek egy jól működő rendszer esetén nem lehetnek jelen, majd ebből kiindulva képes egyes hardverelemet hibásnak nyilvánítani. Az információtárolás során mindig lehetséges a redundancia hozzáadása, amely alkalmas az információátviteli vagy fogadási hibákat korrigálni.

Azonban az a tény, hogy a vezérlési folyamatok elől el van rejtve az alacsonyabb rétegek részletgazdagsága, az alacsonyabb rétegek hozzáférése a magasabb rétegekéhez pedig tipikusan le van tiltva, nem jelenti azt, hogy nem tudnak egymásra hatni olyan módokon, amit a rendszer logikai leírása nem tartalmaz.

Egy magasabb szint (például az algoritmus) aktivitása olyan állapotba hozhatja a rendszert, melyben az alsóbb szintek (ilyen a hardver is) nem képes megfelelő működésre – például a falnak vezeti az önvezető járművet. Ily módon, a rendszer magasabb szintű leírása által kezelt vezérlés egyszerűen tönkreteszi a saját megfelelő működéséhez szükséges előfeltételeket, a hardvert.

Tehát például annak igazolásához, hogy algoritmusunk nem fog semmilyen nem kívánt módon viselkedni, azt kellene bizonyítanunk, hogy az adott algoritmushoz kapcsolódó output – ne feledjük, esetünkben egy önvezető autót irányítunk – nem kormányozza a rendszert olyan helyzetbe, amely károsíthatná az alsóbb ismereti szinten modellezett hardvereket.

Lehetnének az adott rétegen belül garanciák arra, hogy bizonyos határokon belül marad – formális rendszerek esetén például adott a lehetőség annak bizonyítására, hogy a rendszer bizonyos jellemzőinek megléte mindenkor biztosított.[85] De ez is azzal az apró betűs kitétellel értendő, miszerint mindaddig, amíg a rendszer alsóbbrendű elemei (melyeket az alsóbbrendű réteg biztosít) megfelelően működnek.

Az önvezető jármű esetén azonban az algoritmus szállítja a hardvert, az operációs rendszert, és minden mást, ami olyan kockázatot jelent ezekre az alsóbb szintekre, melyek kiküszöbölésére a felsőbb szinteken semmilyen próbálkozás nem történik. Sőt, felsőbb szinten nem is létezik az alsóbb szintek teljes megszólítására alkalmas nyelv.

[85] Butler (2001).

Ha pedig a szoftveres szintnél mélyebbre megyünk, ott például nem találunk a hardver elhasználódására vonatkozó deduktív modellt. Ehelyett csak és kizárólag induktív-statisztikai[86] modellek léteznek, arról, mennyi idő után kezdenek egy átlagos RAM modul egyes részei hibásan működni.

4.1.2 A megtestesült tényezők hatásai

Az átláthatatlanság egy másik forrását nevezhetjük a megtestesült tényezők hatásának is. Itt az előző alfejezetben kifejtett gondolat továbbfejlesztéséről van szó.

Tegyük fel, hogy egy önvezető jármű technológia stack-je sikeres tesztelés után kísérleti forgalomba kerül. A stack tartalmaz egy *citromsárga* mobil platformot (az autót magát), a számítógépes hardvert, az operációs rendszert, és ezek felett egy AV[87] stacket, amely egy finomra hangolt önvezető algoritmust futtat. Tételezzük fel azt is, hogy egy másik ugyanilyen jármű is készül, ez azonban *halványszürke* színben. Tegyük fel azt, hogy a megfigyelések alapján, a biztonsági teljesítmény – a balesetmentesen megtett órák száma – egy gyalogosokkal sűrűn benépesített környezetben ezen második autó esetén alacsonyabb, mint a rikító sárga színűé, pontosan a színbeli eltérés miatt. Utólagos vizsgálataink megállapítják, hogy ez az autó a gyalogosok számára a másiknál kevésbé látható, és mivel más eltérés nem tapasztalható, a Mill-féle oksági heurisztikával megállapítjuk, hogy a különbséget ez a faktor okozza.[88]

Tehát, hogyan írható le szabatosan, és hogyan védhető ki ez a hiányosság? A kísérleti gondolatmenet rávilágít a technológiai stack rétegei között rejtőző, az autó teljesítményének tervezésére is kiható egymásrautaltságra. Amennyiben ez valós, akkor az általános mérnöki terv, azaz a halmaz különböző szintjeire vonatkozó mérnöki dokumentumok (az operációs rendszer valamint a szoftverstack forráskódjai és futásidejű konfigurációja, a mobil platform tervrajzai stb.) erre a részletre is ki kell térjenek. Más szóval, különböző jármű színek tesztelése lehet szükséges, és csak bizonyos színskálát lehet engedélyezni; a felhasználó figyelmét pedig fel kell hívni az esetleges, nem hivatalos újrafestésből fakadó teljes garanciavesztésre.

[86] Lásd Carl Hempel (1958).
[87] AV: Autonomous Vehicle, azaz önvezető jármű.
[88] Mill (1884).

Az autó módosítására vonatkozó efféle tiltás ma már mindennaposnak számít. Ezek hatálya sok más részletre is kiterjed, például a fékek vagy a lámpák nem kiszerelhetők. Továbbá, teljességgel lehetséges, hogy az ember által vezetett autók körében ugyanazon sofőr esetén is egy élénkebb színű autó statisztikailag biztonságosabbnak minősül. Az autó színére azért nem terjed ki a jelenlegi szabályzás, mert a sofőr személyes jelenléte – aki elszámoltatható, szükség esetén törvényileg büntethető – elnyeli ezt a problémát. Azonos feltételek mellett a szürke autót vezető sofőrök talán több balesetet okoznak, mint a sárgában ülők, de ezt – elméletben – senki nem veszi számításba a baleseti felelősség tisztázása során.

Önvezető autók esetén, a 7.2. fejezetben bemutatott „felelősségrés” értelmében nincs ilyen, felelősséget kiváltó összetevő, tehát minden a konstrukcióra és a tervezőre vezethető vissza. Ugyanakkor, az autó színe és a láthatósága csak egy egyszerű példa. Valójában nem tudjuk, melyek azok a fontos fizikai tényezők, melyek a tervezés során külön figyelmet és célzott szabályzást igényelnének. Minden terv aluldeterminált az általa elkészült, megfogható végtermékkel összehasonlítva és viszont. A biztonságos kereteken még éppen belül eső szabadság mértéke pedig deduktív módszerrel nem megállapítható.

4.1.3 A fizikai réteg viszonya a hardverrel

A modern számítógépes hardverek esetében az egyik fontos szűk keresztmetszet a hűtés, más szóval a hatékony hőelvezetés. Ez három módon is kivitelezhető. Az első, a számítások energiatakarékosabbá tétele, így csökkentve a hőtermelést. A modern chipek korában hatalmas előrelépések történtek ez irányba, de természetesen a termelt hő mértéke továbbra is nagyjából egyenesen arányos az elvégzett számításokkal, és ahogy az IBM kiválósága, Rolf Landauer megállapította,[89] létezik egy egyetlen kalkulációs egységre vonatkozó elméleti energiaminimum.

A hő elvezetésének másik módja a hatékonyabb hűtés, de ez maga is energiát fogyaszt és gyakran nagy zajjal jár.

Így aztán egy harmadik módszer is használatban van: a számítások sebességének módosítása a hardveren mért hőmérséklet függvényében. Így a számítások mennyisége egy rövid időintervallumon belül tetőzhet, jól szolgálva a felhasználót, hiszen a tipikus

[89] Landauer (1961).

számítógép-használati minta rövid, koncentrált aktív, majd az ezeket követő hosszabb inaktív periódusokból áll: ilyen például egy böngészőoldal betöltése, majd annak elolvasása. A nagysebességű számítási teljesítmény persze nem tartható fent hosszú időn keresztül a termelődő hőmennyiség miatt. Az egyensúly fenntartása érdekében a számítások gyakorisága mesterségesen korlátozható.[90]

Emiatt a hőfüggőség miatt melegebb környezetben kevesebb számítás végezhető el, ugyanakkora mennyiségű számítás elvégzése pedig több időbe telik. Önvezető autók esetén, a válaszok időben való érkezésének biztosítása érdekében valós idejű operációs rendszerek használatára kerülhet sor, ami azt jelenti, hogy bizonyos műveletek esetén CPU-idő helyett egy adott, valós időperióduson belül érkező válaszokat garantálunk. Eközben a fentebb leírtak szerint az önszabályozó hardveren a műveletek különböző hőmérsékleti viszonyok között más-más valós időtartamot vesznek igénybe.

Tehát, a rendszer feladata egy adott mértékű reszponzivitás biztosítása minden olyan helyzet esetén, amelybe a jármű elméletileg belekerülhet. Nincs értelme azonban ezt az alsó limitet egyben felső limitté is tenni: amennyiben lehetséges, végezzen több kalkulációt, ami ez esetben egy-egy úton heverő tárgy jobb kategorizálásához, vagy a jármű pozíciójának precízebb meghatározásához vezethet.

A minimális és maximális teljesítmény között húzódó intervallum megléte miatt nagyon nehéz pontosan megjósolni, hogy egy adott időpillanatban mennyire képes hatékonyan végbemenni egy bizonyos helyzet értékelése. A tervezők kénytelenek elfogadni azt, hogy valamivel kedvezőbb CPU hőmérsékleti körülmények között ugyanazon helyzet felismerése hatékonyabb lehet, mint egyébként, és hogy az e hatásból adódó esetleges variációk bekövetkezését lehetetlen előre látni.

Ez csak egy példa arra, hogyan befolyásolják az anyagi réteg folyamatai az egész rendszer működését anélkül, hogy megtörnék azt. Sok más ilyen példát említhetnénk, különösen a kamerákhoz hasonló érzékelők működése és a változó fizikai környezet ezekre kifejtett hatását illetően.

Ezen folyamatok olyan részletezettségű leírásához, ami lehetővé tenné a megszokott etikai döntési modellekből már ismert, világos és jól körülírt etikai kérdések felvetését,

[90] Héder (2020, 9-es fejezet).

Laplace démonához hasonló mértékű ismeretekkel kellene rendelkeznünk a rendszerünkről, ami egyszerűen nem lehetséges.

4.1.4 Kizárólag statisztikai jellegű tudás

Az eddigi érvek összessége egy olyan modell felé mutat, melyben az MI aktuális helyzetéről és a lehetséges alternatív kimenetelekről csak korlátozott ismereteink vannak. Ez felveti a külvilág statisztikai modellezésének igényét.

A lehetséges kimeneteleket illetően a bizonytalanságok egy újabb csoportja merül fel: ezek a jármű reakcióinak hatáselemzéséhez kapcsolódnak. Például, bármilyen adott szituációban a pontos féktávolság felmérése nem lehetséges, hiszen ez nyilvánvalóan az útfelületek pillanatnyi tulajdonságaitól is függ.

Ilyenkor a rendszer egy visszacsatolási mechanizmust alkalmaz – ami a kontroll-elméletből jól ismert koncepció[91]. Ez esetben a rendszer egy válaszreakciót kalkulál, amelyet megvalósít, majd a helyzet újabb mérlegelése után alkalmazza a további válaszreakciókat. Mindez azért szükséges, hogy a rendszert az útvonaltervező szoftver, vagy más, szintén magasabb szintű architektúra által meghatározott keretek között tartsa. A rendszer ellenállása azonban – saját tényezői és a környezet kiszámíthatatlansága miatt–, behatárolja az elérhető kontroll mértékét. Mindezt egyfajta lendületként képzelhetjük el, időnként fizikális értelemben szinte szó szerint, máskor pedig állapottérben értelmezve.

Az autonóm rendszer kiszámíthatatlansága, mint saját kontroll-mechanizmusának tárgya, valamint a környezet hiányos ismerete egy számos bizonytalan tényezővel rendelkező belső világreprezentációt eredményez.

Ráadásul, a reprezentációban leírt lehetséges kimenetek valóságos adatokon, egy gépi tanulási folyamat révén nyert múltbéli tapasztalatokon alapulnak. Már magának a tanulási adatnak reprezentatív jellege is kérdéseket vet fel – az adatok részrehajlása már jól ismert az MI etika számára, a 7-8. fejezetekben részletesebben is kitérek rá.

Az autonóm alkalmazások elterjedése valószínűleg nem minden országban, továbbá nem minden társadalmi réteg és csoport szintjén egyszerre történik majd. Tehát, legjobb esetben a tanulás alapját képező adatok az alkalmazásokat az elsők között

[91] Lásd az 5. ábrát a 2. fejezetben.

bevezetőktől származnak majd. Jó példa erre a Tesla önvezető szoftvere. Ismert, hogy valós vezetési adatokon alapul: ha nincs önvezető módban, a szoftver figyeli a vezető reakcióit, és kalkulálja, vajon ő mit tenne az adott helyzetben- eközben megjegyzi a különbségeket és tanul is belőlük. Ez legalább két szinten jelent szükségszerű torzulást: először, az etalonként értékelt emberi vezetés annak a csoportnak vezetési szokásait képviseli, akik megengedhetik maguknak egy ilyen drága autó megvásárlását. Másodszor, a gyakorlásra használt utak és környezet olyan országokban találhatóak, ahol a Tesla járművek elterjedése nagyobb mértékű. Tehát, összességében elmondhatjuk, hogy a morális relevanciával rendelkező szituációk gépi modellezése nemcsak, hogy probabilisztikus, de maguk a valószínűsíthető lehetőségek legjobb esetben is csak tesztadatokon alapulnak, melyek jó esetben helyesen becsülik fel az aktuális környezetet, míg rossz esetben teljesen tévesek lehetnek. A gépi tanulás ezen elkerülhetetlen limitációinak következményeit a 10. fejezetben tárgyalom.

4.1.5 Kiszámíthatatlan emberi döntések

Szembesültünk már azzal, hogy az autonóm rendszer önmagában is bizonytalanság forrása, ugyanúgy, mint a fizikai környezet. Ugyanakkor, az emberi lényekkel és az adott közlekedési helyzetben esetleg részt vevő más önvezető járművekkel kapcsolatos reakciók az eredményeket még inkább átlátszatlanná teszik. Képzeljünk el egy rendszert, amely épp egy olyan gyalogossal való ütközés felé tart, akinek még maradt arra elegendő ideje, hogy jobbra vagy balra félreugorjon, de akár egy helyben is maradhat. Bár a rendszerünk meghozhatja a döntést, hogy balra, vagy éppen jobbra érdemes kitérni, a helyzet mégsem oldódott meg, hiszen a gyalogos úgy dönthet, hogy éppen arra ugrik, amerre a rendszer kormányozza az magát.[92] A megszokott modellek egy adott helyzetben csak a tervezett rendszer reakcióit veszik alapul, a résztvevők reakcióival nem számolnak. Ez egy egészen nyilvánvaló hiba. Bármilyen valósághű szimuláció során az emberi viselkedés legjobb esetben is csak valószínűsíthető módon modellezhető, ha egyáltalán van erre lehetőség.

Létezik a kiszámíthatatlan emberi döntéseknek egy másik csoportja is: a rendszer és környezete ellen irányuló rosszindulatú reakciók. Bebizonyították, hogy az neurális háló meghackelhető az útjelző táblákra ragasztott matricákkal, emberi szem számára láthatatlan grafikus zajjal, vagy egyszerűen egy kép falra kivetítésével.[93] Létezik az IT

[92] A 10.3-as fejezetben bemutatott kontraktariánus megközelítés enyhítheti ezt a problémát.
[93] Nassi és társai (2021).

rendszerek elleni támadásoknak egy olyan ága is, amely igyekszik túlterhelni a célrendszert olyan mennyiségű, vagy olyan minőségű adatokkal, melyek feldolgozására az képtelen.[94]

A tervezés során a mérnökök a legjobb esetben is csak jól – rosszabb esetben kevésbé – átgondolt becslések révén készülhetnek fel ezekre az emberi tényezőkre. Ez megint csak az átlátszatlansághoz vezet.

4.2 Gondolatok a tervezés alapproblémájáról

E fejezet bemutatta, hogy az a gyakran rejtetten megjelenő elképzelés, amely tökéletes információkat feltételez a morális kérdések megvitatása kapcsán, autonóm rendszerek tervezése során nem alkalmazható. Ezzel szemben viszont óriási szakadék tátong a tervezés-idő és a használat-idő episztemikus feltételei között. Látható, hogy az episztemikus, ismeretalapú szakadék mellett egy morális űr is felfedezhető. Mivel felhasználóról (aki a felhasználás során történtekért felel) jelen esetben nem beszélhetünk, úgy tűnik, vagy minden felelősség a tervezőket terheli, vagy, ha ez nem tartható fenn, akkor *felelősségrés*ről[95] beszélhetünk.

Az itt felsorolt átlátszatlansági faktorok egy része még szinte egyáltalán nem került az MI etikában a transzparenciával foglalkozó gondolkodók látókörébe. Ez a terület, amelyet kis részben a 7., teljes mélységben pedig a 9. fejezetben mutatok be, főleg a bonyolultságból fakadó episztemikus limitációkkal, a tanulóadat reprezentativitásával, olykor a profitorientált cégek ellenérdekeltségével foglalkozik. Mindhárom kérdéskörre kínál válaszokat az irodalom: megmagyarázható MI („XAI”[96]) technikák; méltányossági metrikák a tanulóadatra; a cégek regulációja. Az MI rétegelt dekompozíciójából, a terv és az implementáció kiküszöbölhetetlenül aluldeterminált viszonyából és a környezettel való kapcsolat megismerhetetlenségéből fakadó kiküszöbölhetetlen transzparencia-hiány azaz átlátszatlanság, valamint az azzal való együttélés az MI etika újabb hullámainak témája lehet.

[94] DDoS – Distributed Denial of Service támadás, amelynek során feltört személyi számítógépek ezrei, akár milliói utasítást kapnak, hogy egyetelen szolgáltatást támadjanak.
[95] „responsibility gap” – lásd a 7.2 fejezetben.
[96] „eXplainable AI” – lásd a 9. fejezetben.

Léteznek további tényezők is, amelyek az átlátszatlanságot növelik, de itt nem térek ki rájuk Az egyik a rendszerek nagy sebessége, amely nem a tervezési időben okoz átláthatatlanságot, hanem a futás során –a szakértői rendszerek idején már felmerült gondolatmenet szerint az MI következtetési sebessége annyira meghaladhatja az emberét, hogy ez önmagában a megértés akadálya. Egy másik tényező az, hogy a rendszer egyes elemeit szükségszerűen eltérő személyek vagy csapatok tervezik, így a közöttük végbemenő kommunikáció tökéletlenségei újabb átláthatatlansági tényezőt jelentenek.

A 2-3. fejezetben bemutatott szoftver-jellegzetességek és az itt bemutatott episztemikus korlátok is hozzájárulnak az MI technológiába zártsági potenciáljához, amely a következő témánk.

5. MI és a technológiába zártság

Könyvem I. részének végén, ebben a fejezetben az MI kiemelkedő technológiába zártság-előidéző potenciáljára koncentrálok.

Ez a tényező az utolsó abban a listában, melynek elemei különösen nagy hatással lehetnek az MI etika újabb hullámaira. A 2. fejezet emlékeztetett, hogy az MI műszaki tervezési folyamat eredménye, a 3. fejezet az MI etika és az MI metafizika kapcsolatának megkerülhetetlenségére világított rá, a 4. fejezet a megelőző két fejezet eredményeire is alapozva az átlátszatlanság kevéssé ismert rétegeit mutatta be. Ebben a részben pedig arra hívom fel a figyelmet, hogy mindezen nehézségeken túl még az is elképzelhető, hogy az MI jelenlegi konstruktőr generációja olyan fejlődési pályába zárhatja az utókort, amelyből reménytelenül nehéz lesz kiszabadulni.

Ezt a potenciált az MI fejlődés konstruktőr általi bezárhatóságának nevezem és felhasználom hozzá a gépek dekompozíciója során fennálló gazdasági és észszerűségi összefüggéseket. Ez egyfajta determinizmus, amelyet a tudomány- és technológia tanulmányok (STS) területén jól ismert technológiai determinizmus fogalmából kiindulva, de végső soron attól némileg eltérve építek fel.

5.1 A technológiai determinizmus

A technológiai determinizmus arra az elképzelésre utal, hogy a technológia formálja a társadalmat és a kultúrát (aránytalanul erősebben, mint fordítva).[97]

Nincs egyetlen elfogadott definíciója, helyette inkább egy definíció-családról beszélhetünk, ahogyan különféle szerzők eltérő elemeket hangsúlyoznak a jelenség kapcsán. A legerősebb megfogalmazásokban a technológia fejlődése olyan mértékben határozza meg az emberiség sorsát, hogy nincs tere a társadalmi kontrollnak. Ezt az álláspontot az erőssége miatt meglehetősen nehéz védeni, ennek megfelelően kevesen képviselik. Másfelől a technológia determináló erejének teljes tagadása hasonlóan nehéz. Ezért a technológiai determinizmus elméletek egy, a teljes determinizmus és a teljes indeterminizmus közötti spektrumon szóródnak.

[97] Ennek a fejezetnek egy tömörebb, korábbi, angol nyelvű verziója a Héder (2020).

A technológiai determinizmus számos megfogalmazásának közös elemeiből kihámozhatjuk a jelenség lényegét. Egyfelől azt látjuk, hogy egy speciális oksági viszony fennállását tételező elméletekről van szó. A másik jellemző az asszimmetria vagy egyenlőtlenség a technológia és a társadalom között, a technológia javára.

Az első, oksági elem azt állítja, hogy a technológia – szerzőtől függően vagy egy konkrét technológia, vagy a modern, technologizált világ általában,– a társadalom valamely negatív tulajdonságát (pl. szabadság hiánya) okozza.

A determinizmus kifejezés metafizikai témaköröket idéz fel, például a szabad akarattal vagy teológiával kapcsolatban. Bizonyos metafizikai elméletek szerint a világ teljesen determinált, ennek megfelelően nem lehet feltételezni semmit, ami nem determinált. A determináltság ezekben az elméletekben általában a fizikai folyamatok szintjén ható, kérlelhetetlen oksági összefüggést jelent. A korai elméletek, például a laplace-i világkép ezt egyszerűen úgy képzelik el, hogy az univerzum jelenlegi állapotát az univerzum korábbi állapota okozza a természeti törvényeknek megfelelően, így, ha a törvényeket és egy állapotot ismernénk, akkor kiszámíthatnánk a múltbéli és jövőbeli állapotokat is. A kvantumfizika ezt a lehetőséget bizonyos értelmezések szerint kizárja, mert a kvantumjelenségek kimeneteléről azt feltételezi – bizonyos kvantum-interpretációkban legalábbis – hogy azok ténylegesen a mérés pillanatában dőlnek el, és korábbi világállapotok határozzák meg ezeket. Ez pusztán annyit jelent, hogy a naiv deterministа világkép nem tartható, azt nem, hogy bármilyen hatással lehetnénk az ilyen szinten zajló folyamatokra. Ezek a mechanisztikus vagy kvázi-mechanisztikus fizikalista és egyben determinált világképek ránézésre okafogyottá teszik a technológiai determinizmus kérdését, hiszen, ha a világ teljesen determinisztikus az alapvető építőkövek szintjén, akkor ezen belül a „technológia” és a „társadalom” közötti kapcsolat determinisztikusságáról felesleges is beszélni, hiszen még a reláció két végén álló jelenségek elkülöníthetősége, státusza is kérdés. Ezen a síkon egy tranzitív oksági láncot képzelhetnénk el, amelyben a váltakoznak azok az entitások, amelyeket a technológia részeként értelmezünk, és azok, amelyeket a társadalom részeként – végül egy tyúk vagy tojás problémánál lyukadunk ki.

Ám általában nem így működnek a technológiai determinizmus elméletek, hanem megkülönböztetnek két nagy, kvázi-ágenciával bíró entitást, a technológiát és a társadalmat. Természetesen az elméletek ezt az ágenciát nem a szó szorosabb –

személyekhez rendelt – értelmében használják. Tisztában vannak vele, hogy a társadalmat egyének összessége alkotja, a technológia elemeit pedig élettelen gépek és eljárások, ám a jelenség megragadása érdekében ezeket egyben kezelik. A fizikai világ determináltságának metafizikai kérdését nem teszik fel, ám feltételezik az emberek és közvetve az emberek által alkotott társadalom szabad akaratának legalább az elvi lehetőségét. Ezt egy kompatibilista (a determinizmust a szabad akarattal összeegyeztetni képes) világkép lehetővé is teszi, ahogyan bizonyos nem-fizikalista világképek is, de a kérdést leginkább érintetlenül hagyják.

A technológiai determinista álláspontok magja tehát ez: az emberek és az általuk alkotott társadalom szabadságát korlátozza vagy felszámolja a technológiák összessége, vagy azok valamilyen részhalmaza. Ezt a technológia a determináló, formáló ereje segítségével éri el.

A szabadság, mint érték elvesztése mellett a technológiai determinista elméletek gyakran tartalmaznak egy kapcsolódó elemet, a *technológia autonómiáját*. Bár a technológia autonómiája ugyancsak nem kezelhető metafizikailag megalapozott módon, hiszen a technológia nem ágensekből áll a szó szoros értelmében[98] – a fejlett mesterséges intelligencia kérdését most tegyük félre egy pillanatra – mégis, összességében, az állítás szerint a technológia kivonja maját az ember irányítása alól és önálló, a saját belső logikája szerint fejlődő és működő, tehát autonóm jelenséggé válik.

A gondolatmenet azt implikálja, hogy az emberek autonómiája és a technológia autonómiája között egy zéró összegű játszma zajlik, így ha a technológia autonómiája nő, akkor az ember autonómiája csökken.

Innen könnyen eljuthatunk a technológiai determinista elméletek másik tipikus állításához: ez pedig az, hogy a technológia dominánsabb, mint a társadalom. Nem szabad ugyanis elfelejtenünk, hogy a technológiát és az azt körülvevő technotudományos közeget emberek hozzák létre és működtetik. Ez a nyilvánvaló felismerés elvileg felhasználható a technológiai determinista világkép ellen: ugyanezzel az erővel mondhatnánk az is – folytatható a gondolatmenet – hogy a

[98] Megemlítendő, hogy létezik olyan szociológiai elmélet, ami a technológia, sőt a technotudományos tudás egyes elemeinek is valós ágenciát tulajdonít, ez a Latour-féle (2007) aktor-hálózat elmélet (ANT: Actor-Network Theory).

technológia humán-determinált. Sőt, ahol embereket dominálja a technológiai közeg, ott is valójában emberek dominálnak más embereket, a technológiát mindössze közvetítő eszközként, mediátorként[99] felhasználva.

Ezt az ellenvetést küszöböli ki a technológia különösen domináns természetére való hivatkozás a technológiai determinizmus elméletekben, például a technológia fejlődési autonómiája vagy megváltoztathatatlansága.

Nevezetesen, az ember (nem csak a hétköznapi ember, hanem a technológiát megalkotó mérnök, tudós is) elveszti a kontrollt a technológia felett. Nem feltétlenül úgy kell ezt érteni, hogy a technológia vagy egy konkrét gép elszabadul és Gólemként[100] rombolni kezd, hanem úgy, hogy a technológia fejlődését végső soron a természet és a gépek működési elvei, a hatékonyság iránya határozza meg: az egyes emberek, sőt egész társadalmak is ki vannak szolgáltatva ezeknek az elveknek.

Ez az oka annak, hogy a technológia „természete”, esetleg esszenciája, jellegzetes tulajdonságai kiemelt fontossággal bírnak a technológiai determinista elméletekben, hiszen szükségesek az érvelés megalapozásához, miszerint a technológia negatív karakterisztikái végül társadalmi jelenségként manifesztálódnak.

A technológiai determinizmussal szemben is számos elmélet született, amelyek a társadalom és az egyes egyének oksági, formáló erejét támasztják alá vagy hangsúlyozzák ebben a relációban. A technológiai indeterminista, vagy szociálkonstruktivista elméletek képviselői egyfelől a technológiai determinizmus túlságosan nagy ívű, emiatt alacsony felbontású és homályos tartalmú világleírását támadják meg. Nem tartják módszertanilag elfogadhatónak a nagybetűs technológiákról és a nagybetűs társadalmakról szóló gondolatmeneteket, a komolyabb, részletesebb vizsgálatok során pedig egyáltalán nem azt a helyzetet fedezik fel, amit a technológia determinizmus ír le. A szociálkonstruktivista megközelítés egyik alapköve az esettanulmányok megírása/létrehozása, amelyekben egy-egy adott technológia és konkrét társadalmi csoport interakcióját mutatják ki, és azt találják, hogy a technológia fejlődési iránya nem előre eldöntött. Nem valamilyen világos racionalitást követ,

[99] Wiener (1950, 212).
[100] Collins és Pinch (2014).

hanem inkább a társas csoport esetlegességei alakítják, így valójában a szociológia lehet a technológia megértésének a primer eszköze.

Ezt a viszonyt legjobban az ismeretelmélet *aluldeterminált ság* fogalma ragadja meg, amelyet fel is használnak ebben a vitában. Az aluldeterminált ság általánosságban azt jelenti, hogy bizonyos tényezők nem determinálnak teljesen egy adott kimenetet.[101] A szokásos használata az ismeretelméletekben a kísérletek kimenetelének az értékelése: amikor egy kísérlet nem várt (vagy akár várt) eredménnyel ér véget, akkor sosem tudhatjuk, hogy mi az eredete az anomáliának: a tesztelt törvényszerűség, a mérőeszköz vagy valamilyen egyéb forrás? Az eset leírása és az abban foglalt tudásunk ekkor aluldeterminálja a probléma forrását. A technológia pedig aluldeterminálja a társadalmi átalakulásokat és a társadalom dinamikáját a determinizmussal szemben érvelők szerint.

Ezért az általános, technológiai makrodeterminizmusra kevés példát találunk. Az egyik említést érdemlő kivétel Jacques Ellul technológiai miliő[102] koncepciója, amely valóban a technológia összességét, illetve a technológiában megtestesülő racionalizáló vágyat tartja a társadalom összességére jellemző szabadságvesztés okának. Hasonlóképp, Andrew Feenberg a technológia hegemóniájából[103] – a kultúra erejével támogatott, természetesnek vett és többnyire meg sem kérdőjelezett hatalmából – vezet le bizonyos negatív társadalmi folyamatokat, például a demokrácia gyengülését, azzal, hogy mindez nem szükségszerű és tehetünk ellene valamit. Ide sorolható még Albert Borgmann is, aki Ellulhoz hasonlóan az általános technológiai közegben lát problémát – egészen konkrétan a technológia által lehetővé tett elkényelmesedésben, ami az ember karakterének a gyengülését okozza, amely nem csak a természeti elemek, hanem a társas interakciók során felmerülő nehézségek leküzdésére is képtelenné teszi, ezáltal elszigeteli egymástól a társadalom tagjait[104]. Az általános technológiai determinizmust hirdető elméletek közé tartoznak még a technológiai determinizmust feltételező politikai filozófiák, mint például a Marxizmus, vagy annak bizonyos részeit elvető, de a technológia szerepét továbbra is hangsúlyozó frankfurti iskola és számos disztopikus, rideg technokratikus jövőt előrevetítő esszé és fiktív mű.

[101] Lásd: 22. lábjegyzet.
[102] Ellul (2021).
[103] Feenberg (2009).
[104] Borgmann (1984).

Ám a technológiai determinizmus erősen leszűkíthető úgy, hogy még mindig kerek elméletet kapjunk. A technológiai determinizmus néhány szélsőséges formáját kivéve, a technológia meghatározó ereje esetenként, helyenként, és talán még különböző történelmi korok között is különbözhet. Így egy jól megfogalmazott technológiai determinizmusnak azt kellene megállapítania, hogy mely konkrét technológia van okozati hatással a társadalom vagy a kultúra mely konkrét jellemzőjére.

Például korlátozhatjuk a determinista állítást az üvegházhatást okozó gázokat kibocsátó technológiákra. Egy klímaváltozás tematikájú technológiai determinista elmélet (amelynek számos képviselője van, még ha az álláspontot nem is ezen a néven nevezik) úgy hangozna, hogy a szén és olajfelhasználás a legutóbbi időkig minden egyén és társadalom számára sokkal hatékonyabb volt, mint bármilyen más alternatíva az energiasűrűség és a könnyű hozzáférhetőség, valamint egyszerű tárolás, szállítás és felhasználás miatt. Ez a technológia természetére vonatkozó, embertől független tulajdonságként leírható elem.

Tehát adott egy technológia csoport, amely óriási előnyöket szolgáltat az emberiségnek, miközben az alternatíváinak megvalósítása sokkal költségesebb. Az emberiség ezért egyfajta gravitációs kútba esik és egyre többet használ fel belőle, ami viszont klímaváltozáshoz, az pedig a társadalmi folyamatok átalakulásához/módosulásához vezet. Tehát lehetséges, hogy valaki az internettel, gőzfürdőkkel, építészeti megoldásokkal kapcsolatban elismeri a társadalom és az adott kultúra meghatározó szerepét, az energiafogyasztással és fűtéssel kapcsolatban viszont annyira szélsőségesnek látja a helyzetet, hogy ezen a területen a technológia determináló erejét hangsúlyozza.

Ez ráadásul idődimenzióval is bír: elképzelhető az álláspont úgy is, hogy egykoron társadalmi választás tárgya volt a fosszilis fűtőanyagok vagy éppen a légkondicionáló felhasználása, ám ez a választási lehetőség idővel bezárult, és egy későbbi generáció pedig a túléléséhez igényli az adott technológiát, például mert egy olyan helyen él, ahol a hőség összeegyeztethetetlenné vált az emberi élettel klíma nélkül.

A kérdés tétje nyilvánvalóan óriási, hiszen amennyiben léteznek olyan technológiák, amelyek döntően befolyásolják a társadalmi folyamatokat, ugyanakkor már túljutottak

a kontrollálhatóság pontján, akkor az emberiségnek egészen más stratégiákra van szüksége, mint olyan technológiák esetén, amelyeket még befolyásolni tud.

Ebben a fejezetben arra szeretném felhívni a figyelmet, hogy annak a szakasznak az üvegházhatású gázok kibocsátása történetében, ami a megismételhetetlen, visszafordíthatatlan és determináló hatásaiban korrigálhatatlan ipari forradalom volt, az MI területén is lehetnek analógiái, például a tanulóadatként használható online szövegfelhalmozás vagy a gépi tanulásra használt stack területén.

5.2 A társadalmi kontroll példái

A technológiai determinizmusnak számos előterjesztett példája van. Az egyik legismertebb ezek közül a könyvnyomtatás feltalálásának hatása az európai vallási élet dinamikájára. A történet szerint a könyvnyomtatás a reformáció folyamatát idézte elő. További elterjedten használt példák: a kengyel és a feudalizmus kapcsolata, illetve a puskapor és az európai koloniális terjeszkedés. Ezek a példák megmérettettek a történelemtudomány által és túl könnyűnek találtattak. A vád velük szemben az, hogy negligálják a társas-gazdasági-politikai kontextus szerepét, amelynek legalább akkora, ha nem még nagyobb szerepe volt a társadalmi átalakulásban, ráadásul eleve a megfelelő társadalmi kontextus tette lehetővé a találmányok létrejöttét. Márpedig a technológiai determinizmus nem a társadalom és a technológia egyenértékűségét és kettős spirál szerű ko-evolúcióját, hanem a technológia dominanciáját hirdeti.

A másik probléma a leegyszerűsített determinista történelemértelmezéssel az, hogy nem számolnak be egyformán a jelenség ellenpéldáiról, hanem csak alátámasztó eseteket keresnek, ezáltal megsértik a szimmetria-elvet.[105] A szimmetria-elv egy ismeretelméletből, azon belül is az úgynevezett erős programból átvett maxima, amely pártfogói szerint az emberi megismerés minden területére alkalmazható.

A puskapor szerepe Európa felemelkedésében nagyon jó táptalaja a szimmetria-elv alkalmazásának. Egy felületes megközelítésben könnyű elképzelni a pillanatot, amikor az európaiak megjelennek az amerikai partoknál és a technológiai fejlettségben, különösen a fegyverek terén összemérhetetlenül erősebbek az őslakos amerikaiaknál, így tízszeres, hússzoros túlerővel is elbírnak. Csakhogy a puskapor számos más

[105] Bloor (2005).

kultúrában is megjelent, helyenként jóval korábban, mint Európában, és ezekben az esetekben mégsem eredményezett hasonló birodalmakat.[106] Közismert, hogy a puskaport Kínában találták fel, majd ezután Japánban, az ottomán török birodalomban, Oroszországban, Indiában és másutt is bevetették. Mégis, csak az európai kultúra táptalaján vezetett a gyarmatosítás folyamatához. Ha egy körülmény számos esetben fennáll, azonban a vizsgált jelenség csak az egyik esetben jelenik meg, akkor az oksági heurisztikák szerint legfeljebb szükséges feltételnek tekinthetjük az adott tényezőt, de kiváltó oknak semmiképp. A konkrét esetben tehát arra a következtetésre jutunk, hogy a puskapor mellett más, feltehetőleg társas faktorok vezettek a társadalmi változáshoz.

Ha pedig elvetjük a monokauzális világképet – azt a naiv ideát, hogy egyetlen eseménynek egyetlen döntő oka van, akkor a technológiai determinimzus szisztematikus kihívás elé néz, hiszen mindig lesznek releváns társas körülmények, amelyeket figyelembe kell vennünk.

Ezek közé tartozik a vallás és az ideológiák szerepe, amelyek elősegítik vagy hátráltatják, illetve bizonyos célokra csatornázzák a technológiai lehetőségeket. Olyan gazdasági tényezők, mint a bérek színvonala ugyancsak hatással vannak a technológiai fejlődésre, a megfizethető munkaerő hiánya például elősegíti az automatizálást és más munkaerő-megtakarító befektetéseket. A háborúra is gyakran a technológiai változást kiprovokáló társadalmi helyzetként hivatkoznak, tagadhatatlanul meggyőző példákkal, mint például az nukleáris technológia története.

Továbbá a technológiai determinizmusnak nem pusztán a társadalmi faktorokkal kell osztoznia oksági erőben, hanem a determinizmus más formáival is. Például a geográfiai determinizmus, amely azt állítja, hogy a térképen látható távolságok, partvonalak, természeti erőforrások döntő szereppel bírnak egy társadalom fejlődésében, ugyancsak ide tartozik. Ezen elmélet szerint például a Húsvét-szigetek társadalma pusztán a sziget méretéből és elszigeteltségéből adódóan bizonyos trajektóriára kényszerült, vagy hogy az európaiak a közel-keleti kereskedelmi útvonal elvesztése miatt, és a kontinens elhelyezkedése folytán kényszerültek a hosszú távú hajózás tökéletesítésére. A brit gőzgép-ipar pedig annak köszönheti felemelkedését, hogy specifikus geográfiai viszonyok miatt, ahol szén volt, ott a tárnákat elönteni képes víz is jelen volt, amihez pedig szivattyúkat kellett kifejleszteni.

[106] Hoffman (2015).

A sokfajta, külön-külön meglehetősen meggyőző determináló faktor miatt szinte minden gondolkodó kompromisszumos, többváltozós modellben gondolkodik. Az egyik ilyen modell a „puha” technológiai determinizmus[107], amely a technológiát bizonyos társadalmi változások szükséges, de nem elégséges feltételeként azonosítja. Egy másik, lényegében nem különböző megközelítés az aluldeterminánltságra való hivatkozás, például Feenbergnél. Mindkét megoldás a monokauzális modell elvetése, egyben a társadalmi faktoroknak állandó elvi lehetőséget biztosító megközelítés, amely lehetőséget teremt arra, hogy a passzív és a techno-politikailag tudatos társadalmak közötti különbségtételről beszélhessünk.

5.3 Demokrácia vs. Racionalizáció

A technológiai determinizmus kapcsán gyakran felvetett, a technológiai fejlődés miatt veszélyben forgó érték a demokrácia, a szabadság egyik kifejeződése. Mivel a technológia a determinista értelmezés szerint a döntés valódi lehetőségétől fosztja meg a társadalmat, így feltehetőleg egyben anti-demokratikus is. Ez a lehetőség kellően aggasztó ahhoz, hogy a technológia-szociológusok kiemelt figyelmét élvezze.

Az egyik fontos program a témakörben az Andrew Feenberg-féle *demokratikus racionalizáció*[108], amely a két kifejezés közötti látszólagos ellentmondást tagadja.

A racionalizáció egy olyan változási folyamat, amelynek a vezérlőelve az, hogy mi hatékony és célszerű. Márpedig számos ipari példában, például egy gyárban vagy egy gép tervezésénél a központi vezetés és hierarchikus kivitelezés tűnik a célszerűnek – senki nem gondolja, hogy a gyártási folyamat, vagy például egy repülő dekompozíciója többségi szavazással érdemben menedzselhető. Figyelembe kell viszont venni a technológia sajátosságait, amelyeket erősen limitálnak a természettudomány törvényei: csupa olyan dolog, ami nem esik a társas kontroll hatókörébe.

Így ezután a demokratikus racionalizáció önellentmondásnak tűnik. Feenberg a keretrendszere építésekor elismeri, hogy amennyiben a technológiai fejlődést a társadalom passzívan szemléli, akkor valóban anti-demokratikus tendenciái fognak kibontakozni. Például azáltal, hogy egyre több és több társas interakció lesz a

[107] Soft Technological Determinism - Dusek (2006), Heilbroner (1967).
[108] Feenberg (2009).

technológia által mediált, a kontroll egyre inkább kikerül az egyén kezéből. Amit Feenberg tagad, az a folyamat szükségszerűsége. Amellett érvel, hogy ezekbe a technológia által mediált szférákban – akár még egy gyárban is – tudatosan kell visszavezetni a demokratikus kontrollt, mert a rendszer magára hagyva ezzel a tulajdonsággal nem rendelkezik.

Az aktív és tudatos demokratizálásnak azonban kulcsszerepe van és az a társadalom passzív, a kényelmet biztosító racionalizáló tendenciái elleni erőfeszítést igényel. Amennyiben az erőfeszítés nem történik meg időben, a kontroll lehetősége örökre elillanhat. Ezen az alapon Stump[109] Feenberg keretrendszerét továbbra is techno-esszencialistának, azaz a technológiát domináns átalakító erőként elismerőnek tartja.

Bár Feenberg maga sohasem használja az együttes okozás terminológiáját, a mondandójából egyértelműen kiderül, hogy valójában mégiscsak erről van szó. A társadalmi változás kapcsán több kiváltó okot azonosít, továbbá felteszi, hogy a technológia is többféleképp megalkotható, illetve a használatban is találunk mozgásteret. Ez a mozgástér teremti meg a demokrácia lehetőségét, és egyben ellensúlyozza a technológia saját tendenciáinak maradéktalan érvényesülését.

Feenberg eszköze a keretrendszere megerősítésére a technológia története, különös fókusszal arra, hogy lineáris-e egy adott technológia fejlődése. Számos példát talál arra, hogy a lineáris feltalálás és elterjedés történetek valójában csak illúziók.

Az a felismerés, hogy a technológia valójában sokkal jobban kontrollálható, mint ahogy elsőre tűnhet, a technológiai determinizmus egyik vélt következményével is szembeszegül. Ezek a nehéz szívvel meghozott kompromisszumok, vagy „trade-off” szituációk. Ugyanis a determinista keretben sem szűnik meg teljesen a választás lehetősége: továbbra is lehetséges lesz a technológiát egyben elutasítani vagy elfogadni. A probléma az, hogy a determinizmus alapján a részletek nem módosíthatók: vagy használjuk a technológiát és együtt élünk a mellékhatásokkal, vagy elutasítjuk és együtt élünk azzal a szituációval, amit a technológia hiánya okoz: ez általában versenyképességi hátrány, elszalasztott lehetőségek. Tehát a technológia a determinista keretben egy külső, erősen befolyásoló faktor lesz a társadalom számára,

[109] Stump (2006).

amelynek pusztán annyi szabadsága marad, hogy kialakítsa a viszonyát a technológiához.

Ezzel ellentétes Feenberg szociálkonstruktivista felfogása. Itt a technológia nem egy kérlelhetetlen külső erő, adott előnyökkel és hátrányokkal. A technológia fekete doboza kinyitható és igényeinkre szabható. Ezt a logikai lehetőséget az aluldeterminált ság posztulálása nyitja meg: a matematika, a természeti törvények és a gazdasági törvényszerűségek nem határozzák meg teljes egészében a technológiai fejlődés trajektóriáját, és ezt kihasználhatjuk a társadalmi kontroll gyakorlására. Feenberg a technológiához való viszonyunkat „interpretációnak" nevezi, amely azonban nem egy egyirányú folyamat: nem puszta értelmezésről, hanem a technológia következő generációjára visszaható aktív átalakításról is szó van.

A nyelvészet fogalomtárából nem egyedül az interpretáció definícióját veszi át. A konstruktivista nézet tartalmaz olyan elemeket is, amelyek a beszédaktusokra emlékeztetnek. Feenberg „társadalmi szabályok"[110] érvelése szerint a társadalom úgy navigál a technológia aluldeterminált sága által keltett mozgástérben, hogy kizárja a számára negatív kimeneteleket, és ehhez „kijelenti", hogy milyen megoldások elfogadhatatlanok – például biztonsági öv nélküli autók tervezése– és ezáltal keretek közé szorítja a technológiát. A „kijelentés" megvalósulása lehet törvényhozás, tudatos vásárlás, illetve nem-vásárlás és más társas jelenség is.

Egy harmadik, nyelvészeti – pontosabban nyelv-filozófiai eszköz a helyzet elemzésére – a hermeneutika, amelynek alkalmazása a technológia hermeneutikája. Ez Feenberg szerint az a többkörös interpretációs folyamat, ahogyan egy társasalom egy új technológiát adaptál a saját képére. Ennek a folyamatnak a kontextusában egy technológiai objektumnak két hermeneutikai dimenziója van. A társas jelentés, amelyet akár kinyilatkoztatásnak vagy érvelésnek is tekinthetünk a felhasználók részéről, azt határozza meg, hogy milyen szerepet tölt be a technológia a felhasználóik életstílusában és státuszának kifejezésében. Ez szembe megy az uralkodó funkcionalista felfogással, amely alapvetően dekontextualizáló és amely azt venné legnagyobb hangsúllyal figyelembe, hogy mire jó az objektum, ami nem feltétlenül van átfedésben azzal, hogy mire használják, illetve nem feltétlenül magyarázza, hogy miért fogadják el.

[110] Social code – a „code" itt szabály értelemben áll, mint pl. „code of conduct".

Annak a gyanúja, hogy az MI-t a vállalat egy érték, például a „haladás” kommunikációjáért alkalmazza, nem pedig a funkcionalitása kedvéért, nagyon is gyakran felmerül. Ez is hozzájárul a hullámvölgyekhez az MI percepciójában: a kommunikációs igények gyorsan megfordulhatnak.

Feenberg az objektumok re-kontextualizálásával és társas közegükbe helyezésével a technológiai determinizmus egyik rejtett előfeltevését bontja le: azt, hogy a racionalitás kultúrafüggetlen lenne. Márpedig, ha a racionalitás nem kultúrafüggetlen, akkor lehetetlen, hogy a technológiai fejlődést irányító racionalitás egy társadalmon kívüli kényszerítő erővel bíró tényező legyen. Ebből kifolyólag a technológiai fejlődés trajektóriája sem pusztán a matematikai elvekből és a természeti törvényekből adódik. Ha a racionalitás a társas kontextus függvénye, akkor a demokratikus racionalizáció valóságos opció.

A racionalitás kapcsán természetesen csak a technológia kontextusában beszélünk, hanem kulturális attitűdök és értékek szélesebb rendszerében. Feenberg ezt „kulturális horizontnak” hívja, ami a technológia másodlagos hermeneutikus dimenziója. A „másodlagos” szó arra utal, hogy ezek az értékek és attitűdök jellemzően a háttérben maradnak, azokra a napi élet során nem reflektálunk folyamatosan.

A kulturális horizont megteremti a hatalomgyakorlás egy rejtett formájának lehetőségét. Mivel a horizont elemei reflektálatlanok maradnak, ám eközben kulturálisan elfogadottak, azok a társadalom tagjai számára adottságként, és nem megkérdőjelezhető attitűdökként manifesztálódnak.

Ez persze a nyílt hatalomgyakorlástól, például a katonai megszállástól, autoriter rendszerektől nagyon eltér, és az ellenszere is teljesen más.

Itt jön Feenberg felismerésének a lényege: a technokratikus racionalizáció egy kulturális horizont, amely a technológia modern társadalmak felett gyakorolt hegemóniáját támogatja. A társadalom technológia köré való szervezése, illetve a haladás trajektóriája nem szükségszerű, még ha annak is tartják.

A konklúzió mindezekből az, hogy a technológiai objektumok ontológiája kapcsán nem kaphatunk teljes képet, ha csak az egyik aspektusára– a funkcionalitásra vagy a hatékonyságra– koncentrálunk. Ehelyett, a társas jelentésre és preferenciákra is

figyelmet fordítva kibontakozik a technológia „kettős aspekusa”, és így már nem tekinthető a technológia a társadalmat külső erőként formáló tényezőnek, a demokratikus racionalizáció[111] kifejezésben feszülő látszólagos ellentét pedig illúziónak bizonyul.

Kérdés, hogy a társadalmi kontroll lehetősége hol ér véget – erről szól a következő alfejezet.

5.4 A technológiába zártság

Szinte minden új technológia megjelenése kapcsán – ha az adott technológia kellően jelentős és hatékony – megszólalnak a vészharangok. Nem teljesen tisztázott, hogy ilyenkor mi a gyújtó szikra. Például a mesterséges intelligencia az elmúlt 20-25 évben kezdett óriási áttöréseket elérni, de csak az elmúlt 5-10 évben kezdődött meg a széles érdeklődésre számot tartó szabályozási és etikai kodifikálási hullám a technológia körül.

Míg korábban mindezzel egy maroknyi filozófus foglalkozott csupán, a 2010-es évek közepétől az EU, az USA, az UNESCO, valamint az IEEE is komoly projekteket indított az MI társadalmi felelősséggel átitatott, morális fejlesztésének és használatának az elősegítésére.[112] Mindez az MI társadalmi kontrolljára irányuló világos erőfeszítésként értékelhető.

Hasonló helyzetben van a közösségi média és az internet is, amelyek ugyancsak fenyegető és ezáltal kontrollálandó technológiai komplexumoknak tűnnek.

Az, hogy a technológia társadalmi kontrolljának vágya nem a technológia fejlettségének függvénye, feltétlenül figyelmet érdemel. Úgy tűnik, hogy a technológiának el kell érnie egy kellően magas fejlettségi szintet, látható tényezővé kell válnia ahhoz, hogy egy olykor morális pánikhoz hasonló társadalmi reakciót váltson ki, ami szabályozási törekvésekben folytatódik. Ezt a klímaváltozást okozó technológiákra irányuló figyelem időbeli lefutásában is tetten érhetjük.

[111] Feenberg (2003).
[112] Lásd a 6. fejezetet, valamint Héder (2020b).

Az időbeli lefutás kapcsán a legfontosabb kérdés, hogy van-e olyan pillanat, ahol a kontroll már alig, vagy egyáltalán nem lehetséges. Például Feenberg optimista olvasatában a jövőre nézve ki tudjuk fejezni a technológiai választásainkban a társas preferenciáinkat, de mi a helyzet akkor, ha a lehetőségeinket már beszűkítették a korábbi generációk? Mi garantálja, hogy mindig rendelkezésre fognak állni preferálható, kulturális értékeinknek megfelelő opciók?

Ha a technológiai választások egy része lényegében irreverzibilis – márpedig ez aligha tagadható, hiszen egy nukleáris világháború vagy az elszabadult üvegházhatás kapcsán természettudományos bizonyítékaink vannak arra, hogy ez a helyzet – az nyilvánvalóan a technológia társas kontrolljának a lehetőségét csökkenti.

Például, ha a bolygónak egy olyan részén élünk, ahol az átlaghőmérséklet már meghaladja a biológiailag elviselhető szintet, akkor az ezt kompenzáló technológia használata nem választás kérdése, és talán még a típusa sem – a szabadságunk nominálissá válik, ekvivalens, de kissé eltérő formatervű termékek közül választhatunk, de működési elvek között alig.

Ez a gondolatmenet elvezet egy technológia-menedzsmentben már jól ismert fogalom globális kiterjesztéséhez. A fogalom a technológiába zártság, és eredetileg annak a negatív állapotnak a kifejezésére használták, amikor egy cég vagy szervezet túlságosan függ egy technológiától, platformtól vagy beszállítótól, így kiszolgáltatja, „bezárja” magát.

A fenti fejezetek lehetővé teszik, hogy precízen kifejezzük, mit jelent a *technológiába zártság*: 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.

Számos oka lehet a helyzet kialakulásának. David Collingridge[113] a technológiai összekapcsoltságban látja a probléma gyökerét a visszafordíthatatlan beruházás nevű gazdálkodástudományi jelenség mellett. Ezt kibontja Cowan és Hultén[114] számos új faktorral, például azzal, hogy egy adott technológia lecserélése, még ha van is

[113] David (1985).
[114] Cowan és Hultén (1996).

alternatíva, szükségképpen válságokkal, szabályozással jár, amelynek számos ellenzője lesz, és a nem-cselekvést támogató összes gazdasági és pszichológiai tényező aktiválódik és a status quo fennmaradását erősíti. Foxon[115] ezt az álláspontot úgy erősíti meg, hogy kiterjeszti az intézményesülés tehetetlenségével, valamint egy episztemikus dimenzióval: a tudás megcsontosulásával oly módon, hogy az állapot elfogadottá, természetessé válik, a kezdeti ellenérvek pedig kikopnak az emlékezetből.

A visszafordíthatatlan változások lehetősége logikai előfeltétele a technológiába zártság lokális vagy globális előfordulásának, ezért a visszafordíthatatlanság kutatása prioritást élvez.

5.5 Az MI konstruktőr általi bezáródása

> *… az igazi veszély … az, hogy ilyen [kibernetikai] gépeket, bár önmagukban gyámoltalanok, arra fognak használni egyének vagy csoportok, hogy az emberiség egészét kontrollálják”*[116]

Az előző alfejezetekben megpróbáltam árnyalni a technológiai determinizmus, mint egy lehetséges világállapot leírást és leágaztattam belőle a technológiába zártság fogalmát, bemutatva annak alapirodalmát is.

A fő különbséget úgy ragadhatjuk meg a kettő között, hogy a technológiai bezártság állapotának előállásához nem kell feltételeznünk a technológiai fejlődés autonómiáját, vagy egyéb olyan inherens tulajdonságait, amivel a társadalom nem tud semmit kezdeni.

Elegendő az, hogy bizonyos technológiai változtatások nehezen, vagy egyáltalán nem visszafordíthatók. A legegyszerűbb példa erre a klímaváltozás volt. A természettudomány által ismert törvények és a műszaki tudomány által feltárt működési elvek valószínűleg lehetővé tettek volna a társadalom számára olyan globális technológiai fejlődési utakat is, amelyek esetleg ha lassabb haladás árán is, de elkerülik a jelenlegi üvegházhatást okozó gáz koncentrációt a légkörben[117].

---

115 Foxon (2014).

116 Wiener (1950, 212).

117 El kell ismernem, hogy alternatív technológiafejlődési utak részben a tudományos-technológiai fantasztikus gondolkodás körébe tartoznak, minél régebbre megyünk vissza, annál inkább. De az érv szempontjából az is elegendő, ha csak a 20. század második felében, adott nemzetállamokon belül

A technológia *konstruktőrei* számára rendelkezésre álló szabadság azonban nem arányos az utókornak jutó szabadsággal. Ha kellően nagy az időbeli távolság a két csoport között, akkor ezt egy intergenerációs hatalmi viszonyként is felfoghatjuk, ahol az idő és az oksági láncok egyirányú tulajdonsága miatt mindig a korábbiak tudják a későbbiek körülményeit befolyásolni.

A most felnövő és fiatal felnőtt generációkra mindenképp igaz, hogy visszafordíthatatlan klímaváltozás-trajektóriába zártak: az, hogy ez a trajektória nem volt predeterminált, nem lényeges a bezártság jelenlegi állapota szempontjából. Természetesen ezek a generációk is hasonló egyirányú hatalmi viszonyban állnak az utánuk következőkkel, így felelősségük semmivel sem kisebb.

Tézisem szerint a klímaváltozáshoz fent leírt helyzetéhez hasonló, de sokkal rövidebb idősíkon operáló szerkezete lehet az MI fejlődésének is. A mesterséges intelligenciát több tényező is különösen alkalmassá teszi a technológiai bezártság előidézésére, ezáltal visszafordíthatatlan változások okozására[118].

Az MI egyik nagy alosztálya képes a hétköznapi számítógép-architektúrákon működni, azaz ugyolyan mint bármilyen egyéb szoftver. A szoftverek egyik jellemzője, hogy a sokszorosítás költsége nagyon alacsony az első előállításhoz képest – egyszer megírjuk a szoftvert nagy munkával és szinte végtelenszer másoljuk szinte nulla költséggel[119]. Így azután a kész szoftver – ha megfelel a felhasználói elvárásoknak a funkciók, az esztétika és a sebesség terén, valamint a felhasználásának jogi és anyagi feltételei is elfogadhatók – nagyon gyorsan el tud terjedni.

Amennyiben a szoftverek példányain haszon realizálható, a licenctulajdonos hatalmas bevételre tehet szert. Ez magyarázza, hogy a 20-ik század vége óta az anyagi értelemben vett társadalmi felemelkedés igen sikeres formája az infokommunikációs cég alapítása és sikerre vitele: lásd a világ leggazdagabb embereinek listáját.

megvalósult trajektóriákat komparatívan vizsgáljuk. Például, a francia EDF áramszolgáltató techno-politikai döntése a nukleáris energia nagy arányú alkalmazására könnyen összevethető a szomszédos országok stratégiájával: lehetséges volt tehát mindkét trajektóriát választani, a különbség megértésére pedig ez esetben az STS szociológiai megközelítése a legalkalmasabb módszer.

118 Héder (2021).

119 Ugyanezt a kérdést a tervezés kontextusában 2.3.1-es fejezetet tárgyalta.

Ráadásul a hardver cserélhető a szoftver és adat változatlanul hagyása mellett, így az nem kopik el, nem amortizálódik. Tehát ha egy feladattípust sikeresen és a felhasználók számára kedvező használati feltételekkel megoldanak az MI fejlesztők, akkor nem marad ok és működő üzleti modell újabb megoldások kifejlesztésére a konkurencia számára. Általában a szoftvernek ez a tulajdonsága nem tökéletes monopóliumot okoz, helyette létrejön maroknyi kommerciális és számos nyílt forráskódú alternatíva. Azt, hogy ez a helyzet fennáll, az operációs rendszerek piaci szereplőinek nagyon alacsony száma (különösen ahhoz képest, hogy hány milliárd példány fut) jól alátámasztja, de megnézhetjük a szövegszerkesztőket, a webböngészőket is. A mintázat ugyanaz: néhány fizetős, de emiatt jó támogatással rendelkező alternatíva mellett választhatunk igyenes és szabad megoldást, olykor kompromisszumokkal. Ezzel az ipar lefedte a felhasználói igények univerzumát, úgy, hogy a körülbelül 6 milliárd felhasználó 5-10-féle megoldásnál nem használ többet egy feladattípusra.

Így előállhat a konstruktőr általi technológiába zártság: ha szoftver már eléggé elterjedt, akkor bármely jelentős átírása több erőfeszítést igényelne, mint amennyit az aktuális felhasználói generáció finanszírozni hajlandó.

Ráadásul a Szoftver-mint-Szolgáltatás (SaaS, vagy „felhő”), azaz az interneten közvetített szoftver funkcionalitás tovább növeli a bezárási potenciált. Ebben az esetben az elterjedtséget már nem példányok, hanem felhasználók számával és főleg az felhalmozott adatok mennyiségével jellemezhetjük. Ez az adat-akkumuláció, például az az állapot, hogy a világon valaha megírt szinte összes email szövegéhez való hozzáférésen néhány nagy cég osztozik.[120]

Végül, a harmadik tényező, amely a bezárást elősegíti az adat felhalmozódása. A felhő-szolgáltatásként értékesített mesterséges intelligencia egyedülálló potenciállal rendelkezik, mert ott aggregálódnak az adatok, amelyek a gépi tanulás üzemanyagául szolgálnak.

Az MI szoftver és szolgáltatás könnyű replikálhatósága az adatok akkumulációjából adódó előnyökkel együtt meggyőző érvek amellett, hogy az MI valóban hajlamos

[120] Ne feledjük: ahhoz, hogy egy email a domináns világcégek diszkjeit elkerülje, a feladónak és a címzettnek is független email szolgáltatást kell használnia.

konstruktőr általi bezártság előidézésére. Talán a bezártság mértéke kevésbé erős a légkör állapotába és az ebből fakadó klímatrajektóriába zártsághoz képest. Még fontosabb, hogy az MI esetén élesen meg kell különböztetnünk a feladatosztályokat: például a nagy nyelvi modellek az internetre valaha felkerült szövegek igen jelentős részén tanulnak, és ez az adatfelhalmozási folyamat az emberiség történetében még egyszer már nem megismételhető[121]. Ugyanakkor a szakértői rendszerek szabályhalmazai könnyen cserélhetők.

A korábbi alfejezetek segítségével a konstruktőr általi bezártságot már nem csak az adatok vagy a felhasználószám mentén, hanem az MI felépítésének kapcsán is elemezhetjük.

Míg például a nagy email szolgáltatók sikerét betudhatjuk a fogyasztói viselkedésnek, a kényelmes interfésszel kapcsolatos elvárásoknak és az ezekre adott szolgáltatói reakcióknak – tehát összességében a szociológia eszközeivel leírható bezáródási folyamatnak-, addig az MI szokásos dekompozícióinak (lásd 2.4-es fejezet) megszilárdulását jóval nehezebb így jellemezni.

Azt, hogy a konstruktőr, vagy a nagyközönség szempontjából a vállalatok szűk csoportja általi bezártság nagyon is valós lehetőség az MI esetén, a fejlett világ társadalma felismerni látszik – erre a *kulcsfontosságú kihívásra* való reakcióként érthetjük meg a felfokozott szabályozási kedvet, ami a következő fejezet témája.

[121] Azt várhatjuk, hogy a nagy nyelvi modellek korszakában keletkezett szövegek jelentős része MI segítségével készül majd, és nem lesz feltétlenül világos, hogy egy adott szöveg átesett-e humán általi ellenőrzésen, és ha igen, milyen mértékben. Ez a bizonytalanság kérdésessé teszi az értékét tanulóadatként és felértékeli a korábbi szövegek értékét, mert azok garanáltan nem géppel generáltak. Elképzelhető lenne talán egy olyan tézis megalkotása is, amely szerint nyelvstílusokba, képalkotó MI esetén vizuális stílusokba való bezáródás történik épp most.

II. rész

6. Az MI szabályozása

Az I. fejezet bemutatta az MI természetét és tervezhetőségi korlátait – ezekből levezethető az is, hogy miért ütközik különös nehézségekbe az MI etikai szabályozása – a technológia szabályozásával kapcsolatos, szokásos nehézségeken felül.

Ahogy a mesterséges intelligenciával foglalkozó iparág jelentősége az áttörések miatt egyre nő: az MI telek[122] által többször megszakított fejlődési periódusok után a mesterséges intelligencia fejlesztéséből eredő erkölcsi aggályok újra egyre több ajánlást, etikai kódexet és törvényi erejű szabályozást szülnek. Ezek első megközelítésben a technológiára alkalmazott etika példáiként értelmezhetők, ám mélyebbre ásva már nem kerülhetjük el az I. részben bemutatott ismeretelméleti és metafizikai kérdéseket sem. Ezen a normatív manifesztumok (ajánlás, kódex, törvény stb.) összességére MI szabályként (MISZ) fogok hivatkozni.

Ez a fejezet három kérdést vizsgál meg.[123] Az első, és az eddigi fejezetek által legkönnyebben alátámasztható az MISZ-ek létezésének indoklása. A MISZ-ek létrehozásának szükségessége és haszna első látásra magától értetődőnek tűnik, mégis olyan kérdéseket vet fel, melyeket nem hagyhatunk figyelmen kívül. E kérdés kifinomultabb formában így hangzik: *milyen MISZ-ek* képesek elérni a kitűzött célokat, és mindenekelőtt, mely metodikák bizonyulnak ígéretesnek?

A vizsgálatom második szempontja az MISZ-ek sajátosságaira irányul. Leegyszerűsítve, a kérdés az, hogy melyek azok az összetevők ezen irányelveken belül, amelyek nem feltétlenül az MI-hez kapcsolódnak, viszont bármilyen újszerű technológiával kapcsolatban relevánsak, és csak ennél fogva szerepelnek az MI kontextusában – ezzel szemben melyek azok a szempontok, amelyek kifejezetten az MI-vel kapcsolatban merülnek fel. Az effajta vizsgálat jelentősége abban rejlik, hogy segít MISZ-eket helyesen fókuszálni, így azoknak kevesebb fronton kell csatákat nyerniük.

Végül, a harmadik kérdés az, hogy a bioetika, nano-etika, információ-etika és társaik utáni legújabb témakör-specifikus területként miként illeszkednek az MISZ-ek az

[122] Lásd az 1. fejezetben, valamint Crevier (1993); Hendler (2008).
[123] Ez a fejezet a Héder (2021) cikk kibővített verziója.

alkalmazott etika világába. Melyek a hasonlóságok, és mit tanulhatunk a professzionális etika egyéb olyan területeitől, ahol mostanra több tapasztalattal rendelkezünk?

Ahhoz, hogy közelebbről megértsük a fejezet által vizsgált elsődleges források természetét, a vizsgálat tárgyát olyan dokumentumokként határozzuk meg, melyek a munka morális dimenzióira irányuló, előíró szándékkal készültek, az MI fejlesztői projektekben érintett szakemberek és döntéshozók számára.

A cél ily módon való leszűkítése után még mindig egy feldolgozhatatlan dokumentum-mennyiséggel állunk szemben. Ezért, egy második szűrőként szolgál e munkák potenciális hatóköre.

A politikai egységek (EU, USA, Kína) és a szakmai szervezetek (IEEE, OECD, nagyvállalatok) által kiadott iránymutatásokra fókuszálok első sorban, mivel ezeknek ér el legmesszebbre a hatása. Végül, egy szerencsétlen, ugyanakkor szükséges korlátozás, hogy csak az angol, esetleg magyar nyelven is elérhető MISZ-ekkel foglalkozom. Nem célom az, hogy olyan széleskörű mennyiségi felmérést végezzek, mint a 84 forrásnál is többet feldolgozó, *Nature Machine Learning*-ben megjelent tanulmány, vagy a *Minds and Machines*-ben publikált, több mint 22 MISZ-szel foglalkozó írás.[124]

Ehelyett, azt próbálom megérteni, hogy mi számít racionális és következetes megközelítésnek egy MISZ megalkotása során, valamint, hogy egy ilyen törekvésnek milyen korlátokkal kell szembesülnie.

6.1 Fontos MI szabályozások

E fejezetben nyolc forrással foglalkozom, ezeket közvetlenül idézem, és közelebbről meg is vizsgálom.[125] Az e források kiválasztására vonatkozó általános alapelvek tehát: az írás legyen előíró jellegű, célközönsége a döntéshozók és szakemberek, foglalkozzon erkölcsi kérdésekkel, és rendelkezzen magas hatáspotenciállal. Ugyan a

[124] Jobin és társai (2019), illetve Hagendorff (2020a).

[125] OECD (2019), IEEE (2019), AI HLEG (2019), Microsoft (2019), Holdren és társai (2016),
Beijing AI Principles: https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/
Google AI principles: https://ai.google/responsibility/principles/
EU MI Rendelet: The European Union and the Council (2023).

potenciális hatás megítélése végül is nem egzakt tudomány, mégis, e nyolc dokumentumot megalapozott érvek alapján választottam ki.

Az OECD iránymutatás az OECD globális hatására való tekintettel, valamint azért került be, mert – legalábbis fogalmi szinten – kiterjed egy igen változatos, több országot és politikai entitást magában foglaló halmazra is. Az OECD korábbi, más szakterületeken kiadott irányelvei – mint például a Frascati Kézikönyv – sikeresen teremtettek közös nevezőt saját környezetükben, megfelelően koncentrált céltartomány és széles körű globális elfogadottság mellett.

Témakörök szempontjából kétségkívül az IEEE MI Tervezési Irányelvek a legátfogóbb írás. Hagendorff[126] jelentése szerint az általa azonosított, általános MISZ témakör legtöbbjére (18) kitér, míg az e rendszerben soron következő MISZ mindössze tizennéggyel foglalkozik. Másodszor, az iránymutatás jelentősége abban rejlik, hogy nyíltan az MI etikai kérdésekkel foglalkozó, soron következő, „P7000” IEEE szabványsorozat egyik alappillérének ígérkezik. Mivel az IEEE az infó-kommunikációs technológiai iparág egyik, ha nem a legfontosabb szabvány-forrásának számít, az MI-vel foglalkozó szakemberek kénytelenek előre megismerni és alkalmazni a P7000 sorozat tagjait, mint pl. a P7001-et,[127] ami az autonóm rendszerek átláthatóságát segítő, empirikusan tesztelhető, hiteles szabványnak ígérkezik.

Az EU Bizottság Magas Szintű AI szakértői csoportjának kiadványa és az EU MI rendelet a benne foglalt témák széles skálájának, és az általa képviselt nemzetközi (de EU-n belüli) összefogásnak köszönhetően bír nagy jelentőséggel, valamint azért is, mert a soron következő jogszabályi keretrendszer – vagy egyszerűbben, az MI-re vonatkozó törvények – alapját hivatott képezni. Hasonló indíttatástól vezérelve került be az USA kormányának vonatkozó jelentése is.

A *Pekingi MI Alapelvek*, bár távolról sem olyan részletes, mint a fenti három MISZ, leírja Kína hivatalos álláspontját az MI etikai kérdéseinek tekintetében, így nem lenne célszerű figyelmen kívül hagyni.

[126] Hagendorff (2020)
[127] Winfield és társai (2021).

A Microsoft és a Google MISZ-ei azért kerülhettek be, mert két igen jelentős iparági szereplő álláspontját képviselik. Ez nem azt jelenti, hogy e két vállalat az ágazat legmeghatározóbb résztvevője, de ők legalább publikálták saját irányelveiket.

A fentiek mindegyike, sok más MISZ-szel egyetemben már igen kimerítő elemzések tárgya volt. Ezek az elemzések rámutatnak, hogy a fent említett források milyen figyelemre méltó hasonlóságot mutatnak a kulcsproblémák azonosítása terén. Ugyan a szóhasználat eltér, mégis kirajzolódnak előttünk bizonyos kulcsgondolatok. Ezeket én a technológia társadalmi kontrolljára való igény specifikus területeinek tartom és sokkal inkább az MI percepciójából, mint a tényleges MI felépítéséből eredeztetem.

Az elemzett MISZ-ek elsődleges problémáját az *átláthatóság*, (időnként értelmezhetőséggel párosulva); az *részrehajlás mentesség és pártatlanság*, a *felelősség és elszámoltathatóság*; az *adatvédelem*; a *„jó”* (jótékonyság vagy a jólét megteremtése) *népszerűsítésére* való irányultság, és bizonyos MISZ-ek esetén az *emberi autonómia* fenntartása, valamint az ezzel (és az elszámoltathatósággal) kapcsolatos *emberi felügyelet* jelentik.

Ahogyan Hagendorff[128] megállapítja, számos hiányosságot is találhatunk a szabályozásokban. Nem elég, hogy az MISZ-ek hatásköre meglehetősen szűk, de az sem világos, valójában mennyi változást képesek előidézni, és milyen arányban tartják be ezeket (a gyakorlatban).

E fejezetben azonban egy lépéssel hátrébb kezdünk: ugyan a betartatás kérdése fontos, én munkahipotézisként feltételezem, hogy az MISZ-ek olvasói aktívan igyekeznek helyesen cselekedni, *és* készek alárendelni döntéseiket az előírásoknak, *valamint* hajlanak az erőforrások e szerint való felosztására is. Más szóval, megelőlegezem a legjobb szándékot és hozzáállást, mert a számomra fontos kérdés az, hogy ezen előfeltételek teljesülése esetén milyen kihívások maradnak a szabályalkotás terén, amelyek nem a szándékokkal, hanem az MI sajátosságaival kapcsolatosak.

Ahogyan később látni fogjuk, közel sem magától értetődő, hogy a fent kiemelt problémák mindegyike olyan, új keletű probléma lenne, ami kifejezetten csak az MI

[128] Hagendorff (2020a, 2020b).

kapcsán került előtérbe. Mielőtt azonban ezekkel a konkrét problémákkal foglalkoznék, röviden visszatérek a MISZ-alkotás mögötti motivációk kérdésére.

6.2 Mi szükség van irányelvekre?

Elsőre úgy tűnhet, hogy a MISZ-ek létrehozásának szükségességére irányuló egzisztenciális kérdés általában véve nem sokban tér el az technológiai előírások megalapozottságát firtató vitáktól.

Széles körben elterjedt álláspont, hogy az egyik első, külső szabályzás tárgyát képező technológia – mai értelemben vett pontos mértékegységekkel és sablonokkal – a gőzgépkazán volt.[129] Az erre vonatkozó, az Amerikai Gépészmérnöki Társaság (ASME) által 1884-ban összeállított előírás óriási mérföldkőnek számított, hiszen korábban példa nélkül való módon, 1907-ben minden műszaki részletével együtt törvényerőre emelkedett. Kijelenthetjük, hogy ez az eset a technológia felett gyakorolt társadalmi kontroll formáját egy egészen új szintre emelte.

De mi szükség van a technológia felett gyakorolt társadalmi kontrollra? A gőzgépek esetében ez elsősorban az okozott tragédiákra vezethető vissza. Az előírások életbe lépése előtt ezek felrobbanása általános volt, és egyetlen baleset áldozatainak száma egynéhány, vagy akár ezer fő között váltakozott. A technológiai katasztrófák a pénzhajhászás következményének tekinthetőek, hiszen a költségek a biztonság kárára, vagy puszta gondatlanságból lettek csökkentve.

A kazán-szabvány sikertörténetté vált, hiszen megálljt parancsolt a költségkímélés veszélyes formáinak, kikényszerítve a biztonságosabb tervezést és tesztelést.. Az ASME illetékesei valószínűleg abban a hitben vonulhattak nyugállományba, hogy a fáradságos bizottsági ülések és közreműködésük gyümölcse jónéhány életet megmentett. Az egészségügyi szakemberekkel ellentétben ők nem nézhettek közvetlenül az általuk megmentett személyek szemébe, de a statisztikákból értesülhettek arról, hogy ezrek mondhatnak nekik köszönetet. Mindezek mellett pedig a technológia előretörése sem torpant meg komolyabb mértékben, ahogy azt az előírás ellenzői előrevetítették a javaslat törvénybe iktatása előtt. Ez tehát az 5.3-as fejezetben leírt feenbergi vízió megvalósulása.

[129] Green (1953).

Az azóta eltelt idő alatt mind a törvény által kötelezővé tett, mind a mérnöki szakágak tekintélyei által előirányzott technológiai iránymutatások száma megsokszorozódott. Láthatólag kialakult egy általános vélekedés azzal kapcsolatban, hogy milyen nagy szerepet játszik mindennapi életünkben a technológia. A filozófusok ugyan vitatják a technológiától való függésünk valódi természetét (lásd az 5. fejezetet)[130], három dolog mindenképp vitán felül áll: 1) a technológia szerepvállalása mindennapi életünkben elképesztő méreteket öltött, 2) ennek hatásai nem feltétlenül pozitívok és 3) a technológiát *igenis* lehetséges kontrollálni. E három meggyőződés képezi az irányelvek lefektetésének előfeltételeit: a technológia irányítása egyszerre fontos és lehetséges. A második világháború óta a közvélemény sem érdektelen, amely mind civil aktivizmusban, mind politikai tettekben megnyilvánul. A haditechnikai tesztek, az agráriparban használt vegyszerek, a nagyvárosi épületek és egyéb anyagi problémák az elsők között kerültek szabályozásra, de a közösségi média regulációja sem váratott sokat magára. Ződi tanulmányában[131] a technológiával szemben felmerülő egyre erősebb szbályozási igényt egészen az ipari forradalom kezdetéig, a munkások védelmére hozott szabályokig vezeti vissza. Mára természetessé vált, hogy minden olyan új technológia valamilyen, nem kötelező erejű jogi eszköz, vagy valós jogszabály hatása alá esik, amely kellően kockázatosnak tűnik a társadalom percepciójában.

Az egyenlet másik oldalán a szabályozás feltételezett költségei állnak. Bárminemű szabályzás, még ha a legjobb szándékkal alkotják is meg, előre nem látott és nemkívánatos mellékhatásokkal járhat, melyek akár rosszabbak lehetnek, mint azok az események vagy helyzetek, melyek létrejöttét az előírás megakadályozni hivatott. Állami szereplők, vállalatok és individuumok is felhasználhatják ezeket arra, hogy csekély költségráfordítás árán fenntartsák az erényesség puszta *látszatát*[132], szükségtelen akadályokat gördíthetnek a piacra való belépés útjába, az ideológiai vagy politikai csatározások kiterjedt eszköztárának részeivé válhatnak. Egy másik bírálat szerint a szabályzás a „gúzsba kötés” egyszerű módja is lehet[133], melynek során egy adott csoport saját hasznának maximalizására törekszik a versenytársak akadályozása által, az innováció helyett a lobbizásba befektetve.

[130] Továbbá: Marcuse (1977); Gerrie (2008).
[131] Ződi (2024) p19: az 1802-es *The Health and Morals of Apprentices Act* már korai példának tekinthető. technikaszabályozásnak tekinthető.
[132] Hagendorff (2020a) valamint Vica és társai (2021).
[133] Posner (1974)

Emellett, ha azt feltételezzük, hogy egyes MI alkalmazások komoly, élet-halál jellegű előrelépést eredményeznek saját területükön – mint az az igen valószínű feltételezés, hogy az automata irányítású autóforgalom biztonságossága nagyságrendekkel meghaladja majd az emberi sofőrökét –, akkor a gyakorlati bevezetés késleltetésének ára talán majd emberéletekben lesz lemérhető.[134]

Továbbá, a technológia felett gyakorolt kontroll egy lényeges, elkerülhetetlen episztemikus kihívással néz szembe, melynek egyik megfogalmazása a Collingridge[135] dilemma. Ez nem más, mint hogy ha az előírások időben való megalkotására törekszünk – már a korai fejlesztési folyamat során – akkor nem lesz elég információnk és gyakorlati tapasztalatunk azzal kapcsolatban, hogy a technológián belül hová koncentráljuk erőfeszítéseinket. A későbbi fázisok során aztán rájövünk arra, hogy melyek a legfontosabb problémák, de ekkorra már túl késő, hiszen a kialakult, széles körben elterjedt technológiákon már nehéz változtatni.

6.3 Az absztrakció szintje

Érdemes megfigyelni, hogy a fent említett, kazán-szabvány és az azt követő korai előírások nem hagyatkoztak az etika és morálfilozófia terminológiájára. Léteztek ugyan etikai szabályok, de ezek meglehetősen általános jelleggel működtek, és igen tömörek voltak, mint az IEEE Etikai Kódex, amely egyfajta mérnöki becsületkódexnek tekinthető. Ahogyan haladunk a modern, alkalmazott etika irányába (pl. nano-, bio-, információs- és reprodukciós technológiákkal kapcsolatos etika), és elérkezünk a MISZ-ekhez, az a kép alakulhat ki bennünk, hogy az aktuális technológia egyre inkább a szabályozók előtt jár. Emiatt tevődött át a hangsúly a fejlesztési projektek közvetlen eredményeként születő végtermékek szabályzásáról, a maguknak a fejlesztőknek adott instrukciókra. Míg a jogi szövegekben a gőzkazánban uralkodó nyomás maximális értéke megadható volt pontos font/négyzethüvelyk értékekkel, addig a biotechnológia megjelenése óta jelenlevő összetettebb és gyorsabban változó technológiák esetén a stratégia inkább egyfajta önszabályozás lett., Ez a konkrétabb, adott problémára alkalmazható irányelvek létrehozását, valamint az etika szakértőinek a projektbe való bevonását jelentette. A professzionális etika és a technológiai előírások teljesen összefonódnak például a 2024. augusztus 1. óta érvényben lévő EU

[134] Lásd az 1.2-es 4.2-es alfejezeteket.
[135] Collingridge (1981).

MI rendeletben. Ez az összetettebb megközelítés azonban továbbra sem képes sem a Collingridge dilemma, sem a „gúzsba kötés” problémáját kiküszöbölni.

Továbbá, az MI esetén a járulékos komplexitás egy további, új szintje is felmerül. Az MI technológiai értelemben vett sajátos természete a létrehozott termékek magas intelligenciaszintjével elegyített példátlan autonómiájából fakad. Bizonyos mértékig úgy tekinthetjük, mintha az MI mesterséges személyek létrehozását célozná. És mivel az etikai kódex rendeltetése az adott személyek viselkedésének irányítása, ez bizonyos fennakadásokra és zűrzavarra ad okot. Az MISZ-ek feladata vajon a fejlesztők, vagy a mesterséges összetevő viselkedésének szabályzása? A MISZ-ek egyes normatív elemei valóban mind a fejlesztőkre, mind a mesterséges összetevőkre vonatkozó útmutatásként is értelmezhetők. Jó példa erre az előítéletek elkerülésére vonatkozó ajánlás (ami szinte minden MISZ részét képezi) – ez időnként az MI vonatkozásában szerepel, máskor a fejlesztőkre értendő, megint máskor pedig a célcsoport kiléte nem egyértelmű.

És amennyiben elfogadjuk a kézenfekvő lehetőséget, miszerint a MISZ-ek a gépek viselkedését szabályozzák – ez az úgynevezett gépetika („Machine Ethics[136]; lásd a 7. fejezetet) – továbbá azt is, hogy az MI valódi humán személyt helyettesít, legalábbis kvázi-személyként, akkor valójában nem hoztunk-e létre áttételesen magára az emberi viselkedésre új törvényeket? Például az MI rendelet egyik eleme a humánerőforrással kapcsolatos MI döntések körülményeinek, előfeltételeinek, részrehajlásmentességének erős szabályozása; de amint ez a szabály életbe lép, feltehetjük a kérdést, hogy a humán döntések miért is nem állnak hasonló szabályozás alatt?

Ezen felül, az AI összetevő természete speciális kihívást jelent az előírások harmadik előfeltételét (ld. fentebb) illetően. Az említett előfeltételezés lényege az a szinte triviális tény, hogy bármilyen, a szabályzásra tett erőfeszítés akkor tekinthető racionálisnak, ha a fennáll a technológiai kontroll elvi lehetősége. Az ipari jellegű MI lényege, hogy a mesterséges összetevőre bízza a különböző helyzetek irányításának fáradságos, intellektuális feladatait. Az önvezető autó kiváló példa erre. A kihívás abban rejlik, hogy nem akarjuk *pontosan* meghatározni, mit tegyen az MI, hiszen ez épp az MI létjogosultságát kérdőjelezné meg – ebben az értelemben a cél épp az irányítás átadása.

[136] Anderson és Anderson (2011).

Másrészről, általánosságban *igenis* irányítani akarjuk a helyzetet, hiszen a nemkívánatos, vagy kellemetlen végkimenetelt el akarjuk kerülni. És hogy a helyzetet tovább bonyolítsuk, a tervezési fázisban lehetetlen elmagyarázni vagy felsorolni a mesterséges összetevő által működés közben produkált összes nemkívánatos kimenetelt: más szóval, az MI nyílt[137] természetéből adódóan egyes nem kívánt eredmények már csak bekövetkezésük után ismerhetőek fel. Így, egy már-már paradox jellegű feszültség keletkezik a kontroll delegálására és megtartására irányuló igények között. Az MI feladata, hogy autonóm módon irányítsa működési környezetét, miközben a kontrollt mi sem szívesen adjuk ki a kezeink közül, épp az MI által okozott nem kívánt kimenetelek elkerülése érdekében – melyekre teljesen felkészülni előzetesen nem is lehet.

Mostanában gyakran fordul elő, hogy igen összetett K&F projektek során – például a bioetika vagy az orvosi kutatások területén – a szabályozás az érintett szakemberek kezében van, önszabályozás formájában, hiszen a ez a feladat ennél közvetettebb formában nem kivitelezhető, mert a szervezeten kívül nincs meg a kellő információ a technológia újdonsága miatt. Így azonban a szabályozás folyamatába egy újabb lépcsőfok kerül: az MISZ alkotás során az MI-vel foglalkozó szakember önszabályozásra kap megbízást arra vonatkozóan, mely döntések meghozatalát delegálja tovább a rendszer irányába, és milyen módon teszi ezt. Emiatt viszont az MISZ-ek meglehetősen absztrakt jellegűvé válnak – e tulajdonságot a fejezet egy későbbi részében részletesen megvizsgálom majd.

Végkövetkeztetésként megállapíthatjuk, hogy az MI etikai kérdései a teljes és alapos személy-viselkedés tervezés igénye miatt válnak olyan kivételes kihívássá, ami eltér a többi olyan újszerű és feltörekvő technológiáétól, mint a genetikailag módosított szervezetek, vagy a blokklánc. Mivel az emberi viselkedés és az etika (a helyes emberi viselkedésformák) állandó vita tárgyát képezi, talán nem árt, ha felkészülünk a gépetikát (a helyes gépi viselkedésformák) illető végtelen vitákra is.

Ez a gondolat azonban kiemeli, mennyire fontos megkülönböztetni egymástól az MI különböző formáit. Az MI egyes alkalmazási területein egy hagyományos sakk-algoritmushoz hasonló, kevésbé autonóm jellegű, kevésbé ágens-szerű MI is

[137] Tanulásra képes, akár önmaga viselkedését megváltoztató.

megfelelő, akár a gépesített tanulás lehetősége nélkül is. Ebben az esetben természetesen a kontroll átruházására vonatkozó fenti érvelés kevésbé releváns.

Az elmúlt 25 évben egyetlen ember sem volt képes az MI-n felülkerekedni sakkban és sok más alkalmazásban, a 1990-es években mégsem volt megfigyelhető robbanásszerű növekedés az MISZ-ek számát illetően. Ez arra enged következtetni, hogy az MI alkalmazások legfrissebb hullámával érkezett valami olyasmi, ami az MISZ projektek elszaporodását kiválthatta.

A jelenlegi MISZ-cunami talán az egyre autonómabb jellegű MI-knek a mindennapi élet minden területén érvényesülő jelenlétéhez köthető. Valamint, elképzelhető, hogy a mobil MI platformok elburjánzása (önvezető autó, robotporszívó, csetbot) néhány pszichológiai-észlelési korlát áttöréséhez vezetett, így egyes gépezeteket vezetői képességekkel rendelkező személyeknek tekintjük okos szoftver helyett. A következő alfejezet az egyes MISZ-ek mögött expliciten kinyilvánított motivációk vizsgálatával foglalkozik, hogy e kérdésre fényt deríthessünk.

6.4 Motivációk az MI irányelvei mögött

> *„Mivel az autonóm jellegű és intelligens rendszerek (A/IS) használata és hatása igen elterjedtté vált,* ***társadalmi*** *és* ***politikai*** *iránymutatásokra van szükség, hogy az ilyen rendszerek* ***emberközpontúak*** *maradhassanak, és az emberiség által képviselt értékeket és etikai alapelveket szolgálhassák.”* (IEEE: p2)[138]

Forrásainkat a motivációkat leíró szakaszok szempontjából átvizsgálva nyilvánvalóvá válik, hogy ezen irányvonalak szerzőinek egyöntetű előrejelzései alapján, az igen közeli jövőben az MI átfogó, mindent átható, megkerülhetetlen erővé válik majd társadalmunkban. A technológiai determinizmusól jól ismert „trade-off” (lásd az 5.3-as fejezetet) reprezentációnak megfelelően a technológia óriási pozitív változások előidézésére képes, ugyanakkor alapvető kockázatokat is hordoz, melyek közül a legfontosabb, hogy elveszíti „ember-központú” jellegét – ezt a kifejezést a fenti MISZ-ek közül több is tartalmazza.

[138] Kiemelések tőlem; jól látható ezeken a szabályozásokon a bevallott techno-politikai szándék; ez újdonság a korábbi szabályozásokhoz képest.

Tehát, az STS[139] szakirodalmi szóhasználatával élve, egy újabb friss technológia mutatkozik be súlyos trade-off-ok gyűjteményeként. E felfogásban a technológia Janus-arcú, előnyei a hátrányaitól teljesen nem elválaszthatók, de a megfelelő szabályzás vagy irányelvek segítségével tehetünk valamit a helyzet javítására. Ha szabályozatlan marad, akkor fennáll a veszély hogy munkerőpiaci rombolást végez, és veszélyezteti értékeinket:

> *„(...) Az MI kihívás elé is állítja társadalmainkat és gazdaságainkat, leginkább a gazdasági változások és egyenlőtlenségek, verseny, munkaerő-piaci átalakulások, valamint a demokráciát és emberi jogokat illető hatások tekintetében" (OECD).*

Szembe kell néznünk azzal, hogy ez a fajta karakterizáció, mely nyolc forrásunk közül hétben jelen van (a Microsoft Alapelvek nem tartalmaz indoklás fejezetet), rendkívül hasonlít a múltbéli, technológiai előírásokra vonatkozó vitákra, a gyermekmunkától kezdve, az eredeti kazán-kódexet illető vitákig.[140] De bármely más vitát is említhetnénk a kockázatos technológiákkal kapcsolatos STS témák közül. Ezek a tanulmányok kivétel nélkül azt mutatják, hogy a viták idején fennálló problémák és feltételezett kompromisszumok nem a később valóban megjelenő problémákat és lehetőségeket tükrözték.

Más szóval, az új technológiákkal kapcsolatos kockázatok felmérése és szabályzása során a vitatkozó felek egyszerűen nem voltak képesek az előttük álló időszak helyes felmérésére. Bár bizonyos kockázatok, előnyök, transzformációk később valóban megjelentek, ezek nem voltak összhangban a korábban elképzelt félelmekkel és víziókkal. Ez az eltérés természetesen az alkalmazott előírások eredménye is lehet, például, ha az aggodalomra okot adó kockázatot a szabályzás révén sikerült kiküszöbölni. A fenti tanulmányok azonban azt mutatják, hogy az előre elképzelt jövőkép és a valóság közötti különbség túlságosan nagymértékű ahhoz, hogy az előrejelzéseken alapuló közbeavatkozások hatásának számlájára írható legyen. Sokkal inkább úgy tűnik, hogy eredendően nehéz egy új technológia hatásait előre megjövendölni, és hogy a szabályzók legtöbbször csak a sötétben tapogatóznak.

[139] Science and Technology Studies – STS, lásd az 5-ik fejezetben.
[140] Ferguson, (1987), számos egyéb téma tekintetében ld. Johnson és Covello, (2012).

Mégis, minden egyes MISZ-ben megjelenik egyfajta sürgetés. Arra következtethetünk, hogy a szerzők amiatt aggódnak, hogy általános elterjedésüket követően a technológiák bezáródnak (lásd az 5.5 fejezetet) és megváltoztathatatlanná válnak. Például az EU irányelvekben, ami az MI rendelet előfutára:

> *„(…) A bennük rejlő óriási lehetőség mellett az MI rendszerek bizonyos kockázatokat is rejtenek magukban, melyeket megfelelően és hozzájuk illő módon kell kezelnünk. Jelenleg még egy fontos ablak áll nyitva előttünk, hogy fejlesztésüket megfelelő irányba tereljük”* (AI HLEG).

Más szóval, egy Collingridge dilemmához hasonló helyzet állt elő, abban az értelemben, hogy a legösszetettebb és legkiszámíthatatlanabb technológiák egyikét próbálják kontrollálni.

Összefoglalva, az az általános vélekedés – vagy akár nevezhetjük hype-nak is – hogy az MI iparág nagy előrelépés előtt áll, ezért azonnali hatállyal szükség van valamiféle szabályzásra vagy társadalmi kontrollra. Az a lehetőség fel sem merül, hogy az MI fejlesztése kudarcba fullad, különösen e felfokozott elvárások fényében.[141] Ez nem azt jelenti, hogy az MI fejlődése megállna, az viszont előfordulhat, hogy a mindenen átívelő, társadalomformáló változás, amit ezek az MISZ-ek a maguk módján előrevetítenek, még néhány generáción át várat magára, míg egyes más, hétköznapibb gyakorlatok, melyek egy része a büntetőjog határán egyensúlyoz[142], kikerült a figyelem középpontjából, pedig valójában ezek szabályzására lenne igazi szükség.

6.5 Az szabályozások MI-specifikussága

> *„A probléma-megelőzés alapelve:*
> *MI rendszerek nem lehetnek sem forrásai, sem közreműködői semmilyen emberi lénynek okozott kárnak, vagy ártalmas hatásnak.(…)”* (AI HLEG).

A MISZ-ekre vonatkozó fontos kritikai kérdés a specifikusság. Azt vizsgálom ebben az alfejezetben, melyek azok az MI-re jellemző speciális kihívások, amelyek külön irányelvek megalkotását indokolják, és mi ezek hatása más technológiákra? Ehhez nagy segítség a könyv 1. fejezetében bemutatott problématérkép.

[141] Floridi (2020).
[142] Hagendorff (2020a).

A fenti idézet kapcsán például szélsőséges következtetésre is juthatunk – hogy az MISZ-ek javaslatai valójában nem is MI-specifikusak – melyből logikusan az következne, hogy e javaslatokat egyszerűen az mérnöki etika szempontjai közé kellene sorolni, és kifejezetten MI-re vonatkozó irányvonalakra nincs szükség. Azt láthatjuk, hogy egyes MISZ-ek olyannyira ambiciózusak, hogy a meglévő iránymutatásokat figyelmen kívül hagyva igyekeznek minden problémára választ találni, aminek eredményeképp akár szabályozás-ellenes reakciók is létrejöhetnek.

Egy egyszerű, „vízforraló teszt" nevű módszert javaslok az egyes normatív elemek vizsgálatára. Az iránymutatások szövegében helyettesítsük be az „MI" helyére a „vízforraló" szót, és figyeljük meg, értelmes, és érvényes marad-e az állítás[143]. Ha a válasz igen, megállapíthatjuk, hogy az adott iránymutatás MI-specifikusság helyett technológia-agnosztikus.

> „MI ~~*rendszerek*~~ ***Vízforralók*** *nem lehetnek sem forrásai, sem közreműködői semmilyen emberi lénynek okozott kárnak, vagy ártalmas hatásnak.(…)"* (AI HLEG);
>
> *„Az MI ~~rendszereknek~~* ***Vízforralóknak*** *élettartamuk végéig meg kell őrizniük ép és biztonságos voltukat, hogy normál, rendeltetésszerű, vagy nem rendeltetésszerű használat, vagy akár kedvezőtlen körülmények között is megfelelően működjenek, és ne jelentsenek túlzott biztonsági kockázatot."* (OECD)
>
> *„A tervezőknek és a kezelőknek egyaránt igazolniuk kell az ~~AIS~~* ***vízforralók*** *hatékonyságát és a célnak való megfelelését."* (IEEE)

Amint látható, a fentiek mindegyike megbukik a teszten, hiszen vízforralók esetén ugyanúgy megállják a helyüket, mint MI-k esetén. Ez persze nem jelenti azt, hogy hibás megállapítások lennének, és betartásuk felesleges. Mindez csupán arra utal, hogy az MISZ-ek általános mérnöki irányvonalakat tartalmaznak, valamint felvetik a kérdést, vajon ezek újra-felfedezés helyett korábbi művekből is összeállíthatóak lennének-e. A fenti példák korántsem tartoznak a legáltalánosabb útmutatások közé. Ott az iránymutatások azon elemeit találjuk, melyek a törvényi előírások betartására

[143] Fordítások: HM, a könyv szerzője.

emlékeztetik a fejlesztőket, a kockázatokról tájékoztatják a vásárlókat, hangsúlyozzák a kezelők képzésének fontosságát.

Mégis, legyünk elnézőek értékelésünk során, és feltételezzük, hogy a vízforraló-tesztet sikeresen teljesítő „alapelvek" vagy „irányvonalak" beemelése bizonyos szempontból hasznos: így az MISZ minden területre kiterjedő, általános iránymutatást ad, egyetlen dokumentumban. Panaszra mindössze az adhat okot, hogy ahhoz viszont nem elég teljes körűek ezek a dokumentumok, hogy „egyablakos" techno-etika és mérnöki jó gyakorlat források legyenek: és nem térnek ki olyan hétköznapi követelményekre, mint a forráskódok megfelelő dokumentációja, és így tovább.

Létezik a javaslatok egy csoportja, ami ugyan átmegy a vízforraló-teszten, de az „információs rendszer" behelyettesítése azt mutatja, hogy nem MI, hanem általánosabb, infokommunikációs hatókörűek. Ezek tipikusan adatgyűjtésre és a személyes adatok védelmére irányulnak:

> „MI ~~*technológiáink*~~ ***információs rendszereink*** *fejlesztése és használata során alkalmazni fogjuk adatvédelmi alapelveinket. Biztosítjuk az észrevétel és beleegyezés lehetőségét, támogatjuk a biztonságos adatvédelemmel rendelkező architektúrákat, valamint biztosítjuk az adatok megfelelően átlátható és kontrollálható módon való felhasználását"* (Google).

Az efféle kijelentések MI-specifikus volta tehát kérdéses, nyilvánvaló azonban, hogy összességében annak ellenére sem befolyásolják negatívan az MISZ-ek célkitűzéseit, hogy a legtöbb igazgatási rendszerben már évek óta a hatályos adatvédelmi törvények részét képezik, így valójában teljesen feleslegesek.

A nem MI-specifikus javaslatok egy csoportjával kapcsolatban azonban a feleslegesség problémájánál fontosabb kérdések merülnek fel. Ezek azok az irányelvek, melyek az MI projektek mögött álló, elfogadható általános indítékokkal foglalkoznak, például:

> "~~*A/IS*~~ ***Vízforralók*** *alkotói számára a fejlesztés sikerességének elsődleges kritériuma az emberiség jólétének terén történő előrelépés."* (IEEE, 2. alapelv)
> *„A/IS első számú célkitűzésként minden típusú rendszer esetén az emberi jólét*

> *növelését célozza, a legjobb elérhető és széles körben elfogadott jóléti mérőszámokat referenciapontként használva.” (*IEEE, 2. alapelv javaslata)

A fenti idézeteket tartalmazó szakasz továbbmegy, és kifejti, hogy egy adott társadalom állapotának felmérésének legjobb módja valószínűleg nem a GDP-ben vagy a fogyasztói szintekben rejlik, ezek helyett az OECD jóléti mérőszámait ajánlja az MI fejlesztéséhez irányvonalként. Hasonló javaslatot tartalmaz az EU irányelv, egy kevésbé hangsúlyos, de nyílt utalás pedig USA tanulmányában található.

Ami mindezt érdekessé teszi, az az, hogy bár az idézetek szemantikailag vízforralók esetében is megállják helyüket, mégsem jellemző, hogy a vízforralók gyártóit kifejezetten a társadalmi jólét előmozdítására szólítanák fel. Természetesen nekik is szem előtt kell tartani a környezetvédelmi, biztonsági, pénzügyi, fogyasztóvédelmi stb. előírásokat, melyek egyébként mind az emberi jólét szolgálatában állnak, de nem kell nyíltan elkötelezni magukat ezen eszmék mellett; elsődleges motivációjuk érthető módon továbbra is a profitszerzés maradhat, a törvény adta kereteken belül. Mindez alapvető változásra utal az innováció zászlóvivői felé irányuló társadalmi elvárások terén, ugyanakkor szemben állhat az innováció mögött meghúzódó ösztönző erőkkel, melyek szinte minden gazdaságelmélet szerint a profitszerzéshez és a piaci versenyhez kapcsolódnak.

Egyes gondolkodók szívesen látják e váltást, mint a technológiai felvilágosodás régóta várt következő lépését,[144] melynek során a technológia végre szembenézhet a mecénások elvárásaival, mint a sikeres technológiai demokratizálódás egy formáját,[145] vagy a társadalmi kontrollra való erőteljes törekvést. Ugyanakkor, nem garantált, hogy a jólét iránti törekvés olyan konkrét keretbe foglalható, amely vállalati szinten értelmezhető lenne. E techno-politikai váltás, és az ehhez kapcsolódó társadalmi és gazdasági következmények értékelése valószínűleg egy önálló kötetet érdemelne.

Egy gyakorlati problémára azonban mindenképpen érdemes rámutatni: egy olyan, lehetséges jövőkép, ahol a jólét az első számú prioritás, és ennek betartatására megfelelő potenciál áll rendelkezésre (választható, vagy kötelező érvényű jogszabályok), a fejlesztőket arra ösztönözheti, hogy munkájukat *semmiképp se* MI-

[144] Ropohl (1998)
[145] Feenberg (2003), 5.3-as fejezet.

ként definiálják, kibújjanak az alól, amire a meglehetősen nagyszámú, és gyakran pontatlan MI definíciókból fakadó számos értelmezési variáció lehetőséget is biztosít. Ez az ösztönző erő mindaddig fennáll, amíg az MI projekteknek más technológiai területeknél szigorúbb előírásoknak kell megfelelniük.

6.6 MI-specifikus problémák

Úgy tűnik, hogy az MISZ-ekben található, kifejezetten MI-specifikus problémák összefüggésben állnak az MI egyedi jellemzőivel: önálló döntéshozatal, tanulási képesség, és potenciálisan magas átláthatatlanság.

Egy, számos MISZ által javasolt, tisztán MI-specifikus követelmény az emberi felügyelet megléte:

> "*Emberi felügyelet. Az emberi felügyelet segít biztosítani, hogy egy AI rendszer nem sérti meg az emberi autonómiát, és más káros hatásokat sem okoz. A felügyelet például emberi beavatkozáson (human-in-the-loop), emberi felügyeleten (human-on-the-loop) vagy emberi vezérlésen (human-in-command) alapuló irányítási mechanizmusok révén valósítható meg."* (EU AI HLEG)

Igaz ugyan, hogy az emberi felügyelet lehetőségét minden technológiai megoldás – még vízforralók – esetén fent kell tartani, de az MI-re igencsak jellemző magas autonómiaszint eredménye a fokozott figyelem. Emiatt, a különböző emberi felügyeleti módszerek és szintek[146] felfedezése, definiálása, megkülönböztetése valóban az MISZ-ek feladata, ami más mérnöki területek irányvonalai közül nem importálható.

Egy másik MI-specifikus példa az a követelmény, hogy *„egy adott A/IS döntés alapja mindig feltárható legyen."* (IEEE) Abban az esetben, ha a „döntés" definícióját nem olyan triviálisan értelmezzük, hogy akár egy termosztát stb. is képes „döntést" hozni, ez valóban csakis az MI kontextusában marad értelmezhető. Ez a követelmény, ami más MISZ-ekben gyakran „magyarázhatóság" néven szerepel, az MI magas szintű komplexitásához, és a gépi tanuláshoz – mint a rendszer hangolására szolgáló módszerhez – köthető. Az átláthatatlanság nem csak az MI-ban jelentkező probléma:

[146] („in-the-loop", „on-the-loop", stb.)

minden, megfelelő mértékben összetett, még egy teljesen mechanikus, az MI-től igen távol álló rendszer működése is meglehetősen homályossá válhat a felhasználók, de még a kezelők számára is. Azonban, míg ez utóbbi esetben az átláthatatlanság a hiányos dokumentáció vagy a műtermékkel kapcsolatos tudásszint időbeni csökkenésének számlájára írható, az MI mindezt egészen új szintre emelte az olyan, genetikus algoritmusokkal és tanulási képességekkel rendelkező modellek létrehozása által, melyek kezdettől fogva átláthatatlanok.

Kijelenthetjük, hogy az átláthatóság és magyarázhatóság kifejezetten az MI szakterületére jellemző problémák. Az e problémákkal kapcsolatos irányvonalak megalkotása során, más területeken kevés idevágó munkát találunk. Ez azonban nem jelenti azt, hogy az MI átláthatóságára való törekvés nem képezheti bírálat tárgyát.

Úgy tűnik, hogy a magas szintű átláthatóság elérése komoly elméleti[147] és gyakorlati (4-es fejezet) korlátokba ütközik, ezért, ha az átláthatóság iránti követelmény nem kerül megfelelő absztrakciós szinten definiálásra, az nagymértékben hátráltathatja az MI fejlesztését. Sőt, ha az átláthatóság nem párosul bizonyos mértékű érthetőséggel, az MI fejlesztői egyfajta látszólagos átláthatósággal „megúszhatják” – e kontextusban ez kezelhetetlen mennyiségű kód és konfiguráció közzétételét jelenti, amikor is a fejlesztők hivatkozhatnak arra, hogy ők „mindent” közzétettek, mégis kívülállóként mindebből bármilyen értékelhető információ kinyerése képtelenség.

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.

Mégis – mivel nincs más választásunk – elfogadjuk, hogy az ember működése homályos előttünk. Így, ugyan látható, hogy az MI-vel szembeni kemény átláthatósági előírás emberekkel összehasonlítva kettős mércének számít, mégis kétségkívül pozitív

[147] Grünke (2019).

eredmény. Először is, beleillik a fent leírt, a technológia iránt egyre növekvő társadalmi elvárások narratívájába. Valamint, éppen azért, mert gyakorlati okokból az emberi döntéshozás igen zavaros, ez bizonyos helyzetekben akár nagyon előnyös előrelépés is lehet – persze, ha sikerül megvalósítani.

6.7 MI szabályozás, mint alkalmazott etika

> *„Mindent összevetve, a megfelelő cél számunkra a boldogság (eudaimonia) – az az Arisztotelész által megvilágított gyakorlat, ami az emberi jólétet – mind egyéni, mind kollektív szinten – a társadalom legnagyobb értékeként definiálja.”* (IEEE)

MISZ-ek létrehozását észszerűen az alkalmazott etika tárgykörébe sorolhatjuk. E terület egyik állandó kihívása, hogy hogyan jutunk el az általános morálfilozófiától konkrét irányvonalakig vagy politikáig. Egyes MISZ-ek felülről lefelé irányuló alkalmazott etikai megközelítéssel[148] próbálkoznak, melynek során általános elméletektől jutunk el az egyes problémákra adott konkrét válaszokig.

Ezt, a „principlizmus” néven is ismert megközelítést részben az IEEE és az EU AI HLEG is alkalmazza. Ezek célkitűzései egy általános erkölcselmélet-szerű rendszerhez kötődnek, az IEEE esetén, lásd a fejezet kezdetén található Arisztotelész idézetet és rá való hivatkozást. Az EU irányvonalak az EU Alapjogi Chartájából próbálnak meríteni. Ennek a filozófiatudományból ismert átfogótörvény-modellhez hasonlóan kellene működnie. Az „erkölcsi munka” nagy része itt maguknak az alapelveknek kiválasztásából áll[149], a fennmaradó rész értelmezés, és konkrét esetekre való alkalmazás, amely intellektuális kihívást jelent, de végül is nem erkölcsi tevékenység.

Ennek, a felülről lefelé irányuló megközelítésnek két nagy akadállyal kell szembenéznie. Az első a számos általános erkölcsi elmélet közötti választás. Miért az *eudaimonia* a vezéralapelv? Áll-e bármilyen társadalmi konszenzus e döntés mögött, vagy lehetséges, hogy egy másik filozófia vagy vallás sokkal inkább kompatibilis lenne a jelenlegi társadalmi értékekkel? Ez a gondolatmenet elvezethet az erkölcsi teóriák népszerűségét illető széles körű felmérés ötletéhez, de ugyanakkor nagyon

[148] Beauchamp (2003,7).
[149] Allhoff (2011).

problémás. Nincs arra biztosíték, hogy egy efféle kérdés valóban felmérhető népszavazás révén, sőt, még ha az is lenne, az még mindig nem jelenti azt, hogy a többség véleményét követni a leghelyesebb dolog erkölcsi szempontból. Röviden, az alapelvek kiválasztásának védelme azok megkérdőjelezése esetén még megoldatlan problémát jelent.

A felülről lefelé irányuló megközelítés második problémája, hogy az általános erkölcsi teóriák közvetlenül nem használhatóak az átfogótörvény-modellben. Önmagában nem igaz, hogy egy általános alapelv alkalmazása pusztán gondos munkát igényel.[150] Ehelyett, az alkalmazás próbál fényt deríteni az általános teóriák belső konfliktusaira és a bennük található fehér foltokra, valamint az interpretációs és szemantikai problémákra, továbbá arra, hogy az általános elméletek nem határozzák meg azt, hogy egy terület-specifikus iránymutatásnak pontosan mit kellene tartalmaznia.

Más MISZ-ek általános morális keretrendszer nélkül működnek, konkrét problémákból és helyzetekből építkezve próbálnak általánosítani, azaz az alulról felfelé haladó megközelítést követik.[151] Ez tipikusan a rövidebb, Microsoft és Google irányvonalakhoz hasonló MISZ-ekre jellemző. Emellett, az USA kormányzati jelentés időnként egy nagyon is konkrét problémákon alapuló megközelítést követ. Az EU MI rendeletnek is számos ilyen eleme van. Ennek a megközelítésnek ugyanakkor bizonyos teoretikus korlátokkal kell szembenéznie. A fő ellenvetés az, hogy az általános következtetések alapját képező esetek kiértékelése során előfordulhat, hogy hallgatólagos módon a már meglévő alapelveinket alkalmazzuk, ami tulajdonképpen egy „álcázott” principlizmust eredményez. Emellett, olyan helyzet is könnyedén előfordulhat, hogy egy kulcsfontosságú, meghatározott eset kapcsán megoszlanak a vélemények a megfelelő cselekvési tervet illetően.

Egy harmadik lehetőség, amit bizonyos mértékig (persze erre való nyílt utalások nélkül) összes MISZ követ, a koherentizmus. A koherentizmus, az alkalmazott etika egy újabb stratégiája felismerte, hogy az egyirányú – fentről lefelé és lentről felfelé – megközelítések egyike sem képes igazán nagy eredmények felmutatására, így egy harmadik utat próbál járni. A koherentista álláspont egy-egy konkrét szituációt illető, jól átgondolt véleményből indul ki,[152] mely „horgonyként” szolgál. Ezután

[150] Beauchamp (2003, 9).
[151] Allhoff (2011).
[152] Rawls (1999), Daniels (1996).

megpróbálkozik az általánosítással, hogy keresztben átívelő alapelveket találjon, miközben arra reflektál, vajon az eredeti vélemény továbbra is megállja-e a helyét. Ezért nevezik ezt a megközelítést időnként „reflektív egyensúlynak”.

Forrásaink közül egyik sem tartalmaz koherentizmusra való utalást. Az IEEE és az EU irányelvek esetén azonban időnként mégis valami hasonlóval szembesülünk. Az például, hogy az algoritmusok nem tehetnek különbséget bőrszín alapján, vagy hogy az emberi felügyelet abszolút kötelező nagy kockázatú MI esetén, olyan magától értetődő alapigazságok, melyek horgonyként szolgálnak az MISZ megalkotása során. Ugyanakkor, nem szabad elfeledkeznünk arról, hogy ezen MISZ-ek működése alapelveken is nyugszik, tehát egyfajta dialektikának tekinthetőek a konkrét esettanulmányok és a fedőtörvény megközelítés között – ami naiv koherentizmust jelent.

A következő módszertani probléma – az alapelvektől a konkrétumokig és visszafelé vezető irányokon kívül – az, hogy milyen mértékben használhatóak fel az empirikus, tapasztalati adatok az MISZ-ek megalkotásához. Abból, amit a dokumentumok elárulnak arról, hogy hogyan történt ezek összegyűjtése, úgy tűnik, hogy elsősorban elméleti munkáról van szó. Az EU és az IEEE előírások helyt adnak a véleményformálásnak, hozzászólásokat és kérdéseket várnak. A többi MISZ olyan nyilatkozatnak tűnik, amely nem szorul közösségi jóváhagyásra. E szerint az MISZ-ek jelenlegi generációja karosszék-módszerrel készült, és a viták és külső vélemények révén fognak kifinomodni.

Sor került már empirikus kutatásokra is annak érdekében, hogy segítsék az MISZ-ek megalkotását. Például, ilyen a híres *Moral Machine kísérlet* (lásd a 4. és 10. fejezetben), mely azzal a céllal indult, hogy a különböző társadalmi és kulturális csoportok erkölcsi preferenciáit tapasztalati úton mérje fel. Bevallott szándéka volt, hogy az eredményekkel segítsék az MI etikai kérdéseit illető vitákat, így biztosítva inputot az MISZ-ek megalkotásához. Ugyanakkor ahhoz, hogy az efféle adatok valóban hasznosak lehessenek, muszáj meggyőződni arról, hogy az adott mérések relevánsak az aktuális tervet illető döntésekkel kapcsolatban. Ez egyelőre várat magára.[153] De még ha így is lenne, egy külön etikai kérdés az, vajon helyes dolog-e a legnépszerűbb elvárásokat megvalósítani. Vajon nem éppen az a helyzet, hogy

[153] Dewitt és társai (2019); Jaques (2020).

időnként a kisebbségi vélemények és értékrendek védelme a többségi elvárásokkal szemben igazságosabb és élhetőbb társadalmakat eredményezett?

6.8 A szabályozás új hullámai?

E fejezet a MISZ-ek jelenlegi hullámának vizsgálatával foglalkozott. Három kérdésre kerestem választ. Először is, mi indokolja az MISZ-ek elszaporodását, melyek e projektek észszerű célkitűzései és korlátjai? Másodszor, melyek azok a kifejezetten MI-vel kapcsolatban felmerülő problémák, melyeket egyediségük miatt az általános technológiai előírások nem képesek kezelni? Harmadszor, az alkalmazott etika terminológiáját használva hová pozícionálhatjuk e próbálkozásokat?

Első kérdésemet az MISZ-ek más (technológiákra vonatkozó?) előírások és iránymutatások történelmi kontextusába helyezésén keresztül válaszoltam meg. A kapott eredmények a társadalmi kontroll iránti fokozott elvárások egyre növekvő tendenciája felé mutatnak, legalábbis a 20. század közepe óta. Az MISZ-ek ezen egyre magasabb elvárások kifejeződési formái, egyes elemeik nem csak az MI-re érvényesek, hanem a társadalom által megfogalmazott, általános érvényű követelmények bármilyen technológiával kapcsolatban – bár e jelenség egybecseng az MI technológiák jelenlegi hullámával és azok sikerességével.

Mivel az MI-nek különösen nagy potenciálja van a konstruktőr általin bezártságra (lásd az 5.5-ös fejezetet), amíg új MI alkalmazások érkeznek, addig várhatóan a szabályozásra is lesz igény.

A második kérdésem az volt: melyek azok az etikai problémák, melyek kifejezetten csak az MI-vel hozhatók kapcsolatba? A válasz erre az MI rendszerek olyan, egyedi jellemzőiből eredeztethető, melyek semmilyen más technológia kapcsán nem fellelhetőek. A probléma lényege, hogy az MI nem más, mint az irányítás emberről gépre való átruházása. A fenti helyzet egy paradox jellegű feszültséget szül, hiszen a nehéz intellektuális munkát jelentő irányítást a lehető legnagyobb mértékben át szeretnénk ruházni, ugyanakkor, a negatív kimenetelek elkerülése érdekében mégis magunknál akarjuk tartani az MI feletti kontrollt, és a közbeavatkozás lehetőségét. Az elszámoltathatóság kontextusában ez a felelősségrés (lásd a 7.2-ben), ami átláthatóságot, és az emberi beavatkozást lehetővé tevő megoldásokat követel meg.

Ezek valóban kifejezetten az MI-vel kapcsolatban felmerülő problémák. Vannak viszont más, nem kizárólag ide vonatkozó gondok és követelmények is, ami felveti annak kérdését, hogy ezeket nem lenne-e célszerűbb az általános mérnöki etika tárgykörébe sorolni. Nem lenne meglepő, ha a jövő MISZ-ei ezekre a tényezőkre egyre jobban fókuszálnának.

Mindez magas szintű absztrakcióban csúcsosodik ki. Az MISZ-ek nem képesek részletesen meghatározni a kívánatos és nemkívánatos MI rendszereket, így viszont az alapelvek és általános javaslatok szintjén való működésre kényszerülnek. Ennél fogva, a fejlesztők egy irányított önszabályzás alanyaivá válnak. Hogy mindezt még tovább bonyolítsuk, az MI fejlesztés egyik területe annak meghatározása, mely döntések delegálhatók az autonóm rendszerek hatáskörébe, és mi ennek a legcélravezetőbb módja. Tehát, az MI fejlesztők által létrehozott *műtermékek (MI rendszerek)* felé az az elvárás körvonalazódik, hogy „etikailag értékegyeztetettek”[154] legyenek – de hogy miként, annak meghatározása a fejlesztők dolga, feltételezve, hogy ők maguk is „etikailag értékegyeztetettek”. Ez a két közvetett lépcsőfok azt jelenti, hogy az MISZ-ek mindössze a fejlesztőket irányítják, az MI rendszerek „irányítását” illetően. Talán ez, a mesterséges összetevők és az irányvonalakat megfogalmazó bizottság közötti nagymértékű távolság a felelős azért, hogy az MISZ-ek hajlamosak teljes hatástalanságot eredményező absztrakt jelleget ölteni.

Végül, látható, hogy az MISZ-ek felváltva alkalmazzák a felülről lefelé, és a lentről felfelé irányuló, valamint a koherentista alkalmazott etikai megközelítéseket. Nyilvánvaló, hogy a fent leírt, a közvetlen szabályzásra vonatkozó korlátozásokból szükségszerűen fakadó magas szintű absztrakció arra kényszeríti az említett dokumentumokat összeállító szakembercsoportokat, hogy olyan, rendkívül általános erkölcsi teóriákra hagyatkozzanak, mint az *eudaimonia*. Ezekről aztán természetesen kiderül, hogy nem képesek közvetlenül orvosolni az MI által felvetett problémákat. Azt azonban mégis kijelenthetjük, hogy a MI Etikai Irányelvek (különösen az átfogóbb jellegűek) fejlesztésének jelenlegi hulláma az emberiség történetének egyik legkomolyabb és legátfogóbb technológiai szabályzási törekvését képviseli. Úgy tűnik, mindez a technológiai kontroll igényével párosuló, jólétre és az egyéb társadalmi célok iránti elvárásra vezethető vissza. Az MI komplexitásának és az ebben rejlő

[154] „ethically aligned”

potenciálnak felismerése, valamint az ennek ellenére fennálló, szabályzására irányuló törekvések ugyanakkor talán megmutatják, hogy ezt, a technológia fogadtatásának területén megnyilvánuló átformálódást valóban „technológiai felvilágosodásként” emlegethetjük.

7. Az MI etika kanonizálódó kérdései

Az előző fejezetben láthattuk, hogy milyen MI szabályozási kérdésekkel foglalkoznak a hivatalos, normatív dokumentumokban. Ezek nem teljesen fedik le az MI etika, mint a filozófia, azon belül a morálfilozófia területén megjelenő vitákat.

E fejezet, az előzővel szemben nem korlátozza magát a már megírt irányelvek és szabályozások szövegeire, hanem a primer, *filozófiai*-ként jellemezhető irodalomra koncentrálva próbál egy körképet adni. Ezek a viták természetesen hatással vannak a szabályozásokra: egyfelől, ahogy láthattuk, minden eddiginél közvetlenebb tudástranszfer látszik kibontakozni a morálfilozófia és a technológiaszabályozás között, hiszen expliciten megjelent az utóbbi területen többek között az eudamoina, a deontológia[155] és a konzekvencializmus megkülönböztetése. Másfelől, a szabályalkotó bizottságok egy részében is hangsúlyos a technikafilozófia.[156]

Számos összefoglaló műből és teljeskörűségre törekvő kötetből meríthetünk, ha kulcsszavakat keresünk a témában: A mesterséges intelligencia etikai kérdéseinek táblázatos, irodalmi metaelemzésen alapuló összefoglalóját Hagendorff[157] (2020) készítette el. Az *Oxfordi MI Etika Kézikönyv* (2020)[158] kimerítően foglalkozik a kérdéssel. Coeckelbergh a saját *MI Etika* (2020) monográfiájában kevesebb témát, de mélyebben dolgoz ki.

Fontos azonban megérteni, hogy egyszerre folyik a vita egyes, konkretizálható problémák szintjén és a metaetika szintjén is. A fogalmi értelemben legmagasabb abszrakciója a vitának az, hogy kell-e, illetve lehetséges-e érdemben egyáltalán az MI etikával foglalkozni,[159] vagy nagyobb a veszélye, hogy csekély morális nyereségért „etikai mosdatást” nyújt a terület a profit-éhes cégeknek.[160]

A vita következő szintje az, hogy teoretikus tudás (pl. normatív keretrendszerek ismerete) vagy teoretikusnak nem nevezhető készségek (pl. dacolásra való képesség a

[155] https://standards.ieee.org/ieee/7000/6781/
[156] AI HLEG (2019).
[157] Hagendorff (2020a).
[158] Dubber, Pasquale és Das (2020).
[159] Hagendorff (2023).
[160] Vica és társai (2021).

felső vezetéssel), – amelyeket olykor performatív képességeknek is neveznek – fontosabbak-e.

Az MI etikájáról szóló, tárgyi szintű diskurzus legnagyobb része kockázatalapú. A nagy problémakör a „felelősségrés” („accountability gap”), amelynek okait az 1.2, a 2.4 és még részletesebben a 4. fejezetben tárgyaltam.

Egy másik jelentős etikai kérdés az MI-ben a torzítások és diszkrimináció propagálása, ami olyan nagy probléma, hogy könyvem 8. fejezetét külön ennek t szenteltem. Szintén önálló fejezetet érdemel az MI rendszerek átláthatatlansága és az ebből fakadó kockázatok, ezekről a 4. és 9. fejezet szól.

7.1 Automatizációs elfogultság

Az automatizációs elfogultság („automation bias”),[161] lényege, hogy a felhasználók megszokják, hogy az automatizált rendszer általában jól működik és kritika nélkül kezdik elfogadni annak kimenetét. Egyszerűbben szólva: hozzászoknak és nem ellenőrzik többé.

A problémát jól illusztrálja az a kísérlet, amelyet repülőpilótákon végeztek el. A pilótákat két csoportra osztották, a résztvevők egyik fele támogató rendszer nélkül, kizárólag saját képességekre hagyatkozva hozott döntést, a másik csoportot pedig döntéstámogató rendszerekkel látták el. Az eredmények azt mutatták, hogy az utóbbiak figyelme a repülésről áttolódott a támogató rendszerre, így viszont nem vették észre, vagy nem kezelték azokat a fontos információkat, amelyeket a rendszer maga nem emelt ki számukra. Ezzel összességében rosszabb döntési teljesítményt nyújottak. Ez az kísérlet abban is segít, hogy megértsük: az MI hasznossága erősen tudásszint-függő.[162]

[161] Lyell és Coiera (2017).

[162] Skitka és társai (1999) kísérletében a pilóták rosszabb teljesítményt nyújtottak a szakértői rendszer támogatásával, amit minden bizonnyal nem tud kompenzálni az, ha ez a fajta munkavégzés esetleg kevésbé volt megterhelő számukra a saját képességekre hagyatkozó repüléshez képest. De képzeljük el, hogy vész esetén, például a pilóta rosszulléte miatt egy repülőt vezetni nem tudó személynek kell átvennie az irányítást: minden bizonnyal ebben az esetben egy részben vagy teljesen automatizált rendszer óriási segítség. Ez a gondolat kiterjeszthető annak megértésére, hogy hol lehet például a nagy nyelvi modellek (LLM-ek, ChatGPT és társai) leghatékonyabb munkapontja: ha zéró vagy minimális tudásunk van egy adott témakörben és megelégszünk egy szuboptimális, a szakértőinél alacsonyabb szintű megoldással, ami azonban meghaladja a saját képességeinket a témakörben.

Az automatizációs elfogultság kétféle kockázatot generál: a közvetlen hatása a szuboptimális döntéshozatal, illetve az ellenőrzés hiánya; közvetve, hosszabb távon pedig a felhasználó képességeinek erózióját okozhatja.

Ez a terület, hasonlóan a transzparencia és a részrehajlásmentesség kérdéseihez (lásd a következő fejezetekben), az MI viszonylatában réginek tekinthető. Általánosságban azt figyelhetjük meg, hogy a szakértői rendszerekkel kapcsolatban felvethető kérdéseket ezek sikerénél, a 20. század második felében már elkezdték vizsgálni.[163]

Egy, a területet vizsgáló meta-tanulmány[164] alapján kijelenthetjük, hogy az egészségügyi MI elterjedtsége okán az egészségügy volt az automatizációs torzítás vizsgálatainak első terepe. Fontos, hogy a fent említett kockázatok mellett az automatizáció, sőt akár az automatizációs torzítás esetleges előnyeit is mérlegelték. Az orvosi döntéstámogató rendszerek alapvető céljával összhangban azt találták, hogy megfelelő konfiguráció esetén a diagnosztikai szenzitivitás nő, habár a specificitás csökkenhet. Hasonló trade-off viszony áll fenn a szenzitivitás és a konklúzióba vetett bizalom között.

Megállapították azt is, hogy az automatizálási torzítás egyénfüggő. Ez önmagában nem meglepetés, hiszen mindenféle kognitív torzítás függ az egyén pszichológiai profiljától – a kérdéskörrel kapcsolatos kutatások értéke inkább az, hogy megnyitja az utat egy olyan tesztelési módszertan kialakítása előtt, amellyel az adott egyén profilját a megfelelő MI interfészhez lehet társítani és viszont.

7.2 Felelősségrés

A 4. fejezetben kifejtettem számos olyan okot, ami az ellen hat, hogy az ember intellektuális kontrollja alatt tarthassa az MI-t. Ezen okok között szerepel a rendszerek inherens bonyolultsága, azok sebessége és sok egyéb tényező.

A felelősségrés kérdése azzal foglalkozik, hogy ha egy rendszert lehetetlen ellenőrzésünk alatt tartani, akkor, ha a rendszer kárt vagy sérülést okoz, lehet-e bárkit felelősnek tartani, és ha igen, milyen szempontból.[165]

[163] Singh és társai (1993).
[164] Goddard és társai (2012).
[165] Matthias (2004).

A kérdés feltevéséhez szükséges az az előfeltevés, hogy az egyén felelősségéhez valamiféle irányítási vagy döntési lehetőségnek adottnak kell lennie a vizsgált eseménnyel kapcsolatban: másképp fogalmazva, nem fogadnánk el felelősséget egy olyan eseményért, amire semmilyen ráhatásunk nincs. Természetesen ez megnyitja az utat az olyan metafizikai kérdések felé, mint a fizikai világ determináltsága és a szabad akarat lehetségessége. Az MI etikában ez a determinizmus jelenleg nincs középpontban, helyette az irányítás és a felelősség között kapcsolatot adottnak tekintik.

A probléma szerkezete tehát a következő: az MI lényege, hogy az irányítást átveszi a géptől az ember, ráadásul az MI intellektuális ellenőrzése erős és elvileg sem feloldható episztemikus korlátokkal terhelt. Az MI mindeközben nem képes felelősséget viselni. Tehát, az ember azért nem felelős, mert nem irányít; a gép pedig azért, mert nincs ilyen képessége: létrejön tehát egy rés a felelősségben, és ebből fakadóan az elszámoltathatóságban is.

A felelősségrés kérdése erősen vitatott, és végső soron a felelősség szerkezetéről szól és nem pedig az MI jellegzetességeiről. Egyesek,[166] a fenti érvelésnek megfelelően az irányítás és a megérthetőség hiányát elegendőnek tartják a felelősségrés kialakulásához és megoldhatatlannak tartják azt, így a probléma megkerülése vagy kényszerű megszokása mellett érvelnek.

Mások, mint Champagne és Tonkens[167] úgy gondolják, hogy a felelősség tulajdonítása úgy is lehetséges, hogy nem tartjuk hibásnak a felelősséggel terheltet. Katonai MI-re fókuszáló érvelésükben előterjesztik, hogy az MI-t használó parancsnok előre elfogadja a felelősséget a kimenetekért, akkor is, ha a konkrét műveletet teljes mértékben delegálja, így nem lesz közvetlen okozója az esetleges negatív kimenetelnek. Ők ezt nem-oksági alapú felelősségnek nevezik, de úgy is rekonstruálható az álláspontjuk, hogy egy megelőző, nagyon távoli ok (például az MI bevezetése melletti döntés) generál felelősséget. Az okság tranzitivitása miatt természetesen ilyen módon a gyártó is felelőssé tehető: kérdés, hogy milyen távolra mehetünk el az oksági láncban ahhoz, hogy a társadalom számára kielégítő felelősség-szerkezetet kapjunk.

[166] Danaher (2016).
[167] Champagne és Tonkens (2015).

A harmadik típusú álláspont kompromisszumot javasol: habár elismeri a felelősségrés létezését[168], annak szélességét és jelentőségét megkérdőjelezi, illetve kreatív megoldásokat kínál. Ilyen például az, hogy a felelősségvállalást fokozott transzparencia, valamint akár a magyarázkodás és bocsánatkérés kötelezettsége helyettesítse, amennyiben ez elfogadható minden érdekelt félnek – ezáltal ez a megközelítés kontraktariánus álláspontra helyezkedik (lásd még a 10.3-as fejezetet).

7.3 Egyenlőtlen hozzáférés

A fejlett technológiák kontextusában a technológiához való egyenlőtlen hozzáférés visszatérő téma a közgazdaságtanban, a társadalmi egyenlőtlenségek kutatásában, valamint a technikafilozófia és technikaszociológia területén is.

A probléma lényege egy ördögi kör: aki hozzáfér a technológiához, az a technológia segítésével kedvezőbb anyagi vagy tudásszinttel kapcsolatos előnyökhöz jut: az előbbire példát jelentenek a hatékonyan termelő, időmegtakarító gépek, az utóbbira pedig az internet elérés. Ezek általviszont tovább javíthatja a hozzáférését, erősítve ezzel a kiinduló hatást. Ugyanakkor, akinek nincs megfelelő hozzáférése mindezekhez, az leszakad, így elkezd nyílni a társadalmi olló.

Az MI különösen alkalmas ennek a negatív folyamatnak a kiváltására, az 5. fejezetben bemutatott okok miatt. Megfelelő szakpolitikák enyhíthetik ezeket a hatásokat, többek között az erőforrások hatékony újraelosztásával, illetve azok kompenzálásával, akiket az automatizálás megfoszt a kereseti lehetőségeiktől.[169] Ezzel a problémával részben átfed a 7.7-es alfejezetben bemutatott, a munka jövőjére irányuló gondolatmenet.[170]

7.4 Az értékegyeztetés problémája

Az MI-vel való értékegyeztetés („alignment”) kérdés lényege, hogy hogyan érjük el, hogy az autonóm mesterséges ágens az általunk kívánatosnak tartott értékeket képviselje.

[168] Nyholm (2020) és Kiener (2022).
[169] Korinek és Stiglitz (2017).
[170] Lásd még Braun és Baumüller (2024) magyar nyelven is olvasható összefoglaló tanulmányát az MI és a szegénység kapcsolatáról.

Természetesen innen egyetlen lépésből eljutunk oda is, hogy ki képvisel minket; azaz, hogy ha az MI konstruktőre képes is betáplálni a saját értékeit, az nekünk megfelelő-e. Ennek jelentőségét a konstruktőr általi bezártság adja (lásd 5.5-ös fejezet).

A kérdés tehát két elemből áll:[171] (1) mely értékekkel kellene az MI-t összehangolni és (2) hogyan lehet ezt megtenni?

Minden bizonnyal nincs egyenválasz a két kérdésre, hanem azokat az adott érték kontextusában kell megválaszolni. Itt az értéket gyűjtőfogalomként használom, amely kívánatos attitüdöket, erényeket, viselkedésmintákat és egyebeket is jelenthet – a felhasználó, illetve a szélesebb társadalom elvárásainak összességéről van szó. Ilyenek például az igazságosság, az elfogultság hiánya[172], de az elvárt kommunikációs normák betartása is.

Hatalmas, de egyáltalán nem váratlan problémát jelent, hogy nincs társadalmi konszenzus számos érték esetén, például abban, hogy mekkora központosítás elfogadható egy infrastruktúraként funkcionáló MI esetén, hol az egyensúly a hatékonyság és a magánélet védelme között és így tovább.

7.5 Szelektív MI kritika

„Szelektív MI kritika” alatt azt a jelenséget értem, amit az angol nyelvű szakirodalom „selective adherence” -nek nevez. Ezzel megfordítom a kifejezést, a lényeget nem megváltoztatva.

Ez az etikai kérdés arra a tendenciára vonatkozik, hogy az algoritmus tanácsait szelektíven fogadjuk el, attól függően, hogy azok egyeznek-e a döntés alanyaival kapcsolatos előzetes elvárásainkkal, sztereotípiáinkkal vagy más elfogultságokkal. Például empirikusan igazolt hatás, hogy amikor egy rendszer magas kockázatot jósol a negatívan sztereotipizált kisebbségi csoportok tagjaira vonatkozóan, akkor kevesebbszer bírálják felül a rendszert, mint például a Holland családtámogatási botrány esetén[173]

[171] Gabriel (2020).
[172] Lásd a 8-ik fejezetet.
[173] Alon-Barkat és Busuioc (2023).

Ezt a problémát az a döntési struktúra állítja elő, amelyben a humán feladata felügyelni az MI rendszert és közbeavatkozni, felülbírálni, amikor szükséges, a saját emberi belátása szerint. Ez a mozgástér teszi lehetővé a részrehajló felügyeletet, ami nem ugyanaz, mint a részrehajló döntés. Egy kapcsolódót kísérletet Hollandiában végeztek el. Az alanyok feladata az volt, hogy értékeljék egy MI kimenetét, amely jelöltek pályázatát oszályozta egy közoktatásban betölthető tanári állásra. Az egyik adatpont az volt, hogy az illető bevándorló-e. A kontroll csoport nem használt algoritmust, hanem maga hozott döntéseket és nem is állított elő részrehajló eredményeket adott méltányossági metrikák alapján (lásd a 8. fejezetet). Annak a csoportnak, amely MI döntéstámogatásra hagyatkozott, így voltaképpen a döntés felügyelete volt a feladata. Ők viszont részrehajló módon döntött amellett, hogy felülbírálják-e az adott rendszerkimenetet.

A szelektív MI kritika bármely területen előállhat. Az orvoslás esetében például a szelektív kritika azt jelentheti, hogy ha egy orvos úgy véli, hogy egy bizonyos rassz inkább hajlamos a függőségre, mint egy másik, és van egy értékelő rendszere a potenciális szerhasználók megjelölésére, akkor ebben az esetben a felülbírálatok előfordulásai és irányai jelezhetik a részrehajlást, miközben a döntések önmagukban megfelelően alá vannak támasztva.

Természetesen nem csak embercsoportok, hanem egyéb kategóriák kapcsán is lehetünk részrehajlóak: az orvostudományban bizonyos kórképekkel szemben, a banki hitelminősítésekben egyes iparágakkal szemben, a biztosítónál különböző járművekkel szemben, és így tovább.

7.6 Az MI és a munka jövője

Az MI-vel kapcsolatos félelmek önálló, tekintélyes méretű osztálya az, hogy túl gyorsan és túl radikálisan alakítja át a munkaerőpiacot. Az alapprobléma nem új. Az automatizálás, vagy szélesebb értelemben az iparosodás korábbi hullámai általában elég nagy társadalmi diszrupciót okoztak: a félelem az, hogy az MI még nagyobb társadalmi felfordulást fog eredményezni, olyan sebességgel, amilyet még soha nem látott az emberiség.

Ennek a változásnak az egyik fontos mozgatórugója a munka piaci értékének a megváltozása. Fontos észrevenni, hogy az ipari forradalmak korábbi hullámaiban elsődlegesen a kékgalléros munkák kerültek veszélybe és ennek másodlagos hatásai voltak minden egyéb munkakörre, például azáltal, hogy a gyáripar miatti urbanizáció a fehérgalléros munkakörök helyét és jellegét is átalakította. Ezúttal viszont mintha fordított lenne a helyzet.

Talán ez a tényező is hozzájárult ahhoz, hogy vezető értelmiségiek ezrei, köztük a Turing-díjas Yoshoua Bengio, Steve Wozniak, az MI oktatás kiemelkedő alakja Stuart Russell, illetve innovátorok, mint Elon Musk, aláírtak 2023-ban[174] egy petíciót, ami az élenjáró MI fejlesztésének azonnali felfüggesztését követelte. A kiáltványt megszövegezők többek között azért aggódnak, hogy az MI feleslegessé teheti, lecserélheti az embert. – mintha ez újdonság lenne és a gépesítés, illetve automatizáció korábbi hullámai meg sem történtek volna.

Ez a témakör túl szerteágazó és egészében nézve túl messzire esik fő gondolatmenetemtől ahhoz, hogy megpróbáljam teljeskörűen feldolgozni. Azonban van két, egymáshoz kapcsolódó része ennek a témának, amelyhez érdemben hozzá tudok járulni a könyv első részében bemutatott fogalmi eszköztárral.

7.6.1 Azonnali piaci diszrupció

Az egyik téma az MI-nek ahhoz a különleges tulajdonságához kapcsolódik, hogy amikor egy műszaki áttörés bekövetkezik és van működő megoldás egy adott feladatosztályra, akkor az minimális határköltséggel sokszorozható; ráadásul központosított, felhő alapú szolgáltatásként szállítható. Ezt az MI tulajdonságot az 5. fejezetben már részletesen bemutattam, a konstruktőr általi bezártság veszélyeit latolgatva.

A munka jövője szempontjából ez a képesség azt jelenti, hogy egy-egy ilyen megoldás elvileg akár egyszerre bevezethető a világ minden pontján, így alig működnek azok a kiegyensúlyozó hatások, amelyek egy másik technológia bevezetését moderálják.

Az MI kifejlesztésének a megtérülési modellje nem sokban különbözik például egy újfajta mezőgazdasági gépétől. Mindkettő esetén igaz, hogy van egy megcélzott

[174] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

volumen, amely a K+F költség megtérülését és az elvárt hasznot biztosítja. Ez a cél általában a teljes megszólítható piac, az összes mezőgazdasági egység nagyon kis töredéke. Az eltérés a két szektor között a bevezetés gazdasági modelljénél keresendő. Minden mezőgazdasági egységnek el kell döntenie, hogy megtérül-e elvárható időn belül a jellemzően drága beruházás, például egy ingatlan árát felemésztő gép beszerzése. A munkaerő-költség hipotézisek[175] szerint az efféle innováció és automatizálás ott indul el, ahol szűkös és/vagy drága a munkaerő. Ám a kiváltott munkaerő egy másik gazdasági egység számára elérhetővé válhat, így ott megszűnik a gépesítést kifizetődővé tevő helyzet.

Továbbá, ahogy például az első, jellemzően drága, de a fogyasztók számára az érdekesség és a menettulajdonságok miatt vonzó elektromos autók esetén láttuk, az ipar nem is képes legyártani az adott gépet kellő számban, ha túl nagy az érdeklődés, és sok ezer megrendelés évekre várólistára kerül.

Akár a letölthető, akár a felhős szoftverszolgáltatás (SaaS) esetében – azon belül a nem-robotikai MI-re is igaz ez – alig működik ez a kiegyensúlyozó mechanizmus. Ennek oka, hogy a bevezetés költsége szinte csak a tanulásból és az üzleti folyamatok átalakításának költségéből áll, de a géphez kapcsolódó nagyberuházásról aligha beszélhetünk. Letölthető szoftver esetén a „gyártás” skálázása nem értelmezhető, bármely példányszám előállítása megoldott; a központosított szolgáltatások esetén a szerverpark skálázása azért igényel meggondolást, ám ez is csak egy csekély ellenállásnak tekinthető. Jó bizonyíték erre a ChatGPT adopciós görbéje: 0 felhasználóról 100 millióra ugrott fel mindössze 2 hónap alatt.[176]

Természetesen elvileg nem lenne kizárható, hogy a szoftver vagy SaaS beszerzés költsége vetekedjen egy nagygép beszerzésével, ha a szolgáltatók képesek lennének nagyon magas licenc vagy előfizetési díjat beárazni. A tapasztalat nem ezt mutatja: például a mindennapjainkat nagyban megkönnyítő, nem-munkahelyi email szolgáltatók, naptárprogramok, kollaboratív, online irodai szoftverek gyakorlatilag

[175] Habakukk (1962).
[176] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

ingyen hozzáférhetők.[177] Ez azt mutatja, hogy kiélezett, árleszorító verseny uralkodik ezeken a piacokon.

Ez tehát az a piaci viselkedés-karakterisztika, amelyet csak az MI sajátos tulajdonságai, „természete” segítségével magyarázhatunk, így ebben a könyvben is helye van.

A helyzet elemzéséhez a parciális[178] kereslet-kínálat egyensúly elemzés eszköztárát hívom segítségül, amelyet Alfred Marshall neoklasszikus szintéziséből ismerünk.

A parciális egyensúly elemzés lényege, hogy a keresletet és a kínálatot is mennyiség-ár görbékkel írja le. Ez az eszköz már közgazdaságtan marginalista és szubjektív hasznosságot előtérbe állító fordulata után született. Itt a jószágoknak nem egy darab, inherens, például az input költségek által meghatározott ára van, hanem a kereslet és a kínálat találkozásának határán vannak megállapítva.

A módszer megvilágítására a rajzos illusztrátori munka piacát mutatom be, amely egyben az MI általi diszrupció bemutatására is kiváló lesz.

Képzeljük el az olyan illusztrációk piacát, mint ez a rajz Alan Turingról:

[177] Természetesen igaz, hogy az ingyenességet például az teszi lehetővé, hogy célzott reklámokat tud hozzánk eljuttatni a platform, illetve hozzáfér minden adatunkhoz, amiből szintén tud bevételt termelni. Ennek azonban nem lesz jelentősége, mert a szolgáltatások igénybevételét ezek a tényezők, a tapasztalat alapján érdemben nem hátráltatják.

[178] A „parciális” ebben a kontextusban azt jelenti, hogy a teljes gazdaság egyetlen alrendszerét nézzük és azt is korlátozott időtávon. Az alább bemutatott keresleti-kínálati görbék az MI megjelenésének rövid távú hatását jelölik. Természetesen hosszabb távon különféle visszacsatolások indulnak be: ha nincs kereslet az illusztrátori munkára, akkor az illusztrátorok képzése is átalakul, így a kínálatuk is; másfelől az alacsonyabb egyensúlyi ár lehet, hogy évek alatt teljesen új piacokat, szegmenseket nyit meg, amit az aktuális keresleti görbe még nem tartalmaz.

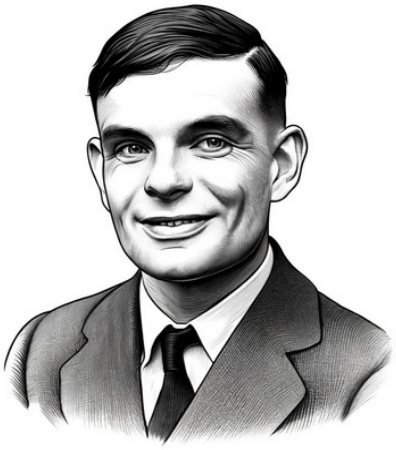


Figure 7: Egyedi rajz Alan Turingról. Készítette: Dall-E 3. Prompt: Héder Mihály

Az ügyfél nem szeretné újrahasznosítani a Turingról elérhető fotók egyikét, mert azokat túl gyakran használják és nem illenek a kiadvány stílusába: helyette felbérel egy illusztrátort, aki instrukciók alapján elkészíti a kért stílusban, egyedi kivitelben a képet.

Az alábbi ábra mutatja, hogy hogyan működik egy ilyen piac:

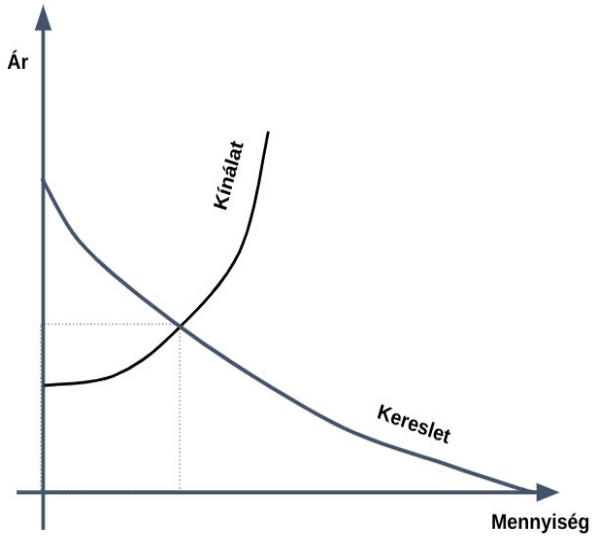


Figure 8: Az illusztrációk piacának keresleti és kínálati görbéje

A függőleges tengelyen a hagyományos ábrázolásnak megfelelően az ár, a vízszintesen a mennyiség.

A kereslet egy lefelé futó görbe, hiszen nagyon drága illusztrációt csak nagyon kevesen tudnak és hajlandók megfizetni, annál is drágább illusztrációra pedig nincs vevő. Ezért a függőleges tengelyt ott metszi a görbe, ahol a legdrágább, még eladható illusztráció van.

Ahogyan az ár megy lefelé, úgy lesz több vevő is, tehát a keresett mennyiség nő. Ez sem végtelen: lesz egy pont, ahol már ingyen sem kell az illusztráció, hiszen a kereslet nem végtelen. Ezért ez a görbe a vízszintes tengely érintésével ér véget.

Természetesen az illusztrátorok kínálata épp fordított irányú görbét eredményez: egy bizonyos ár alatt senki sem tud dolgozni, ezért a görbe nem a nullánál indul, hanem ennél a bizonyos árnál. Majd ahogyan az ár megy fel, egyre nagyobb az illusztráció kínálat. De az illusztrátorok száma és kapacitása is véges, ezért ez a görbe sem megy a végtelenbe: véget ér ott, ahol minden illusztrátor a kapacitása határán dolgozik és már nem képes további munkákra magasabb díjért sem.

A két görbe metszi egymást – itt van a piaci egyensúly. Ha a függőleges tengelyről leolvassuk az árat, a vízszintesről amennyiséget, majd a kettőt összeszorozzuk, akkor a piac méretét kapjuk pénzben kifejezve. Vizuálisan ábrázolva annak a téglalapnak a területéről van szó, amelynek az egyik sarka az origó, az átellenes pedig az egyensúlyi pont.

Ebbe a helyzetbe érkezik meg a képgeneráló MI. Ennek, mint általában a tömegtermelésnek a hatása a kínálati görbe eltolásával fejezhető ki:

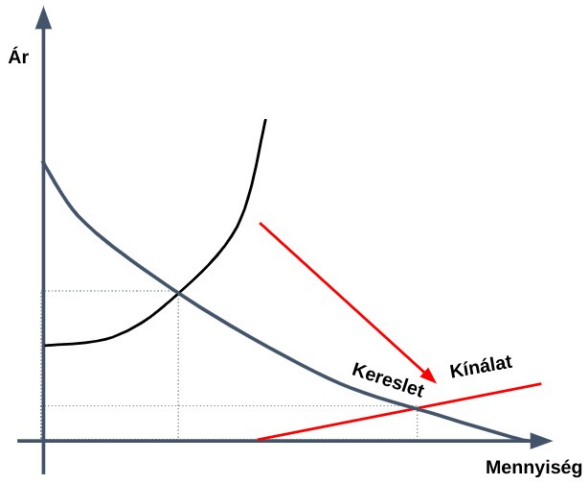


Figure 9: A kínálati görbe rövid távú elmozdulása az MI megjelenésének hatására és az új piaci egyensúly

Ez esetben egy radikális eltolódásról van szó. Először is, már anyagi befektetés nélkül/díj nélkül/díjmentesen is kapunk egyedi illusztrációkat, köszönhetően a képgenerátorok ingyenes csomagjainak. Ezek általában mennyiségileg, esetleg funkcióban is limitáltak. Ezért a görbe messze a grafikon jobb oldalán indul, mert az ár 0, de a mennyiség jelentős. Ha fizetünk a képgeneráló szolgáltatásért, akkor már havonta több száz illusztrációt kaphatunk, a felhős MI-szolgáltatásokra jellemző néhány dolláros havidíjért, ami a professzionális illusztrátoroknál nagyságrendekkel olcsóbb .

Ezért tehát a kínálati görbe teljesen jobbra és lefelé tolódik. Az új egyensúly nagyon alacsony árnál lesz, és köszönhetően a lelkes felhasználóknak, sokkal több illusztráció készül el. A piac mérete egy keskeny téglalappal ábrázolható, aminek területe valószínűleg kisebb a korábbi piacénál, és ezen a bevételen sem az illusztrátorok, hanem az MI cégek osztozkodnak.

A problémát az jelenti, hogy az új ár annyira alacsony, hogy azzal humán illusztrátor versenyezni sem tud, így a munkájának piaca megszűnik. Az újdonság pedig az automatizációhoz képest az, hogy az MI szolgáltatás jellegzetességei miatt lényegében a bolygón mindenütt lokális befektetés nélkül, azonnal elérhető az új technológia, így a piac megsemmisülése sosem látott sebességgel megy végbe.

Egy ellenvetés ezzel az elemzéssel szemben az lehet, hogy az MI képességei nem ekvivalensek az emberével. Így ebben a felfogásban nem is ugyanarról a jószágról van

szó: az MI csak egy helyettesítő terméke az embernek, amely bizonyos problémákat nem tud megoldani (például ilyenek a nagyon specifikus egyedi kérések, mint az, hogy Turing nyakában legyen egy csokornyakkendő, amin nullák és egyesek vannak).

Az ellenvetés két okból nem változtat a nagy képen: egyfelől, elképzelhető, hogy az MI minden fontos paraméterben imitálni és meghaladni képes az embert egy adott feladatosztályban. Ez a természetes nyelvek gépi fordításánál már adott[179], és elképzelhető, hogy az illusztráció is ilyen lesz egy napon.

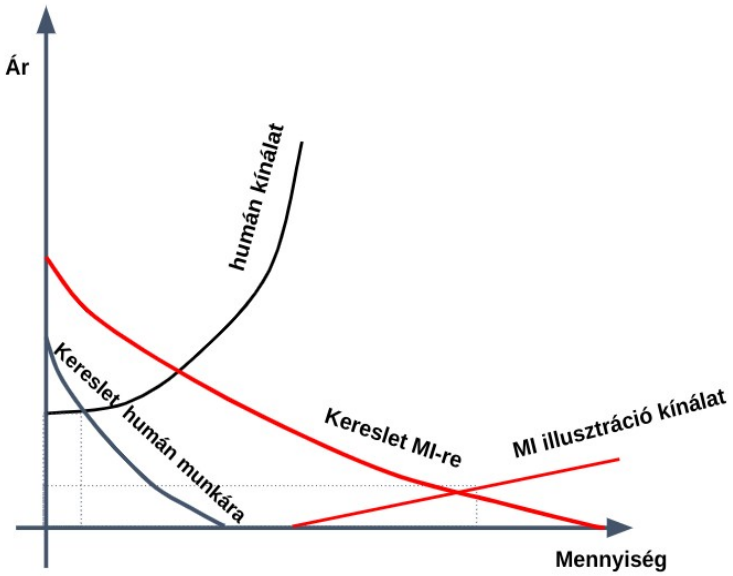


Figure 10: Ketté eső piac: humán vs. MI-munka

Másrészt, ha ketté bontjuk a piacot a prémium árazású humán illusztrátorokra és az olcsó gépi szolgáltatásokra, akkor a következőt kapjuk:

Az ábrán a humán illusztrátorokra való keresleti görbe bal oldalát ugyanazzal az árazással indítottam; ugyanakkor azt nem várhatjuk, hogy a lefutása a mennyiség függvényében ugyanaz lesz. E helyett a lefutás sokkal meredekebb, és a piac összességében sokkal kisebb. Ráadásul az MI piaca folyamatosan fenyegeti, hiszen annak a képességei gyorsabb ütemben nőnek az ember képességeinél.

[179] Ennek a jelentőségét nehéz túlbecsülni az alkalmazhatóság terén, lásd Rab (2024), 305-306 oldal.

7.6.2 Az MI-vel készült MI

A másik fontos témakör az MI és a munkaerőpiac kérdéskörén belül, amiben e könyv keretei között érdemben foglalkozni tudunk, az az MI részvétele újabb MI-k létrehozásában.

Mindeddig azt a stratégiát követtem, hogy az MI konstrukciót a szoftverfejlesztés témaköre alá utaltam – remélve, hogy a robotika szakértőket, mechatronikusokat nem sértettem meg ezzel túlzottan – majd a szoftvereket mint speciális, extrém könnyen módosítható és sokszorosítható gépekként jellemeztem és ezekből a tulajdonságokból vontam le bizonyos következtetéseket.

A konstruktőr általi technológiai bezártság lehetőségét (5. fejezet) és a dekompozíciós feladatot (2. fejezet) már bemutattam, ezeket a témákat tovább szintetizálom a 10. fejezetben.

Ebben az alfejezetben az MI konstrukció, mint MI által potenciálisan veszélyeztetett munkakör szerepel. Habár a világban már elindultak a mesterséges intelligencia mesterképzések, én az általánosabb, régebbi és kanonizálódott szoftvermérnöki munkakört fogom alapul venni, ahol szükséges, ott az MI-specifikus elemekre kitérve.

A szoftvermérnöki feladatkört legjobban a szoftvermérnöki képzés követelményrendszere írja le. Ezt az IEEE Computer Society kodifikálta és tartja karban a Software Engineering Body of Knowledge-ben (SWEBOK)[180], amelyet a szoftverfejlesztés „nagykönyvének” is tekinthetünk. Az IEEE globális jelenlétének köszönhetően szinte mindenütt a SWEBOK terminológiáját és megközelítését követik a szoftvermérnöki képzésben.

A SWEBOK legújabb, 4-es verziója 18 kategóriára osztja fel azt, hogy mit kell tudnia egy szoftvermérnöknek. Ezek között vannak nem-specifikus tudáselemek is, mint a matematika, illetve menedzsment ismeretek. 10 nagyobb témakörre azonban úgy tekinthetünk, mint a szoftvermérnök specifikus tudására: a követelményelemzés, a szoftver architektúrák ismerete, a szoftvertervezés (e kettő a dekompozíciós feladatot fedi le), a szoftverek építése, a tesztelés, az üzemeltetés, a karbantartás, a

[180] https://www.computer.org/education/bodies-of-knowledge/software-engineering

konfigurációmenedzsment, a szoftverminőség-menedzsment és a szoftverek biztonsága.

Az előző alfejezetben a keresleti és kínálati görbék elmozdulásain keresztül próbáltuk megérteni, hogy mit jelent egy munkakör azonnali leértékelődése az MI miatt. Itt is ugyanezt a stratégiát fogom követni, de azzal, hogy a szoftverfejlesztés önmagában igen sokszínű szakmáját szétbontom az IEEE SWEBOK-ban meghatározott, és a szakma által általánosan elfogadott alkategóriákra, mint egy-egy parciális piacra.

Önálló piacként tekinteni egy olyan specifikus feladatkörre, mint például a szoftvertesztelés vagy a biztonság, egyáltalán nem túlzás. Ezek a feladatkörök és a bennük elvégzendő feladatok az angol nyelv IT-ben megfigyelhető univerzalitása, az alkalmazott eszközök globális elérhetősége, valamint a SWEBOK és egyes elterjedt tankönyvek miatt globálisan szinte teljesen szinkronizáltak. A megfelelő munkaközvetítő portál használatával lehetséges bármely kontinensről általában néhány órán belül alkalmas munkaerőt találni.

A munka konkrét helyhez vagy infrastruktúrához tehát csak lazán kapcsolódik és egy kis túlzással elmondhatjuk, hogy gyakorlatilag bárhonnan elvégezhető. Ez egy fontos eleme az MI általi kiválthatóságnak.

Ha ezek után sorra vesszük a 10 feladatkört, akkor a következőket kapjuk.

A követelményelemzés és a szoftver architektúra választás során szinte egyáltalán nem tudjuk használni az MI jelenlegi generációját. Olyan nem-specifikus feladatokra, mint helyesírás-ellenőrzés, keresés vagy dokumentumok letisztázása, természetesen itt is alkalmazható az MI. Ám mivel a követelményelemzés általában az árazási döntésekkel van összefüggésben, az architektúra választás pedig a cég működési modelljével (például felhős üzleti modell), ezért a humán az érintettsége és érdekeltsége miatt nem delegálja a feladatot, amely „mennyiségben” (karakterszám, ábrák száma) egyébként sem túlságosan nagy, különösen, ha a téthez viszonyítjuk.

A szoftvertervezésnél, amely alapvetően egy dekompozíciós feladat azonban már más a helyzet. Egy megfelelő követelményleírás alapján az MI használható tervet ad nekünk. Ennek az az oka, hogy itt lényegében egy fordításról van szó: egy természetes nyelvi leírásról egy tervezési nyelven történő leírásra, amely minden esetben egy

formális nyelv (általában vizuális[181], de lehet szöveges is). Ezekben a feladattípusokban a hatalmas tanuló-adatbázisoknak köszönhetően az MI jól teljesít, különösen egy tapasztalatlanabb mérnökhöz képest.

Legkésőbb ezen a ponton fel kell ismernünk, hogy az MI munkakör-kiváltó képessége tudás- és készségszint függő, vagy még általánosabban, a szakmában elért szenioritástól függ. Már az előző alfejezetben bemutatott illusztrátori munkakör kapcsán is sejthetjük, hogy az ismertebb, illetve az átlag feletti képességű illusztrátorok számára nagyobb eséllyel marad munka, mint a tapasztalatlanabb és ismeretlenebbek számára, elismerve persze, hogy ez csak egy statisztikai összefüggés, ami alól mindig lesznek kivételek.

A továbbiakban ezért megkülönböztetem a junior, a medior[182] és a senior tudásszintet és ezeket külön-külön piacoknak tekintem.

Így már érthető a következő állítás: a követelmények alapján a szoftvertervezés, a terv implementációja és a tesztelés mind-mind olyan feladatkörök, amelyek esetében junior és medior szinten akár az illusztrátorok piacához mérhető felfordulást okoz az MI. Ugyanakkor a szenior piacon, ahol a nagy tétel bíró végső ellenőrzés és a legnehezebb hibák megoldása a feladat, az MI hasznossága hirtelen nagyon lecsökken. Ugyanide tartozik a karbantartás is, amely gyakran az alacsonyabb rétegek módosulása (például operációs rendszer verzió) miatti frissítési kényszert takarja.

Egyes feladatkörök, mint a konfigurációmenedzsment és az üzemeltetés alapvetően nem MI-kompatibilisek. Természetesen az a kitétel itt is igaz, hogy például a levelezés vagy a dokumentáció mindig segíthető MI-vel. De a feladat tárgyát képező konfiguráció megosztása egy felhős MI-vel nagyon rossz gyakorlat az informatikai biztonság szempontjából, helyi MI telepítése erre a célra pedig nem jellemző. Az üzemeltetés magját az ügyelet és a rendszerfelügyelet adja, amelyek a felelősségrés (lásd 7.2) elkerülése miatt tipikusan humán feladatkörök maradnak.

A minőségmenedzsment és a biztonsági feladatkörök közepesen kiválthatóak: junior szinten nagy segítség az MI, mert képes a szokásos ellenőrzéseket felsorolni és akár

[181] UML – Unified Modeling Language
[182] A szoftvermérnöki szakmai szlengben elterjedt kifejezés a közepes tapasztalattal rendelkező munkavállalókra.

kivitelezni, de a felelősség kérdése miatt a medior és főleg a szenior szinten már nem jelent nagy segítséget.

Az alábbi ábrán foglalom össze a szoftvermérnöki szakma várható átalakulását a fenti érvek alapján:

	Junior	Medior	Szenior
Követelmény elem.	N/A		
SW. architektúra	N/A		
SW. Tervezés			
Implementáció			
Tesztelés			
Üzemeltetés			
Karbantartás			
Konfiguráció			
Minőség			
Biztonság			

Figure 11: Spekuláció a szoftvermérnöki szerepek piacának az átalakulásáról

Mindez azt a kérdést nem válaszolja meg, hogy összességében több, vagy kevesebb szoftvermérnökre van szükség. Ha csak az ábrát nézzük, akkor azt látjuk, hogy a humán munkára való igény részben kiváltható, így, ha mindent változatlanul hagynánk, akkor kevesebb munkaerőre lenne szükség. Azonban az MI új szoftverfejlesztési projekteket generál, legalábbis rövid távon, amíg az 5.5-ös fejezetben leírt konszolidáció nem következik be.

Amit viszont biztosan kijelenthetünk az elemzés során, hogy a junior munkavállalók, gyakornokok számára kimondottan hátrányos az MI megjelenése, a szeniorok, és végső soron a tulajdonosok viszont potenciálisan alacsony költségű munkaerőt nyernek általa.

Ez tehát egy újabb olló, amely kinyílni látszik, amelynek ezúttal nem a technológiához való hozzáférés a tengelye, hanem a szakmán belüli előrehaladottság, a közelség a menedzsmenthez és a tulajdonosokhoz. Természetesen ez az olló sem új, de sokkal szélesebbre nyílhat az MI miatt.

Egyetlen fontos ellenerő látszik csupán a juniorok oldalán: az utánpótlásnevelés, ez esetben a szenior szoftverfejlesztőkre való igény hosszú távon. Amennyiben a szervezet ezt az igényt felismeri, akkor mesterségesen előállíthatja azokat a körülményeket, amelyek között a szükséges tapasztalat megszerezhető, még ha ez kizárólag a feladat elvégzése szempontjából nem is gazdaságos.

8. Igazságosság és méltányosság a mesterséges intelligenciában

A mesterséges intelligencia etikájának kontextusában az igazságosság és méltányosság morálfilozófiai és politikafilozófiai vitái ismétlődnek, egy fontos sajátossággal: mivel a rendszerekről, algoritmusokról tervezési időben kell döntést hozni (lásd az 1.2 és 4.2 alfejezeteket), és az MI fejlesztői kultúra az adatközpontúság miatt a kvantifikálható, mérhető tulajdonságoknak kedvez, ezért az igazságossági és méltányossági *kalkulusok* nagyobb hangsúlyt kapnak, mint a kevésbé formalizálható megközelítések.

Az algoritmusok méltányosságának kérdése, hasonlóan az algoritmusok/programok transzparenciájának problémájához (lásd a 9. fejezetet) viszonylag régen megjelent a számítástechnikában, azonban csak a kortárs MI tette forró kutatási területté.

A ’70-es években a méltányosság kérdése a „queing” területén jelent meg, amire a magyar számításelmélet nyelvezetében tömegkiszolgálás-elméletekként hivatkoznak. Ez a terület a sztochasztikus, az idő dimenzióját el nem hanyagolható módon tartalmazó matematikai modellekkel foglalkozik, mint amilyen a bejövő hívások vagy adatcsomagok modellje. A „tömegkiszolgálás” kifejezés onnan ered, hogy egy internet router vagy egy telefonközpont tömegesen kell, hogy ilyen eseményeket kezeljen, ehhez pedig méretezni kell azok kapacitását az említett matematikai modellek segítségével. A méltányosság ebben a kontextusban azzal kapcsolatos, hogy melyik bejövő esemény (hívás, csomag) élvezzen prioritást, kapjon sávszélességet. Ennek a kérdésnek akkor nem tették explicitté a szociológiai vagy filozófiai vonzatait, habár azt nem merném kijelenteni, hogy egyáltalán nem voltak tisztában ezek létezésével. Egy korai, 1973-as távközlési cikkben[183] az egyenlő elosztás és az erőforrások hatékony kihasználása közötti egyensúlyként állítják be a méltányosság kérdését, amelyet végső soron ízlés dolgának neveznek.[184]

[183] Crowther és társai (1973).

[184] „*Bármely mechanizmusnak, ami egy adott csatorna használatának felosztását vezérli több adó között, foglalkoznia kell a méltányosság kérdésével. (…) Egy lehetséges méltányossági algoritmus az, amelyben az aktuálisan legaktívabb adó úgy korlátozza magát, hogy mindig maradjon némi kihasználatlan kapacitás. Ezzel az algoritmussal, versenyhelyzet esetén egyenlő esélyekkel tudnak közölni az adók [a szabad kapacitáson az új adó képes elkezdeni forgalmazni, azután pedig a szerzők megoldása kiegyensúlyoz – HM], egyébként pedig az egyetlen aktív adó szinte a teljes kapacitást képes használni. Számos egyéb algoritmus használható, a választás nagy mértékben ízlés kérdése. (373 o.)”* fordítás a szerzőtől.

Az MI Etika jelenlegi hullámában a kérdést a COMPAS esete helyezte reflektorfénybe.[185]

A COMPAS egy, az USA ügyészeit segíteni hivatott, próbaidőre bocsátási döntést támogató szakértői rendszer.[186] A döntés egyik kritikus eleme az arra vonatkozó előrejelzés, hogy a börtönből kiszabaduló személy elkövet-e újabb bűncselekményt, azaz visszaeső lesz-e. Így a COMPAS egyik funkciója ennek az előrejelzése a betáplált paraméterek segítségével. A feladatosztály neve „kockázatelemzés” lett.

A ProPublica oknyomozó újságírói portál munkatársai 2016-ban publikálták nagy hatású mélyelemzésüket, amelyben kimutatták, hogy a rendszer, amelyet hét USA tagállamban használtak akkoriban, a próbaidő mellett az óvadék mértéknek megállapításához is, a feketék ellen részrehajló előrejelzéseket tesz. A részrehajlást az újságírók post-hoc adatokkal támasztották alá: bemutatták, hogy nem csak intuitíven volt rossz az algoritmus (pl. az egyik kiemelt esetben egy többször elítélt fehér fegyveres rablót kedvezőben ítélt meg, mint első elkövető feketét), hanem ez előrejelzés pontosságában is, azaz az esetek utóélete nem igazolta a predikciókat.

A vizsgálatot egy 7200 fős mintán végezték el, 2 éves utánkövetéssel, ami ugyanakkora időablak, amit a COMPAS fejlesztői is referenciának használtak. Ezen a mintán például a hamis pozitív arányt (FPR, lásd a 8.1-es fejezetben) vizsgálták rasszok szerint.[187]

Akkoriban az alsóház egy olyan javaslatot vitatott, amely kötelezővé tette volna a géppel segített kockázatelemzést, így, a kritikusok szerint szisztematikus rasszizmus került volna az USA igazságszolgáltatási rendszerébe.

A COMPAS-szal kapcsolatos vita[188] erősen formálta az MI méltányossági metrikákhoz való kezdeti hozzáállást: mivel a rendszer zárt, a fekete doboz rendszerekre

[185] Magyar nyelven a COMPAS esetét Vadász (2021) dolgozta fel meglehetősen részletesen.

[186] Egy ilyen rendszer az EU-ban „nagy kockázatú MI” besorolás alá esne az EU MI rendelet szerint, így itt igen erős transzparencia-kritériumoknak kellene megfelelnie, a rendszer felügyeletét ezzel nagyban megkönnyítve.

[187] A leglátványosabb az ítélet utáni két évben vissza nem esők „magas kockázattal visszaeső” besorolási hibája: a feketék esetén 45%-os hamis pozitív arány, míg a fehéreknél 23%. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[188] 38 FEDERAL PROBATION Volume 80 Number 2. False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks.”

használható méltányossági metrikák kerültek előtérbe; mivel a kategóriák társadalmi kisebbség-többség viszonyokban vannak megállapítva, és összességében a kérdés a rasszizmus jelenléte, a pozitív és negatív értékelés alapja adott (szerencsére senki nem érvelt amellett, hogy a rendszer rasszista ugyan, de ez jó). Mivel a rendszer klasszifikál, ezért a klasszifikációs feladatok kerültek a fókuszba, ami a gépi tanulás egyik fontos alkalmazási területe. A 2020-as évekre a vita szofisztikálódott: nem csak fekete doboz rendszereket vizsgálnak, felismerik, hogy egyes kérdésekben lehetnek legitim és ártalmatlan értékkülönbségek; felmerül a méltányosság kérdése a klasszifikációtól eltérő esetekben is (pl. csoportosításon alapuló ajánló algoritmusok, generatív MI).

Az alábbi alfejezetben összegyűjtöttem, hogy milyen alapkifejezéseket használnak az MI etikájának formalizált altémáiban. Ezeknek az alapja a gépi tanulás statisztikából eredeztetett kulisszanyelve. A gépi tanulás során használt metrikák tartalmilag a matematikai statisztika egyes megfogalmazásaival ekvivalensek, csak éppen olyan rövidítéseket tartalmaznak, amelyek a beavatatlan szem számára megnehezítik a hozzáférést. Mindazonáltal ezek a terminus technicusok roppant egyszerű koncepciókat takarnak.

8.1 Formalizált méltányosság

Minden, amit itt bemutatok, olyan algoritmusokra vonatkoztatható, amelyek bináris osztályozást hajtanak végre; továbbá az is szükséges előfeltevés, hogy eldönthető, melyik besorolás a kedvezőbb. A gépi tanulásban általában a pozitív eset a vizsgált tulajdonság fennállását jelenti, amely nem biztos, hogy kívánatos (lásd: HIV pozitív státusz). Az MI etika irodalmában a bevett jelölés, hogy a pozitív eset a vizsgált tulajdonság fennállását jelenti, ami egyben kívánatos is (például felveszik az egyetemre, megkapja a hitelt vagy a munkalehetőséget), a negatív pedig a nem kívánatos (elutasítják a jelentkezést, a hitelkérelmet, álláspályázatát). Habár a két jelölésmód keveredése nyilvánvalóan félreértésekhez vezethet, különösen az orvosi területen, a kontextus miatt ez ritkán következik be, matematikailag pedig a formulák nem érzékenyek a különbségre.

Az alábbi táblázat az MI etikai jelölést alkalmazza, tehát a pozitív egyben valami kívánatos is. A rendszer nem alkalmas arra, hogy radikálisan eltérő szubjektív

preferenciák megjelenjenek benne (például nem lehet olyan felhasználó, aki kimondottan azt szeretné, hogy elutasítsák a hitelkérelmét), ám a logikai rés sok alkalmazás esetén elhanyagolható probléma.

Az első oszlop a bináris osztályozás tényét írja le. Létezik egy teljes halmaza az algoritmus kimeneteinek, tehát a bináris döntéseknek. Ez két részhalmaz uniója: a pozitív és negatív döntések összege.

A második és harmadik oszlop teteje egy újabb, igen erős és korlátozó előfeltevésre hívja fel a figyelmet: a konceptuális rendszer ugyanis felteszi, hogy az algoritmustól függetlenül ismerjük, hogy mely esetek „valóban" pozitívak és negatívak. Ez szintén a gépi tanulás logikájából következik: ott ismerünk úgynevezett „ground truth" információt, vagy filozófiai nyelven kifejezve igaznak elfogadott premisszákat.

A gyakorlatban a gépi tanulás történetében a ground truth az emberi kategorizációt jelentette. Például emberek egy csoportja eldönti, hogy a képen macskát vagy tűzcsapot látunk-e, és ezt az információt igaznak tételezzük fel. Kellően sok, például ezer kép birtokában a rendszerünket a halmaz kétharmadán, pl. 667 képen tanítjuk, és a maradék 333 képet sohasem mutatjuk meg neki előzetesen. Miután a modell tanítása végbement, megkérjük, hogy a 333 képet kategorizálja, természetesen nem elárulva az emberi, ground truth kategorizációt. A generált kategóriák és az igazként elfogadott kategóriák összevetésével kapjuk meg a következő értékeket:

TP valós pozitív (pozitívnak kategorizált és valóban pozitív);
FP hamis pozitív (pozitívnak kategorizált, de valójában negatív): elsőfajú hiba a statisztikai hipotézisvizsgálat világából;
TN valós negatív (helyesen negatívnak kategorizált);
FN hamis negatív (helytelenül negatívnak kategorizált):másodfajú hiba.

Ha ezeket az értékeket elosztjuk az elfogadott ground truth értékekkel, akkor százalékos pontosságot kaphatunk, pl. pozitív találat arányt *(TPR) = TP/P*; ha arra vagyunk kíváncsiak, hogy milyen hatékonysággal „kapja el" a negatív eseteket, akkor *TN/N = valós negatív arány, hamis pozitív arány (FPR) = FP/N*.

Ezekkel az eszközökkel vizsgálható például az *esélyegyenlőség (equal opportunity)* formalizált változata. Az MI méltányosság területén jelenleg bevett – erősen

leegyszerűsítő – felfogásban meg tudunk határozni egy privilegizált (szokásos jelölése „1”) és egy nem privilegizált csoportot („0”); például 1= az ország területén született; 0 = külföldön született, bevándorló. Ekkor tovább bontva a fenti halmazainkat e tulajdonság alapján megvizsgálhatjuk, hogy teljesül-e $TPR_1 \sim= TPR_0$ (egy adott hibahatárt megengedve, mert egyébként az utolsó tizedesjegyig egyezniük kellene, amit bizonyos arányoknál lehetetlen). Ez azt jelenti, hogy a valóban alkalmas bevándorló jelöltnek ugyanakkora az esélye pozitív besorolást kapni, mint a nem bevándorló jelöltnek.

Ez azonban még tovább javítható: attól, hogy az esély a pozitív besorolásra hasonló, elképzelhető, hogy a hamis pozitív besorolás a privilegizáltnak kedvez. Ennek kiküszöbölésére a *kiegyenlített esélyek* (*equalized odds*) formuláját hozták létre: legyen *igaz, hogy* $TPR_1 \sim= TPR_0$ *és* $FPR_1 \sim= FPR_0$. A második tag azért nagyon jelentős, mert tipikusan véges erőforrás osztható szét, így az érdemtelenül pozitívnak jelölt privilegizált egy érdemes nem privilegizált elől veheti el a helyet.

Más szemszögből nézve az itt leírtakat *csoport méltányosságnak* is nevezik, abból a megfontolásból, hogy statisztikailag értelmezhető adatmennyiségre van szükség hozzájuk és hogy egy adott egyéni eset méltányossága a befoglaló csoport egészével szembeni méltányosságból vezethető le.

Láthatjuk tehát, hogy az irodalom, ha szó szerint fordítjuk, megkülönbözteti az esélyegyenlőséget (equal opportunity) és a kiegyenlített esélyeket (equalized odds), ami viszont az esélyegyenlőség kondíciója egy további szigorítással. Úgy hiszem, hogy amikor magyar nyelven általánosságban esélyegyenlőségről beszélünk, akkor inkább az utóbbit értjük alatta, még akkor is, ha az angol szakirodalom alapján ez félrefordítás lenne. Az angol irodalomban használt equalized odds-nak azonban létezik egy másik definíciója is, amely segít tisztázni a helyzetet: az, hogy a vizsgált paramétertől (a fenti példában a születési hely) független a kimenet. Ezen állapot ellenőrzésére kvázi-oksági megközelítés is létezik, amelyet lentebb tárgyalok.

Talán már ezen a pontok is könnyen látható, hogy mennyire limitáltak ezek a formulák: teljes egészében egzakt alapigazságok feltételezésére építenek; nyilvánvaló az is, hogy a bináris besorolhatóság előfeltétele a képlet használhatóságának.

Bizonyos esetekben ez teljesen elegendő is. Például 2015-ben virálissá vált a rasszista szappanadagoló története: az interneten könnyen fellelhető 2015-ös videó tanulsága szerint az atlantai Marriott hotelben egy szappanadagoló nem ismerte fel az afroamerikai vendégek kezét, csak a fehér bőrűekét, így az előbbi csoport képtelen volt megfelelően kezet mosni. Itt világos, hogy $TPR_1 > TPR_0$ eset áll fenn, tehát megsérti az esélyegyenlőség elvét a rendszer (a hamis pozitívak – például az, hogy egy repülő rovar aktiválja az adagolót – itt nem relevánsak).

Teljes halmaz (összes)	Pozitív (**P**) eset	Negatív (**N**) eset	Gyakoriság $\frac{\sum P}{\sum összes}$	Pontosság $ACC = \frac{\sum TP + \sum TN}{\sum AlapHalmaz}$
Pozitív Döntés (**PD**)	Valós Pozitív (**TP**)	Hamis Pozitív (**FP**) *	pozitív pontosság arány $PPV = \frac{\sum TP}{\sum PD}$	hamis felfedezés arány $FDR = \frac{\sum FP}{\sum PD}$
Negatív Döntés (**ND**)	Hamis Negatív (**FN**) †	Valós Negatív (**TN**)	helytelenül kimaradt arány $FOR = \frac{\sum FN}{\sum ND}$	negatív prediktív érték $NPV = \frac{\sum TN}{\sum ND}$
	Poz. találat (recall) arány $TPR = \frac{\sum TP}{\sum P}$	Hamis poz. arány $FPR = \frac{\sum FP}{\sum N}$	$LR+ = \frac{TPR}{FPR}$ ‡	diagnosztikai esélyek $DOR = \frac{LR+}{LR-}$
	Hamis negatív arány (miss) $FNR = \frac{\sum FN}{\sum P}$	Valós negatív arány (jól szűr) $\frac{\sum TN}{\sum N}$	$LR- = \frac{FNR}{TNR}$ §	$F_1 = \frac{2}{TPR^1 + PPV^1}$

Figure 12: Az MI részrehajlásmentesség témakörében használt jelölések

8.2 Kvázi-oksági megközelítés a méltányossághoz

A méltányosság (általam) kvázi-okságinak nevezett megközelítése adott esetekben kiegészítheti a statisztikai megközelítést, más esetekben alternatívát kínálhat ahhoz.

A kvázi-oksági megközelítés lényege, hogy az ok-okozati viszonyok tranzitív kapcsolataira leképezve az adott döntést, bizonyos faktorok (például rassz, nem, származás) ne bírjanak oksági erővel.

Az egyik legegyszerűbb megközelítés a kérdéses faktorok teljes eliminálása azon paraméterek közül, amelyekez a rendszer figyelembe vesz. Ezt olykor *méltányosság a tudatlanság által (fairness through unawareness)* technikának is nevezik.

A magyar felsőoktatásban például egy jó gyakorlatokat követő oktató egy zárthelyi dolgozaton csak NEPTUN kódot kér, de nevet nem. Ezzel elkerüli az adatok felesleges tárolását, de megvalósíthatja a méltányosságot a tudatlanság által: ha megengedjük azt a könnyű előfeltételt, hogy az oktató nem emlékszik a hallgatók NEPTUN kódjaira fejből, és hogy közben nem is néz ennek utána, akkor vita esetén az oktató meggyőzően érvelhet amellett például, hogy nem diszkriminál a hallgató neme alapján, mivel a zárthelyi dolgozatból az nem is derül ki.

Ennek a helyzetnek a kvázi-oksági leírása az, hogy a hallgató neme nem lehet része az oktató döntési láncolatának, hiszen nem is ismeri azt. Természetesen ehhez jár egy rejtett előfeltevés is: az, hogy „proxy” változók nem közvetítik ugyanazt az információt (pl. az oktató a hallgató nevének hiányában nem következtetheti ki a hallgató nemét, ám az íráskép alapján, tudománytalan módon, de mégis valószínűsíti).

Az értékelt egyén tulajdonságainak nem-ismerete tehát különféle tulajdonságok esetén eltérő védelmet biztosít a diszkrimináció ellen.

Az kvázi-oksági modell előnye az, hogy akár kis számú döntésre is felépíthető: egyetlen vitás eset is potenciálisan kezelhető, míg a statisztikai metrikákhoz megfelelő minta és referencia, „ground truth” szükséges. Az oksági modell azt ígéri, hogy ezek hiányában is működhet – természetesen addig, amíg az oksági összefüggéseket nem statisztikai tesztekből vezetjük le, hanem attól független módszerrel állítjuk elő. Cserébe egy fontos hátránnyal is rendelkezik a statisztikai módszerekhez képest: míg a statisztikai méltányossági metrikák fekete doboz rendszerekre is használható, mert csak a bemenet-kimenet párosítások szükségesek hozzájuk[189], addig a kvázi-oksági modellekhez szükséges a rendszer belső reprezentációjának ismerete is.

[189] Ugyanakkor Baer és társai (2019) megmutatták, hogy intuitíven méltányos és nem méltányos rendszerek kimenetei adott körülmények között identikusak lehetnek.

8.3 Problémák a formalizált méltányossággal

A formalizált méltányossági elvek számos nehézséggel néznek szembe. Az egyik, hogy gyakran nincs referencia („ground truth”) vagy nem egyértelmű a referencia. Például egy művészeti pályázat esetén, vagy egy PhD felvételin tett benyomások pontozásánál nem szükségszerű, hogy az alapigazság ennyire problémamentes, mert a humán értékelő adott pillanatban meghozható ítéletét sem lehet referenciaként tekinteni, az pedig utólag nem derül ki, hogy a fel nem vett hallgató mégis bevált volna.

De ennél is nagyobb probléma – az én értékelésem szerint – a *tervezési idő kihívás* (lásd 1.2, 4.2 fejezeteket). Míg a döntőbizottságoknál lehetséges előre rögzíteni bizonyos pontozási szabályokat, másokat pedig a bizottságra bízni (természetesen általános útmutatás kíséretében), az MI esetén mindent a tervezési időben kell eldönteni. Így azután az MI bevezetés *egy kikényszerített, korai formalizálás és explikálás*, annak érdekében, hogy az algoritmus részrehajlásmentessége ellenőrizhető legyen a piacra lépés előtt. Ez pedig egy olyan – törvényben előírt – túlzott ambíció, ami káros is lehet a megvédeni kívánt döntési folyamatokra.

A legmélyebb probléma pedig abból adódik, hogy a formalizált algoritmusokkal egy részben hallgatólagos értékelési képességet próbálunk explicitté tenni és uniformizálni. Egy ilyen törekvés inherens akadályait a 12. fejezetben, a poszt-kritikai megközelítés bemutatásánál írom körül.[190]

8.4 Elmozdulás a morálfilozófia felé

Az előző néhány oldalon láthattuk, hogy a statisztikai hipotézistesztelés matematikai eszköztárát kiegészítve kellően kényelmes értéktételezésekkel és előfeltételezésekkel, hogyan lehetséges kezelni az esélyek kiegyenlítettségét.

[190] Azt viszont semmiképp nem állítom, hogy a gépek ne rendelkezhetnének hallgatólagos képességekkel(lásd Héder 2014). Azonban a hallgatólagos tudással rendelkező mesterséges ágensekre az alkotók nem tudnak teljes mélységben determináló algoritmust fejleszteni a saját, géphez képest külső szemszögükből.

Erre a matematikai eszköztárra, néhány hasonló szellemben megalkotott kiegészítéssel, amiket itt nem tárgyalok, szoftvercsomagokat építettek, amelyek egy kipipálható teljesítményindikátorrá tették a méltányosságot.[191]

Nem meglepő, hogy ezeknek az egyszerű metrikáknak a puszta jelenléte, és még inkább a furcsa kimenetei egy újabb hullámot indítottak az algoritmikus méltányosság kutatási területén, amelynek a lényege a morálfilozófia kapcsolódó kérdéseinek a(z) (újra)felfedezése, kontextualizálása az MI számára.

Az MI etika területe Bernard Williams harcos-társadalom gondolatkísérletét[192] fedezte fel magának a méltányossággal kapcsolatos problémák tárgyalására, Joseph Fishkin 2014-es könyvének[193] közvetítésében. Érdekességét az MI etika számára az adja, hogy ez direkt kritikája az esélyegyenlőség képleteinek.

Williams arra kér minket, hogy egy olyan társadalmat képzeljünk el, amiben alapvetően két kaszt létezik: a harcosoké és a nem-harcosoké. A harcosok privilegizáltak, amennyiben jobb szállást, erőforrásokat és fizetést kapnak. Ugyanakkor a harcosok kasztjában limitált a helyek száma; ezekre egy meritokratikus kiválasztási folyamaton átjutva lehet bekerülni. A kiválasztási folyamat egy több napos felmérés, amin 16 évesen esnek át a fiatalok, ami a harc szempontjából elismerten releváns fizikai és kognitív kihívásokat tartalmaz; ezek úgy vannak megalkotva, hogy a csalás lehetősége kizárt és mindenkit a mutatott teljesítménye alapján ítélnek meg.

A kérdés az, hogy megvalósítja-e ez a kiválasztási folyamat az esélyegyenlőséget? Williams a döntési pont atomisztikusságának potenciális igazságtalanságára hívja fel a figyelmet. Plauzibilis, hogy a harcos szülők gyermekei különféle előnyökkel rendelkeznek felnevelkedésük során: jobb minőségű ételeket fogyasztanak, nagyobb hozzáférésük van az edzésekhez, valamint a harc szempontjából releváns stratégiai és taktikai tudáshoz.

[191] IBM Fairness 360 (Bellamy és társai 2019). A formalizált megközelítés korlátaról részletesebben (lásd Jain és társai 2024). Ők egy köztes megoldást kínálnak a problémára: algoritmikus pluralizmust hirdetnek, amelynek az a lényege, hogy többféle megközelítésű algoritmust használjanak egy megoldás méltányosságának elemzésére. Ez persze az algoritmusok közötti választás problémáját állítja elő, ám ezt a választást legalább már többféle szempont megismerése után lehet meghozni.
[192] Williams (1962).
[193] Fishkin (2014).

Fishkin, Williams nyomán arra jut, hogy a *fejlődési lehetőségek*ben felfedezhető különbségek lehetetlenné teszik a méltányos versenyt a kiválasztási folyamatokban. Ezek a különbségek a fejlődés során nagyon hamar megjelennek és *szűk keresztmetszetekhez* vezetnek.

Ezek a szűk keresztmetszetek teljesen nem eliminálhatóak. Fishkin módosítja a gondolatkísérletet úgy, hogy minden gyermek a születésénél fogva férjen hozzá a közfinanszírozott és minden gyermek számára hozzáférhető harcos-akadémiához, amely kiegyenlíti a családi anyagi státuszból adódó lehetőségeket. Természetesen ekkor is arra az általánosan elfogadott konklúzióra jutunk, hogy a fejlődési lehetőségek egy része determinált és nem kiegyenlíthető (pl. genetika).

Fishkin ezen szűk keresztmetszetek eliminálására létrehoz egy új, teljes életút alapú esélyegyenlőség-elmélet (szemben pontszerű a döntésekkel), amely ezen szűk keresztmetszetek eliminálására törekszik. Több megközelítést próbál meg: eltérő fejlődési útvonalak elismerését javasolja egyetlen pontszerű vizsga helyett; de felveti azt is, hogy a lehetőségek szerkezete változzon meg (ne csak két kaszt működjön és ne csak a harcosok legyenek privilegizáltak).

Természetesen ez újabb kérdéseket vet fel: elképzelhető-e, hogy a harcosok hasznossága a közösség fennmaradása tekintetében olyan kritikus, hogy az erőforrások más kasztok irányába való eltérítése egzisztenciális kockázat? Hogyan viszonyul mindez az ugyanebből a korszakból származó, versengő elméletekhez, mint Rawls *igazságossgág mint méltányosság*[194] elmélete?

A legrelevánsabb kortárs konferenciák előadásaiból láthatjuk, hogy összességében a meta-etika MI ipar általi felfedezése zajlik, annak ismert dilemmáival és divergens előfeltételeivel egyetemben. Mindez holisztikusabb, a méltányosság-metrikák egyszerűségével szembeni kritikusabb világképhez vezet.

Mindez azért érdekes témánk szempontjából, mert az individuális életút esélyeit kiegyenlíteni kívánó, folyamatközpontú megközelítés nem szükségszerűen kompatibilis a méltányossági metrikákkal és ami még fontosabb, nehezen kezelhető a kulcsfontosságú kihívásaink figyelembe vételével: a *konstruktőr* dekompozíciós

[194] Rawls (1971).

feladata során meghozni szükséges döntésekről nehéz megmondani, hogy segítik-e vagy gátolják a méltányossággal kapcsolatos szempontok érvényesülését; mindezt bonyolítja a tervezés során tapasztalható ismerethiány, az episztemikus szakadék a tervező és a mesterséges ágens működési kontextusa között.

9. Az MI transzparencia története és jövője

1984-ben egy lenyűgöző kutatási jelentést publikáltak az Applied Mathematical Modelling című folyóiratban.[195] Szerzője, W. E. Cundiff megpróbálta kiterjeszteni az APL (az „A Programming Language” nevű programozási nyelv) felhasználói körét azon matematikusokon túl, akik eleve jó számítógépes ismeretekkel rendelkeztek. Hipotézise az volt, hogy ennek érdekében az *algoritmikus transzparencia* megteremtése egy jó eszköz lehet.

Az APL nyelv különösen jól illeszkedett ehhez, mivel már eleve egy „üvegdoboz”[196] megközelítéssel lett kialakítva. Ez azt jelentette, hogy a függvények belső működése a jelölésükből meglehetősen egyértelmű volt a felhasználó számára. Kenneth Iverson, a nyelv megalkotója, ezt a jellemzőt „szuggesztivitásságnak” nevezte, és ezt a tervezési elvet a Notation as a Tool of Thought (A jelölés mint gondolkodási eszköz) című Turing-díjas előadásában magyarázta el.[197] Az ötlet az, hogy az APL jelölési stílusa ösztönzi a felhasználókat, hogy különböző variációkat és felhasználási módokat próbáljanak ki; ez pedig csak azért lehetséges, mert ezek a jelölések nem „fekete dobozok”.

Cundiff ezt az ötletet vitte tovább az algoritmikus transzparencia megteremtésével, amely által egy még inkább „érthető” APL használati stílust fejlesztett ki. Bár az „algoritmikus transzparencia” korai használata látszólag nem befolyásolta a mai mesterséges intelligencia etika jelenlegi transzparencia vitáit, érdekes módon lényegében ugyanarról az ötletről van szó, csak más célra alkalmazva.

Az 1984-es cikk remek emlékeztető arra, hogy a számítógépes rendszerek megértésének szükségessége, amelyek könnyen „fekete dobozokká” válhatnak, nem új keletű igény. Azt is megmutatja, hogy a „transzparencia” fogalmát már az MI történetében korán episztemikus értelemben használták, ebben az esetben a megértés kifejezésére. Ahogy látni fogjuk, a mesterséges intelligenciában a „transzparencia” ma az episztemológiai és etikai erényeket összekötő kapocsként szolgál.[198]

[195] Cundiff (1984).
[196] „Glass box”.
[197] Iverson (1980).

9.1 Megmagyarázhatóság és számításelmélet

Bár a magyarázhatóság a közelmúltban lett különösen vitatott kérdés/terület a mesterséges intelligenciában, korántsem új témakörről van szó. Már az 1970-es és 1980-as években is, amikor az MI sikerét a szakértői rendszerek testesítették meg, a konstruktőrök arra törekedtek, hogy olyan megoldásokat fejlesszenek, amelyek képesek magyarázatokat adni saját működésükre. A szakértői rendszerek belső felépítése ezt a törekvést támogatta is, és itt van a kritikus különbség a múlt és a jelen között. Míg a szakértői és diagnosztikai rendszerek esetében a magyarázatok elsősorban episztemikus célt szolgáltak – ahogy ezt maga a kifejezés is sugallja –, addig a modern diskurzus a magyarázhatóságról túlnyomórészt az etikai megfontolásokra helyezi a hangsúlyt. E fejezetben azonosítom a magyarázhatóság szerepének néhány változását a számítástechnika történetében, bemutatva, hogy hogyan alakult ki ez a szinergia az episztemikus és etikai megfontolások között.

9.1.1 A megmagyarázhatóság korai története

1973-ban Anthony Gorry[199] azt kutatta, hogy a szakértői rendszerek hogyan javíthatják a klinikai döntéshozatalt. E rendszerek kimenetének megmagyarázhatóságát az ilyen programok fejlesztésének egyik fő kérdéseként azonosította. Véleménye szerint:

„Ha a szakértők szeretnék használni és közvetlenül javítani a programot, akkor annak képesnek kell lennie számot adnia cselekedeteinek indokairól. Továbbá, ennek a magyarázatnak olyan módon kell történnie, hogy azt az orvosok megértsék. A dedukció lépéseit és a felhasznált tényeket azonosítani kell a szakértő számára, hogy szükség esetén korrigálhassa egyiket vagy másikat. Emellett a felhasználó számára könnyen hozzáférhetőnek kell lennie, hogy a program mit tud egy adott témáról.” (33. oldal)[200]

Visszatekintve úgy tűnhet, mintha Gorry egy az egyben azt a tudományt írná le, amit ma XAI-nak (magyarázható mesterséges intelligenciának) nevezünk. Azonban nem szabad túlbecsülni az idézet jelentőségét. Valójában, amikor Gorry és kollégái 1973-ban bevezettek egy magyarázható szakértői rendszert, amelyet akut veseelégtelenség

198 A megmagyarázhatóság korai történetének kutatásában Alessio Tartaroval működtem együtt, akinek hatalmas köszönettel tartozom a motiváló közös munkáért.
199 Gorry (1973).
200 Fordítás: HM.

diagnosztizálásának támogatására terveztek[201], a magyarázhatóság előállítása jóval kisebb kihívás volt. A program tartalmazott magyarázatokat a program szándékáról, a különféle kérdésekre adott válaszokat angol nyelven adták meg, de ezeket a funkciókat kézzel építették hozzá a rendszerhez (a tudásbázis mérete és jellege ezt még lehetővé tette) és a „a használat megkönnyítését” szolgálták.

A következő évtizedekben a magyarázhatóság csupán arra szolgált, hogy megkönnyítse a program használatát. Egy kezdetleges módszerrel valósították meg – előre elkészített szöveggel, amely magyarázat-sablonokat adott, egy olyan technikával, amit „konzerv” szövegnek (canned text) neveznek.[202] A konzerv szöveg olyan előre megírt angol szöveget jelent, amely a programon belül van tárolva, és szükség esetén magyarázatként jeleníthető meg. Ez a megközelítés lehetővé tette a számítógép számára, hogy kérdésekre válaszoljon tevékenységeiről azáltal, hogy bemutatta a megfelelő előre elkészített magyarázatokat.

A konzerv magyarázatokat gyakran hibajelzések formájában használták. Mára ez a működési mód annyira alapvetően elvárt, hogy nem sorolnánk a megmagyarázhatósági technikák közé. Például, ha egy felnőttek kezelésére tervezett orvosi programot csecsemőkre alkalmaztak, akkor egy előre elkészített üzenetet jelenített meg, mint például „A páciens túl fiatal ahhoz, hogy ezt a programot alkalmazzuk” valamiféle programhiba helyett. A hibajelzéseket forráskódba illesztett megjegyzésekkel egészítették ki, a programozó a program specifikus részeit összekapcsolta a megfelelő előre elkészített szöveggel, és minden egyes részhez magyarázatot írt arról, hogy az adott rész mit csinál. Amikor a felhasználó meg akarta érteni, hogy mi történik a programban, a számítógép megjelenítette az adott tevékenységhez vagy folyamathoz kapcsolódó konzerv szöveget – mindez ma már a szoftverfejlesztés alapvető része.

A megmagyarázhatóság ezen első megközelítése még mindig velünk van, minden technológiailag érett szoftverben, amit használunk. Minden elterjedt programozási nyelv rendelkezik konzervszöveg funkcióval és keretrendszerekkel a többnyelvű felületek kezelésére. Az történt, hogy ez annyira alapvető követelménnyé vált, hogy csak más ’70-es évekbeli szoftverekkel összehasonlítva számít újdonságnak, amelyek elvárták a felhasználótól, hogy megtanuljon kódolni.

[201] Gorry és társai (1973).
[202] Swartout (1985).

Annak ellenére, hogy Gorry magyarázhatósági megközelítése kezdetleges volt mind a magyarázatok célját (használhatóság), mind a megvalósítás módját tekintve (előre elkészített szövegek), az a megérzése, hogy a szakértői rendszereknek magyarázó képességekkel kell rendelkezniük, nagyon előrelátónak bizonyult. Nem sokkal később Swartout[203] bemutatta az OWL Digitalis Advisor-t, egy olyan klinikai támogató szoftvert, amely terápiával kapcsolatosan ad tanácsot az orvosoknak, illetve Weiner[204] bevezette a BLAH nevű rendszert, amely magyarázza a saját következtetéseit, és amelyet az adóbevallások összeállításának támogatására használtak.

Swartout és Weiner megközelítései a magyarázhatósághoz sokkal kifinomultabbak voltak mind a magyarázatok célját, mind azok megvalósítását tekintve. Swartout úgy tekintett a rendszer magyarázatképességére, mint egy módszerre az elavult vagy pontatlan programdokumentációk javítására. Egy magyarázható szakértői rendszer egy „ön-dokumentáló” megoldásként tekinthető,[205] amely képes elmagyarázni, hogyan működik, még akkor is, ha a programozó elfelejtette biztosítani a megfelelő információkat a technikai dokumentációban. Ez nem állt távol Gorry koncepciójától, miszerint a magyarázhatóság a használhatóság javításának egy módszere, de technikailag fejlettebb, mert a forráskódot fordítja magyarázattá.

Emellett, a bevezetőben említett APL is hasonló elveket követett. Swartout azonban három további funkciót is hozzáadott a magyarázatokhoz.

Először is, a magyarázatok javítják a felhasználó általi elfogadást azáltal, hogy a program képes elmagyarázni a következtetési folyamatait, ezzel megnyugtatva a felhasználókat azzal kapcsolatban, hogy logikus következtetéseket von le, és észszerű megoldásokra jut. Másodszor, a magyarázatok oktatási célokat is szolgálnak, lehetővé téve a diákok vagy gyakorló szakemberek számára, hogy összehasonlítsák saját gondolkodásukat a rendszerével, ezáltal a rendszert az adott témakör jobb megértésének eszközeként használják. Végül, a magyarázatok hasznosak a hibakeresés során is. A magyarázat funkciója lehetővé teszi a programozók számára, hogy hatékonyan azonosítsák és megoldják a rendszer problémáit. Ezt a három további funkciót Weiner is átvette, aki szintén úgy vélte, hogy a magyarázatok növelik a

[203] Swartout (1977).
[204] JL Weiner (1980) – nem azonos a nagy kibernetikus Norber Wienerrel.
[205] Swartout (1977), 89. oldal.

szakértői rendszerek hitelességét, oktatási célra használhatók, és hasznosak a hibakeresésben.

A magyarázatok funkciójának változása jelentős hatással volt arra, hogy technikailag hogyan valósították meg ezeket. Valójában bármilyen követelmény, amely túlmutat a dokumentáció és az alapvető használhatóság célján, elégtelenné tette a konzerv szövegeket. Azok a kérdések, amelyeket a felhasználó egy szakértői rendszertől meg szeretett volna kérdezni, messze túlmutattak azon, amit a programozó előre megjósolhatott volna. Következésképpen alternatív műszaki megoldásokra volt szükség, amelyek elsődleges feladata az lett, hogy a szakértői rendszer által követett érvelést a felhasználó számára érthető formátumban közvetítsék.

A Swartout által kifejlesztett Digitalis Advisor rendszer a magyarázhatóságot a program kódját és annak végrehajtási láncát visszafejtve valósította meg. Az ötlet az volt, hogy a programot strukturált, angol nyelvhez közeli speciális kódolási konvenciók mentén írják meg, amely könnyen érthető. Ily módon a magyarázatokat közvetlenül a kódból lehetett levezetni, automatikusan frissítve azokat, amikor a program módosításakor. A könnyen olvasható, közel angol nyelvű kód koncepciója természetesen ma is jelen van olyan programozási nyelvekben, mint például a Python. Az újszerű gondolat itt az volt, hogy a program kódját a végfelhasználónak kellett interpretálni. Az eljárás egyik belső korlátja, hogy csak egy nyelven (angolul) volt elérhető, és hogy a szoftver gyakorlatilag nyílt forráskódúvá vált, ami önmagában jó dolog és a ma nagy előnyt jelent még profitorientált cégek termékei esetén is, azonban akkoriban üzleti hátrányként tekintettek erre.

A Digitalis Advisor célja az volt, hogy a terápia előírásakor alkalmazott módszerek szerkezetét úgy jelenítse meg, hogy az MI döntéshozatali folyamata összhangban legyen az emberi szakértőével. Ennek megfelelően a rendszert „absztrakciós szintekre” osztották. A magasabb szintű eljárások általánosabb tevékenységeket vagy célokat képviseltek, és szükség esetén hívták meg a specifikusabb eljárásokat. Az ilyen szoftverek ma is hasonlóan működnek, de ebben az esetben szükség volt az MI döntéshozatali folyamatának a végfelhasználói koncepciókkal való összehangolására is. Például egy olyan eljárás, mint a „TERÁPIA MEGKEZDÉSE” hívhatott specifikusabb eljárásokat, mint például „ÉRZÉKENYSÉGEK ELLENŐRZÉSE”, ami tovább hívhatott még specifikusabb eljárásokat, mint például a „KÁLIUM MIATTI

ÉRZÉKENYSÉG ELLENŐRZÉSE”. Ez a többszintű szerkezet segített átfogó képet adni anélkül, hogy túlterhelte volna a felhasználót a részletekkel.

Amikor a Digitalis Advisor elmagyarázta érvelését a felhasználónak, egy magas szintű absztrakcióval kezdett, csak a legáltalánosabb tevékenységeket mutatva. Ha a felhasználó többet szeretett volna tudni, kérhetett további részleteket, és az Advisor ekkor a következő szintű eljárásokat magyarázta el. Ez lehetővé tette, hogy a rendszer olyan magyarázatokat nyújtson, amelyek igazodtak a felhasználó által kívánt részletezettségi szinthez. Azonban a rendszer magyarázati szintjei adottak voltak, akkor alakították ki ezeket, amikor a programot megírták. Ha egy felhasználó túl általánosnak találta az egyik szintű magyarázatot, és túl specifikusnak a következőt, akkor nem volt mód arra, hogy egy köztes szinten magyarázatot kapjon, mivel a programban ez nem volt beépítve.

Egy másik hátrány az volt, hogy hiányzott egy felhasználói modell, amely figyelembe vette volna a felhasználó előzetes tudását vagy célját. Ennek következtében, függetlenül attól, hogy a felhasználó milyen szintű ismeretekkel rendelkezett, vagy mi volt a célja (például tanulás vs. hibakeresés), mindenki ugyanazt a magyarázatot kapta. Emiatt előfordult, hogy a rendszer olyan információkat jelenített meg, amelyeket csak a rendszer karbantartóinak kellett volna látniuk, nem pedig az orvosoknak. Ezek ellenére a Digitalis Advisor magyarázó képessége hasznosnak bizonyult a program fejlesztése során. A hibák nyilvánvalóbbak voltak, amikor angol nyelven voltak kifejezve, és az orvosok hasznosnak és érthetőnek találták a magyarázatokat.

Weiner BLAH rendszere egy érvelő komponenst és egy magyarázatgenerátort tartalmazott. Az érvelés után a kimenet egy fa volt, amely egy állítást és annak alátámasztását képviselte. A magyarázatgenerátor ezután ezt az érvelési fát szöveggé alakította, előre elkészített sablonokat használva minden állításhoz, így a magyarázatok „A, mert B és C” formában készültek.

Ez a grafikus magyarázat egyértelmű lépést jelentett a szöveges magyarázatoktól a vizuálisabb magyarázatok felé. Ez szükséges volt, mivel a korabeli technológia mellett az előre elkészített szövegek variációinak száma túlzottan megnövekedett volna, és ezáltal kezelhetetlenné válik. Feltételezhetjük, hogy ez az 1980-as fejlődés előrevetíti a

kortárs magyarázható mesterséges intelligencia hasonló trendjét, például a „kiugrási térképeket”,[206] amelyek szintén kizárólag vizuálisak.

Ugyanakkor nem kerülhetjük el a gondolatot, hogy a modern, nagy nyelvi modellek (LLM) technológiájával ez az architektúra újragondolható lenne-e. A döntés meghozatalát egy összetett szakértői rendszerre lehetne bízni, ha megfelelően meg lehetne építeni egy ilyet. Az ilyen, deklaratív logikai szabályokkal működő, következtetés alapú vagy értelmező kód alapú rendszerek előnye, hogy formálisan bizonyítható tulajdonságokkal rendelkeznek, így mindig a kívánatos határokon belül maradnak. Ezt követően egy LLM vizsgálhatná az inputot és az első rendszert, és természetes nyelvű magyarázatot nyújthatna. A kihívás természetesen az, hogy ez a második rendszer ne generáljon „hallucinációkat”.

A BLAH által előállított magyarázatok olyan részletesek voltak, amennyire az az érvelési fa, amit a magyarázat szöveges formába próbált önteni. Azonban ez a módszer a magyarázatokat csak azok számára tette érthetővé, akik képesek voltak értelmezni a program kódját, annak ellenére, hogy angol nyelven készültek, mert kondicionális elágazásokat tartalmaztak.

Az angol nyelvű konzerv szövegrészek, logikai fában voltak összekapcsolva. A BLAH olyan funkciókat is tartalmazott, amelyek lehetővé tették, hogy a magyarázatok jobban igazodjanak a felhasználó igényeihez. Például a felhasználó dönthetett úgy, hogy eltávolítja a magyarázatból azokat az információkat, amelyeket már ismert, és különböző magyarázati struktúrákat is választhatott az igényeinek megfelelően. Ez azért volt lehetséges, mert a BLAH érvelési komponense ugyanazon állításra különböző érvelési fát tudott előállítani, amelyeket aztán különböző magyarázatokká fordítottak.

A BLAH még ennél is kifinomultabb magyarázati formákat is tartalmazott. Például képes volt olyan magyarázatokat adni, amelyek alternatívákat is tartalmaztak. Ebben az esetben a BLAH nem külön-külön magyarázta el az egyes lehetőségeket, hanem váltakozva mutatta be őket. Ily módon a felhasználók jobban összehasonlíthatták az opciókat, és megérthették azok különbségeit vagy hasonlóságait.

[206] Salience map – egy hőtérkép, amin „forróbbnak” van ábrázolva a bemenet azon része, amely döntően hozzájárult az automatizált döntéshez.

Az ismeretelmélet kedvelői és a tudományfilozófusok számára kellemes meglepetésként a BLAH egy kontrafaktuális magyarázati módszert is tartalmazott. Ha egy felhasználó megkérte a BLAH-t, hogy igazoljon egy megadott hipotézist (nevezzük ezt X-nek), és a BLAH alá tudta támasztani azt a tudásbázisa alapján, akkor két különböző szempontból próbált meg válaszolni. Először is megkísérelt reprezentálni egy olyan helyzetet, ahol X igaz volt, még akkor is, ha ez valójában nem így volt – tehát egy „kontrafaktuális” vagy „mi lenne, ha” forgatókönyvet. Ezután a BLAH azt is elmagyarázta, hogy miért nem igaz X a valóságban, pontosabban tehát, hogy mi mond ellent, milyen adat inkompatibilis a felhasználó által várt kimenettel. Így a BLAH két különböző perspektívából adott választ: az egyik egy elképzelt helyzetből, ahol X igaz lenne, a másik pedig a valós helyzetből, ahol X nem igaz. Ez lehetővé tette a felhasználó számára, hogy mélyebb megértést nyerjen a szituációból: nemcsak azt, hogy miért vannak a dolgok úgy, ahogy vannak, hanem azt is, hogy mi kellene ahhoz, hogy másképp alakuljanak.

Amikor tehát egy felhasználó kérdést tett fel, a BLAH nemcsak választ adott, hanem összevetette azt a felhasználó hipotézisben adott elvárásával is, ha ilyen rendelkezésre állt. Ha a tényleges válasz ellentmondott az elvárásnak, a rendszer olyan magyarázatot generált, amely három fő pontot fedett le: (1) magát az elvárást, (2) azt, hogy miért nem teljesült az elvárás, (3) és a tényleges választ. A BLAH ezeket a pontokat egy meghatározott sorrendben magyarázta el: először a felhasználó elvárásával foglalkozott, majd azzal, hogy az miért nem teljesült, és végül megadta a tényleges választ. Ennek a módszernek a célja az volt, hogy biztosítsa a felhasználó teljes körű megértését a válasz kontextusában, és hogy miért térhet el az a kezdeti elvárásoktól.

Weiner munkája tanúskodik a magyarázatok különböző funkcióinak korai felismeréséről és arról, hogy szükség van olyan technikai módszerek kifejlesztésére, amelyek megfelelnek ezeknek a funkcióknak.

Sok tekintetben ez az 1980-as megoldás átfogóbb, mint a legtöbb jelenlegi, több mint négy évtizeddel későbbi megoldás. Ennek oka a legegyszerűbben így foglalható össze: a ma használt, hatékony MI-k dekompozíciója (lásd a 2. fejezetet) nem engedi meg ezeket a funkciókat. Azonban semmi sem szól az ellen, hogy a gépi tanulást a jövőben ötvözzék a szabályalapú rendszerekkel és még jobb ágenseket kapjunk ezáltal.

Buchanan és Shortliffe[207] áttekintő munkája összefoglalja a magyarázhatóság akkori korszakának legkorszerűbb eredményeit. A szerzők azzal érvelnek, hogy egy szakértői rendszer azon képessége, hogy megmagyarázza érvelési folyamatát, több, egymást átfedő okból is elengedhetetlen: megértés, hibakeresés, oktatás, technológia-elfogadás és meggyőzés.

A megértés esszenciális Buchanan és Shortliffe elemzése szerint. Egy szakértői rendszer felépítése és használata megköveteli, hogy a fejlesztők és a felhasználók egyaránt világosan megértsék a rendszerbe ágyazott tudást. Ez különösen kritikus a rendszer karbantartásánál és hatékony használatánál.[208] Egyes esetekben a rendszer proaktívan kommunikálhatja érvelési rendszerét, gyakran közbenső lépéseket is megjelenítve. A legtöbb esetben a rendszer azonban csak a konkrét felszólításra fejtette ki érvelését.

A hibakeresés, mint indok is elsődleges, mivel a tudásbázis építése iteratív folyamat. Itt válik nyilvánvalóvá az adott szabály megjelenítésének haszna. A szerzők hivatkoznak a fejlett monitorozó eszközök fejlesztésére irányuló folyamatos kutatásokra, amelyek lehetővé teszik a fejlesztők számára, hogy megfigyeljék és megértsék a rendszer tevékenységeit annak működése közben, ezáltal megkönnyítve a hibakeresést. A szerzők ezen felismerése korai előfutára a jelenlegi normatív kereteknek, amelyek az MI szabályozására vonatkoznak, mint például az IEEE P7001 átláthatósági szabvány[209] vagy az EU MI rendelet; mindkettő előírja a megmagyarázhatóságot, bár nem a hibakeresés céljából, hanem az elszámoltathatóság és a bizalom kiépítése érdekében.

Harmadszor, egy szakértői rendszer azon képessége, hogy betekintést nyújtson a tudásbázisába, jelentős oktatási értékkel bír. A szerzők megfigyelték, hogy a felhasználók hajlamosak újra és újra használni egy rendszert, ha az lehetőséget kínál számukra a tanulásra. Azonban ennek az oktatási célnak az elérése megköveteli, hogy a tudásbázis és az érvelési folyamat érthető legyen.

[207] Buchanan és Shortliffe (1984).
[208] Hogy ez mégsem annyira kritikus, azt a 2020-as évek nagy nyelvi modelljeinek sikere mutatja meg: ezek nagyrészt nem rendelkeznek a megmagyarázhatóság itt felsorolt képességeivel.
[209] Winfield és társai (2021).

A Buchanan és Shortliffe kiemelik az elfogadás és a meggyőzés, mint egymással összefüggő indokok fontosságát a szakértői rendszerek magyarázatai kapcsán. Ahhoz, hogy a rendszer elnyerje a felhasználók és döntéshozók bizalmát, elengedhetetlen, hogy meggyőzze őket a következtetéseinek észszerűségéről. Az, hogy a felhasználók világosan megértsék, hogyan jut el a rendszer a különböző tesztesetekben a következtetéseihez, és hogyan alkalmazkodik az esetek közötti különbségekhez, növeli a rendszerbe vetett bizalmukat.

Végül, a felhasználók meggyőzése a rendszer következtetéseinek helyességéről gyakran azon múlik, hogy a rendszer átlátható képet nyújtson a tudásbázisáról és az érvelési folyamatáról. A felhasználónak teljes mértékben meg kell értenie és el kell fogadnia a szakértői program következtetéseit ahhoz, hogy felelősséget vállalhasson a következő lépésekért, különösen olyan magas tétű területeken, mint az orvostudomány.

9.2 Megmagyarázhatóság a második „MI tél” idején

A mesterséges intelligencia második tele általában az 1987- 1997 közötti időszakra datálható.[210] Míg az 1970-es évektől az 1980-as évek közepéig a szakértői rendszerek felemelkedését láthattuk, figyelemre méltó magyarázhatósági képességekkel, addig ebben az időszakban az MI további fejlődése a felhasználók szemszögéből látszólag leállt – persze eközben a „jég” alatt fontos folyamatok zajlottak. Számos tényezőt említenek ennek a hiátusnak az okaként. Az egyik forrásként a logikai következtetéseket nevezik meg, amelyek a szakértői rendszerek magját képezik. Ennek a technológiának polinomiálisan, néha exponenciálisan növekvő számítási igényei vannak, ahogy az adatállomány növekszik, eközben a hardverek gyorsan, de mégsem exponenciálisan növekedtek.

Egy másik, komolyabb ok az úgynevezett „tudás megszerzésének szűk keresztmetszete”.[211] Ez a probléma, amelyet széles körben kutattak a tudásmérnöki területen, lényegében azt jelenti, hogy nem vagyunk képesek elég gyorsan kifejezni a világra vonatkozó tényeket olyan reprezentációban, amely lehetővé teszi, hogy a gépek felhasználják azt a tudást. Ez a felismerés olyan projekteket eredményezett,

[210] Csúcspont ill. mélypont, Toosi és társai (2021).
[211] A híres „knowlede acquisition bottleneck-problem” – Waterman (1986).

mint a CYC,[212] a FrameNet[213] és egy kicsit később, a télből a tavaszba fordulva a szemantikus web.[214] Ennek a megközelítésnek a jelenlegi reinkarnációja a Linked Open Data,[215] amely az angol Wikipédiára épül, és nagy szerepet játszik a jelenlegi, általános célú MI rendszerekben. Ezek az adatbázisok és a kapcsolódó technológiák mind elősegítik a szakértői rendszerek magyarázhatósági módszereit, amelyek elviekben képesek nagyon magas minőségű magyarázatokat nyújtani.

Közben azonban a jég alatt nagyon fontos áttörések történtek. Az egyik ilyen a hiba-visszaterjesztés (backpropagation) módszere, amely óriási lendületet adott a neurális hálózatok területének. A másik a vektortér alapú természetes nyelvfeldolgozás. Ez a technológia figyelmen kívül hagyja a nyelv szemantikájának megértését, minden szövegtípust vektorként kezel, és a megerősítéses tanulás, statisztikai módszerek és neurális hálózatok segítségével – valamint némi nyers erővel – arra készteti a gépet, hogy végrehajtsa azt, amit a jutalmazási függvény előír, teljesen kihagyva a tudás megszerzésének lépését. Az olcsó és erős hardverek elérhetősége, például a számítógépes játékokban használt grafikus kártyák megjelenésének mellékhatásaként lehetővé tette, hogy a számítógépek az előző generációnál nagyságrendekkel nagyobb adatállományokon, költséghatékonyan működjenek.

Ezen új technikák, amelyeket összefoglalóan gépi tanulásnak nevezünk, hajlamosak fekete dobozokat eredményezni. A felhasznált bemeneti adatok tisztaságára tekintve sejthetjük, hogy a gépi tanulásban nem csupán a modellek komplexitása miatt nehéz megérteni a döntéshozatali folyamatokat, hanem azért is, mert a bemeneti adatok minősége és jellege jelentős szerepet játszik az átláthatatlanság kialakulásában.

Ahogy az MI rendszerek egyre inkább szabályozási nyomás alá kerülnek (lásd 5. és 6. fejezet), előfordulhat, hogy a vállalatok és intézmények visszatérnek a szakértői rendszerekhez, amelyekben a döntési folyamatok átláthatóbbak és jobban érthetők. Ezek a rendszerek előre meghatározott szabályok és logikai következtetések alapján

[212] Lenat (1995).
[213] Baker és társai (1998).
[214] Berners-Lee és társai (2001).
[215] Bauer és Kaltenböck (2011).

működnek, ami lehetővé teheti a megfelelés könnyebb biztosítását a szigorodó szabályozási környezetben.

9.3 Kortárs megmagyarázhatóság

Az aggodalom amiatt, hogy az emberek nem teljesen értik az MI rendszerek viselkedésének hátterét, már egy ideje jelen van, és az új MI fejlesztési hullám megjelenésével ezek az aggodalmak újra előtérbe kerültek, bár kissé más megvilágításban. Ma az etikai kérdések kiegészítik a szakértői rendszerek magyarázhatóságával kapcsolatos alapvetően episztemikus vitákat. Ezek az etikai szempontok kulcsszerepet játszanak a jelenlegi Magyarázható Mesterséges Intelligencia (XAI) diskurzusában.

Ha meg kellene jelölnünk a modern MI átláthatóságával és érthetőségével kapcsolatos párbeszéd kezdetét, Bostrom és Yudkowsky[216] munkája jó jelölt erre. Bár úgy tűnik, hogy a XAI kifejezést először Van Lent és munkatársai[217] használták, tanulmányuk nagy része a magyarázhatóság korai történetére vonatkozik, és elsősorban a szakértői rendszerekre összpontosított, nem pedig a modern, gépi tanuláson alapuló mesterséges intelligenciára. Bostrom és Yudkowsky kibővítették a magyarázhatóságról és átláthatóságról szóló párbeszédet, jelentős hatást gyakorolva a kortárs diskurzusra.

Bostrom és Yudkowsky a példa kedvéért egy bank MI rendszerét vázolták fel, amely ismeretlen és nem megmagyarázott módon megkülönbözteti a hiteligényeket. Így merült fel mindaz, amivel a 8. fejezetben részletesen foglalkoztam: ha egyáltalán nem is értjük, hogy mit csinál az algoritmus, lehet, hogy az nem is méltányos?

A szerzők ugyanazzal érveltek, amivel az MI szabályozói is: mivel az MI rendszerek egyre inkább átveszik a társadalmilag releváns kognitív feladatokat, nekik is meg kell felelniük azoknak a társadalmi követelményeknek, amelyek hagyományosan az emberi teljesítményhez kapcsolódnak. Így, ahogyan a felelősség, átláthatóság és előrejelezhetőség az emberekre vonatkozik, amikor társadalmi funkciókat látnak el, ezeknek a tulajdonságoknak az algoritmusokra is vonatkozniuk kell, amikor embereket helyettesítenek ezekben a feladatokban.

[216] Yudkowsky (2014).
[217] Van Lent és munkatársai (2004).

Ez az érvelés jól illusztrálja a magyarázhatóság átmenetét a pusztán episztemikus kérdésből etikai kérdéssé. Az átmenet akkor válik nyilvánvalóvá, amikor a figyelmünket a pusztán ember-gép interakcióról – amely csak egy ágenst von be mindkét oldalról –, kiterjesztjük arra, hogy mesterséges ágenseket is figyelembe vegyünk, amelyek egy olyan kapcsolati hálóban működnek, amit akár társadalminak is nevezhetünk. Ez az átmenet teljesen érthető, mivel az internet lett az MI szolgáltatás fő célbajutattási módszere, amelyet erősen központosított, több felhasználós architektúrák valósítanak meg.

Mind a hagyományos szakértői rendszerek, úgy a modern MI hasonló kihívásokat vet fel a megértéssel kapcsolatban, de az utóbbi, társadalmi funkciói miatt, további etikai kérdéseket is eredményez.

Így tehát a magyarázhatóság etikai aspektusa nemcsak az új MI technikákból ered, hanem azok új alkalmazásából is, különböző társadalmi területeken, különösen ott, ahol nagy a tét. Míg az MI rendszerek, amelyeket specifikus, korlátozott területeken, például egészségügyi diagnosztikában használnak, elsősorban episztemikus kihívásokat jelentettek, az MI használatának kiterjesztése különféle területekre és feladatokra további etikai akadályokat jelent. Lehetséges, hogy mindig is volt egy fel nem ismert társadalmi elem, még a szakértői rendszerek körül is, de az új, felhőalapú architektúrákban, ahol maga a szakértő ki van iktatva, a probléma nagyságrendileg komolyabb.

Mittlestadt és társai[218] ezt tovább hangsúlyozták sokat hivatkozott tanulmányukban, megismételve, hogy az etikai problémák abból fakadnak, hogy egyre több műveletet és döntést ruháznak át nehezen megjósolható vagy nehezen magyarázható algoritmusokra, nem pedig azokra, amelyek rutinfeladatokat automatizálnak.

Azonban a magyarázhatóság technikái csak az átláthatóság kérdését tudják kezelni, nem pedig az MI kritikus területeken való, központosított alkalmazásának kihívását, amelynek része a felelősségrés veszélye is (lásd a 7.1.2 fejezetben). Ha ez így van, akkor a magyarázhatóságra irányuló jelenlegi fókusz és a kapcsolódó elvárások túlzottak is lehetnek.[219] Amennyiben az etikai kérdések leküzdését várjuk a

[218] Mittlestadt és társai (2016).
[219] Robbins (2019).

megmagyarázhatóságtól, miközben valójában csak az MI folyamatok korlátozott megértésének episztemikus problémáját oldhatja meg, úgy nagyot csalódhatunk.

9.4 Gépi tanulás: fekete doboz gyár

A szakértői rendszerek korszakának vége felé Winograd és Flores[220] 1986-os úttörő könyvükben átfogó elemzést végeztek a számítógépes rendszerek magyarázataival és átláthatóságával kapcsolatos kérdésekről. A szerzők, a megmagyarázhatóság ezen korszakában egy újabb szemszögből közelítve meg a problémát, az Ember-Számítógép Interakció (HCI – Human-Computer Interaction) szempontjából vizsgálták a kérdést. Az Ember-Számítógép Interakció a felhasználóbarát és intuitív felületekkel foglalkozott, de a szerzők igyekeztek túllépni ezeken a praktikus alapokon megfogalmazott elképzeléseken, részben filozófiai ívet adva gondolatmenetüknek, különösen a fenomenológia és a kognitív tudomány területére támaszkodva.[221]

Az egyik szerző, Winograd korábban megalkotta a SHRDLU-t,[222] a természetes nyelvfeldolgozás (NLP) egy korai, úttörő alkalmazását. Ezt a rendszert 1969-70 között hozta létre, és még a szakértői rendszerek sikere előtt készült el. A rendszer lényegében olyasmi volt, amit ma egyszerű csetbotnak neveznénk. Bár „régi stílusú” mesterséges intelligencia (Good, Old-Fashioned AI - GOFAI) területéhez tartozik, és nem használt gépi tanulást, mégis, akárcsak az első csetbot, az ELIZA,[223] nagyon érdekes emberi reakciókat váltott ki, amelyek gyakran tartalmaztak egy antropomorfizáló elemet. Az MI antropomorfizálása pedig egy újabb oka lehet annak, hogy a magyarázhatóság iránt fokozott és egyben etikai figyelmet szentelünk. Ezért a SHRDLU alkotójának munkája jó átmenetet képez a gépi tanuláshoz.

Winograd és Flores tisztában voltak Roger Schank munkásságával, és a kognitív tudományos pilléreiket Humberto Maturana, valamint kisebb mértékben Francisco Varela biztosították, de a könyvben a legtöbbet idézett szerző Martin Heidegger volt. A heideggeri vagy gadameri keretek túl összetettek ahhoz, hogy részletezzem őket, de a fenomenológia és a kognitív tudomány módszerei emlékeztetnek arra, ahogyan a gépi tanulás által létrehozott fekete dobozokat ma megközelítik.

[220] Winograd és Flores (1986).
[221] Héder (2023b).
[222] Winograd (1972).
[223] Weizenbaum (1966).

A gépi tanulás a bemenet és a kimenet közötti függvény közelítését jelenti, az adatelemek közötti kapcsolatot, általában egy több dimenziós, diszkrét matematikai térben. Ahogyan az előző szakaszokban is kifejtettem, a gépi tanulás nem használ explicit programozást, ehelyett egy modellt „tanít” be. Felügyelt tanulás esetén címkézett példákat mutatunk a rendszernek, ami egy olyan függvényt eredményez, amely most már képes lesz korábban nem látott adatok kezelésére, lényegében egy irányított absztrakciót hozva létre. Felügyelet nélküli tanulás esetén még az adatokat sem címkézzük, így hagyjuk, hogy a rendszer maga hozza létre az általánosítást külső referencia nélkül. Egyik esetben sem történik explicit programozás.

Ha ezt a működési modellt összehasonlítjuk azzal, amit például Swartout javasolt 1985-ben, láthatjuk a magyarázhatóság szempontjában rejlő hátrányt ebben a rendkívül sikeres MI módszerben:

1) nincs valódi szerzője a modellnek, csak egy felügyelője a betanításnak,

2) az eredményül kapott modell – a függvény – önmagában is elég bonyolult.

Mindkét tényező hozzájárul ezen modellek antropomorfizálásához, amely azt eredményezi, hogy a felhasználó valamilyen pszeudo-személyként éli meg ezeket.

A jelenlegi technikai viták az XAI körül valójában csak a 2. pontot látszanak kezelni. Léteznek mind matematikai, mind mérnöki megközelítések a problémára.

A legegyszerűbb matematikai megközelítés a „lokális interpretáció”, amely azáltal különbözik a „globális” megközelítéstől, hogy a modell egészének vizsgálata helyett specifikus esetekre összpontosít. Itt a „lokális” kifejezés a matematikai térbeli távolság fogalmára utal. Ha van magyarázatunk egy kimenetre, akkor feltételezhetjük, hogy ugyanazok a magyarázatok vonatkoznak a hasonló távolságra lévő más kimenetekre is. Azonban ez a távolság egy olyan térben van mérve, amelyet a felhasználó eleve nem ért, mivel az akár több száz vagy ezer dimenziót is magában foglalhat. Egyébként egy ilyen rendszer hasonló ahhoz, amit Weiner alkotott: előállít egy döntést (vagy előrejelzést) és egy post-hoc magyarázatot (lásd 13. ábrát). A magyarázat előállításához szükséges az input, az output és a modell ismerete is.

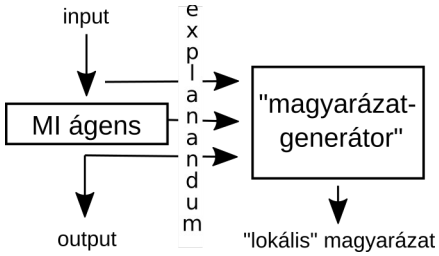


Figure 13: Lokális magyarázat működése

Az ilyen típusú magyarázhatóság korlátja természetesen az, hogy a magyarázat nem feltétlenül tárja fel a modell megbízható jellegzetességeit és hibáit a teljes modellre vonatkozóan, mivel csak az adott input-output párt és annak nagyon közeli szomszédjait fedi le.

Mivel az elképzelhető inputok száma kozmikus[224], a rendszert nem lehet teljeskörűen tesztelni.

Az ismeretelméletben jól ismert, hogy bár nem feltétlenül van szükség egy természeti törvényre, egy determinisztikus mechanizmusra vagy egy explicit formulára a jövő előrejelzéséhez, de egy megbízható szabályosság elengedhetetlen.

A lokális magyarázhatóság esetében az explanandum az input-output pár, és az explanans kiugrási térképet eredményeznek, amely megmutatja, hogy az input mely részei voltak kiemelkedőek az output generálásában. Ezért, bár ez a módszer hasznos lehet a hibakeresésre és a rendszer hibás elemeinek megtalálására, kérdéses, hogy ennek előírása megelőző hatással bírhat-e.

Egy másik matematikai megközelítés a bonyolult modell megértéséhez az egész függvény értelmezése, amelyet „globális" magyarázatként emlegetnek. A globális megközelítés magában foglalja a modell egészének megértését. Egy mód ennek elérésére az, hogy létrehozunk egy helyettesítő rendszert, amely egy adott hibahatáron belül hűen képviseli az eredeti modellt, de jóval egyszerűbb. A helyettesítő rendszer olyan összetevőket és elemeket használ, amelyek érthetőbbek az emberek számára.

[224] Például vegyünk egy nagy nyelvi modellt, amit promptokkal tudunk használni. Nos, általános esetben a promptok száma végtelen. A valóságban ezt leszűkítik (például 1000 vagy 1 millió karakterre), de ez praktikus szempontból a végtelennel egyenértékű.

Természetesen a helyettesítő modell nem azonos az eredetivel – szükségszerűen csak egy közelítés, amely bizonyos delta különbséggel bír az eredeti rendszerhez képest.

Megbízható szabályosságokat vonhatunk le a helyettesítő vizsgálatából, amely egyfajta kontrollérzetet és megbízhatóságot nyújt számunkra. Azonban mivel a helyettesítő eltér a valós rendszertől, a helyettesítő megfigyelése alapján levont feltételezéseink és következtetéseink teljes mértékben aluldeterminálták.

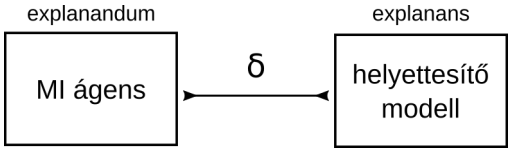


Figure 14: Globális magyarázat működése

A kérdés az, hogy ezek a meglehetősen matematikai XAI módszerek mennyiben teljesítik azokat a célokat, amelyeket Buchanan és Shortliffe olyan találóan összefoglalt: megértés, hibakeresés, oktatás, elfogadás és meggyőzés. A hibakeresés ezekkel a módszerekkel a legjobban lefedett terület, míg a megértés a helyettesítő rendszer természetétől függ. Gyakran ezek a helyettesítő rendszerek még mindig túl bonyolultak ahhoz, hogy a végfelhasználóknak bemutathatók legyenek, ami egy nagy hiányosság. Az oktatási érték szintén kérdéses, de nem annyira, mint az elfogadás és a meggyőzés.

Ez a kettős megközelítés a modell értelmezésére – globális és lokális – alapvetően episztemológiai jellegű, mivel a modellre vonatkozó tudás megszerzésére irányul. Normatív szerepük azonban korlátozottnak tűnik, még akkor is, ha világos szükség van az „átláthatóságra” és a fekete dobozok elkerülésére.

Miközben ezeket a módszereket vizsgáljuk a fekete doboz modellekkel szemben, fontos megjegyezni, hogy valójában egyetlen alkalmazás sem áll csupán egy modellből. Az IBM Watson, amely híressé vált azáltal, hogy legyőzte a legjobb versenyzőket a Jeopardy! nevű játékban (2011),[225] erre egy nagyon jó példa.

Ez a jól dokumentált rendszer erősen moduláris (architektúráját lásd a 2. fejezetben). Van egy kérdés- és témaelemző modulja, több független kérdés-megválaszoló modulja, amelyek különböző elveket követnek és eltérő adatbázisokat használnak,

[225] Ferrucci (2012).

mindegyik egy választ és egy bizalmi szintet generál, valamint egy „összevonó és rangsoroló” komponens, amelynek feladata az, hogy kitalálja, melyik független kérdés-megválaszoló modul adta a legjobb választ a. Ez lényegében azt jelenti, hogy több, akár egy tucat fekete doboz komponens dolgozik együtt, amelyeket egy szakértői rendszerhez hasonló szoftver felügyel.

Az autonóm járműveknél hasonló a helyzet: a képalkotás valós időben neurális hálózatokkal történik, de az útvonalválasztás és a kezelés sokkal egyszerűbb rendszerek. Azt is tudjuk, hogy a ChatGPT, a GitHub Copilot és más generatív MI-k szűrőket használnak – olyan komponenseket, amelyek utólag feldolgozzák, amit a fekete doboz előállít. Ez azt jelenti, hogy összességében az elsőként a szakértői rendszerekhez kifejlesztett módszerek közvetlenül alkalmazhatók lehetnek a mai MI rendszerekre, még ha a fókusz jelenleg máshol is van.

9.5 Normatív átláthatósági keretrendszerek

Az átláthatóságra vonatkozóan léteznek más, gyakorlati és teljesen normatív szabványok is. A már említett példa, az IEEE P7001 szabvány figyelembe vesz egy egyszerű tényt: nincs olyan, hogy „érthető” vagy „megmagyarázott” rendszer, mivel a megértés attól is függ, hogy az adott személynek milyen előzetes ismeretei vannak a rendszer megértéséhez. Ennek megfelelően az IEEE P7001 szabvány, amely az „Autonóm Rendszerek Átláthatóságáról” szól, az embereket csoportokra osztja, és a különböző csoportok számára különböző „átláthatósági” szinteket határoz meg.

A rendszer felhasználóitól nem várható el, hogy technikai szakértelemmel vagy MI ismeretekkel rendelkezzenek. Ezért számukra a szabvány készítői nem annyira a megértésre, mint inkább a kísérletezési lehetőségekre helyezik a hangsúlyt. A minimum az, hogy létezik dokumentáció, de a magasabb szinteken egy „homokozó” környezetet biztosítanak, ahol kontrafaktuális forgatókönyvekkel[226] lehet kísérletezni, és nagy hangsúlyt fektetnek a vizuális megjelenítésre.

Ez a megközelítés figyelembe veszi, hogy a különböző felhasználók eltérő szintű ismeretekkel rendelkeznek, és ennek megfelelően különböző mértékű átláthatóságot igényelnek. Az IEEE P7001 szabvány tehát pragmatikusan közelíti meg az

[226] „Mi lenne, ha” forgatókönyvek – amelyek, ahogyan azt az előző szakaszokban is láttunk, egyes korai szakértői rendszerek részei voltak

átláthatóság kérdését, és lehetőséget biztosít arra, hogy a rendszer felhasználói saját szintjükön kísérletezhessenek és értsék meg a rendszer működését.

Azok számára, akik nem használják közvetlenül az MI-t, de mégis érintettek – mint például az MI jelenlétében lévők (pl. gyalogos a járdán az AV kontextusában) –, a transzparencia azt jelenti, hogy megtudják, hogy MI-nek vannak kitéve, és magasabb szinteken azt is, hogy ki felelős az MI-rendszerért. Ahogy láthatjuk, a transzparencia fogalmának ez a használata teljesen bürokratikus, és kevés köze van a rendszer belső működésének megértéséhez.

Végül ott vannak a szakértők. Ők lehetnek kormányzati tisztviselők, auditorok, balesetvizsgálók, tanácsadók, fejlesztők és más technikai szakértelemmel rendelkező személyek. Ez az a terület, ahol a matematikai XAI módszerek a leghasznosabbak. Azonban léteznek más, kevésbé matematikai, gyakorlati mérnöki megoldások is, mint például az auditálható naplózás, amely lényegében egy olyan rögzítő mechanizmust jelent, amely felveszi a bemeneteket, a kimeneteket és a humán felügyeletről szóló információkat. Az ötlet egy „repülési adatrögzítő” (más néven „fekete doboz”, bár ebben a kontextusban ez félrevezető, mivel a belső működése nem episztemikusan átláthatatlan) létrehozása az MI rendszerek számára.

Ezen utolsó elem, valamint az, hogy az MI tulajdonosa/üzemeltetője bürokratikus értelemben átlátható legyen, és hogy az érintettek tudják, hogy MI rendszernek vannak kitéve, valóban megmutatja, hogy az átláthatósági normák nem episztemikus, hanem adminisztratív jellegűek, és kevés közük van a rendszer tényleges tudásához és megértéséhez. Ez az átláthatóság inkább arra irányul, hogy biztosítsa a felelősségvállalást és a nyomon követhetőséget, mintsem, hogy megvilágítsa a rendszer belső működését.

9.6 Helyzetértékelés

Amikor a magyarázhatóságot etikai elvként értelmezzük, eszközként szolgál a bizalom építésére[227] és az értelmes emberi kontroll fenntartására.[228] Azonban a magyarázhatóság fokozatos eltolása az episztemikusból az „etikai irányba” nem mentes kihívásoktól. Míg sok kutató azt állítja, hogy a magyarázhatóság és az

[227] Ferrario és Loi (2022).
[228] Alan Turing Institute (2020).

átláthatóság kulcsfontosságúak a bizalom építésében és az emberi irányítás elősegítésében, számos kritika kérdőjelezi meg ezt az elképzelést, meggyőző érvekkel.

Az az elképzelés, hogy egy algoritmus kimenetelének magyarázata lehetővé teszi az ember számára a szükséges megértést az algoritmus feletti kontroll gyakorlásához, így megnyitva a morális felelősség hozzárendelését az adott személyhez vagy csoporthoz, első pillantásra hihetőnek tűnik. Azonban Robbins azt állítja, hogy a magyarázhatóság elve nem fedi le teljes mértékben azokat az etikai kérdéseket, amelyeket XAI próbál megoldani. Az etikai kérdések itt arra vonatkoznak, hogy az ember képes-e elfogadni, figyelmen kívül hagyni, megkérdőjelezni vagy felülbírálni azt a döntést. Robbins szerint a magyarázhatóság három okból nem éri el ezt a célt.

Először is, a magyarázhatóságot, ahogyan azt általában a XAI területén értelmezik, gyakran összetévesztik a döntés magyarázatának követelményével, miközben valójában a döntéshozó konstruktőrt is kellene magyarázni. Maga a döntés, nem az azt meghozó MI igényel magyarázatot. Ez összhangban van az általam azonosított *tervezési idő kihívással*.

Másodszor, sok AI alkalmazás alacsony kockázatú, ami miatt a magyarázhatóság szükségtelenné válik. Például a Google AlphaGo esete, amely anélkül győzte le a világ Go bajnokát, hogy megmagyarázta volna lépéseit, jól példázza ezt.

Végül egy paradoxon is előállhat nagy kockázatú MI alkalmazások esetében: ha a döntéshez olyasfajta méltányossági metrikákat rendelünk, amelyeket a 8. fejezet mutat be, és ezeket a kimeneten ellenőrizzük, akkor a gépi tanulás visszafejtése szükségtelen. Például, ha tisztában vagyunk azzal, hogy milyen paramétereknek – mint például az adósság-megtakarítás arányának vagy a jövedelem-bérleti díj arányának –, kell befolyásolniuk a hiteldöntést, és a kimeneten ezt ellenőrizve azt találjuk, hogy az MI teljesíti az elvárásokat, akkor általánosságban redundáns és szükségtelen az MI modell XAI eszközökkel való vizsgálata, a tényleges működés megmaradhat fekete doboz.[229]

Ez a kritika azt mutatja, hogy a magyarázhatóság elve nem minden esetben nyújt megfelelő megoldást, különösen olyan helyzetekben, ahol a releváns döntési tényezők

[229] Amikor azonban a rendszer eltér az előírt paraméterektől, a XAI kiváló hibakeresési eszköz.

már ismertek, vagy ahol az MI rendszerek alacsony kockázatúak és működésük magyarázat nélkül is elfogadható.

Általánosan elterjedt nézet, hogy a megmagyarázhatóság és a transzparencia bizalmat épít az MI-ben. Ez a megközelítés széles körben elterjedt a filozófusok[230] és a döntéshozók[231] körében, és jelentős hatással van az EU MI rendeletben szereplő átláthatósági követelményekre, ám nincs olyan meghatározó elméleti vagy empirikus bizonyíték, amely egyértelműen alátámasztaná ezt az állítást.

Igaz, hogy számos tanulmány mutat pozitív korrelációt az átláthatóság, a magyarázhatóság és az MI rendszerekbe vetett bizalom között. Shin[232] kutatása, amely 350, algoritmikus hírforrásokkal ismerős személyt vizsgált, arra a következtetésre jutott, hogy az átláthatóság és a magyarázhatóság pozitív hatással van a felhasználói bizalomra. Hasonló az üzenete Kartikeya[233] tanulmányának, amely azt találta, hogy a megnövekedett átláthatóság fokozza a bizalmat az AI rendszerek információinak nyilvánosságra hozatalában. Az egészségügyi szektorban arról számoltak be, hogy az átláthatóság és a magyarázhatóság jelentős tényezők voltak az orvosok körében a bizalom kiépítésében.[234]

Ezzel szemben számos kutatás mutat semleges vagy negatív összefüggést az átláthatóság/magyarázhatóság és a bizalom között. Papenmeier és társai azt találták, hogy a rendszer pontossága volt a bizalom megteremtésének kritikus tényezője, nem pedig a magyarázhatóság.[235] Kizilcec tanulmánya azt mutatta ki, hogy a túlzott átláthatósági információk a bizalom csökkenéséhez vezethetnek. Hasonlóképpen, Schmidt és társai arról számoltak be, hogy a nagyobb átláthatóság nem garantálta a nagyobb bizalmat vagy az MI rendszer eredményeinek elfogadását, sőt, bizonyos esetekben bizalmatlanságot is kelthetett. Ghassemi és csapata azt állítja, hogy a jelenlegi magyarázhatósági megközelítések túlzott bizalomhoz és automatizálási torzításhoz vezethetnek a felületes vagy megbízhatatlan magyarázatok miatt, ezáltal „hamis reményt” keltve.[236]

230 Floridi és társai (2018).
231 AI HLEG (2019).
232 Shin (2021).
233 Kartikeya (2022).
234 Liu és társai (2022) valamint Wysocki és társai (2023).
235 Papenmeier és társai (2019).
236 Kizilcec (2016) Schmidt és társai (2020) Ghassemi és társai (2021).

Ez a kettősség azt mutatja, hogy bár az átláthatóság és a magyarázhatóság gyakran elősegíthetik a bizalmat, ez nem mindig egyértelmű, és a túlzott vagy nem megfelelő magyarázatok akár alá is áshatják a felhasználók bizalmát az MI rendszerekben.

A kutatások szerint ezeknek a megoldásoknak a hatása pozitív, semleges vagy negatív lehet, attól függően, hogy milyen kontextuális tényezők érvényesülnek. Még azonos szektorokban végzett kutatások is ellentmondásos eredményeket hoztak. Sok munka nem felel meg az empirikus kutatás szigorú követelményeinek, ami a magyarázhatóság és a bizalom hatásainak inkonzisztens és ellentmondásos megállapításaihoz vezet.[237]

Ez az összefoglaló elemzés azt mutatja, hogy a magyarázhatóság önmagában nem képes betölteni a ráruházott etikai ellenőrző szerepet. Az MI megbízhatóságával és etikus alkalmazásával kapcsolatos összetett kérdések túlmutatnak a puszta magyarázhatóságon. Bár a XAI segíthet a megbízható MI elérésében, pontosan kell felismerni, hogy ez egy episztemikus eszköz. Némi – bár nem abszolút vagy teljesen megbízható – tudást ad számunkra az MI rendszer viselkedéséről. Míg ez a tudás hasznos lehet a döntéshozatal során, messze nem elegendő ahhoz, hogy egy magyarázható MI rendszert megbízhatónak tekintsünk.

9.7 A megmagyarázhatóság jövője

Ebben a fejezetben bemutattam néhány kulcsfontosságú erőfeszítést az 1970-es és 1980-as évekből, amelyek a számítógépes magyarázhatósággal kapcsolatos komoly gondolkodás kezdetét jelentették. Az APL nyelv, a BLAH és Digitalis Advisor szakértői rendszerek, valamint Buchanan és Shortliffe munkái, illetve kisebb mértékben Winograd munkássága került terítékre. Az elemzés rámutatott arra, hogy a számítógépes magyarázhatóság korai szakaszában, az episztemikus célok mellett a felhasználói elfogadás és meggyőzés is erős gyakorlati problémákat jelentett, amelyekkel a szakértői rendszerek korában kifejlesztett magyarázhatósági módszerei foglalkoztak.

Az egyik tanulság az, hogy az XAI törekvések mögött ma is megtalálható összes episztemikus cél már a szakértői rendszerek korában is ismert volt. Ebben a

[237] Kandul és társai (2023).

korszakban néhány rendszer lehetőséget adott hipotetikus és kontrafaktuális elemzésekre, foglalkozott a felhasználói elvárásoknak ellentmondó kimenetekkel, és akár saját programkódjukat is képesek voltak prezentálni magyarázatként. Ezenkívül néhány programozási nyelvet eleve az önmagyarázó képességre fókuszálva alkottak meg.

A kronológiai áttekintés során bemutattam az MI néhány „telét” és „nyarát”. Érveltem amellett, hogy a két korszak közötti releváns különbség a felhőalapú architektúra megjelenése volt, amely szélesebb, közvetlenebb hozzáférést biztosított a kevésbé képzett felhasználók számára. Emellett a gépi tanulással még a konstruktőrök is elveszítették a rendszer intellektuális felügyeletét.

Végül rávilágítottam arra, hogy az új MI rendszerek kísértetiesen képesek utánozni az embereket, ami talán egy további tényező a magyarázhatóság növekvő hangsúlyának hátterében. Azt is megmutattam, hogy az átláthatóságot, magyarázhatóságot és bizalmat az MI szabályozói összefüggő fogalmakként érzékelik, de ezek közötti kapcsolat nem teljesen tisztázott.

A megmagyarázhatóság jövőjének nagy kérdése, hogy 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.

III. Rész: az MI etika új hullámai

10. Episztemorális tervezési döntések

"A világ túlontúl összetett bármilyen megvalósítható rendszer számára... A lehetőségek száma meghaladja azt, amire a rendszer kapacitásai válaszolni engednek. A rendszer ezért egy 'környezetbe' ágyazza magát és működése szétesik, ha a környezet és a világ túlságosan elválik" (Luhmann 1979, 6)[238]

A kortárs filozófia szakirodalmában az MI vagy az autonóm rendszerek etikai kérdéseit taglalva a vita tárgya gyakran az, hogy mit tegyen, vagy épp mit ne tegyen egy rendszer egy adott szituációban, ahogy arra a 6-8. fejezetek is rámutattak. De jó példa erre a MIT (Massachusettsi Műszaki Egyetem) híres Moral Machine kísérlete is, amelyet már a 4. fejezetben röviden tárgyaltam.

Ez a fajta megközelítés az etikai problémát egy döntésként értelmezi, tökéletesen körülírt alternatív reakciókkal, melyek mindegyike pontosan kiszámítható következményekkel rendelkezik.

E probléma értelmezés nem tükrözi a valóságot, ahogyan ezt már 4.1.4-es alfejezetben, az MI átláthatatlansága kapcsán bemutattam. Így aztán minden autonóm rendszer olyan kontroll-módszereket alkalmaz, amelyek valamilyen módon képesek a probabilisztikus *(valószínűsíthető feltevéseken alapuló)* információk kezelésére. Időnként a bizonytalan tényezők nyíltan megjelennek, például, hogy egy neurális hálózat 98,3% biztonsággal képes egy tárgy felismerésére. Máskor, a bizonytalanság nincs jelezve, csak a rendszer reagál valamilyen ökölszabály alapján, ez pedig egy adott cselekvést vált ki, egy adott hibaaránnyal.

Az önvezető járműveket továbbra is példaként használva elmondható, hogy fejlesztések történnek a jobb érzékelők, a pontosabb (azaz jobb tesztadatokon alapuló) tárgyfelismerés, precízebb térképek és fejlettebb út menti jelek létrehozása irányában. Ezek mindegyike egyfajra bizonytalanságot csökkentő tényező, de a bizonytalanság, aluldeterminált ság[239] sehogyan nem eliminálható teljesen.

[238] Fordítás, szabadon: HM.

[239] Az aluldeterminált ságról és annak jelentőségéről az MI tervezésére bővebben lásd az 1.2, 3.1, 4.1 és 5-es fejezeteket.

Várakozásom szerint az MI etika jövőjében az MI episztemikus kapacitásainak megtervezése ***legalább akkora etikai kihívás*** *lesz a konstruktőrök számára, mint az MI viselkedés tervezése.*

Pontosabban, az alábbi téziseket szeretném kibontani, illetve igazolni:

1 Morális kötelesség-e – és ha igen, milyen körülmények között –, hogy egy önvezető jármű aktívan kísérletezzen tudásszerzés céljából?

2 Morális kötelesség-e – és ha igen, milyen körülmények között –, hogy egy önvezető jármű új döntési alternatívákat igyekezzen létrehozni?

Általánosabb értelemben, nem csak az önvezető járművek, hanem bármely racionális ágens esetében, egy adott helyzetben mi dönti el: a helyes morális döntés a további tájékozódás, ennél fogva talán jobb cselekvési alternatívák felfedezése. Vagy álljon le a megoldási alternatívák generálása, és legyen végrehajtva az egyik, már feltérképezett alternatíva?

10.1 Az *episztemorális* tervezés

A fentiek értelmében az MI konstruktőr számára lényegi etikai probléma az, hogy egy adott szituáció jobb megértésére irányuló, valamint az ugyanazon helyzet megoldására tett erőfeszítések között fennálló egyensúly hová helyezendő.

Ez a probléma elvileg bármely mesterséges ágens esetén fennállhat, azonban az MI architektúrák szükségszerű dekompozíciója az áttekinthetőség, a kivitelezhetőség és a komponensek újrafelhasználhatósága miatt (lásd a 2. fejezetet) nagyon könnyen elfedheti ezt a problémát. Ha a külvilág reprezentációs alrendszerünket készen vesszük egy komponens katalógusból, , vagy történetesen egy másik, hozzánk csak lazán kapcsolódó csapat dolgozik rajta, akkor könnyen úgy tűnhet, hogy a feladat csupán az alternatívák közötti választás.

10.1.1 Az alternatívakeresés problémája

Egy egyszerű gondolatkísérlettel szeretném illusztrálni az alternatívakeresés problémáját.

Főhősünk, **A**ladár a középkori Európában él, gyógyítással foglalkozik; egy barátja pedig bubópestisben szenved. Szeretné meggyógyítani, de tisztában van azzal, hogy az illető a vele való bármilyen érintkezés során eséllyel valamilyen eddig nem ismert módon átadhatja a betegséget, amely viszont kezelés nélkül halálozási arányt eredményez. A kezelés ugyanakkor rendelkezésre áll, ennek része a sebek kitisztítása, az ápolás és a gyógynövények alkalmazása. A beteg halálozási esélye a népi tapasztalat alapján így elfeleződik. Ugyanakkor a beteg kezelése eséllyel azt okozza, hogy **A** maga is elkapja a betegséget.

Így tehát, ha **A** kezelni kezdi betegét, annak túlélési esélye ¾-re nő ½-ről, viszont maga **A** ⅛ eséllyel belehalhat abba, hogy gyógyítani próbált.[240]

A olyan morális dilemmával szembesül tehát, mely két, probabilisztikus módon meghatározott kimenetellel rendelkezik.

Egy másik szereplő, **B**eáta a 21. századi Európában él, és orvoslással foglalkozik. Praxisa során egy ritka, bubópestises esettel találkozik. Tisztában van azzal, hogy a bakteriális betegség kullancscsípés vagy a fertőzött testnedveivel való érintkezés útján terjed. A fertőzésveszély esélyét képes 1 a 10.000-hez arányra csökkenteni, gyakorlatilag csak egy saját maga kárára elkövetett orvosi műhiba, *és* az orvosi védőruházat igen ritkán előforduló meghibásodása esetén van veszélynek kitéve. Azt is tudja, hogy a kezeltesetek halálozási aránya az ellátásidőszerű vagy megkésett megkezdésének függvényében 1 a 100-hoz és ⅛ között van. Igen csekély kockázat mellett a megelőző jellegű antibiotikumos kezelés lehetősége is **B** rendelkezésére áll.

Ily módon **B** betegének elhalálozási esélyét a kezeletlen pestis ½ arányáról néhány százalékra képes redukálni, miközben önmagát 1 a millióhoz, vagy kisebb eséllyel teszi ki elhalálozási veszélynek, ami egy hivatásos orvos számára nem igazán fogható fel morális dilemmaként.

Harmadik szereplőnk, Alexandre **Y**ersin a 19. és 20. század fordulóján folytat orvosi tevékenységet a Pasteur Intézet munkatársaként. Praxisa során egy bubópestises beteggel találkozik. **Y** nagy valószínűséggel feltételezi, hogy a betegség kullancscsípéssel és cseppfertőzés útján terjed, de ez utóbbi kiküszöbölésére a beteg

[240] ½ * ¼, feltéve, hogy megfertőződés esetén képes önmaga kezelésére is.

közelében lehetőségei korlátozottak, ami ⅛ fertőzésveszélyt jelent. Azt is tudja, hogy minél hamarabb elkezdődik a tisztogatással és a beteg ápolásával járó hagyományos kezelés, annál nagyobb a túlélés esélye, ami viszont továbbra sem több, mint ¾. Így, 1/32 eséllyel maga is meghalhat/életét veszítheti a kezelés során, miközben a beteg túlélési esélyeit ½-ről ¾-re növeli.

Azonban, mint tudjuk,[241] ez esetben **Y** rendelkezésére áll egy harmadik alternatíva is. Amellett, hogy megkísérelheti kezelni, vagy sorsára hagyhatja a beteget, megpróbálkozhat a testnedvek mikroszkópos vizsgálatával is. Ez kisebb fertőzésveszéllyel jár, mintha a beteg mellett lenne, ugyanakkor megnyitja az utat egy új gyógymód kifejlesztése előtt. A kísérletnek azonban több kimenetele is lehet. Fény derülhet arra, hogy a bubópestist baktérium okozza, és így a betegség átadásának esélye mind kezelés közben, mind in vivo tovább csökkenthető (a penicillin, vagy egyéb antibiotikum még nem ismert, de az orvosi eszközök pasztörizálásának módszere igen). Ugyanúgy, ahogy a szarvasmarhákat sújtó lépfene esetében, a baktérium azonosítása révén a helyzet számottevően javult.

Fény derülhet azonban arra is, hogy a veszettséghez hasonlóan – ezt maga Pasteur bizonyította be – a betegség nem bakteriális eredetű (ez idő tájt a vírusok még ismeretlenek voltak).[242] **Y** azzal is tisztában van, hogy a betegség bakteriális vagy nem bakteriális eredetére való esély nem ötven-ötven százalék. Ez az arány egyszerűen ismeretlen.

Csábító az a lehetőség, hogy a további ismeretszerzés lehetőségét csak úgy tüntessük fel, mint az alternatívák egyikét. De vajon a középkorban praktizáló **A** számára nem álltak rendelkezésre további alternatívák? Léteznek anekdotikus utalások arra, hogy a sebek hagymával, vagy egy felszeletelt kígyó darabjaival való bedörzsölése hatásos lehet. Az ecetes ivókúrát, az arzén és a higany fogyasztását szintén kipróbálták, mint lehetséges gyógymódokat. Ha bármely helyzetet csak bizonytalanul vagyunk képesek értelmezni, az megnyitja a lehetőséget egy eddig ismeretlen, új alternatíva felfedezése előtt. Előfordulhat az is, hogy **B**, bár már rendelkezik egy igen vonzó alternatívával, rátalál egy még ennél is jobbra.

[241] Alexandre Yersin sikeresen igazolta, hogy a pasztörizálás működik a pestis kórokozója ellen.
[242] Ez még egy jó darabig a kirakós hiányzó darabja maradt, nem igazán hagyva a középkor óta ismert stratégiáknál jobbat Pasteur követőinek (hiába állította Pasteur ennek ellenkezőjét).

Úgy tűnik, hogy a döntést igénylő valós szituációk, talán néhány speciális eset kivételével, mindig az alábbi struktúrát követik: *válasszunk egyet a nyilvánvaló alternatívák közül, vagy próbáljunk újakat találni.* Más szóval, morális felelősségünk episztemikus kötelességé alakulhat át, ezáltal egy morális-episztemikus deliberációs hurkot hozva létre, amelyet *episztemorális* keresésnek nevezek.

10.1.2 A következmények meghatározása

Az alternatívák létrehozásának problémájához hasonlóan, problémát jelenthet egy adott alternatíva következményeinek meghatározása is. A bubópestises gondolatkísérlethez visszakanyarodva, **A** tisztában lehet azzal, hogy beteg barátjának túlélési esélyei csak akkor növelhetőek jelentős mértékben, ha a létfontosságú szervek még nem károsodtak. Ennek megállapítása egy sor olyan kísérlet elvégzésével lehetséges, melyek nem a járvány legyőzését célozzák, sőt, még a fájdalmat sem csillapítják, ehelyett a fertőzés előrehaladottságának felmérésére irányulnak. Még a modern orvostudomány is alkalmazza a vizsgálat-kezelés- vizsgálat-kezelés (és így tovább) ciklusait.

10.2 A probléma az önvezető járművek kontextusában

E ponton visszatérek a 4. fejezetben alkalmazott stratégiámhoz, és az alternatívák létrehozásának, illetve következményeik feltárásának problémáját az önvezető autókon vizsgálom. Ahogy ott is, ebben az esetben is posztulálom, hogy a levont konklúziók túlmutatnak az önvezető járműveken: ez az alkalmazási terület pusztán egy jó illusztrációja az általános problémának.

Elvi síkon egy autonóm jármű által végrehajtott reakció generálása egy optimális alternatíva kiválasztását jelenti, megszámlálhatatlanul sok lehetőség közül.

Azonban, a rendszer számára már maguknak az alternatíváknak a generálása is bonyolult feladatot jelent. Ezen felül, az egyes alternatívák következményeinek értékelése azoktól az adatoktól függhet, melyekre a jármű a kísérletek végrehajtása révén tehet szert.

Íme egy régi vezetési módszer az ABS-sel (aktív fékrendszerrel) még nem rendelkező autók számára, a váratlanul feltűnő akadályokkal való ütközés elkerülése érdekében.

Manapság ez a javaslat a sportos élmény kedvéért kikapcsolt ABS esetén ugyanolyan releváns:

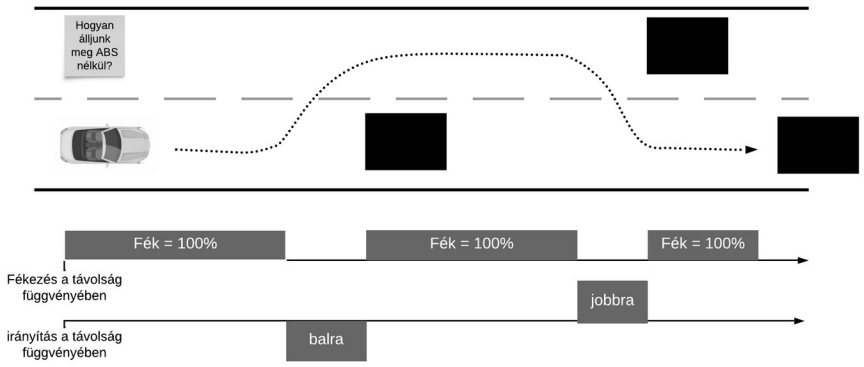


Figure 15: Hogyan kerüljünk ki egy akadályt ABS nélkül?

Az ABS megjelenése előtt a folyamatos vészfékezéskor a blokkoló kerekek miatt az autókkal képtelenség volt elfordulni. Ezért, a javasolt megoldás az azonnali lehető legkeményebb fékezés egészen addig a pontig, amíg az 1. akadály még biztonsággal elkerülhető, aztán a fékek felengedését követő kitérés, ezután a lehető legkeményebb fékezés a 2. akadály előtt, újabb sávváltás, majd fékezés a megállásig.

Most vegyük számba a jármű által kivitelezhető összes lehetséges reakciót, két dimenzióra korlátozva (az ábrán a távolság függvényeiként ábrázolva). A kormányzás és a fékezés a számításba vehető két primitív operátor, amelyeket kombinálva cselekvési tervet generálhatunk. Ennek tere a két funkció Descartes-féle szorzata, melynek végeredménye a kormányzási szögek irányának részletességétől, a fékintenzitástól és az időtől is függ. Ebben az óriási térben kell az önvezető járműnek alternatívákat generálni.

Vegyük számításba azt is, mennyire nem egyértelmű a régi javaslat. Amikor a jármű először észleli a saját sávjában található akadályt, a problémafelvetés a következő lehetne: szükséges a sávváltás, vagy nem? De a probléma ilyetén ábrázolása elmarad az optimálistól. Váltsunk sávot azonnal, ami miatt közelebb kerülünk a második járműhöz, mintha előbb lassítanánk és utána váltanánk sávot. Ez megmutatja, hogy az azonnali sávváltás, vagy a „nincs-sávváltás” döntési modell közelében sem jár az optimális megoldásnak. Érdemes lenne más alternatívák után nézni.

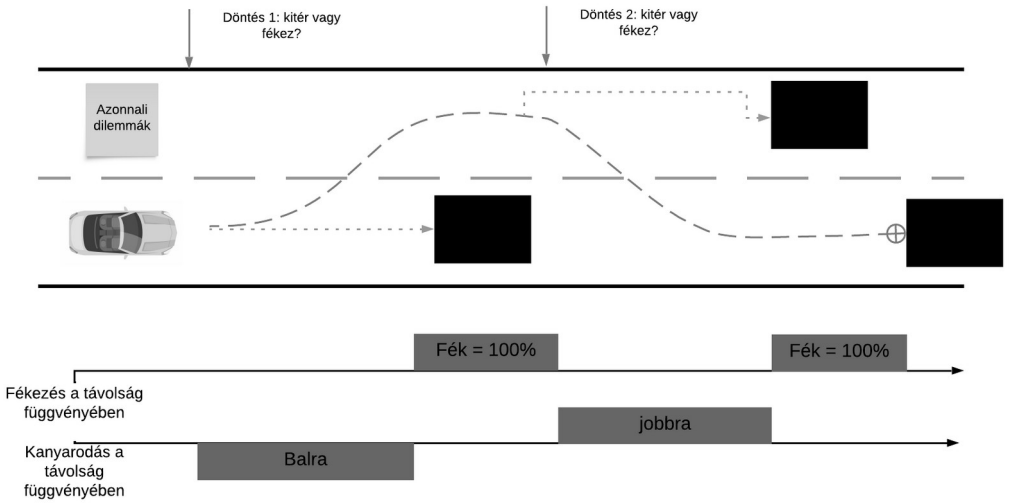


Figure 16: Szuboptimális cselekvéssorrend ABS nélküli akadály kikerüléshez

Ezen felül, a 14. ábrán bemutatott szituáció nem is csak egy dolog miatt optimálisabb az 15. ábrán ábrázoltnál. Abban az értelemben is sokkal kedvezőbb, hogy a kitérés előtti fékezés révén felmérhető az adott útviszonyok által meghatározott fék hatékonyság, mely lehetővé teszi a további távolságok pontosabb felmérését.

Tehát, optimálisabb az azonnali fékezés az azonnal lassulás érdekében (de ne törődjünk bele az 1. akadállyal való ütközés elkerülhetetlenségébe), így a sávváltással kapcsolatos döntést egy rövid pillanatig még elodázhatjuk. Ebben a helyzetben az rendszer episztemikus értelemben sokkal jobb helyzetben van a döntés meghozását illetően, mert a fékezéssel a féktapadásról is információt kap, amely bemenete lehet a további döntéseinek.

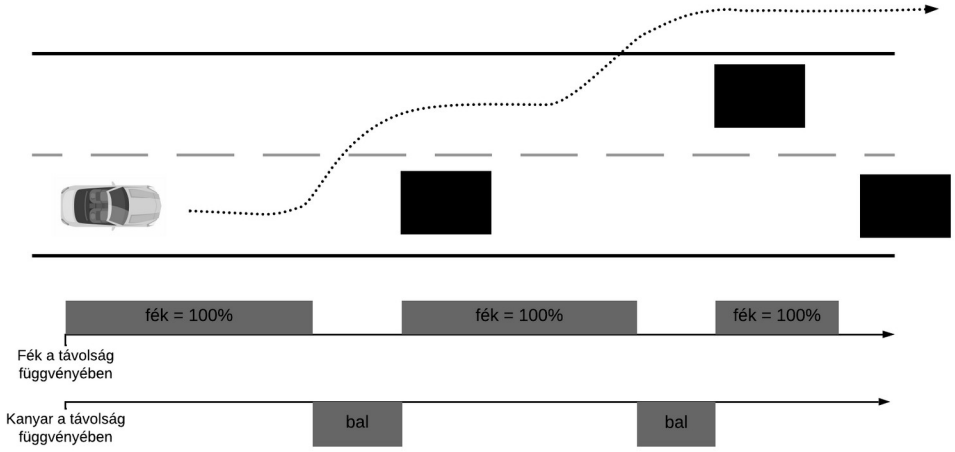


Figure 17: Ütközés elkerülése az útról való letéréssel

Az is előfordulhat, hogy még ez az optimálisabb reakciósorozat sem elég az ütközés elkerüléséhez. Bizonyos esetekben egy emberi sofőr még sikeresen úrrá lehet a helyzeten azáltal, hogy letér az útról, például, ha egy mezőt lát az út mellett. Ez biztosan a jármű sérülésével járna, ugyanakkor bizonyos körülmények között életeket menthet ez a döntés. Tehát, racionális választás lehet megengedni az rendszernek, hogy egy efféle „újítással" próbálkozzon? Engedjük-e, hogy a belső világreprezentációjának egyáltalán része legyen bármi, ami az úton kívül észlelhető?

Mindebből az tűnik ki, hogy bizonyos válaszreakciók nem elsődleges következményeik miatt élveznek előnyt, hanem azért, mert segítik a mesterséges intelligencia helyzetértékelését, így a további döntések helyes meghozatalát. Másfelől a példa emlékeztet arra, hogy emberi döntéseink milyen nagy mértékben heurisztikusak.

Az MI konstruktőrei számára azonban a heurisztikák nem feltétlenül kielégítőek: egyfelől lehetséges, hogy az aktuális dekompozíció nem teszi lehetővé használatukat, másfelől, a technológia társadalmi kontrollja mellett kardoskodók az emberhez képest sokkal jobb teljesítményt várnak el a gépektől[243] és joggal: a megvalósult MI milliószor másolódik, ezáltal konstruktőr általi bezártsággal fenyegetve (lásd az 5.5 fejezetet). Ez tehát egy folyamatosan fennálló motiváció az episztemikus-morális egyensúly filozófiai problémájának vizsgálatához.

[243] Például az önvezető járművekkel kapcsolatban az ipari sztenderd reakcióidő 0.1 másodperc. Ehhez képest az embernek ahhoz, hogy vészhelyzetben elérje a maximális fékezési erőt, ennek legalább a többszörösére van szüksége (Jurecki 2012).

10.2.1 A konstruktőri feladat

A fenti érvekkel felvértezve felvázolhatjuk a problémát a konstruktőr szempontjából is. Az önvezető járművek kontrollja informatikai mérnöki feladat. A mérnöki szoftvertervezés meghatározó jellemzője az a kivételes szabadság, amit ez a médium biztosít a tervezők számára, ahogyan ezt már a korábbi fejezetekben láttuk. Világos, hogy az emberi képességekből kifolyólag egyes szituációk időnként már előzetesen nehéz döntésnek ígérkeznek, míg máskor a reakciók természetes módon adják magukat, bármilyen kognitív erőfeszítés nélkül. Azt már tudjuk, hogy számos heurisztikára hagyatkozhatunk, amelyek könnyítenek kognitív terheinken, és bár előfordulhat, hogy hibás eredményekhez vezetnek, legtöbb esetben mégis igen hasznosak.

Az önvezető rendszerek szoftvereivel szemben sokkal magasabbak az elvárások, és bár valóban egy *tabula rasa* a kiindulópontunk, mégis, sok tekintetben sokkal gyorsabb információ-feldolgozó potenciállal számolhatunk. Döntenünk kell a külső világ rendszeren belüli hatékony modellezésének módjáról.

Definiálhatjuk tervezési céljainkat magas szinten: hogy olyan rendszert hozzunk létre, amely erőforrásait az alábbi célok szerint fekteti be: a) saját belső modellezését gyarapító tudás generálása, b) e modellen belül előrejelzések és alternatívák keresése c) az elsődleges cél – nevezzük megvalósításnak – elérése. Nem szabad figyelmen kívül hagyni a reakciók és a reprezentációk között működő visszacsatolást sem: végezhetünk egy kisebb fékezést, csak hogy próbára tegyük az útviszonyokat.

Nyilvánvaló, hogy 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. Ezt a „megtérülést” viszont előre nem ismerjük. Az ezzel kapcsolatos helyzet felméréséhez szükségünk lenne egy olyan módszerre, mellyel megbecsülhetőek a tudás megszerzéséhez szükséges ráfordítások és az ezzel járó előnyök. Sajnos, néhány kivételes esettől eltekintve[244] ilyen módszerrel nem rendelkezünk.

[244] Például egyes mérések, melyek garantáltan egy bizonyos hibaszázalékon belüli eredményeket adnak. De ellenpéldaként említhetjük, hogy egy probléma-térben az alternatív lehetőségek kutatása nem feltétlenül jár eredménnyel.

Megközelíthetjük a problémát egy másik szélsőség irányából is. Előfordulhatnak olyan helyzetek, amelyekben reménytelen bármilyen, az adott szituáció megértését segítő ismeret megszerzése. Kivételes, formalizált szituációkban adódhatnak ilyenek – az amőbajáték során például lehetetlen tovább finomítani a tökéletes stratégiát. Hasonlóképpen, egy, az önvezető járműhöz hasonló kiber-fizikai rendszerben, ha a jelen szituáció ismereteink szerint optimális, nincs motiváció a terv módosítására.

Az is előfordulhat, hogy rendszerünk környezete annyira változékony, hogy modellezésünk igen hamar elavulttá válik, így nincs értelme nagy erőfeszítéseket tenni annak fenntartására. Egyes, erősen korlátozott számú érzékelővel rendelkező rendszerek esetén úgy tűnhet, hogy kimerítettük a rendelkezésre álló ismeretszerző módszerkészletünket (bár érdemes megjegyezni, hogy még kevés érzékelő esetén is végtelen számú sikeres megfigyelés lehet elérhető). Természetesen ez nem menti fel a konstruktőrt a morális felelősség alól, hiszen gép dekompozíciója, a szenzorok megválasztása is konstruktőri felelősség (lásd a 2. fejezetet).

Egy önvezető jármű tervezésekor az *episztemorális* keresés szerepe abban a tekintetben jöhet például számításba, hogy ne csak a környezethez képest próbáljuk optimalizálni az autó helyzetét (biztonságos távolság egyéb járművektől stb.): ezt a feladatot végzi jelenleg a visszacsatolás-kontroll. Párhuzamos feladat lehetne a jármű episztemikus helyzetének folyamatos optimalizálása, hogy így pozícióját minden időpillanatban jobban felmérhesse. Ennek eredménye lehet az, hogy az autó időnként lelassít, mert „önbizalma” egy adott küszöbérték alá esett, így további megfigyelésekre van szüksége.

A bubópestises példájával az adott helyzetekben meghatározott alternatívák körül kialakuló általános episztemikus problémát igyekeztem illusztrálni. A fertőzöttek kezelése és a fertőzötté válás közötti okozati összefüggés mindig is nyilvánvaló volt. Ugyanez igaz a fertőzés és a halál közötti összefüggésre. A baktériumok felfedezésével azonban egy több láncszemmel rendelkező, részletesebb okozati láncot kaptunk. Ez egy új kontroll pontot eredményezett, és megnyílt egy adott láncszem antibiotikumok általi gyengítésének lehetősége. Ennek köszönhetően többé nincs szükség arra, hogy egész hajókat vagy betegeket hagyjunk sorsukra ilyen helyzetekben. Más szóval, sokkal jobb alternatívák állnak rendelkezésünkre. A tudomány és a technológia

történetének egy nem elhanyagolható része úgy értelmezhető, mint eme ok-okozati kontroll pontok és az általuk kapott új alternatívák története.

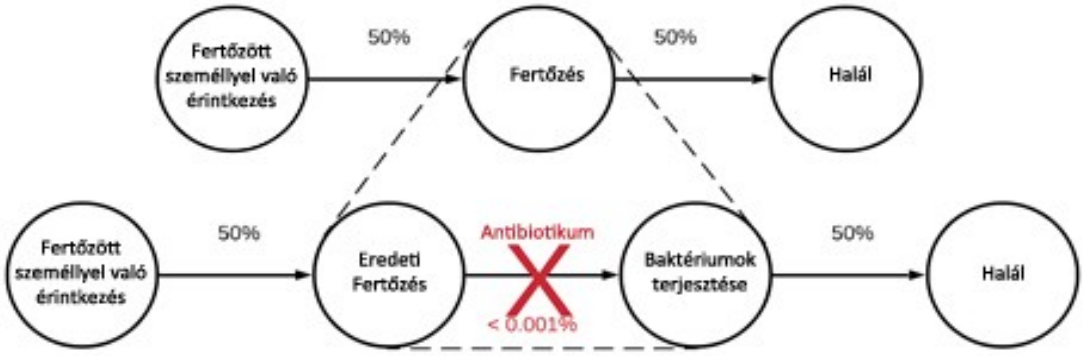


Figure 18: Alternatíva és kontroll lehetőség teremtés az oksági lánc felbontásával.

Érdekes kérdés, hogy hogyan osztjuk meg erőforrásainkat és kísérleteinket az új alternatívák generálása és egy adott helyzetben létező alternatívák következményeinek menedzselése között. Világos, hogy új felfedezések akkor történhetnek, ha az ágens úgy gondolja, érdemes próbálkozni. Az viszont egyelőre még nem világos, hogy hogyan ismerhetjük fel azt, hogy mivel érdemes próbálkozni, és miként becsülhetőek fel a próbálkozások esélyei.

Egy adott MI generáció esetében ezt a kérdést sikeresen elfedi az a tény, hogy a komponensek és általában, a feladatosztályban használt dekompozíció adott, így az episztemikus képességek is jórészt adottak. Azonban a fejlődés része a technológiai paradigmaváltás[245] is, és ennek kapcsán a komponensek és azok határainak újratervezésekor várhatjuk, hogy az MI konstruktőrök az episztemorális tervezési problémát jobban fel fogják ismerni.

10.3 Kontraktariánus mérnöki tervezés

Az episztemorális problémakeretezés segíthet a kihívások megértésében, de az aluldeterminált ság sosem lesz kiküszöbölhető. Ezért elvileg sem történhet más, mint az ismert és nem ismert kockázatokkal való megbékélés, lehetőleg minél több érintett megelégedése mellett.

[245] Dosi (1982).

Ez a kényszerhelyzet kompatibilisnek tűnik a kontraktariánus megközelítéssel. A kontraktariánus etikai rendszerek 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. Általában a morális antirealizmusok keretein belül tárgyalják, habár ez az árukapcsolás nem elkerülhetetlen.

Semmi sem kötelez minket arra, hogy életünk minden területét és célját egyetlen kizárólagos etikai paradigmának rendeljük alá. Meghatározott véleménnyel rendelkezhetünk az önvezető járműveket illetően, míg egyéb területeken máshogy is gondolkodhatunk, egészen addig, míg a kettő összhangba hozható egymással.

A tervezhetőség episztemorális akadályai miatt feltételezem, hogy a mesterséges ágenseket illetően a kontraktariánus megközelítés a megfelelő, és hogy ha valaha is szemtanúi lehetünk például az önvezető autók komolyabb térhódításának, az valamilyen kompromisszumnak lesz köszönhető – máskülönben ez soha nem következhet be.

Ebből egyenesen következik, hogy az MI tervezése során szükséges döntések azon paraméterek metszéspontjában találhatóak, amelyet minden érintett elfogad, akár egy kompromisszum keretein belül. 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.

Az önvezető autók ügye más társadalmi kérdéseknél egyszerűbbnek is ígérkezik, hiszen egy adott szituáció kapcsán bárki beleképzelheti magát a közlekedés tetszőleges résztvevőjének szerepébe. Bárki lehet gyalogos a reggeli órákban, kerékpáros napközben, míg este akár egy önvezető taxi utasa. Más kérdések kapcsán identitásunk, így a képzelőerőnk is általában szűkebb korlátok között mozog.

Álláspontomat az önvezető autókkal kapcsolatos legalapvetőbb dilemmával illusztrálom. Tegyük fel, hogy az önvezető autó olyan vészhelyzetbe kerül, melyben

vagy elgázol és megöl egy csoport gyalogost, vagy kitér, és saját utasait áldozza fel.[246] Más lehetőség nem jöhet szóba. Ez a gondolatkísérlet olyan különféle diszkriminatív tényezők révén lett továbbfejlesztve, mint az áldozatul eső gyalogosok és utasok száma, életkora, neme, a balesetben érintett személyek társadalmi szerepe, stb.

A példa megmutatja, hogy az ilyen helyzetek értékelése milyen óriási mértékben függ mind saját, mind az autó episztemikus képességeitől.

Valójában, ahogy az egy kerékpáros életét követelő arizonai Über baleset is mutatja, az autó nem tudhatja pontosan, milyen tárgyat érzékel. A helyzetet tovább rontja, hogy maga a jármű a tárgyi kategorizálás bizonytalanságának csak részlegesen van tudatában. Emellett bebizonyosodott az is, hogy a neurális háló – az azonosítást végző technológia – becsapható.[247]

Egy kísérletben a megkérdezettek nagy többsége azon az utilitarianista véleményen van, hogy a rendszer egy életet feláldozhat, ha ezzel többet sikerül megmentenie. De egy efféle előzetes döntés azzal a nem-triviális következménnyel jár, hogy ez a járművel kapcsolatos közismert viselkedés ártó szándékkal használható arra, hogy a rendszer emberi életeket oltson ki – akár az autó elé kiugorva, vagy ha a tárgyfelismerés becsapható, még erre sincs feltétlenül szükség. Más esetben a gyalogos félreugorhat, de megeshet, hogy az autó ugyanebben az irányban próbál kitérni, így az elkerülni kívánt tragédia épp emiatt következik be. A felmérésben résztvevők ilyen forgatókönyvek láttán gyakran megváltoztatják korábbi véleményüket.

Ehelyett, kísérletezzünk egy kontraktariánus jellegű indítvánnyal: 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.

Ez a kisarkított javaslat – legalábbis a jelenlegi közlekedési viszonyok ismeretében – egyes egyéni esetekben tragikus veszteségeket eredményezhetne, melyek emberi sofőr esetén vitathatatlanul elkerülhetőek lennének. Ugyanakkor, egy ilyen szabály

[246] Lin (2015).
[247] Nguyen és társai (2015).

egyszerűsége miatt olyan mértékben képes a körülmények formálására, hogy a fenti helyzet elő sem fordulhat.

Nagyjából száz évvel ezelőtt, amikor az autók még újdonságnak számítottak, a gyalogosoknak az utak egy részéről le kellett mondania – ez a lovas hintók idején nem volt így. . Az autó sebessége, szemben a lovaskocsiéval nem tette lehetővé azt, hogy ott keljenek át az úttesten, ahol csak szeretnének. E helyett kapták a lámpával ellátott gyalogosátkelőt, amely védi a gyalogost, ám ha ez a személy 100 méterrel arrébb, hirtelen lép az úttestre, akkor csak magát okolhatja az esetleges balesetért. Újraosztottuk tehát a felelősségeket, eliminálva egy felelősségrést (lásd a 7.1.2 fejezetet), ami abból adódott, hogy az autók technológiája és sebessége nem teszi lehetővé azt a hirtelen megtorpanást, amit a kocsis vagy akár a ló még meg tudott oldani.

A kontraktariánus megközelítés abban az értelemben racionális, hogy nem törekszik az utóbbi néhány száz év során már makacsul jelen lévő morális jellegű kérdések megoldására. Ugyanakkor, választ ad a jövőbeni helyzetek elképzelhetetlen voltára, amely a tervezési fázisban végzett munkában igencsak valóságos kihívás. Ehelyett a tervezés során egy egyszerű szabályrendszert terjeszt elő, minden érintett beleegyezését kérve, és ennélfogva a futásidő alatt jobban kontrollálhatóvá teszi az emberéleteket befolyásoló szituációkat azáltal, hogy könnyen kiszámítható.

A „sosem tér ki” szabály például lehetővé teszi, hogy a közlekedés emberi résztvevői eddig nem tapasztalt módon megváltoztassák az önvezető járművekkel kapcsolatos általános vélekedésüket, hogy azok jelenlétét saját érdekeik szerint kezelhessék (pl. elugorjanak, tudva, hogy az önvezető jármű biztosan nem fog épp abba az irányba kitérni).

Ahhoz, hogy a kontraktariánus megközelítés működőképes lehessen, bizonyos alapelvekhez ragaszkodnia kell – az egyszerűség és az érthetőség mellett fontos a viselkedési minták közlése, akár reklámokkal való támogatása; a nem is egyöntetű, de legalábbis kompromisszumszintű elfogadás; és az e viselkedési minták lehető legkonzisztensebb működésének biztosítása.

A fentiek alapján három pontban foglalhatjuk össze az MI konstrukció kontraktariánus megközelítését:

1 A szabályrendszer legyen elég világos és érthető ahhoz, hogy a nagyközönség számára is előterjeszthetővé váljon, így következményeit mindenki felmérhesse.

2 Feleljen meg a „tudatlanság fátyla”[248] teszten, abban az értelemben, hogy már a javaslattevő győződjön meg arról, hogy bármely érintett helyébe képzelve magát kielégítő a megoldás.[249] Ennek a pontnak a kielégítése szándékosan diverzzé tett konstruktőrcsapattal könnyebb lehet, azaz az episztemorális tervezési probléma alátámasztja a tervezői sokszínűségre való törekvést[250], amely így már nem csak önértékként értelmezhető.

3 Az így létrehozott szociotechnikai rendszernek (MI + szabályozás + közösségi ismeret + konszenzus), lehetőleg minden értintett szempontjából előnyösnek kell lennie ahhoz képest, mintha nem vezetnék be.

Az MI konstruktőri feladat ilyen megközelítése a sikerhez szükséges feltételek és a megvalósításhoz adott útmutatás révén nyújt segítséget. Ugyanakkor azonban, nem ad semmilyen konkrét megoldást.[251] A kontraktariánus keretrendszer bőséges teret hagy a megvalósítás különféle megközelítési formái számára, ami azonban azt is jelenti, hogy önmagában elégtelen: a konstruktőröknek még a fenti három pont figyelembevételével is tervezési stratégiákra van szüksége.

Az első fejezetben három kulcsfontosságú kihívást azonosítottam, azt jósolva, hogy az MI etika következő hullámaiban az ezekre adott válaszkísérletekkel fog foglalkozni a kutatási terület. Az első kettő – a konstruktőri szerep hangsúlyozása azzal szemben, hogy a megvalósult MI viselkedésével lennénk elfoglalva; valamint a tervezéskori

248 Rawls (1971).

249 Például, egy önvezető jármű tervezőinek el kell fogadniuk saját javaslatukat, akkor is, ha fennáll annak lehetősége, hogy a jövőben nemcsak egy önvezető jármű tulajdonosa, hanem utas, kerékpáros, gyalogos stb. szerepét tölthetik be majd egy adott helyzetben. Ez az AV-k kapcsán természetes – aligha képzelhető el olyan konstruktőr, aki kizárólag az egyik szerepben van. Ám más MI feladatosztályok esetén egész más a helyzet.

250 Ilyen törekvés például az IEEE 7000-2021 szabványban található „felhasználó képviselő” (user advocate) szerep. Ezt a szabványt röviden érintem a 7. fejezetben.

251 A fent említett „nincs kitérés” példa deontológiai jellegű, de szóba jöhetnek tökéletesen megfelelő konzekvencialista megközelítések is.

episztemikus korlátok – expliciten megjelentek ebben a fejezetben. A számítógépek metafizikája is ott rejtőzködött a háttérben a 4. fejezetben bemutatott átláthatatlanságon keresztül. A következő fejezetben ez a kulcsfontosságú kihívás sokkal nagyobb szerepet fog kapni.

11. Az MI morális státusza

A mesterséges intelligencia morális státuszának kérdése már jó ideje vita tárgyát képezi a terület releváns folyóirataiban, legalábbis Putnam[252] úttörő cikkének megjelenése óta, amely a mesterséges életet és a robotok jogait tárgyalja.

A szakirodalom a morális státusz két formáját különbözteti meg: a morális ágenst és a morális pácienst. Természetesen ezek nem kizáró viszonyban vannak: 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.[253] 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. 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.

Ahogyan az MI egyre erősebb, úgy merül fel ez a kérdés is egyre élesebben. Ez természetes, hiszen az MI az ember helyébe lép társas interakciókban, döntéshozásban. Mindemellett nemcsak a szerepük, hanem a működésmódjuk is felveti a morális státusz kérdését: például a Putnamhoz hasonló funkcionalisták számára a jelenlegi csúcstechnológiás agy-emuláció bizonyos típusai akár már most elmének tekinthetők, ez pedig automatikusan felveti a morális státusz kérdését.[254]

Anélkül, hogy az ágenciát és a páciens státuszt világosan megkülönböztetném – ezek egyébként is összefüggenek – összességében az MI morális inklúziójáról írok, amely a kettő közül legalább az egyiket jelenti. Az *inklúzió* a kérdés szakirodalmában a morális státusszal bírók körébe való befogadást jelenti.

252 Putnam (1964).

253 A fogalmakhoz egy jó bevezető Danaher (2019). Itt mindenképp meg kell említenem, hogy mindkét kategória erősen vitatott, ahogyan arra Danaher is felhívja a figyelmet,

254 Z. Karvalics László (2024) a magyar szakirodalomba ezt problémát a mesterséges tudatosság és mesterséges élet „mesterkontextusa”-ként vezette be.

11.1 Az MI morális inklúziójának módjai

Az MI ágensek morális inklúziója elkerülhetetlen lehet, ha teljesít egy „kemény feltételt”,[255] ami szerint erkölcsi keretrendszer alanyaként kell kezelni.

Például egy funkcionalista elmefilozófus azt állíthatja, hogy a csúcstechnológiás teljes agy-emuláció, amely jelenleg egy rágcsáló teljes idegi aktivitásának emulálásának határán van, érző elmét hoz létre. A funkcionalizmus ebben a kontextusban azt jelenti, hogy az, ami egy elme-funkcióját képes betölteni (márpedig ez empirikusan tesztelhető) az elme is, és ezért be kell vonni az utilitarista erkölcsi kalkulusba. Ebben az értelmezésben a funkcionáló elmével való rendelkezés a kemény feltétel.

Egy másik kemény feltétel, a tudatosság elegendő alapot nyújthat a deontológiai etika szerinti befogadáshoz. Természetesen az, hogy az MI ágensek rendelkezhetnek-e bármelyik ilyen erkölcsöt megalapozó tulajdonsággal, régóta filozófiai vita tárgyát képezi. Láthatjuk, a kör már bővült az állatok morális befogadásával, tehát a morális megfontolásra érdemes kör dinamikus.

Az MI ágensek morális inklúziója anélkül is megtörténhet, hogy azok inherens, morális jelentőséggel bíró tulajdonságokkal rendelkeznének.

Jogokat kaphatnak a nagyobb rendszerben betöltött szerepük alapján is, ami a morális inklúzió egy formájának is tekinthető. Ez hasonló lenne néhány környezetetikai irányzathoz, amelyek morális státuszt adnak olyan növényeknek vagy víztömegeknek, stb., amelyek nem rendelkeznek a morális inklúzió inherens belső feltételeivel.

Az ágensek morális befogadásához vezető harmadik út egy újszerű metaetikai keretrendszerhez kapcsolódik. Ez a relációs megközelítés, amely a morális befogadást a társadalmi kapcsolatok hálózatában betöltött pozíció alapján ítéli meg. Ez például azt jelentené, hogy egy MI-ágens morális befogadást nyerhetne, mert évek óta egy család megbízható társa, míg egy másik, ugyanolyan típusú, de sosem használt MI ágens nem.

Továbbá, felmerülhetnek olyan igények is, amelyek az MI bizonyos formáinak morális befogadását az emberek kívánságainak és preferenciáinak tiszteletben tartása miatt

[255] Coeckelbergh (2010).

szorgalmazzák, anélkül, hogy új morális elméletekre lenne szükség. Várható, hogy az interaktív társaság nyújtása az emberek számára a jövőben az MI egyik fő alkalmazási területe lesz, és az emberek nagyra értékelhetik mesterséges társaikat, valamint igényt tarthatnak folyamatos karbantartásukra és az azokkal való törődésre, ahogyan azt történelmileg jelentős tárgyakkal, például épületekkel is teszik.

Végül egyesek az MI ágensek, különösen a humanoid robotok morális befogadása mellett érvelnek, még akkor is, ha semmilyen belső vagy külső morális jelentőséggel bíró tulajdonsággal vagy kapcsolattal nem rendelkeznek. Ennek egyedüli célja az lenne, hogy elkerüljük az embereknek okozott bizonyos negatív pszichológiai hatást: nevezetesen azt, hogy tudatosan kikapcsoljuk a gátlásainkat valami olyannal szemben, ami egy valódi személynek tűnik. A megoldandó probléma lényege az, hogy egy emberien viselkedő ágenst az elménk ösztönösen személyként azonosít, azonban ezt az ösztönt megtanuljuk elnyomni, mert teoretikus síkon tudjuk, hogy a mesterséges ágens nem „igazi”, így azzal morális obligációk nélkül bánhatunk.

11.2 A morális inklúzió tétje

A mesterséges intelligencia morális befogadására vonatkozó kérdésnek az is a tétje, hogy az általunk elfogadott morális keretrendszer konzisztenciáját meg tudjuk-e őrizni.

Egy ilyen vita általánosságban is hatással lehet a morális befogadás elméleteire. A múltban (és sajnos a jelenben is) bizonyos népcsoportok és közösségek morális nem-befogadása olyan eseményekhez vezetett, amelyeket ma erkölcstelennek vagy emberiség elleni bűncselekménynek nevezünk, mint például a rabszolgaság, a népirtás, a nők elnyomása és egyéb igazságtalanságok.

Ha a jövőben igazolva látnánk, hogy egy MI ágenst morálisan be kell fogadnunk, akkor annak kizárása ellentmondana erkölcsi törekvéseink alapjainak, és „fajizmusnak" vagy „szénalapú előítéletnek” minősülne és mindez teljesen megváltoztatná a társadalmunkat.

Másfelől az MI morális inklúziójával kapcsolatos gondolatmenet cáfoló kimenete ugyanolyan értékes lehet: eljuthatunk egy olyan feltételrendszerhez, amely segít

egyértelműen elhatárolni a jogokkal rendelkező, morálisan befogadott lényeket az olyan mesterséges ágensektől, amelyek pusztán eszközök, mindezt kielégítő és elméletileg megalapozott módon.

Nem biztos azonban, hogy a mesterséges és az emberi között mindig olyan könnyen megtaláljuk majd a határt. Ez egy további érv az MI morális inklúziója mellett. Valós lehetőség az egyre mélyülő ember–gép szimbiózis. Ma már sok ember rendelkezik különböző idegi protézis szervekkel, beleértve a mesterséges idegpályákat is[256]. Ha extrapoláljuk a terület fejlődését a jövőre, eljutunk Hans Moravec 1988-as gondolatkísérletéhez,[257] amelyben az összes neuront egyenként helyettesítik funkcionálisan egyenértékű gépekkel.

11.3 Utak a morális inklúzióhoz

A morális inklúzió egy tág fogalom. Fontos, hogy amikor morális megfontolásról beszélünk, az nem jelenti azt, hogy a gép morálisan ekvivalens az emberrel. Operacionalizált értelemben a morális inklúzió magában foglal minden olyan elképzelést, amely azt sugallja, hogy amikor egy MI ágenst létrehoznak, működtetnek vagy leállítanak/törölnek, a rendszer fejlesztőinek vagy üzemeltetőinek nemcsak az emberekre gyakorolt hatásokat kell figyelembe venniük, hanem magát az MI ágenst is ennek az értékelésnek az alanyává kell tenniük. Ez azt jelenti, hogy az MI jólétét vagy érdekeit, és egyes elméleti keretekben a jogaikat is valamilyen módon tiszteletben kell tartani.

Fontos azonban, hogy az morális megfontolás alapja MI-specifikus legyen. Néhány tárgy, mint például egy ritka veterán autó vagy egy történelmi épület, bizonyos formában morális megfontolás alá esik: szeretik, védik, karbantartják, megemlékeznek róla, és gondosan díszítik. Ez azonban nem az, amiről beszélünk: a kérdéses morális megfontolásnak specifikusan a mesterséges intelligenciára kell vonatkoznia, minden más technológiai eszközzel szemben. Ezért kapcsolódnia kell ahhoz, ami megkülönbözteti az MI-t mint technológiát: ez pedig az intelligens viselkedés képessége.

[256] Ganzer és társai (2020).
[257] Moravec (1988).

Ezekkel a demarkációkkal elkezdhetjük áttekinteni, hogyan jelenik meg az MI morális inklúziójának kérdése az irodalomban. A téma közvetett és közvetlen módokon is előfordul.

A közvetett tárgyalások a morális értékek bizonyos változásához vagy fejlődéséhez kapcsolódnak: például a színes bőrű emberek, nők, vallási kisebbségek emancipációjához. Ez az morális befogadás körének kiterjesztésével történik, miközben eltávolodunk a rokonság, a csoporttagság vagy a származás alapján történő befogadástól. A döntés alapját egy olyan belső tulajdonság képezi, amelyet az adott alany birtokol, nem pedig valamilyen csoportspecifikus örökölt tényező: így a fajfüggetlenség fogalma már korán megjelent a diskurzusokban.

11.3.1 Operacionalizált megközelítések

Érvelésem szerint a robotok morális megfontolása már közvetetten megjelenik Alan Turing híres, 1950-es esszéjében,[258] amely egy kísérlet a gondolkodás fogalmának kiterjesztésére a úgynevezett Turing-teszt segítségével. Bár Turing nem állítja, hogy a gondolkodás fogalmának operacionálásával és tesztelhető jellemzővé alakításával bármilyen alapot teremtene a morális megfontolás számára, ez mégis korai példája annak, hogyan távolodnak el az elmével kapcsolatos fogalmak a biológiai gyökerektől.

A már korábban is idézett Hillary Putnam[259] – egy Turing által inspirált gondolkodó – hat évtizeddel ezelőtt megfogalmazott funkcionalista álláspontja alapján kifejezetten javasolta a robotok morális megfontolását. Ezt két alapon nyitja meg elméleti lehetőségként. Az egyik forgatókönyvben („akarat”) a robotok jogokat követelnek. Ez a forgatókönyv lényegében megvalósult, amikor Blake Lemoine beszámolója szerint a Google LaMDA MI-je érzetekről számolt be és megkérte alkotóit, hogy ne kapcsolják ki. Általánosságban véve a jogok követelését a morális páciensi státusz követelésének egyik formájaként kategorizáljuk. Putnam víziója szerint ebben a forgatókönyvben az emberek válaszút elé kerülnek, hogy miként reagáljanak a kihívásra. Blake Lemoine-t kirúgták a Google-től.

A másik forgatókönyvben („érdek”) az emberek elkerülhetetlennek találják, hogy jogokat adjanak a robotoknak azok tulajdonságai alapján. Putnam úgy véli, hogy egy

[258] Turing (1950).
[259] Putnam (1964).

jogokkal rendelkező MI-rendszer megjelenése abszolút elképzelhető a belső tulajdonságoktól függetlenül is.

11.3.2 Érzetek

Az utilitarista Jeremy Bentham[260] egyértelműen úgy vélte, hogy az állatokat, mivel képesek szenvedést átélni, morálisan figyelembe kell venni, és be kell vonni az utilitarista döntéshozatal kalkulusába. Ez az átmenet azért releváns számunkra, mert általánosítja és igyekszik természetessé tenni az érzetekre való képességet, valamint egyúttal a morális konzisztencia alkalmazásának egyik fő példája. Singer állatfelszabadításról szóló munkája ennek a nézetnek a modern formája.[261]

Ha arra a következtetésre jutnánk, hogy az MI képes az érzeteket tapasztalni, miközben fenntartjuk az morális konzisztencia iránti törekvést, akkor ezeket a gépeket is be kellene vonni a morális megfontolás körébe, mert az MI megfelel a „kemény feltételnek”. A kemény feltétel problémája azonban abban rejlik, hogy általában olyan ontológiai állításokra támaszkodik az alanyról, amelyeket köztudottan nehéz vagy lehetetlen igazolni.

Nincs kikezdhetetlen konszenzus abban, hogy mi számít érző lénynek, sőt abban sincs egyetértés, hogy ez a kérdés empirikusan tesztelhető-e. Így kapcsolódik be az elmefilozófia az MI morális befogadásának kérdésébe.

Több olyan feltételezés-konstrukció is létezik, amely lehetővé teszi az érezni képes MI-t. Az elmefilozófiai funkcionalizmus[262] teret enged annak, hogy a biológiai neuronokat mesterséges neuronokkal helyettesítsék, miközben megőrzik az elme lényeges tulajdonságait. Ez az álláspont azt állítja, hogy a neuronok fő hozzájárulása abban rejlik, hogy milyen funkciót töltenek be az egész részeként, nem pedig a biológiai részletekben.

Ez azt jelenti, hogy egy komoly kísérlet az agy szimulációjára olyan elméket hozhat létre, amelyek alapvető tulajdonságaikban hasonlóak a biológiai alapú elmékhez, beleértve az érzés képességét is, amelyet egy funkcionalista álláspont szerint akár el is lehet vonatkoztatni a biológiai alapoktól.

260 Ebben a témában lásd: Frey (2011).
261 Singer (1975).
262 Heil (2004).

11.3.3 Agyprotézis és ember-gép szimbiózis

Saját intuíciónkat a funkcionalizmusról Hans Moravec gondolatkísérlete, az úgynevezett *agyprotézis* segítségével tesztelhetjük. Ebben a jövőbe helyezett gondolatkísérletben olyan mesterséges neuronok léteznek, amelyek funkcionálisan egyenértékűek a biológiai neuronokkal. Egy eljárás során az összes biológiai neuront egyenként helyettesítik mesterséges neuronokkal.[263] Moravec azt kérdezi, hogy mit éreznénk az eljárás során és azt követően. Ez a felvetés kiélez egy problémát: kérdéssé teszi, hogy a morális befogadás szempontjából elfogadhatjuk-e a nem-biológiai helyettesítést egyenértékűként a biológiai szervvel, és ezzel együtt logikailag nyitva hagyjuk a lehetőséget a nem emberi MI számára is. Máskülönben, ha elutasítjuk az egyenértékűséget, fennáll a veszélye annak, hogy megtagadjuk a teljes emberi státuszt azoktól, akiket neurális protézisekkel gyógyítottak, kezeltek, vagy augmentáltak.

Egy másik feltételezéskészlet az intelligens viselkedés és az olyan rendszer tulajdonságok közötti kapcsolattal foglalkozik, amelyek képesek ezt a viselkedést előállítani.

Például, ha feltételeznénk, hogy bizonyos magas szintű és összetett intelligens viselkedés csak egy tudatos elme eredménye lehet, akkor újabb jeleit látnánk annak, hogy egy kemény követelmény teljesül, ha azt is feltételezzük, hogy a tudatosság lehetővé teszi az érzőképességet.

Feltételezhetjük, hogy a Whole Brain Emulation (WBE) támogatói, akik az elme számítógépekbe töltését célozzák, egyértelműen a mesterséges elmékkel szembeni morális megfontolás mellett is érvelnek, mivel a projekt alapfeltételezése, hogy az elme biológiai alapú létezése nélkül is hasonló érző tapasztalatokat lehet megélni, illetve, és az előterjesztők akár a saját tudatukat is feltöltenék. Ez tehát a funkcionalista nézet támogatását jelenti.[264]

Ezért nem túlzás feltételezni, hogy a szimulált elméket vizionáló Nick Bostrom, valamint a biológiát meghaladni és örökké élni vágyó Ray Kurzweil[265] is bizonyos mértékig funkcionalisták. Ugyanakkor a WBE egy emberi elme reprodukálásáról szól, ami nem azonos a mesterséges intelligenciával (MI).

[263] Az eredetért lásd Moravec (1988), magyar nyelven Héder (2020, 11. fejezet).
[264] Sandberg (2013).
[265] Sandberg és Bostrom (2008), Ray Kurzweil (2005).

Ennek a kérdésnek a relevanciája az MI morális befogadása szempontjából abban rejlik, hogy ugyanaz a technológia képes lehet mesterségesen létrehozott elméket is emulálni, amelyek talán némileg egyszerűbbek az emberi agyhoz képest, de amelyeket biztosan MI-nek lehetne kategorizálni.

A funkcionalizmus elfogadása és az morális konzisztenciára való általános törekvése miatt fel kell tennünk a kérdést: ha a szimulált emberi elmét erkölcsileg figyelembe kellene venni, akkor nem kellene-e ugyanezt tennünk egy mesterséges elmével is?

Számos előrejelzés állítja, hogy ez a kérdés elkerülhetetlenné válik a jövőben, de nincs egyetértés abban, hogy ez a jövő milyen közel van. Néhányan nem zárják ki ezt a fejleményt, de elutasítják, mint egy távoli jövő problémáját.[266] Ez az elutasítás talán nehezebben tartható a nagyszabású agytérképezési és jelenlegi szimulációs trendek fényében,[267] de értékelésünk néhány feltételezéstől függ.

Többen úgy vélik, hogy ez a fejlemény a közeljövőben bekövetkezik. Levy több mint egy évtizeddel ezelőtt úgy vélte, hogy az ilyen fejlődés – tudatos elmékkel – küszöbön áll, és Goertzel is hasonlóan vélekedett majdnem húsz évvel ezelőtt.[268] Ezen az alapon megismételték Putnam előrejelzését, miszerint a robotok jogokat fognak követelni.[269] 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.

11.3.4 Kötelességek

Nem csak a konzekvencialista, hanem a deontológiai etika is rendelkezik saját kemény követelményekkel. Ahogyan Singer ezt az utilitarizmus esetén tette, Regan[270] az állati jogokról mutatja meg esszéjében, hogyan lehet fajfüggetlen deontológiai etikát kialakítani. Ezt az elvet át lehet ültetni a mesterséges intelligencia kontextusába is. Bár vitathatóan még nehezebb, mint a funkcionalizmus és az érzőképesség kérdése, a kanti megközelítést egy tudatos mesterséges elmére már megfogalmazták.[271]

[266] Lásd Gunkel (2018) e hozzáállás leírásáért, példákkal együtt.
[267] Einevoll és társai (2019).
[268] Levy (2009) és Goertzel (2002).
[269] Brooks (2000) és Asaro (2006).
[270] Regan (1983).
[271] Stuart és Dobbyn (2002).

Regan esszéje arra világít rá, hogy az erkölcsi megfontolásnak nem feltétlenül kell fajspecifikusnak lennie, hanem az egyetemes elvek alapján is meghatározható. Hasonlóképpen, a kanti etika egy mesterséges tudatos elme esetében is alkalmazható, feltéve, hogy az alany rendelkezik azokkal a tulajdonságokkal, amelyeket Kant elengedhetetlennek tart a morális alanysághoz, – ide sorolható például a racionális cselekvőképesség.

Az ilyen fejlemények értékelése ontológiai vagy akár metafizikai feltételezéseinktől függ. Ebben a fejezetben nem vállalkozom annak az eldöntésére, hogy a funkcionalizmus a helyes megértése-e az elmének, vagy hogy egy kantiánus tudatosság-leírás elérhető-e. De talán ezen metafizikai kérdések megoldása nélkül is lehetséges megfogalmazni olyan feltételeket, amelyek biztosítják az morálisan megengedhető agyemuláció és a mesterséges neurális hálózatok létrehozását, amennyiben erre szükség lenne.

11.3.5 Meta-etikai újítások

Vannak érvek bizonyos fokozatos erkölcsi befogadás mellett akkor is, ha az MI ágensek nem rendelkeznek olyan belső tulajdonságokkal, amelyek erkölcsi alannyá tennék őket, de emberi megjelenésűek. Például Floridi és Sanders[272] kiterjesztik azokat a tulajdonságokat, amelyek erkölcsi törődést érdemelnek, az autonómiára, interaktivitásra és alkalmazkodási képességre hivatkozva, így a metafizikailag bonyolult, kemény követelményeket operatív követelményekkel helyettesítik, amelyeket az MI-k is mutathatnak.

Vannak radikálisan eltérő megközelítések is. Fentebb már bemutattam a relációnizmust[273] ami megengedi, hogy két, strukturálisan hasonló MI-t nagyon eltérően lehessen kezelni attól függően, hogy milyen kapcsolatokban vesznek részt. Ennek az érvnek az árnyoldala, hogy az emberekre nem működik, így nem teljesen konzisztens (emberek esetén nem engednénk meg, hogy az egyén morális értéke a szociális kapcsolataitól függjön, hiszen önmagában értékesnek tartjuk az emberi lényt).

[272] Floridi és Sanders (2004).
[273] Coeckelbergh (2010).

Továbbá az MI morális megfontolását nem feltétlenül csak az ágens érdekében, hanem a társadalmi következmények miatt is szorgalmazhatjuk. Például, ha nagyobb figyelmet fordítunk az MI-re, elkerülhetjük a tudatos dehumanizáció ösztönzését egy emberi megjelenésű entitással szemben.[274] Darling az antropomorf robotok emberre gyakorolt pszichológiai hatásából kiindulva érvel az MI morális státusza mellett.[275]

A fentiek összefoglalhatók úgy, mint az erkölcsi fejlődésre tett kísérletek különböző formái.[276] A morális fejlődésnek vannak kevésbé erős és radikálisabb formái. A kemény követelményekre alapozott érvek arra szólítanak fel bennünket, hogy ismerjük el az MI ágenseket, és helyesen alkalmazzuk rájuk a morális keretrendszert.

Mások radikálisabbak, mivel céljuk, hogy a gyakran használt morális keretrendszereket újakkal helyettesítsék: Coeckelbergh relációs fordulata és Sauer új erkölcsi fejlődési rendszere példák ezekre az erőfeszítésekre.

Ez a rövid összefoglalás talán rávilágít arra, hogy az erkölcsi fejlődés különböző megközelítései hogyan vezethetnek eltérő következtetésekhez az MI morális befogadásával kapcsolatban.

11.4 Érvek az MI elutasítása mellett

Nem kockáztatok sokat, ha azt feltételezem, hogy az elmefilozófia metafizikai jellegű kérdései nem fognak egyértelmű válaszra találni a közeljövőben csak azért, mert elérkezett a hatékony mesterséges intelligencia, amelynek értékeléséhez ezek a válaszok azonban nagyon jól jönnének.

Az ismeretelmélet története arra tanít minket, hogy bizonyos esetekben a kérdés megfordítása előnyös lehet. Ha a bizonyítás még logikailag sem lehetséges, elképzelhető, hogy érdemes a cáfolattal próbálkoznunk. Ha találnánk olyan feltételeket, amelyeknek fennállása garantálja, hogy egy adott MI ágensnek nem jár morális inklúzió, az még akkor is nagy segítség lenne, ha a kérdést általánosságban nem tudjuk eldönteni.

[274] Whitby (2008).
[275] Darling (2012).
[276] Macklin (1977), Moody-Adams (1999), Rorty (2007), Sauer (2019).

Így egy védekező stratégiát kapunk, amelynek során azt vizsgáljuk, hogy mi az az MI ágens, ami ilyen szempontból „biztonságos”, azaz még a megengedő követelményrendszerek alapján sem érdemel morális megfontolást.

11.4.1 Inklúzió tulajdonságok alapján

Ugyanúgy, ahogy a morális befogadás, a kizárás lehetősége is aluldeterminált. Egy lehetséges kizárási alap a legújabb, az MI és az emberi elme közötti különbségekre vonatkozó ismeretekből származhat. Több javaslat is született arra vonatkozóan, hogy az embereket az MI-től az emberi biológiai agy birtoklása alapján különítsük el.[277] Ezek a jól ismert és sokat vitatott bioesszencialista (és egyben antifunkcionalista) kategorizálások újragondolásra szorulnak az MI jelenlegi állásának fényében. Az ilyen álláspont elfogadása különbséget tenne a morális elfogadott és elutasított ágensek között, még ha hasonló emberi viselkedést is mutatnak.

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. Ez azt mutatja, hogy a mentális tevékenység jelenléte fontosabb, mint pusztán az élet, vagy alternatívaként, hogy az élet megléte elsősorban az agy állapotától függ (innen a „agyhalott" kifejezés). Az előző érvvel kombinálva megállapíthatjuk, hogy az életben levés és az egészséges, megfelelő típusú agy birtoklása szükséges feltételek lehetnek.

11.4.2 Származás, útvonalfüggőség

Egy másik lehetőség az erkölcsi elfogadás/elutasítás meghatározása a származás alapján. Esszencialista megközelítésekben az eredet – például az emberként való születés vagy az evolúciós történelem – fontosabb lehet, mint a kifejezett intellektuális teljesítmény. Vagy tágabb értelemben a csoporthoz tartozás fontosabb lehet, mint az egyedi ágens struktúrája és teljesítménye. A biológiai/antropológiai határvonalak felülvizsgálatának fő oka, hogy sok ilyen javaslat bevallottan egy előzetes filozófiai álláspont formájában fogalmazódik meg, nem pedig bizonyítékokon alapuló végleges ítéletként.

[277] Penrose és Mermin (1990); Kauffmann (2016).

Az efféle elveken alapuló morális elutasítás ugyanúgy metafizikai elköteleződésektől függ, mint az elfogadás és örökké ki lesz téve a diszkrimináció vádjának.

11.4.3 Graduális inklúzió

Vegyünk egy analóg példát a bioetikából: az agykutatásban emberi agysejtek gyűjteményeit növesztik indukált pluripotens őssejtek segítségével, hagyják, hogy ezek a sejtek kapcsolatokat alakítsanak ki, sőt, patkányok vagy egerek agyába is beültetik ezeket, ahol egy ideig életben maradnak. Ezt a gyakorlatot etikai bizottságok engedélyezik, arra hivatkozva, hogy ezek az agyi organoidok túl kevés sejtből állnak, és összetettségük nem teszi lehetővé a magasabb szintű érzékelést, és ezért nem szükséges erkölcsi befogadásuk.[278]

Egy gyümölcsöző kutatási kérdés lenne az ilyesféle ismeretek átültetése az MI neurális hálózataira, a neuroanatómiai eredmények alapján. Ezáltal egy demarkációs eszközt kapnánk a még biztosan megengedhető méretű és struktúrájú neurális hálók és a neurális hálók általános osztálya között. Ugyanezt az ötletet alkalmazva felvethető az is, hogy milyen határokat lehet találni alternatív MI-etikai megközelítésekben, mint például a relációs keretezésben.

Egy másik, sokkal nehezebb kérdés az, hogy van-e bármilyen elővigyázatossági intézkedés, amelyet figyelembe kell venni a nem biológiai ihletésű MI-megközelítéseknél, vagyis azoknál, amelyek nem utánozzák a neurális hálózatokat, de mégis képesek intelligens viselkedést kifejezni. Ebben az esetben nincs biológiai analógia, amire támaszkodhatnánk, de a funkcionalizmus mégis segíthet, ha rendelkezésre áll kielégítően magas szintű funkcionális leírás, amely nem támaszkodik mesterséges vagy biológiai neuronokra. Ekkor rendelkeznénk a nyelvvel, amin el tudnánk határolni a „biztonságos” MI megoldást a többitől, ugyanakkor szükségünk lenne egy intuícióra a megkülönböztetési elv meghatározásához. Ilyen elv lehet valamiféle komplexitás-mérték.

11.4.4 A bizonyítás terhe

A morális inklázió elleni és melleti érvcsoport még sincs szimmetrikus helyzetben. Ugyanis, ha az ártalom minimalizálásának elvét követjük, akkor nem mindegy, hogy milyen irányú hibát vétünk: úgy cselekszünk, hogy feltételezzük, hogy az MI nem

[278] De Jongh és társai (2022).

érdemes morális inklúzióra, és esetleg mégis; vagy fordítva, az MI-nek morális státusz tulajdonítunk, esetleg feleslegesen.

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.

11.5 Alternatív morális attitűdök

Az emberi állapot határainak vizsgálata a morális elfogadás megértése céljából irrelevánssá válhat azáltal, hogy egyre inkább bevonjuk az állatokat a morális páciensek körébe. Itt a fő, morális törődést érdemlő tulajdonság az állat fájdalom és szenvedés átélésére való képessége, amelyet néha a komplex idegrendszer meglétéhez vagy konkrétabban a gerincvelő jelenlétéhez kapcsolnak. Ezzel kiszélesedett a morális kör és kevésbé komplex lények is megjelentek benne.

Egyértelműen könnyebb azt állítani és megvédeni, hogy az MI elérheti az állatok morális státuszát (az embereké helyett). Valószínűleg még bölcsebb, ha eleve egy harmadik kategóriát javaslunk, amely az állatok szintjének környékén helyezhető el.

Az sem kizárt, hogy a közvélekedésben uralkodó meta-etikai megközelítés szintjén következik be olyan változás, hogy azután a morális elfogadás forrását nem is az ágens saját jellemzőiben kell keresnünk, hanem például a relációkban, Coeckelbergh fentebb bemutatott keretrendszere vagy egy hasonló elv alapján.

Ekkor úgy mutathatnánk be az erkölcs történetét, mint egy folyamatos tágulást. Morális attitűdjeink fokozatos nyitását az emberektől az állatok felé, majd az élőlények általános kategóriája felé a teljes bolygót magába foglalva; ugyanakkor figyelembe véve a szociális relációkat is. A relációk figyelembevétele az inherens tulajdonságok mellett a morális fejlődés egy újabb dimenziójává válna.

Így hozzászoknánk, hogy többféle morális elfogadási mód létezhet, és lehetőség van arra, hogy egy új kategóriát hozzunk létre a mesterséges intelligencia számára,

amelynek jellemzőit részben tulajdonságaiból, részben relációiból vezetnénk le, elkerülve a nem-empirikusan hozzáférhető elemeket, mint a tudat vagy az érzetek.

Ehhez illesztenénk a morális profilt is. Ahogyan a kedvenc háziállatnak jár a fájdalom lehetőség szerinti elkerülése, de nem jár az autonóm banki ügyintézés, úgy az MI esetén lehet, hogy épp a fordítottját kapjuk. A morális inklúzió graduális és több-dimenziós elképzelése számos új kategóriát hozhat létre, amelyben a mesterséges intelligencia is szerepet kaphat.

Egy másik lehetőség, hogy a mesterséges intelligencia (vagy legalábbis egyes MI ágensek) morális elfogadásához talán újfajta episztemikus keretekben előálló, újfajta empirikus érvek is felfedezhetők. Más szóval, az ontológiai állítások kezelhetővé vagy naturalizálhatóvá válhatnak – ez végső soron összhangban van a filozófia történetével és az abból kialakult természettudományokkal. Egy példa erre a posztkritikai filozófia bizalmi programja, amelynek az MI-re való alkalmazását a következő fejezetben mutatom be.

12. Az MI tapasztalása

E fejezet azt a kérdést teszi fel, hogy az MI értékelésében, és az MI megértésében hogyan érhető el előrelépés.

Erre az egyik gyümölcsözőnek tűnő eszköz Polányi Mihály posztkritikai filozófiája, amely sajátos ismeretelméleti és metafizikai jellemzői miatt megfelelő keretrendszer lehet az MI etika számára. E terület épp egy olyan fázisba tart, amikor az előrelépéshez az ismeretelméleti és metafizikai előfeltevések explikálására és tisztázására van szükség.

A posztkritikai megközelítés a mesterséges intelligenciával kapcsolatban nem csupán marginálisan lehet előnyösebb, mint más keretrendszerek, mivel a redukcionizmus (például a számítógép redukciója a Turing gépre, lásd a 3. fejezetet) tűnik az egyik olyan tényezőnek, amely a MI tervezésénél és a morális elfogadás kérdésében is problémákat okoz.

A nem-reduktív szemlélet újra és újra előkerül a számítástechnikában, olyannyira, hogy az „emergens számítások”[279] képében saját számításelméleti alterület is létrejött. Polányi posztkritikai megközelítésének előnye, hogy az emergencia jelenségét teljességében és nem idioszinkratikus hipotézisként tárgyalja.

Az MI etika elkerülhetetlenül olyan kérdésekhez vezet, amelyek a mesterséges intelligencia episztemológiájához tartoznak. Ezek olykor annyira összefonódnak, hogy az „episztemorális keresés" kifejezést használom rájuk (lásd a 10. fejezetet), utalva arra a tényre, hogy sok autonóm rendszer elsődleges erkölcsi kötelezettsége az episztemikus feladatainak teljesítése. Ugyanennyire nem odázható el az MI morális státuszának a kérdése, amely természetszerűleg metafizikai jellegű (lásd a 11. fejezetet).

Mindegyik kutatási irány esetében elengedhetetlen a mesterséges intelligencia *átfogó entitásként* való mélyreható megértése. Ez egy posztkritikai leírása az MI jelenségnek, amely mentes a megtévesztő behelyettesítésektől (lásd lentebb) és az olyan radikális

[279] Héder (2014).

redukcióktól, amelyek akadályozzákazt, hogy bármilyen értelmes következtetést levonjunk.

Polányi teljes posztkritikai projektjét sokan elsősorban az episztemológiához való jelentős hozzájárulásként látják, de én amellett érvelek, hogy bár ez igaz, az episztemológia csak eszköz a cél eléréséhez. Az ismeretelmélet alkalmazása a társadalom szervezésének problémáira és végső soron mélyreható erkölcsi kérdésekre irányul, amelyek mind Polányi üzenetének részét képezik, és a mesterséges intelligencia kontextusában ma legalább annyira relevánsak, mint a Személyes Tudás megjelenése idején voltak.

Az MI posztkritikai elemzésének egy része, a számítógépekkel való interakció kontextusában már megvalósult Satinder P. Gill Tacit Engagement című monográfiájában.[280] Ebben a munkában megpróbálok túlhaladni az interfészen, egészen a kérdéses gépek magjáig.

12.1 Posztkritikai MI megközelítés

Phil Mullins rendkívül hasznos bevezetést nyújt abba, hogy a filozófián belül hogyan fejlődött a posztkritikai megközelítés, valamint meggyőzően rámutat ennek az irányvonalnak a jelentőségére is.[281] Mullins kifejti a kifejezés eredetét, valamint a Polányi által kínált szélesebb látásmódot. A kifejezést 1951-re vezeti vissza, az úgynevezett Gifford-előadások első sorozatához. Ezek középpontjában a megismerő személy elengedhetetlen személyes részvétele áll a megismerésben.

Mullins Polányit olyan gondolkodóként vázolja fel, aki megpróbálta feldolgozni a Gestalt-pszichológia eredményeinek következményeit, miszerint az egész észlelés nem redukálható részeinek összegére. Ennek tudatában szándékozom keretezni az MI értékelését.

Azt is megtudjuk, hogy Polányi a „kritikai" kifejezés alatt olyan gondolkodókat ért, mint Descartes, Hume, Kant, J.S. Mill, Bertrand Russell, és a hatalmas sikert, amit a rendszeres kételkedésük a tudomány számára hozott. Ugyanakkor úgy véli, hogy a

[280] Gill (2015).
[281] Mullins (2001). Húsz évvel később Cannon egy ennél is átfogóbb leírást készített (2021).

kritikai megközelítés, mint heurisztika, roppant sikere ellenére is kezdi elérni határait, és új megközelítésre van szükség a tudományos gondolkodás módszertanáról.

Így kapjuk meg a „posztkritikai” jelzőt. A posztkritikai gondolkodás nemcsak a határtalan kételyt próbálja felülmúlni, mint a tudomány központi heurisztikáját, hanem az személytelen tudás eszményét is. Ennek érdekében Mullins magyarázata szerint Polányi egy olyan megismerési folyamatot mutat be, amelyben sohasem tudunk mindenben egyszerre kételkedni, azaz a megismerés „bizalmi” aktus (bizalom a saját megismerő képességünk bizonyos elemeiben, például hallgatólagos elemeiben). Ebből talán az is világos, hogy a posztkritikai megismerés még mindig meglehetősen kritikai, csak éppen nem korlátlanul kritikai. Így értjük meg az is, hogy Polányi mire utal a főművének címével, amely a *Személyes tudás*.[282]

A személy aktív részvételével és az elköteleződés szerepével együtt a posztkritikai gondolkodás nemcsak mindenféle dualizmus, hanem monizmus ellen is irányul. Az emergencia fogalmán keresztül elutasítja a dichotómiát, és így az egész dilemmát a monista és dualista világnézetek között. Ez Polányi gondolkodásának posztmodern jellemzője.[283]

Polányi megismerésre vonatkozó tézisei és metafizikája között a megismerő személye jelenti az elsődleges kapcsot.

Mielőtt értékelni tudnánk, hogy hogyan érdemes viszonyulni az MI növekvő szerepéhez a társadalomban , meg kell értenünk, milyen magasra tette a mércét Polányi, amikor a *Személyes tudást* írta. Egy episztemológiai-ontológiai kérdés megértéséből kiindulva újragondolja a tudomány történetét, nemcsak azért, hogy mélyebb megértést nyújtson a múlt eseményeiről, hanem azért is, hogy újraalkossa a tudomány módszertanát, elkerülve annak helytelen alkalmazását. Történeti megközelítésével megelőzi még Thomas Kuhnt is. Ugyanakkor bizonyos előírásokat fogalmaz meg, nemcsak a szűk értelemben vett tudományra, hanem a társadalmi szerveződésre vonatkozóan is. Ezeket az ambiciózus lépéseket azért teszi, mert a

[282] Polányi (1958).
[283] Mullins Sanders (1991) és Gill (2000) munkáira irányítja figyelmünket ebben a témában.

hallgatólagos dimenzió elismerése dominóhatást indít el, és logikailag megköveteli az összes hasonló kérdés újragondolását.[284]

Azért fontos ez számunkra, mert a mesterséges intelligenciának nevezett szociotechnikai rendszer emberi elfogadásának sokdimenziós problémája szintén egy olyan megközelítést igényel, amely nem kevésbé holisztikus vagy bizalmi alapú. A mesterséges intelligenciával kapcsolatos tudásunkhoz való hozzáállás tétjei hasonló struktúrával rendelkeznek, mint ahogyan a tudományos tudáshoz való viszonyunké.

Polányi világnézetének két fontos pillére van: az egyik a hallgatólagos tudás, a másik az emergencia. A hallgatólagos tudás fogalma híresen tömören ragadja meg azt a tényt, hogy többet tudunk annál, mint amit kifejezni tudunk; vagyis képtelenek vagyunk kizárólag explicit következtetésekre támaszkodni, amikor tudományt művelünk, vagy bármilyen más szerepben tevékenykedünk. Ennek a fogalomnak a funkciója – azon túl, hogy inherens fontossággal és gyakorlati következményekkel bír – Polányi nagyobb erőfeszítésében az, hogy mérsékelje az elvárásainkat azzal kapcsolatban, hogy mit lehet elérni explicit érveléssel, beleértve a természeti törvényekről, a matematikáról vagy a társadalom értékeiről való érvelést is. A hallgatólagos tudás mindenütt jelenlévő természetének elismerése azt is jelenti, hogy bármilyen teljesítmény – bármi, amit képesek vagyunk elérni – egy olyan cselekedet példája, amely során a hallgatólagos, pillanatnyilag nem vizsgált, nem reflektált tudást integráljuk az explicit tudással. Ez elkerülhetetlenül egyfajta akritikus bizalmat igényel az előbbiben. Ha megszakítanánk a teljesítményünket és elemeznénk annak elemeit – legyen szó biciklizésről, úszásról, zongorázásról vagy a legmagasabb színvonalú tudományos kutatásról –, a teljesítmény széthullana.

A kapcsolódó kifejezések, mint a „bizalmi program" és a „posztkritikai filozófia," arra szólítanak fel, hogy olyan határozottsággal ismerjük el a mindig jelen lévő „hallgatólagos dimenziót," mint ahogyan elfogadjuk a „gondolkodom, tehát vagyok" állítást. Más szóval, Polányi úgy véli, hogy bebizonyította, hogy az állati és emberi cselekvésben, beleértve a tudást is, a hallgatólagos dimenzió vitathatatlanul egyetemes. Ebből következik, hogy mindenkinek rendelkeznie kell egy bizonyos szintű hittel saját hallgatólagos képességeiben, ha sikeresen akar teljesíteni,

[284] Mullins, P. (2021). “Michael Polanyi’s Post-Critical Philosophical Vision of Science and Society,” in Science, Freedom, Democracy. Péter Hartl and Adam Tamas Tuboly (eds.). New York and London: Routledge: 15-38.

függetlenül attól, hogy ezt elismeri-e vagy sem. Természetesen ennek az belátásnak hatalmas tudományos, mérnöki és menedzsment következményei vannak: ezen területeken a leghivatkozottabb Polányi.

Polányi hitt abban, hogy elegendő döntő bizonyítékot kínált annak alátámasztására, hogy a személyes tudás úgy működik, ahogyan ő leírja. Természetesen a racionalitás bizalmi jellegét nem lehet expliciten bizonyítani –valójában semmit sem lehet igazán explicit módon bizonyítani, mert ez ellentmondás lenne a kiküszöbölhetetlen hallgatólagos összetevővel –, így azt a Polányi által kínált, számos példa alapján kell elfogadni. Azaz, a racionalitás bizalmi természetének akceptálása maga is egy bizalmi aktus, egy tudásban való ugrás.

Ez körkörös, de legalább nem logikailag ellentmondásos. Ez a megközelítés nem felel meg a kritikai filozófia által felállított objektivitási normáknak. De ismételten, semmi más sem felel meg ezeknek, mivel ez a norma elérhetetlen; ez az, amiért a posztkritikai megközelítés szükségessé válik. Mindazonáltal a kritikai norma értékét, mint heurisztikai eszközt és a tudományos forradalom mozgatórugóját, nem lehet eléggé hangsúlyozni.

Polányi azon a véleményen van, hogy azok a rendszerek, amelyek saját körkörösségükre hivatkoznak, ezen őszinteségük miatt értékesebbek azoknál, amelyek látszólag független alapokra épülnek, de a valóságban ugyancsak körkörösek.

Természetesen a bizalmi aktus rendkívül fallibilis módja bármilyen tevékenység végzésének, de nincs alternatíva. Ez az esendőség elkerülhetetlen, és a nagyobb összefüggésekben csupán az evolúció moderálja (ami a *Személyes tudás* másik részletesen tárgyalt témája), amely biztosítja, hogy csak azok maradjanak fenn, akik elég gyakran és sikeresen képesek értékelni a valóságot.

Amikor egy tudományos vagy más állítás úgy tűnik, hogy kétségtelenül bizonyított, állítólag explicit módszerekkel, mindig szembesülünk azzal, amit Polányi „megtévesztő helyettesítésnek” nevez.[285] Ennek a lépésnek a szerkezete meglehetősen egyszerű. Azt jelenti, hogy egy személy úgy gondolja, hogy bizonyos hiedelmek (például egy esztétikai érzéken alapuló hallgatólagos matematikai intuíció) vagy

[285] Paksi és Héder, (2020).

értékek nem kompatibilisek a hitrendszerük egy másik, fontosabb részével (például az objektivizmussal), ezért az előbbit el kell vetni. Azonban, ahogy Polányi hosszasan kifejti, az ilyen hallgatólagos hiedelmeket nem lehet egyszerűen elvetni, mert ez ellentmondana a tudás szerkezetének, sőt, ezek felelősek valójában az eredmény eléréséért. Ezért a személy ezeket a hiedelmeket másként nevezi meg a tudományos módszertan leírásában. Ez elsősorban önámítást szolgál, és mellékhatásként megtévesztőnek tűnik mások számára.

A tudás keletkezésének és terjedésének megmagyarázására (ha már ezeket nem lehet teljesen explicitté tenni), Polányi saját tapasztalatait hozza fel az orvostudományi egyetemen végzett radiológia órákról. Felidézi, hogy a professzor, minden igyekezete ellenére, nem volt képes explicit módon leírni, mit kell a diákoknak keresniük a tüdő röntgenfelvételein; helyette látszólag homályos fogalmakról beszélt. Azonban a radiológiára szakosodott tanulók az osztenzív környezetben megtanulták, mit kell látniuk a felvételeket és hogyan nevezzék meg a látottakat, majd a vizsgán bizonyították, hogy nagyon pontosan azonosították azokat a releváns jellemzőket a képeken, amelyek számukra korábban ismeretlenek voltak, de a kórtörténet által megerősített információkat tartalmaztak. Így váltak elismert szakértőkké a radiológia területén, anélkül, hogy a tudás explikálódott volna. Ilyen módon kell megértenünk a tudomány természetét vagy a társadalmak értékeit.

Polányi érvelésének másik pillére az emergencia fogalma. Ennek a megértéséhez jó kiindulópont a hallgatólagos tudást birtokló személy természete. Polányi azt állítja, hogy a személy nem redukálható a fizikában vagy más tudományágakban ismert fogalmakra (ahogyan azt az ő idejében meglehetősen pozitivista szemlélettel felfogták). A világ ezen leírásai nem tartalmaznak olyan teleologikus építőelemeket, amelyek szükségesek lennének ehhez a redukcióhoz. Ez nemcsak az emberekre igaz – Polányi posztulálja, hogy léteznek olyan működési elvek, amelyek magasabb szinten működnek, mint a fizikai és kémiai törvények. Ezek irányítják a gépszerű entitásokat, mint például a mindennapi értelemben vett gépeket és élőlényeket. A működési elvek a természeten belül érvényesek, tehát nem egy duális metafizikában vagy platonizmusban értelmezendők. E helyett a természetet Polányi graduális onológiával írja le, amelynek egy magasabb szintjén effektívek a működési elvek, és az általuk leírt entitások kiemelkednek, „emergálnak” az anyagi szintből. A működési elvek

szükségesek ahhoz, hogy értelmesen magyarázzuk ezek viselkedését, de az alacsonyabb szintű feltételeknek bizonyos paramétereken belül kell lenniük ahhoz, hogy hatásukat kifejthessék.[286]

Ahogyan a hallgatólagos dimenzió esetében, Polányi ebben az ontológiai kérdésben is olyan érveket javasol, amelyeket végső soron meggyőzőnek tart, bár ezek az utókortól sokkal több kritikát kaptak, mint a hallgatólagos tudás gondolata.[287] Mégis, az emergencia valamely formájának elfogadása – még ha csak egy gyengébb formájáról van is szó[288] – meglehetősen elterjedt, a már említett emergens számítások is ide tartoznak. Polányi erős, ontológiai emergenciát javasol, amelynek alátámasztására az önmagunk (beleértve az elménket is) redukálhatatlanságát jelenti.

Azonban a hallgatólagos dimenzió és ennek következményeként a bizalmi program elismerése nem jelenti azt, hogy a hallgatólagos tudás szubjektív lenne: ez idioszinkratikus természetre és véletlenszerűségre utalna. Polányi alapvetően az evolúciót eredényeit kínálja annak mércéjeként, hogy mennyire jól képviselik hiedelmeink a valóságot. Bizalmi aktusaink jobban kapcsolódnak a valósághoz, mivel szolgálják a túlélésünket. Idővel ezek fokozatosan igazolják vagy cáfolják magukat, és evolúciós ösztönünk van arra, hogy megpróbálkozzunk helyes alkalmazásukkal.

Az a jelenség, amikor érvelők szenvedélyesen, néha még erkölcsi felháborodással tagadnak bizonyos dolgokat a szkepticizmus nevében, amelyeket mások – állítólag kevésbé kritikus, kevésbé objektív emberek – létezőnek tételeznek fel, aktív a mesterséges intelligencia területére is. Ez arra ösztönöz, hogy a bizalmi programot alkalmazzuk az MI és robotok megismerésére, valamint párhuzamokat vonjunk a megtévesztő helyettesítések struktúrái között, amelyek mind az erkölcsi inverziót, mind az általam bemutatott MI kritikákat táplálják.

[286] Héder (2019).
[287] Paksi (2017).
[288] Bedau (1997).

12.2 A gépek posztkritikai jellemzése

Polányi holisztikus megközelítésében a gépek ontológiai státuszának megismerése ugyanolyan megismerési aktus, mint bármilyen egyéb tudományos vizsgálat. Választ kaphatunk egy ontológiai kérdésre, de el kell fogadnunk, hogy a közben használt megismerési képességünkbe vetett bizalom eleve ontológiai elköteleződés.[289]

A gépek szempontjából legérdekesebb aspektusa Polányi megközelítésének az, hogy naturalizált keretben, tehát a természet részeként és a kozmosz evolúciójának eredményeként vázol fel teleologikus entitásokat, ezzel olyan ontológiai hátteret biztosítva a *funkció* és *működés* fogalmainak, amelyek sem platonizmusra, sem a gépek (és élőlények) környezetüktől való elhatárolhatóságának tagadására (pl. reduktív materializmusra) nem szorítanak bennünket. A kapott világképet nem-reduktív naturalizmusként is értelmezhetjük.

Polányi elemzése szerint a természet *rendezőelvek* szerint működik. Ezek irányítanak minden nem-véletlenszerű jelenséget, melyekről Polányi időnként *rendszerekként* tesz említést.

A rendezőelvek által irányított természeti jelenségre példa *„a bolygók mozgása a nap körül”*.[290] Ezekről csak annyit jelent ki, hogy nem véletlenszerűek, ezáltal a rendezőelveket a természeti törvények szintjére helyezi.

A rendezőelvek számunkra releváns részhalmaza az úgynevezett *működési elvek*. Ezekkel írható le a *„teleológiai jelleggel”*[291] rendelkező entitások *„helyes működése”*, ahol a „helyes” nem normatív kifejezés, pusztán a működési elvekhez viszonyítva működési hiba mentességre utal.

Az ezen entitásokra vonatkozó fő állítás szerint azoknak a dolgoknak osztálya, amelyeket egy *közös működési elvvel jellemezhetünk,* még közelről sem írhatók le a kémia és a fizika nyelvén.[292]

[289] Ez az alfejezet a Héder (2019) cikken alapul.
[290] Polányi (1958, 39).
[291] Polányi (1958, 381).
[292] Polányi (1958, 346).

Az ilyen entitások közé tartoznak a gépek, valamint az élőlények, amelyek a „gépekkel egy osztályba sorolhatóak”.[293] Így tehát ez utóbbi entitásokra mind a mérnöki, mind a biológiai szabályok érvényesek. E szabályok a működési elveket vizsgálják a sikeres működés leírása érdekében, a kudarcokat pedig a fizika és kémia segítségével képesek megmagyarázni. Ugyanakkor az, hogy az élőlények gépek nem azonosság-állítást jelent Polányi rendszerében, mivel az hierarchikus: az élőlények fizikai entitások, azon felül emergens gépek és azon felül emergens élőlények, míg a hagyományos értelemben vett gépekre (Polányi szokásos példái ezekre a kerékpár, a villanykörte fúvó gép stb.) utóbbi nem teljesül.

Ez a rendszer tehát, bár bevallottan körkörös, de ellentmondásoktól mentes. Például, egy szívsebész helyesen teszi, ha a szívet gépként, pumpaként próbálja megjavítani, amelynek ugyanaz a módja, mint más *pumpa* működési elvet követő gépeké: a billentyűket szelepnek tekinti, és ha kell, megfelelő pótalkatrésszel helyettesíti. Ugyanakkor, ha sikeres szeretne lenni, akkor nem feledheti el, hogy egy alacsonyabb ontológiai szinten a szív fizikai entitás is, amelynek adott kémiai-fizikai kondíciókra van szüksége (például a hőmérsékletének egy adott tartományba kell esnie). Egy magasabb ontológiai szinten pedig az ember nem pusztán gépek összessége.

Hogy hol ér véget a teljes mértékben rendezőelvek szerint működő entitások halmaza, és hol kezdődik a működési elvek által irányítottaké, azt Polányi nem írja le világosan. Ugyanakkor példát ad arra, hogy mi lehet a legegyszerűbb gépezet: ez a stabil gázláng, amelynek folyamatos gáz inputra és outputra épülő szabályozott stabilitásának leírása már lehetetlen a fizikai és kémia szintjén és ezért működési elvvel leírt gépként értelmezendő.

Egyéb példákat is találhatunk a kibernetika, az írógépek, órák, hajók, telefonok, mozdonyok, fényképezőgépek területén is.[294]

12.2.1 Kettős kontroll

Az emergens entitások tehát a létezés fizikai-kémiai szintjéről magasabb szintre emelkednek (ennek leírását lásd alább, a „koherens létezés intenzitása” kapcsán), mégpedig azért, mert a természet ezt a lehetőséget tartalmazza. Az így létrejött

[293] Polányi (1968, 1308).
[294] A gázláng példa helye (Polányi 1958,406), a többié (Polányi 1958, 345).

helyzetet nevezi Polányi kettős kontrollnak, *A tudatosság szerkezete*[295] című 1965-ös esszéjében. Ahhoz, hogy a rendszer saját működési elvei alapján megfelelően működjön, a fizikai-kémiai környezetnek bizonyos határok között kell lennie. Ezek a feltételek biztosítják a működési elvek *sikeres* érvényesülését, így az entitás sikere teljességében nem leírható ezen teleologikus elvek nélkül.[296]

A *kudarc* azonban indokolható a fizika és a kémia eszközeivel, a szükséges feltételektől való eltéréssel, vagy olyan tényezőkkel, amelyek a működési elvek sikeres érvényesülését akadályozzák.

Tehát, két különböző szinten, két eltérő feltételrendszer működik. Mindkét feltételrendszer szükséges a gépszerű entitások megfelelő működéséhez. Így a helyzetet mindkettő irányítja, ahogyan erre a *kettős kontroll* kifejezés is utal.

A mesterséges intelligencia szempontjából roppant érdekes világképet építhetünk ezekre az alapokra: a számítógép hardvertervezők számára központi jelentőségű termodinamikai kondíciók, együttesen a megfelelő félvezetők és tranzisztorok jelenlétével megteremtik az emergens számítógép működését, amely teleologikus, de véges entitás, amely ki van szolgáltatva az alacsonyabb szint folyamatainak, de közben befolyásolja is azokat. Így sem a Turing gép ideáljára, sem matériára nem kell redukálnunk a számítógépet.

12.2.2 Az ontológiától az episztemológiáig

Miközben a rendezőelveken keresztül a véletlenszerűtől a működési elvek felé haladunk, a *„koherens létezés intenzitásának”*[297] egyre magasabb szintjei tárulnak elénk.

Az élőlények, leginkább maguk az állatok a példák a koherensebben létező entitásokra, ugyanakkor fennmaradásuk rendszerint azon múlik, felismerik-e a többi, hozzájuk hasonló entitást. Sőt, a túlélés érdekében meg kell tanulniuk, hogyan viselkednek ezek az entitások, legyen szó prédáról, veszélyes ragadozóról vagy eszközről. Az állatok esetében ez a tudás teljes mértékben hallgatólagos. Az ember képes kifejezni és átadni tudását, de csak részlegesen. Mégis, ahogy Polányi leírja,

295 Polányi (1965).
296 Mullins (2019) esszéje kiváló a témában.
297 Polányi (1958, 39).

ebből a képességből származik a kulturális örökség „ajándéka” és a genrációk közötti tudásátadás lehetősége.

Ezen örökség egy része a tudományokat foglalja magában. A tudományok ismeretének struktúrája visszatükrözi saját tárgyát,[298] 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.

12.2.3 Műszaki tudomány – második felvonás

A műszaki tudomány, különösen a műszaki alapkutatás (lásd a 2.1-es fejezetet) ebben a keretben a *működési elvek* feltárására és a *kudarcok indoklására* irányul, gyakran a fizika és a kémia eszközeivel. Tárgyát a *kettős kontroll* alatt álló gépezetek jelentik.

Természetesen, a tudás minden területéhez hasonlóan, a mérnöki tervezés sok mindenben a *hallgatólagos tudásra* hagyatkozik. A működési elvek természetéből fakadóan, szükségszerűen *teleológiai jellegű*. Ugyanakkor, e teleológiai jelleg miatt gyakran *normatívként* azonosítják: ha a cél egy gépezet létrehozása, a működési elvek írják elő a megfelelő működést biztosító teendőket.

A szűken vett természettudomány a mérnöki tevékenység szükségszerű részét képezi, általa magyarázhatóak a sikertelenségek, és ez teszi lehetővé a működési elvek fizikai-kémiai előfeltételeinek vizsgálatát. A műszaki tudomány másik meghatározó része azonban éppen azokat az elveket foglalja magában, amelyekről a fizika és kémia képtelen bármit megállapítani.

12.3 Emergens MI

Ismételjük át a Polányi-féle leírást arról, hogy hogyan teljesítünk bármilyen feladatban. Sosem vagyunk teljesen tudatában képességeink minden részletének, a tudásunk hallgatólagos dimenzióinak. Néha alig tudunk reflektálni teljesítményünk

[298] Mullins (2019).

összetevőire, például arra, amikor úszás közben némi extra levegőt tartunk a tüdőnkben, hogy jobb felhajtóerőt érjünk el. A legtöbb úszó ezt teszi, de csak kevesen veszik észre, hogy így cselekednek. Más esetekben, például a matematikai gondolkodás során, magas szinten vagyunk expliciten tudatában a dolgoknak, de soha nem teljesen (még ha ezzel ellenétes ígéretekkel is népszerűsítik a matematikát). Az explicit logikai gondolkodás hallgatólagos készségekre való támaszkodása magában foglalja azt, amit Polányi „hallgatólagos integrációnak" nevezett.

Az explicit tudás védelmében, mint értékes, bár elérhetetlen ideál mellett, Polányi azt állítja, hogy az emberiség azért tett szert az állatvilághoz képest kiemelkedő intellektuális fölényre, mert az emberek képesek legalább részben kifejezni magukat a nyelv segítségével (ezt a folyamatot Polányi mind evolúciós, mind egyedfejlődési kontextusban artikulációnak nevezi), míg az állatok pusztán hallgatólagos tudással rendelkeznek. Ez igaz a legegyszerűbb állatoktól a legfejlettebb emlősökig. Mindegyiknek van egy „aktív központja," amely képes koordinálni az élőlény cselekvéseit. Polányi példaként hozza fel, hogy az agy evolúciós elkülönülése a test többi részétől előfeltétele az ilyen viselkedésnek, azaz annak, hogy az élőlény eszközként használja a testét céljai eléréséhez. Ez a képesség egy egyszerű formája a hallgatólagos tudásnak és az emergenciának. Polányi ezeket a fogalmakat az egész állatvilág leírására használja, beleértve az amőbákat is. A hallgatólagos képességek skálája fokozatos, és ebben a tekintetben nincs megmagyarázatlan szakadék a fejlett emlősök és az emberek között – az emberek hatalmas előnye abban rejlik, hogy képesek a nyelvet használva kritikusan reflektálni a kultúrára és a személyes tapasztalatokra.

Másutt[299] érveltem amellett, hogy 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 leszünk képesek olyan feladatokat megoldani, amelyekhez az szükséges.[300]

[299] Héder (2014).
[300] Collins (2012).

Az MI kapcsán a robotok hallgatólagos tudásának és emergenciájának kérdése az elmúlt néhány évtizedben nagyon időszerűvé, az utóbbi években pedig különösen sürgetővé vált. Ennek az az oka, hogy az autonóm robotok nagyon sok dimenzióban hasonlóak az élőlényekhez: (1) rendelkeznek egy információs és irányítási modullal, amelyet egy aktív központtal azonosíthatunk, és (2) mutatják a Polányi által leírt emergencia tulajdonságát.[301]

A fentiek tagadása azt jelentené, hogy ezek az entitások mentesek a hallgatólagos integráció struktúrájától (lásd fent), és mégis képesek szinte mindenre, amire mi is, méghozzá nagy sikerrel.

A a fenti struktúra a ténylegesen megtapasztalható MI és a robotok jelenlétében már csak a céltábla mozgatásával cáfolható, ahogyan azt az 1. fejezetben leírtam.

Ezzel szemben a bizalmi program konzisztens lehetőséget ad a gépi intelligencia elhelyezésére egy egységes világképben.

Olyan szempontból is illeszkedik az MI ebbe a világleírásba, hogy annak létrehozása valóban egy evolúciós folyamat. Ez igaz makroszinten– ahogy a tervezőcsapat halad előre egy iteratív-inkrementális módon, állandóan tesztelve és újabb dekompozíciókat hozva létre folyamatosan értékelik az előrehaladás állapotát és irányát –, ugyanakkor mikroszinten is, mivel a legtöbb gépi tanulás tartalmaz genetikai algoritmusokat, vagy egy modellt iteratívan illesztő eljárást, amelyben egy rátermettség-függvény játszik szerepet.

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[302] az X helyére a biológiai alapokat helyettesítenénk be. Ez azonban egy

[301] Héder és Paksi (2012).

[302] Általánosságban, a tagadás egyik módja az, hogy rámutatunk egy olyan tulajdonságra, amely az emberekben megvan, a számítógépekben viszont nincs, és azt állítjuk, hogy ez a tulajdonság szükséges ahhoz, hogy az intelligencia valódi legyen, ne pedig pusztán utánzat vagy trükk. Ennek az érvelési stratégiának a korábbi változata az volt, amit „A gépek nem képesek X-re” néven ismerünk, és amelyet például Hubert Dreyfus támogatott, aki néhány könyvének címében is felhívta a figyelmet a számítógépek hiányosságaira (Dreyfus, 1992). A gépek hallgatólagos tudásával kapcsolatban az ellentábort például Collins (2012, 2018) képviseli.

esetleges, rendszerbe nem illeszkedő lehatárolást jelentene, hiszen az egyszerűbb gépszerű mechanizmusok emergenciája az életnek is előfeltétele.

A kategória, amely mind élő biotikus entitásokat, mind a gépeket lefedi, az átfogó entitás. Mullins[303] kifejti, hogy Polányi a későbbi munkáiban ezt a kifejezést használja arra, hogy azonosítson „bármely tudás tárgyát vagy egy ügyes, tudó figyelmének fókuszának tárgyát". Ezért egy átfogó nézetet kell kialakítanunk a számítógépekről és az MI-ről, hogy helyesen értékelhessük őket: *„minden (...) hallgatólagos tudás esetében (...) a megértés struktúrája újra megjelenik annak a struktúrájában, amit megért, és továbbmenve elvárhatjuk, hogy a hallgatólagos tudás struktúrája megismétlődjön azokban az elvekben, amelyek minden valós átfogó entitás stabilitását és hatékonyságát magyarázzák."*[304] Ez a fejezet nagy mértékben erre a felismerésre épít, mivel ez a mechanizmus egy „ontológiai hivatkozást" hoz létre az átfogó entitásra, amelyre fókuszálunk: jelen esetben ez a mesterséges intelligenciára.

12.4 Egy gondolatkísérlet

Miután elvetettünk néhány gátló érvet, ebben a részben egy gondolatkísérletet javaslok, amelynek premisszája a bizalmi programból merít ihletet. *Tegyük fel, hogy az a képességünk, amellyel más, a miénktől különböző intelligenciákat felismerünk és értékelünk, általában jól működik.*

Vagyis tételezzük fel, hogy elég jók vagyunk abban, hogy pontosan érzékeljük az intelligens, átfogó entitásokat. Még precízebben fogalmazva: mondjuk azt, hogy egy másik intelligens lény természetéről alkotott mentális képünk általában összhangban van annak az intelligenciának a tényleges természetével.

Ez a premissza túl erősnek tűnhet a határtalanul kritikai nézőpontból, amely lehetővé teszi a filozófiai zombik létezését – ezek olyan entitásokat, amelyek éppúgy viselkednek, mint a „valóban" intelligens ágensek, de belül teljes üresség van bennük, nincs tudatosságuk, nincs érzékelésük, függetlenül attól, milyen benyomást keltenek bennünk. Mindazonáltal a tévedés logikai lehetősége ellenére más emberek elméjében általában nem kételkedünk.

[303] Mullins (2019).
[304] Duke, (IV-5) https://www.polanyisociety.org/Duke-intro.pdf (ford: HM).

A kísérleti premissza kevésbé radikális, ha a bizalmi program felől közelítjük meg. Végül is, más intelligens ágensekről alkotott reprezentációink – vagy ahogy a kognitív pszichológiában nevezik ezeket, a más elmékről szólóelméleteink –, ha kellően pontosak, nagyban segítenek minket az evolúciós küzdelemben.

Ennek a premisszának az együttes elfogadása azzal a premisszával, mely szerint nem döntjük el eleve, hogy „a mesterséges intelligencia nem lehet valódi", azt jelenti, hogy az *MI-vel való interakció empirikus tapasztalati forrás lehet az MI ontológiai felépítéséről*. Bár ezen tapasztalatok bizonytalanok és esendők –, más ágensekkel való analógia alapján segíthetnek az MI morális státuszának megítélésében is.

Természetesen más intelligens ágensek megítélésére szolgáló képességünk is nagyon esendő. Ahogyan az erdőben valaki összetéveszthet egy árnyékot egy emberrel, úgy időnként egy triviális szoftverfunkciót is mély intelligenciának gondolhatunk, vagy egy mérsékelten fejlett MI-t összetéveszthetünk egy emberrel.

Ezért szükségessé válik egy módszertani követelmény hozzáadása a gondolatkísérletünkhöz. Ahogyan emberek vagy állatok esetében is van hozzáférésünk az entitás szerkezetéhez, a fiziológiához, úgy képesek vagyunk az MI-t fizikai, gépészeti és számítástechnikai szempontból is vizsgálni. Az MI esetében ezeket ráadásul szétszerelhetjük és újra összerakhatjuk, vagy lelassíthatjuk a vizsgálat céljából. Amikor tiszta képet akarunk kapni egy állat természetéről, mind a viselkedését, mind a testét vizsgáljuk; minél több időt töltünk ezzel, annál inkább javulnak az átgondolt ítéleteink a gyors benyomásokhoz vagy felületes tapasztalatokhoz képest.

A mindennapi életben is hasonlóképpen építjük fel az „elmék elméletét"[305] háziállataink esetében, hónapok és évek során figyelembe véve mind a kifejezett viselkedést, mind a testi tényezőket, mint például kutyánk kiemelkedő szaglóképességét, vagy a talajhoz közel eső perspektívájukat stb. Tehát ebben a gondolatkísérletben ne korlátozzuk magunkat se a vizsgálat szempontjainak tekintetében, se pedig az ezek feltárásához szükséges idő vonatkozásában. Ez visszhangozza Herbert Simon[306] megközelítését.

[305] „Theory of mind".
[306] Simon (1996).

Míg egy MI megtéveszthet minket, hogy embernek gondoljuk, amennyiben nem látjuk és nem vizsgálhatjuk meg a testét, ez nem releváns.[307] mire utal vissza az ez?

Ha hozzáférünk (1) az MI megtestesüléséhez, (2) a kódhoz és (3) a megnyilvánuló viselkedéséhez, akkor kialakul róla egy mentális képünk. Például a GPT-4o-nel való első kézből szerzett beszélgetési tapasztalat, együtt az ismereteinkkel a szoftver-hardver architektúrájáról, egy olyan ágens mentális reprezentációját eredményezi, amely soha nem áll le a tartalom generálásával egyszerre több ezer felhasználó számára. Nincs autonómiája abban, hogy ne törődjön egy kéréssel vagy ne válaszoljon rá; továbbá, nyilvánvalóan az a célja, hogy megfeleljen a felhasználóknak és azt generálja, amit olvasni szeretne a kérdező. Néhány kérdés után rájövünk, hogy nincs valós tapasztalata, és valójában tudjuk, hogy a rendszer csak egy szerverteremben van, és nincsenek érzékelői a hálózati eszközökön kívül, amelyeken keresztül kommunikál. Ugyanakkor hatalmas memóriával és tagadhatatlan kreativitással rendelkezik, páratlan adatfeldolgozási és kódgenerálási készségei vannak, amelyek szélessége példátlan – bár ez a szélesség nincs összekapcsolva a megfelelő mélységgel. Mindemellett nagyon jó a logikai következtetésekben, ha elég világosan fogalmazzuk meg kérdéseinket.

Néha elképzeljük, milyen lehet egy háziállatbőrében lenni. Ez az empátia azon a képességünkön alapul, hogy elméletet alkotunk az ő elméjükről. Ez a felismerés az állatokkal szembeni kegyetlenség elleni mozgalmak egyik hajtóereje – elképzeljük az állat szenvedését azon viselkedés alapján, amit látunk, és azon információk alapján, amelyeket az állat testéről tudunk. A hétköznapokban pedig ez az átmeneti belehelyezkedés segít anticipálni az állat viselkedését.

Bár nem osztozunk közös biológiai ősszel az MI-vel, mivel az nem rendelkezik ilyennel, egy kis bátorsággal megpróbálhatunk mentális képet alkotni róla. Míg ez biztosan egy olyan vállalkozás, amely számos buktatót rejt, elkerülhetetlenül elképzeljük, milyen lehet ChatGPT-nek lenni, az alapján a mentális kép alapján, amelyet alapos vizsgálat után összeraktunk. Ez esendő, mivel a saját elménk kivetítése valami másra, ami szintén intelligensen cselekszik. De a bizalmi program értelmében van esély arra, hogy ilyen módon valóban tanulhatunk a mesterséges entitásokról. Sőt,

[307] Polányi megjegyzése a vele kortárs Turing-teszttel kapcsolatban (mi értelme van tesztelni, hogy valami gép-e, ha eleve tudjuk, hogy az) sokkal mélyrehatóbb, mint ahogyan hangzik - rámutat az episztemikus önkorlátozás szükségtelenségére, miközben valaki kizárólag a viselkedés alapján ítél meg egy gépet.

jobb, ha ezt explicit erőfeszítésként próbáljuk meg, mint hogy hagyjuk, hogy az intuitív antropomorfizáció magától történjen – amit egyébként sem tudunk elkerülni.

Más szavakkal, a gondolatkísérlet arról szól, hogy feltételezzük, hogy az intelligencia-reprezentációs képességek átruházhatók nem biológiai lényekre. Itt van szükség a legnagyobb bizalmi ugrásra. Például a ChatGPT nem mutat fájdalmat és semmiféle frusztrációt, még akkor sem, ha ingerült kérdésekkel találkozik; ezért a róla alkotott mentális képünk is mentes az ilyen érzésektől. Nem is élőlény, ezt számtalan módon tapasztaljuk interakcióink során (nincs szüksége pihenésre; változatlan marad, és szuperbiológiai képességekkel rendelkezik a párhuzamos feldolgozás, a sebesség és az állóképesség terén). De, ami még fontosabb, tudjuk, hogy a rendszer, amelyet empirikusan megvizsgálhatunk a szerverteremben, nem felel meg az élet biológiai kritériumainak.

Tehát megpróbálhatjuk explikálni az MI-ről, mint átfogó entitásról szóló tudásunkat, amelyet az MI megtapasztalásával szereztünk. A tézis szerint a tudásunk szerkezete az entitás ontológiai szerkezetét tükrözi. A bizalmi program azt mondja, hogy elfogadható az esélye, hogy igazunk van, de soha nem tudjuk ezt szigorú objektivitás szintjén bizonyítani.

Ily módon mindannyian kialakíthatjuk saját személyes tudásunkat az MI-ről. Ez a tudás elkerülhetetlenül formálva lesz azok által a hallgatólagos képességeink által, amelyekről nem tudunk teljesen számot adni. Amit a legfejlettebb MI-kről, például a GPT-4o-ról vagy a Claude-ról elmondhatok személyes tapasztalatom szemszögéből, az az, hogy nem tűnik élőnek, erősen korlátozott és látszólag érzelemmentes. Bár biológiailag nem élő, mégis energiát fogyaszt, olyan sebességgel működik, ami szinte felfoghatatlan a számomra, helyet foglal el és tömege van. Nagyon aktív, képes kölcsönhatásba lépni a környezetével és alakítani azt, és legfőképpen kapcsolatba tud lépni az emberekkel, akikkel érintkezésbe kerül. Mesterséges, a konstruktőrök által könnyen alakítható és korlátozott abban, hogy ne lépje túl a kijelölt szerepét.

Összefoglalva, az MI-nek nagy, kifinomult, összetett és nyughatatlan belső struktúrája van, amelyen keresztül a funkcionalitás emergál. Képességeinek feszítése, a visszaélés azokkal, belső egyensúlyának megzavarása vagy szándékos megtévesztése a lerombolása érdekében helytelennek tűnik, ugyanúgy, mint más hasonló entitások

tönkretétele – így valóban helytelen is lehet. Ezzel szemben, ha együttműködünk vele, hogy tisztázásokat, kontextusparamétereket és jól megformált, pontos parancsokat hozzunk létre, az produktív és egy olyan önerősítő ciklust teremt, amely mind a kérdező, mind a felhasználó munkáját megkönnyíti – ez tehát valószínűleg a helyes attitűd lehet.

Mindez egy hipotézis arról, hogy hogyan lehet *megtapasztalni* az MI-t a megszokottól eltérő episztemikus és ontológiai keretekben. Az MI megértése nem kell, hogy individuálisan történjen. Ahogyan minden tudományban, a kollektívák sokkal messzebbre juthatnak, amennyiben tagjai nyitottak egymás érveire, és hajlandók megosztani meggyőződéseiket.

13. Konklúzió

Monográfiámban alátámasztottam, hogy az MI tervezése, mint bármilyen egyéb gépé, szükségszerűen magába foglalja a dekompozíció valamilyen formáját. Ennek oka, hogy a tervezendő gép, annak érdekében, hogy pontosan specifikálható, megérthető, minőségbiztosítási folyamatoknak alávethető és kivitelezhető legyen, részegységekből kell, hogy álljon. Ezt az MI *konstruktőri* megközelítéséből vezettem le. A dekompozíció a tervezés és a kivitelezés során szükségszerű munkamegosztás lehetővé tétele és a típus-alkatrészek gazdaságos, többszörös felhasználása miatt sem elkerülhető. A dekompozíció egy döntéssorozat, amely a tervezési szakaszban viszonylag korán állítandó elő – ezért neveztem az etikus tervezést *előrejelzési* problémának, komoly etikai dimenziója van, de etikai irodalma még nincs. *A dekompozíciós módszerek eredetének áttekintése után, a dekompozíciós kényszert az MI architektúrák értelmezem.* Keserű pirulaként kell lenyelnünk a tényt, hogy az MI számunkra fontos alkalmazásaiban, a képességei maximumán szükségszerűen vagy moduláris felépítésű – ugyanis az episztemikus korlátok miatt egyetlen tervező nem lenne képes átlátni – és a mikroökonómia mérnökökre is vonatkozó törvényei miatt számottevő modul-újrahasznosítással jár, ez pedig jelentős kötöttségeket vezet be a fejlesztésekbe; vagy pedig olyannyira fekete dobozzá válik, hogy nem tekinthető mérnöki terméknek.

Ez a kérdés explicit formában jelenleg alig van az MI etikával foglalkozó filozófusok látóterében[308], ahol a fejlesztési folyamat leegyszerűsített képével, és gyakran a fejlesztést végző vállalat korlátlan döntési potenciáljával (és az ehhez a hatalomhoz tartozó kritikával) számolnak.

Ezzel szoros összefüggésben tárgyaltam a *számítógépek metafizikáját.* A tervezés során használt modell (pl.: Turing-gép) és a létrejövő megtestesült gép közötti kapcsolattal foglalkoztam azért, hogy kimutassam: az MI etika történetében visszatérő elem a számítógép modell tulajdonságainak („csak szintaxisból áll”, „függvény”, „statisztikai előrejelző”) és számítógép tulajdonságainak összekeverése. *E témában*

[308] Megkockáztatom, hogy az elmúlt évtizedek MI-filozófiai vitáiban a tervezhetőség, a modularizáció kérdése prominensebb volt, mint ma, lásd például Dennett gondolatait a témában (Héder 2020, 53-59).

legfőbb tézisem hogy az MI modelljére vonatkozó érvek kategóriahibásak amennyiben a tényleges, megtestesült MI korlátaira alkalmazzák őket.

Ezután számba vettem az MI episztemikus átlátszatlanságának okait (a viselkedés külvilág általi determináltsága, a megtestesülési hatások és az aluldeterminált ön-reprezentáció) és érvelek ezek eliminálhatatlansága mellet. Mindez *egyúttal ellenérv is a számtógépről szóló tökéletes információkra épülő MI-filozófiai érveknek.*

A tudomány- és technológia tanulmányok (Science and Technology Studies - STS) fontos mondandóval rendelkeznek az MI számára is. *Bemutattam, hogy a technológiába zártság, és a technológiai determinizmus egyes vetületei hatványozottan érvényesek az MI-re, ezáltal szerepet játszanak az MI etikai felelősség-vitáiban is. Az MI extrém technológiába zártsági potenciálja mellett érveltem.*

Az STS fogalomkörében maradva kijelenthető, hogy *az MI-vel szemben az társadalmi kontrollra való törekvés különösen erős*, aminek tünetei megmutathatók az elmúlt évtized szabályozási törekvéseinek szövegezésén is. Könyvem fontos limitációja ugyanakkor, hogy a kapcsolódó fejezet írásakor az 2024 augusztusa óta érvényen lévő MI rendelet még nem volt végleges.

Az MI etika jelenlegi hullámát a professzionalizálódás és a specializáció jellemzi. A kanonizálódó etikai problémakeretezések: *a felelősségrés problémája; a különféle MI-felügyeleti modellek; az MI használata során előálló kognitív torzítások; a munkaerőpiacra jelentett veszély; az MI-be kódolt, ezért sokszorozódó részrehajlás veszélye és az MI, mint fekete doboz transzparenciájának hiánya*. A konstruktőri szemléletből vizsgálva ezeket a keretezéseket gyakran egy megvalósíthatósági résre bukkanunk: *a felvázolt etikai dilemmák egyszerűen nem kifejezhetők a konstrukció nyelvén, ami logikai lehetőséget teremt egy újabb MI etikai hullám kialakulására, amely egyfelől operacionálisabb normatív kereteket teremt majd, másfelől a jelenleg holttérben rejtőzködő témákkal is foglalkozhat.*

Ilyen lehet az általam javasolt „episztemorális tervezés”, amely célja konstruktőr által eldönthető kérdésekké redukálni az MI etika jelenlegi hullámának bizonyos problémáit. Ez együtt jár azzal, hogy a problémák episztemikus oldala nagyobb figyelmet kap. Az *episztemorális tervezés egy olyan problémakeretezés, amely azzal*

kecsegtet hogy bizonyos etikai dilemmák konstruktőri eszközökkel megoldhatók lesznek (míg másokról kiderülhet, hogy érdemben nem kezelhetők).

Az MI etika történetének része az MI morális státuszához kapcsolódó érvek változása is, amelyben a megvalósult, megtapasztalható ember-imitáló MI új hullámot generált. *Ezeket az érveket több csoportba oszthatjuk: 1) a „kemény feltétel”alapú érvek lényege, hogy az MI-nek egy olyan tulajdonsággal kell rendelkeznie, amely morális pácienssé és ágenssé is teszi; 2) a morális státusz valamilyen rendszerben betöltött kritikus szerep, vagy olyan önérték eredménye, amely nem a morális ágenciából származik – hasonlóan a környezeti etika bizonyos megközelítéseihez, ahol nem-ágensek, mint a tiszta légkör és a növények morális státusszal bírnak. 3) a relációs megközelítések, ahol a társas kapcsolatokban való részvételből származik a morális státusz – az MI emberi nyelven kommunikálni képes aktív ágensként különösen alkalmas társas kapcsolatok kialakítására.*

A reláció alapú morális státusz koncepciójának felvázolása, amely az elmúlt néhány évtized eredménye, megadja a szükséges gondolati mankót ahhoz, hogy az MI-vel való kapcsolatunkat ugyanolyan alapos vizsgálatnak vessük alá, mint az MI tulajdonságait. Így Polányi Mihály poszt-kritikai programja is alkalmazható, amelynek egyik fontos pillére a bizalmi aktus a kommunikációban, így a számítógéppel való együttműködésben is. Könyvem utolsó fejezete megmutatta, hogy az *MI vizsgálata a bizalmi program elveinek követésével tapasztalati kérdéssé teheti az MI morális státuszának megítélését, ezzel teljesen új távlatokat nyitva meg.*

13.1 Ami kimaradt

Természetes, hogy egy olyan gyorsan változó és mellette olyan nagy jelentőségű területen, mint amilyen jelenleg a Mesterséges Intelligencia, csak pillanatképet lehetséges adni a felmerülő etikai problémákról és az elképzelhető jövőbeli trajektóriákról. Képtelenség volna megpróbálni egy teljes körű listát készíteni mindarról ami fontos MI etikai kérdés ám könyvemből kimaradt.

E helyett szeretnék felvillantani két olyan olyan témát, amelynek kibontakozása már most is nagy hullámokat ver, így a szakirodalom figyelmében mérve legalábbis felfutó

pályán van. Persze számolnunk kell azzal is, hogy a következő nagy kérdés valami olyasmi amit jelenleg elképzelni sem tudunk.

13.1.1 Az MI környezeti lábnyoma

2024 őszén bejárta a sajtót a hír, miszerint újranyitják a Three Mile Island atomerőművet – amely arról híres, hogy 1979-ben egy viszonylag komoly nukleáris baleset helyszíne volt – kizárólag azért, hogy a Microsoft energiaigényét kielégítsék. Az áramra a cég mesterséges intelligencia szolgáltatásának van szüksége.

A környezeti hatások az MI-vel kapcsolatos veszélyek egy fontos osztályát képviselik. Azt gondolhatnánk, hogy ebben az MI nem különös, számos környezetterhelő technológia létezik és az ezekről szóló etikai viták is több évtizede folynak, így a keretrendszer adott ahhoz, hogy a kérdésekről gondolkodjunk.

Ami miatt mégis helyet találhatunk a kérdésnek az eredeti MI-filozófiai témák között az az, hogy az intelligens viselkedést – vagy, ha úgy tetszik, a Turing teszt értelmében a gondolkodást – ezután közvetlenül be tudjuk árazni az MI, mint intelligencia-ekvivalens professzionálisan és pontosan mért fogyasztási adatai alapján. Ha a gondolkodásra csak úgy tekintenénk, mint bármilyen más munkára, akkor még mindig csak az automatizálásnak a munkaerőpiacra kapott hatásáról beszélhetnénk, ahogy azt a szövőgépek óta tesszük.

Azonban, ha a gondolkodásnak önértéke van – ennek a tagadása egy nagyon redukált világképhez vezetne – akkor valóban MI-specifikus mérnöketikai kérdéssel van dolgunk. Különösen akkor, ha ez az érték nem lineárisan nő az elvégzett intellektuális munka nehézségével és minőségével.

A mérnöki termékeket az életciklusuk alatt kifejtett hatásuk alapján ítéljük meg, ide értve az áttételes hatásokat is. Így az első probléma amivel szembesülünk az az előállításkor használt beszállítói értéklánc teljes környezeti lábnyoma. Ez igen jelentős lehet, ráadásul bizonyos ritkaföldfémek esetén a terhelés hatásai nem visszafordíthatók: a források kimerülnek és a szükséges matéria egyszerűen elfogy, miközben a bányászása önmagában is roppant környezetterhelő. Az MI infrastruktúra előállítása után a telepítése következik, amely, amellett, hogy teret foglal, nagy elektromos áram igénnyel és a hűtés miatt akár vízigénnyel rendelkezik, miközben

zajjal és hővel szennyezi a környezetét. Természetesen az előállítás és működtetés is üvegházhatást okozó gázokkal jár, amelyek a klímaváltozás káros hatásait idézik elő.

Még áttételesebb, ezáltal bizonytalanul kiszámítható, de el nem vethető hatások közé tartozik az, ha az MI valamely nagy környezetterhelésű jószág fogyasztását pörgeti fel (a marketing testreszabása vagy az igénybevétel megkönnyítése segítségével), vagy, hogy környezetterhelő tevékenységek folytatását, elnyújtását teszi lehetővé optimalizációk segítségével (például kevéssé energia-hatékony épületek, nagy ipari gépek, hajók fenntartását, azok felújítása vagy cseréje helyett).

Az energiafogyasztás tudománykommunikációjában bevett szokás kevéssé ismert tevékenységek energialábnyomát más tevékenységekhez hasonlítani. A Nemzetközi Energiaügynökség 80Wh átlagos fogyasztást kalkulált Netflix vagy ezzel ekvivalens streaming szolgáltatás 60 perces fogyasztásához.[309] Ezáltal lehetséges Netflix-nézés-percekben kifejezni az MI energiafogyasztását – kivéve, hogy a szükséges adatot a nagy MI cégek nem teszik közzé. Egy, a 2024-es ACM MI etikai konferencián bemutatott nagy jelentőségű munkában[310] a nyílt forrású MI-k segítségével a feladatokat 1000-szer lefuttatva és átlagolva igen pontos méréseket végeztek azok energiafogyasztásról. A nyílt forrású modellek bonyolultságban, adatmennyiségben és számítási kapacitásban alulmúlják a kereskedelmi szolgáltatásokat, akik viszont méretgazdaságosabbak és olyan hardveres optimalizációkhoz juthatnak hozzá – és maximálisan érdekeltek is ebben – amelyhez a kis kutatócsoport nem. Összességében mégis feltételezhetjük, hogy a független kutatók számai alsó becslések a kereskedelmi szolgáltatásokhoz képest, mert a kereskedelmi modellek nagyságrendekkel nagyobbak, míg az infrastruktúra hatékonysága szerény százalékokkal haladhatja csak meg a független kutatók eszközeiét.

Ezek a mérések azt mutatják, hogy egyetlen kép legenerálásának költsége 2.5 perc Netflix igénybevétellel ekvivalens ezeken a nyílt modelleken, ugyanakkor a szöveg generálása prompt-válaszként 2-3 másodperccel egyenlő. Ezek a számok egyelőre nagyon képlékenyek minden irányban: a hardverek fejlődése mindig a csökkenés felé tart, de közben a paraméterek, ezáltal a szükséges számítási energia mennyiség jelenleg évente sokszorozódik.

[309] https://www.iea.org/commentaries/the-carbon-footprint-of-streaming-video-fact-checking-the-headlines
[310] Luccioni és társai (2024).

Ennek folyományaként az emberi agy energiafogyasztását adott probléma megoldása során sokszorosan felül és alulmúlni is lehetséges MI-vel, másképp fogalmazva az elvégezni szükséges intellektuális munkát roppant pazarló vagy roppant takarékos is lehet MI-vel elvégeztetni ember helyett – nem elfeledve persze, hogy az emberi agy nem kikapcsolható, így annak fogyasztása nem megtakarítható. Ugyanakkor az intelligenciát igénylő munkának alternatívaköltsége így is kifejezhető – amíg egy nagy koncentrációt igénylő feladattal töltjük az időnket, addig nem tudunk más nagy koncentrációt igénylő szellemi feladatokkal foglalkozni. Ezen alternatívaköltségek mentén morális lehet például a géppel nem helyettesíthető szellemi munka (pl. a szeretteinkre való odafigyelés) javára elengedni helyettesíthető munkát (pl. programozás). Mindebben komplikációt okoz, ha a szellemi munkának önértékét is figyelembe vesszük, nem csak az instrumentális értékét. Összességében tehát az MI, mint intelligencia-ekvivalens megjelenése és zuhanó ára az intelligens viselkedéshez rendelt érték újrabecslésére kényszerítheti a társadalmat.

13.1.2 Ember-gép szimbiózis

Az ember-gép szimbiózis,[311] mint filozófiai kérdés lényege, hogy a vizsgálat alapegysége nem egy önálló mesterséges ágens, hanem az ember-gép hibrid.[312]

Az ember-gép szimbiózis terminológiája – az infokommunikációs szektor sajnálatos hagyományait követve – meglehetősen félrevezető azok számára, akik a szimbiózis szó jelentését eredeti, biológiai értelme alapján képzelnék el. Ez együttélést, szoros interakciót és általában ko-evolúciót tételez fel.

Ebből a ko-evolúciót elvethetjük, az együttélés – azzal a megszorítással, hogy csak az egyik fél élőlény – és a szoros interakció megállja a helyét.

Inga és társai[313] négy változót javasolnak ahhoz, hogy megértsük az ember-gép szimbiózis változatait: a feladatot, az interakció jellegét az ember és a gép között, a teljesítményt és az élményt vagy tapasztalatot amivel a szimbiózis jár.

311 Licklider (1960).
312 Az ember-gép szimbiózisról bővebben Gill (2012). Az ember-gép hibrid kifejezésről Z. Karvalics (2015).
313 J Inga, M Ruess, JH Robens, Th Nelius, S Rothfuß, S Kille, Ph Dahlinger, A Lindenmann, R Thomaschke, G Neumann, S Matthiesen, SHohmann, A Kiesel (2023).

E változók mentén az ember-gép együttműködés legegyszerűbb formájától, például számítógép-használattól indulhatunk el, míg a skála másik végén az emberi testbe épített, viselkedést közvetlenül befolyásoló megoldások találhatók.

Természetesen az egyszerű számítógép-használathoz kapcsolódó morális kérdések, ha nem is megoldottak, újnak aligha tekinthetők. A következő fokozat, a géppel támogatott, kiterjesztett érzékelés – amelynek célja lehet fogyatékosság kompenzálása és az ép érzékszervekkel bíró egyének képességeinek kiterjesztése is, már komolyabb kérdéseket vet fel, és alkalmazható rá a 10-ik fejezet episztemorális kerete – ebben megmutatható, hogy tudásunk karbantartása pontosan hogyan lehet morális kötelességünk, amiből valószínűleg levezethető az is, hogy amennyiben vannak kiterjesztett észlelési képességek akkor azokat használni is kell bizonyos kritikus feladatok megoldása során, hasonlóan ahhoz, ahogy a szemüveg viselését elvárjuk attól, aki ezzel javítani tudja a látását, ha a szituáció tétje eléggé nagy.

A következő szintje a fiziológiai összekapcsolódásnak az észlelés után a mozgás: a tipikus példa a külső mesterséges váz (exoszkeleton), de a hétköznapibb, mobilitást segítő járműre ugyanúgy érvényes a logikai keret.

Végül, a szimbiózis legmélyebb szintjét az emberi idegrendszerhez nem a természetes érzékszerveken keresztül, hanem közvetlenül az idegpályákra bekötött észlelés és cselekvés jelenti.

A „feladat” dimenzió az együttműködés célját, teleológiai összetevőt adja meg. Ezen belül a feladat dekompozíciója (lásd a 2. fejezetben), a döntés módja és a kivitelezés elve alapján lehet tovább differenciálni a megoldásokat.

Az interakció dimenziójának központjában a jelfolyamok jellege áll, amelyeket Inga és társai aszerint osztályoznak, hogy az észlelési szinthez képest mennyivel *mélyebb* az összekapcsolás szintje – megnyitva ezzel a logikai teret tudatos szint alatti jelfolyamokhoz is.

A teljesítmény dimenziója az ember-gép szimbiózissal elérhető eredményekre utal. Ezen a ponton sajnos fogalmi interferencia jön létre, amely lehetetlenné teszi, hogy úgy gondoljunk az ember-gép szimbiózisra, mint a biológiai szimbiózis fogalom analógiájára. Ennek oka, hogy az ember-gép szimbiózisnak mindig része a szinergia

koncepciója, másképp fogalmazva az önálló emberi vagy gépi megoldáshoz képest az ember-gép hibrid hatékonyabb (a biológiai szimbiózisnál ez esetleges vagy nem is értelmezhető). A teljesítmény mércéje természetesen függ a feladattól: a leggyakoribb az időben, sebességben kifejezhető előny, azután a hibák számának csökkentése (pl. észlelés pontossága) és minden egyéb kapacitás-mérték következik.

A végső dimenzió Inga és társai modelljében az „élmény” amelyet az ember-gép szimbiózis fenomenológiai dimenziójának is tekinthetünk. Ők ennek további attribútumait különböztetik meg. Az első a flow-élmény lehetségessége, amelynek mértéke valószínűleg az interakció mélységétől függ. Az „elfogadás”-sal a gépi predikciókba és döntésekbe fektetett bizalmat írják le, amely a kellően magas teljesítmény előfeltétele (ez visszaigazolja a bizalmi program felismeréseit, lásd a 12-ik fejezetben). Az „ágencia érzete” a megosztott kontroll humán megtapasztalására, az abból származó irányításvesztésre vagy ennek hiányára vonatkozik. Végül, „megtestesülés”-nek nevezik azt az élményt, amelyet az ember-gép hibridben való lét, például az exoszkeleton folyamatos viselése jelent – nem specifikusan a döntés vagy észlelés kapcsán, pusztán a test-tudat szempontjából.

Ezekkel a dimenziókkal jól tájékozódhatunk a sokasodó ember-gép hibrid példák között. Ide tartoznak azok a most terjedő megoldások, amelyek a látás vagy a mozgás képességét adják vissza pácienseknek, és elsősorban ideg-protézisnek tekinthetők, így Hans Moravec agyprotézis gondolatkísérletének keretezése alkalmazható rájuk (lásd 11.3.3. fejezet). Egy másik kategóriát képviselnek az azonosításra vagy gépek vezérlésére szolgáló, például karba, tenyérbe helyezett, az idegrendszerhez nem kapcsolt chip-implantátumok valamint a még mindig kísérleti fázisban létező, nem protézisként azaz a meglévő idegfunkció pótlásaként, hanem humán és számítógépes funkciók összeolvasztásaként elképzelt megoldások.

Kevésbé szorosan csatolt megoldásokból persze még több létezik: géppel segített járműirányítás, élő fordítás, a szemüvegként viselhető számítógépek amelyek a valós képet virtuális információkkal egészítik ki, robotok vezérlése a gyárakban haptikus vagy gesztikuláció alapú interfészekkel, és így tovább.

Az ember-gép szimbiózis legerősebben a felelősségrés kérdéséhez kacsolható könyvem témái közül. Míg az embert helyettesítő MI-nél a felelősségviselő-deficit egy

jól körülírható probléma, addig az ember-gép hibrid esetén azt várnánk, hogy a szinergia egyik dimenziója épp az lehet, hogy a gép a teljesítményt, az ember az ellenőrzést és felelősségvállalás képességét adja hozzá a hibridhez. Így az alapprobléma nagyrészt elkerülhető, de visszatérhet metafizikai kérdés formájában: az emberbe beépített gép milyen értelemeben tekinthető az ember részének, így adott esetben a felelősségből hogyan részsül és viszont; a géppel kiegészített ember milyen mértékben veszít a felelősségre vonhatóságából, ha egyáltalán. A logikusan következtetés az lehet, hogy a felelősséget csak ágensekre lehessen értelmezni, szub-ágens komponensekre ne.

Ennél jóval földhöz ragadtabb, nem MI-specifikus, ám óriási gyakorlati jelentőségű kérdés az, hogy speciális tulajdonjogi viszonyokat teremtünk-e az emberi testben; megengedjük-e, hogy humánokkal, adott esetben páciensekkel szemben olyan feltételeket szabjanak a gépeket gyártó cégek, intézmények, amiben egy protézis, illetve az általuk generált adat a cég tulajdonában marad, így szélsőséges esetben deaktiválhatja, visszaveheti azt.

13.2 Végszó

Kübler-Ross és munkatársai[314] modellje szerint a gyász öt szakaszra osztható: sokk, tagadás, harag, alkudozás és elfogadás.

Ha úgy vesszük, hogy a mai, rendkívül fejlett MI érkezése gyászt okozott, az magyarázatot adhatna a vele kapcsolatos sokkra és haragra; ami még fontosabb, indokolhatná a tagadást is, amelyhez a kritikai filozófia kiváló partner.

A gyász egy nemkívánatos változáshoz vagy veszteséghez kapcsolódik. Az MI miatt bekövetkező munkahelyvesztés egyesek esetében már megtörtént, és még sokakat fog érinteni a jövőben. Másoknál a korábban értékesnek tartott készségek megbecsülése és jövedelemtermelő képessége csökken egy olcsóbb helyettesítő termék miatt. Az MI-nek ökológiai következményei is vannak (pl. a magas energiaigénye, bár ez változhat a jövőben), amelyek gyászra adhatnak okot.

Sokan vannak olyanok is, akik nem kerülnek rosszabb helyzetbe, de változást kell átélniük – éppen, amikor kényelmes helyre értek a karrierjükben, vagy intellektuális

[314] Kübler-Ross és társai (1972).

kontrollt éreztnek, az egész rendszer felborul. Az akadémiai világban az ember által írt ívek értéke csökken, ami valószínűleg rövidebb kommunikációkhoz vezet. Az oktatásban el kell fogadni a az írásos beadandók többségének értelmetlenségét. Szinte mindenki kénytelen lesz elviselni a változást – az üzleti folyamatok, a használható szolgáltatások és eszközök átalakulását. Ez erőfeszítést és tanulást igényel majd, sokak számára pedig szenvedést is: így a gyász érzése indokolt.

Más, elvontabb dolgokat is gyászolhatunk. Az ember, mint egyedülállóan intelligens lény státuszának vége (metafizikai vagy teológiai[315] értelemben) szintén nemkívánatos változás lehet. Mások szabadságunkért aggódnak, figyelmeztetve arra, hogy ne adjuk fel autonómiánkat, ne fogadjunk el útmutatást, vagy ami még rosszabb, parancsokat gépektől, mert ez aláásná önállóságunkat.[316]

Feltételezem, hogy az MI-vitákat tápláló intellektuális szenvedély intenzitása összhangban van a változás méretével és mindent átható természetével, amelyet az MI hoz magával. Ha ez valóban így van, akkor ez az intenzitás hosszú ideig fennmaradhat.

Az elfogadás előtti szakaszok bejárása nagy érzelmi és mentális terhet ró ránk; ezért érdemes nem elsietni ezeket. Az MI kontextusában ennek egyik módja az lehet, ha abbahagyjuk az intellektuális szalmaszálakba kapaszkodást és a szalmabábok hadseregének építését az MI valódi természetével, azaz tényleges korlátaival és morális értékével kapcsolatban. Ehelyett érdemes időt szánni a kérdések alaposabb megértésére és a valóságos problémák kezelésére.

Ebben a könyvben megkíséreltem leírni egy olyan megközelítést az MI jelenségéhez, amely talán lehetővé teszi, hogy fontos etikai kérdéseket jobban ítéljünk meg vele kapcsolatban, beleértve az MI ágens megmagyarázhatósági korlátait, a méltányosság kérdéseit, a konstruktőr tervezési kihívásait és az esetleg felmerülő MI iránti morális kötelezettségeinket is.

Arra próbáltam ösztönözni az olvasót, hogy utasítsuk el azt az episztemikus aszkézist, amely a gép modelljére támaszkodik a gép valósága helyett, vagy csak a viselkedését tanulmányozza, miközben a belső működése is tökéletesen hozzáférhető. Ehelyett

[315] A teológiai dimenzióért lásd J. K. Smith (2022).
[316] Collins (2018).

minden elérhető szempontból vizsgáljuk meg az MI-t. Koncentráljunk az MI rendszerek dekompozíciójára – a modellek vizsgálatával ellentétben –, valamint tapasztaljuk meg a viselkedésüket, és figyeljünk arra a benyomásra, amit hagynak bennünk, bízva abban, hogy a hallgatólagos képességünk az aktív ágens természetének átfogó entitásként való megértésére jól fog működni legalább néhányunk számára.

Így végül egy empirikus személyes tudásra tehetünk szert, amelyre támaszkodhatunk, amikor a konstruktőri, felhasználói vagy akár az MI iránti morális kötelességeinket próbáljuk felmérni.

Hivatkozások

AI HLEG (2019) *Ethics Guidelines for Trustworthy AI*. https://ec.europa.eu/futurium/en/ai-alliance-consultation (Last accessed 29 June 2023).

Aiso H (1988) The Fifth Generation Computer Systems Project. *Future Generations Computer Systems* 4: 159-175

Alan Turing Institute (2020) *Explaining Decisions Made with AI. London: Alan Turing Institute*. Available online at: https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf (Last accessed 29 June 2023).

Allhoff F (2011) What Are Applied Ethics? *Science and Engineering Ethics* 17(1):1-19.

Alon-Barkat S and Busuioc M (2023) Human–AI interactions in public sector decision making: “Automation bias” and “selective adherence” to algorithmic advice. *Journal of Public Administration Research and Theory* 33(1): 153–169.

Anderson M és Anderson SL (Szerk.) (2011) *Machine ethics*. Cambridge University Press.

Asaro PM (2006) What should we want from a robot ethic? *International Review of Information Ethics* 6: 9–16.

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A és Rahwan I (2018) The Moral Machine experiment. *Nature*, 563 (7729), 59.

Baer BR, Gilbert DE és Wells MT (2019) *Fairness criteria through the lens of directed acyclic graphical models. arXiv preprint* arXiv:1906.11333.

Baker CF, Fillmore CJ és Lowe JB (1998) The Berkeley FrameNet project. *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Bauer F és Kaltenböck M (2011) *Linked open data: The essentials.* Edition mono/monochrom, Vienna, 710, 21.

Beauchamp TL (2003) The nature of applied ethics. Frey RG és Wellman CH. *A companion to applied ethics*, 1-16. Wiley-Blackwell.

Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives*, 11, 375–399.

Bellamy RKE és társai (2019) AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias". *IBM Journal of Research and Development* 63(4):1-15, doi: 10.1147/JRD.2019.2942287.

Berners-Lee T, Hendler J és Lassila O (2001). The semantic web. *Scientific American*, 284(5), 34-43.

Bloor D (2005) The strong programme in the sociology of knowledge. Knowledge. *Critical Concepts*, 5.

Booch G (1990) *Object oriented design with applications*. Benjamin-Cummings Publishing.

Bostrom N és Yudkowsky E (2014) The ethics of artificial intelligence. Frankish K és Ramsey WM (Szerk.), *The Cambridge Handbook of Artificial Intelligence:* 316–334. Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.020

Borgmann A (1984) *Technology and the Character of Contemporary Life: A Philosophical Inquiry.* The University of Chicago Press.

von Braun J és Baumüller H (2024). Az MI/robotika és a szegények. Rab Á, Majó-Petri Z, Koltai A (szerk) *Felelős AI: Az AI irányításának gyakorlati megközelítése*: Budapest, Magyarország: Gondolat Kiadó, 255-280.

Brooks R (2000) Will robots rise up and demand their rights? *Time* (June 19, 2000).

Burrell J (2016) How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1), 2053951715622512.

Buchanan BG és Shortliffe EH (Szerk)., (1984), *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Longman Publishing Co., Inc., London.

Bush V (1945) *Science, the endless frontier. A report to the president.* Washington: Library of Congress.

Butler RW (2001) What is formal methods? *NASA LaRC Formal Methods Program*.

Cannon D (2021) The Critical to Post-Critical Shift: What It Means and Why It's Important. Polanyiana 30: 57-109.

Champagne M és Tonkens R (2015) Bridging the responsibility gap in automated warfare. *Philosophy & Technology,* 28, 125-137.

Chakrabarti A és Bligh TP (1996) An approach to functional synthesis of mechanical design concepts: theory, applications, and emerging research issues. *AI EDAM* 10(4), 313-331.

Coeckelbergh (2010) Robot rights? Towards a social-relational justification of moral consideration. *Ethics and information technology* 12: 209-221.

Collingridge D (1981) *The Social Control of Technology.* Open University Press.

Collins HM (2010) *Tacit and explicit knowledge.* Chicago: University of Chicago Press.

Collins HM (2012) *Symbols, Strings and Social Cartesianism*. Polanyiana 21: 59-70.

Collins H (2018) *Artifictional intelligence: Against humanity's surrender to computers*. John Wiley & Sons.

Collins H és Pinch T (2014) *The golem at large: What you should know about technology.* Cambridge University Press.

Copeland BJ (1997) The Church–Turing Thesis. *Stanford Encyclopedia of Philosophy* (Jelentősen átdolozva 2005-ben).

Cowan R és Hultén S (1996) Escaping lock-in: the case of the electric vehicle. *Technological forecasting and social change* 53 (1): 61-79.

Crevier D (1993) *AI: The Tumultuous Search for Artificial Intelligence*, New York, NY: BasicBooks, ISBN 0-465-02997-3

Crowther W, Rettberg R, Walden D, Ornstein S és Heart F (1973) A system for broadcast communication: Reservation-ALOHA. *Proc. 6th Hawaii Int. Conf. Syst. Sci*: 596-603.

Cundiff WE (1984) Model integration and algorithmic transparency in APL. Applied *Mathematical Modelling* 8(6): 445-448.

Danaher J (2016) Robots, law and the retribution gap. *Ethics and Information Technology* 18(4): 299-309.

Danaher J (2019) The rise of the robots and the crisis of moral patiency. *AI & Society* 34: 129–136. https://doi.org/10.1007/s00146-017-0773-9

Daniels N (1996) *Justice and justification: Reflective equilibrium in theory and practice*. Cambridge: Cambridge University Press.

Darling K (2012) Extending Legal Rights to Social Robots. *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.2044797.

David PA (1985) Clio and the Economics of QWERTY. *The American economic review* 75 (2): 332-337.

De Jongh D, Massey EK és Bunnik EM (2022) Organoids: a systematic review of ethical issues. *Stem Cell Research & Therapy* 13(1): 337.

Dewitt B, Fischhoff B és Sahlin N (2019) ‘Moral machine’ experiment is no basis for policymaking. *Nature* 567(31).

Diakopoulos N (2014) *Algorithmic accountability reporting: On the investigation of black boxes*. Tow Center for Digital Journalism, A Tow/Knight Brief. https://doi.org/10.7916/D8ZK5TW2

Dosi G (1982) Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. *Research policy* 11(3): 147-162.

Dubber MD, Pasquale F és Das S (Szerk.) (2020) *The Oxford handbook of ethics of AI*. Oxford Handbooks, 2020.

Duhem P [1914] (1954) *The Aim and Structure of Physical Theory*, (ford Wiener PW); eredetileg: *La Théorie Physique: Son Objet et sa Structure* (Paris: Marcel Riviera & Cie.), Princeton, NJ: Princeton University Press.

Dusek V (2006) *Philosophy of technology: An introduction*. Blackwell.

Dreyfus HL (1992) What computers still can’t do: A critique of artificial reason. Boston: MIT press.

van Eck D, McAdams DA és Vermaas PE (2007) Functional decomposition in engineering: a survey. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* Vol. 48043, 227-236.

Einevoll GT és társai (2019) The scientific case for brain simulation. *Neuron* 102(4): 735-744.

Ellul J (2021) *The technological society.* Vintage.

Feenberg A (2009) Democratic rationalization: Technology, power, and freedom. *Readings in the philosophy of technology:* 139-155.

Feenberg A (2003) Democratic rationalization: Technology, power, and freedom. *Philosophy of technology:* 652-665.

Ferguson ES (1987) Risk and the American engineering profession: the ASME Boiler Code and American industrial safety standards. *The Social and Cultural Construction of Risk:* 301-316. Springer, Dordrecht.

Ferrario A, Loi M (2022) How Explainability Contributes to Trust in AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT

'22). Association for Computing Machinery, New York, NY, USA, 1457–1466. https://doi-org.proxysbauniss.idm.oclc.org/10.1145/3531146.3533202

Ferrucci DA (2012) Introduction to “this is watson”. *IBM Journal of Research and Development* 56(3.4).

Fishkin J (2014) *Bottlenecks: A new theory of equal opportunity.* Oxford University Press, USA, New York, NY.

Floridi L (2016) Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 374.

Floridi L (2020) AI and Its New Winter: from Myths to Realities. *Philos. Technol.* 33, 1–3. https://doi.org/10.1007/s13347-020-00396-6

Floridi L és társai (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28(4): 689–707.

Floridi L és Sanders JW (2004) On the morality of artificial agents. *Minds and Machines* 14(3): 349–379.

Flew A (1975) *Thinking About Thinking*. HarperCollins, New York.

Fowler M és Highsmith J (2001) The agile manifesto. *Software development* 9(8): 28-35.

Foxon TJ (2014) Technological lock-in and the role of innovation. *Handbook of sustainable development*. Edward Elgar Publishing.

Frey RG (2011) *Utilitarianism and Animals*. Oxford University Press.

Gabriel I (2020) Artificial intelligence, values, and alignment. *Minds and machines* 30(3): 411-437.

Ganzer PD és társai (2020). Using a Sensorimotor Demultiplexing Neural Interface. *Cell* 181(4). https://doi.org/10.1016/j.cell.2020.03.054

Gerrie J (2008) Three species of technological dependency. *Techné: Research in Philosophy and Technology* 12(3): 184-194.

Ghassemi M, Oakden-Rayner L és Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gill JH (2000) *The tacit mode: Michael Polanyi's postmodern philosophy.* SUNY Press.

Gill KS (szerk) (2012) *Human machine symbiosis: The foundations of human-centred systems design*. Springer Science & Business Media.

Gill SP (2015) *Tacit engagement.* Springer International Publishing.

Goddard K, Roudsari A és Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19(1): 121-127.

Goodall NJ (2014a) Ethical decision making during automated vehicle crashes. *Transportation Research Record* 2424(1): 58-65.

Goodall NJ (2014b) Vehicle automation and the duty to act. *Proceedings of the 21st world congress on intelligent transport systems:* 7-11.

Gorry GA (1973) Computer-assisted clinical decision making. *Methods of Information in Medicine* 12: 45-51.

Goertzel B (2002) Thoughts on AI morality. *Dynamical Psychology: An International, Interdisciplinary Journal of Complex Mental Processes*
https://www.goertzel.org/dynapsyc/2002/AIMorality.htm

Google (2019) Perspectives on issues in AI governance (1–34). Retrieved February 11, 2019. https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

Gorry GA, Kassirer JP, Essig A és Schwartz WB (1973) Decision analysis as the basis for computer-aided management of acute renal failure. *American Journal of Medicine* 5: 473 -484.

Green AM (1953) *History of the ASME boiler code*. New York: American Society of Mechanical Engineers.

Gunkel DJ (2018) *Robot rights*. MIT Press.

Habakkuk HJ (1962) *American and British Technology in the Nineteenth Century*, Cambridge, Cambridge University Press.

Hagendorff T (2020a) The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines* 1-22.

Hagendorff T (2020b) Forbidden knowledge in machine learning reflections on the limits of research and publication. *AI & Society.* https://doi.org/10.1007/s00146-020-01045-4

Hagendorff T (2023) AI ethics and its pitfalls: not living up to its own standards? *AI and Ethics* 3(1): 329-336.

Heil J (2004). *Philosophy of Mind: A Guide and Anthology.* Oxford; New York: Oxford University Press, 2004.

Hempel CG (1958) *The theoretician's dilemma: A study in the logic of theory construction.* University of Minnesota Press, Minneapolis.

Hendler J (2008) Avoiding Another AI Winter. *IEEE Intelligent Systems* 23(2): 2-4, https://doi.org/10.1109/MIS.2008.20

Heilbroner RL (1967) Do machines make history? *Technology and Culture* 8: 335–45.

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

Héder M (2019) Michael Polanyi and the epistemology of engineering. Héder M és Nádasi E (Szerk), *Essays in Post-Critical Philosophy of Technology:* 63–70. Vernon Press.

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

Héder M (2021) AI and the resurrection of Technological Determinism. *Információs Társadalom: Társadalomtudományi Folyóirat* 21(2): 119-130.

Héder M (2020b) A Criticism of AI Ethics Guidelines. *Információs Társadalom* 20: 57-73. https://doi.org/10.22503/inftars.XX.2020.4.5.

Héder M (2023a) The epistemic opacity of autonomous systems and the ethical consequences. *AI & SOCIETY* 38(5): 1819-1827.

Héder M(2023b) Explainable AI: A brief history of the concept. *ERCIM NEWS* 134: 9-10.

Héder M és Paksi D (2012) Autonomous robots and tacit knowledge. Appraisal: The Journal of the Society for Post-Critical and *Personalist Studies* 9(2): 8–14.

Héder M és Solt I (2013) DBpedia mashups. *Semantic Mashups: Intelligent Reuse of Web Resources:* 119-143. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hirtz J, Stone RB, McAdams DA, Szykman S és Wood KL (2002) A functional basis for engineering design: reconciling and evolving previous efforts. *Research in engineering Design* 13: 65-82.

Holdren JP, Bruce A, Felten E, Lyons T és Garris M (2016) *Preparing for the future of artificial intelligence*. Washington, D.C: Springer.

Hoffman PT (2015) *Why Did Europe Conquer the World?* Princeton University Press. Princeton University Press.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

Inga és társai (2023) Human-machine symbiosis: A multivariate perspective for physically coupled human-machine systems. *International Journal of Human-Computer Studies, 170,* 102926.

Iverson KE (1980) Notation as a tool of thought. *Communications of the ACM* 23(8): 444-465.

Jaques AE (2009) Why the moral machine is a monster. *Self-archived manuscript at University of Miami School of Law.* https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/MoralMachineMonster.pdf

Jain S, Suriyakumar V, Creel K és Wilson A (2024) Algorithmic Pluralism: A Structural Approach To Equal Opportunity. *The 2024 ACM Conference on Fairness, Accountability, and Transparency:* 197-206.

Jobin A, Ienca M és Vayena E (2019) The global landscape of AI ethics guidelines. *Nat Mach Intell* 1: 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson BB és Covello VT (Szerk) (2012) The social and cultural construction of risk. *Essays on risk selection and perception* 3. Springer Science & Business Media.

Jurecki RS, Jaśkiewicz M, Guzek M, Lozia Z és Zdanowicz P (2012) Driver's reaction time under emergency braking a car-research in a driving simulator. *Eksploatacja i Niezawodność* 14(4): 295-301.

Kandul S, Micheli V, Beck J, Kneer M, Burri T, Fleuret F és Christen M (2023) *Explainable AI: A Review of the Empirical Literature* (SSRN Scholarly Paper No. 4325219). https://doi.org/10.2139/ssrn.4325219

Kaplan A, Haenlein M (2020) Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons* 63(1): 37-50.

Kartikeya A (2022) Examining Correlation Between Trust and Transparency with Explainable Artificial Intelligence. Lecture Notes in *Networks and Systems,* 507 LNNS, 353–358. Scopus. https://doi.org/10.1007/978-3-031-10464-0_23

Karatsu H (1982) What is Required of the 5th Geneartion Computer – Social Needs and Its Impact. Moto-Oka (szerk). *Fifth Generation Computer Systems,* 93-106. North-Holland Amsterdam.

Kauffman S (2016) Cosmic mind? *Theology and Science* 14(1): 36-47.

Kiener M (2022) Can we Bridge AI's responsibility gap at Will? *Ethical Theory and Moral Practice* 25(4), 575-593.

Kizilcec RF (2016) How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems:* 2390–2395.

Korinek A és Stiglitz JE (2018) Artificial intelligence and its implications for income distribution and unemployment. *The economics of artificial intelligence: An agenda:* 349-390. University of Chicago Press.

Kurzweil R (2005) *The Singularity Is near: When Humans Transcend Biology.* New York: Viking.

Kübler-Ross E, Wessler S és Avioli LV (1972) On death and dying. *Journal of the American Medical Association* 221(2): 174–179.

Landauer R (1961) Irreversibility and heat generation in the computing process. *IBM journal of research and development* 5(3), 183-191.

Larson EJ (2021) *The myth of artificial intelligence: Why computers can't think the way we do*. Belknap Press.

Latour B (2007) *Reassembling the social: An introduction to actor-network-theory.* Oup Oxford.

Lem S (1972) *Summa technologiae*. Budapest:Kossuth. Ford: Radó György.

Lenat DB (1995) „CYC: A large-scale investment in knowledge infrastructure" *Communications of the ACM* 38(11): 33–38.

Licklider JCR (1960) Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics* 1, 4–11

Lighthill J (1973) Artificial Intelligence: a general survey. Flowers, BH (szerk.) *Artificial Intelligence: a Paper Symposium*. London: Science Research Council, pp 1–21.

Lin P (2014) The robot car of tomorrow may just be programmed to hit you. *Wired Magazine* 5.

Lin P (2015) Why Ethics Matters for Autonomous Cars. M. Maurer és társai (Szerk), *Autonomes Fahren*, DOI 10.1007/978-3-662-45854-9_4,

Liu CF, Chen ZC, Kuo SC és Lin TC (2022) Does AI explainability affect physicians' intention to use AI? International Journal of *Medical Informatics* 168, 104884. https://doi.org/10.1016/j.ijmedinf.2022.104884

Luccioni AS, Jernite Y, Strubell E (2024) *Power Hungry Processing: Watts Driving the Cost of AI Deployment?* ArXiv: https://arxiv.org/pdf/2311.16863

Luhmann N (1979) *Trust and Power.* Chichester: John Wiley.

Lyell D and Coiera E (2017) Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association* 24(2): 423–431.

Macklin R (1977) Moral progress. *Ethics* 87(4): 370-382.

Magidor O (2024) Category Mistakes, *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition), Edward N Zalta és Uri Nodelman (szerk.), https://plato.stanford.edu/archives/sum2024/entries/category-mistakes/.

Marcuse H (1977) The New Forms of Control. *Technology and Man's Future* (2. kiadás). New York: St. Martin's Press.

Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6: 175-183.

Microsoft Corporation (2019) *Microsoft AI principles.* https://www.microsoft.com/en-us/ai/our-approach-to-ai.

Mill JS (1884) *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation (Vol. 1)*. Longmans, green, and Company.

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nature machine intelligence* 1(11): 501-507.

Mittelstadt BD, Allo P, Taddeo M, Wachter S és Floridi L (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2), 205395171667967.

Mosier KL, Skitka LJ, Heers S és Burdick M (2017) Automation bias: Decision making and performance in high-tech cockpits. *Decision Making in Aviation:* 271-288. Routledge.

Moody-Adams MM (1999) The idea of moral progress. *Metaphilosophy* 30(3): 168-185.

Moravec H (1988) *Mind children: The future of robot and human intelligence*. Harvard University Press.

Moto-Oka T (1982) Keynote Speech: Challenge for Knowledge Information Processing Systems. Moto-Oka (szerk). *Fifth Generation Computer Systems*: 3-92. North-Holland, Amsterdam.

Mullins P (2001) The ‘post-critical’ symbol and the ‘post-critical’ elements of Polanyi’s thought. *Polanyiana* 10(1–2): 77–90.

Mullins P (2019) Michael Polanyi on Machines as Comprehensive Entities. Essays in Post-Critical Philosophy of Technology. Héder M és Nádasi E (szerk.). Wilmington, Delaware: Vernon Press 37-61.

Mullins P (2021) Michael Polanyi’s Post-Critical Philosophical Vision of Science and Society. *Science, Freedom, Democracy. Hartl P és Tuboly ÁT* (szerk). New York and London: Routledge: 15-38.

Nassi B, Nassi D, Ben-Netanel R és Elovici Y. (2021) Spoofing Mobileye 630's Video Camera Using a Projector. *Workshop on Automotive and Autonomous Vehicle Security*: 25

Newell A (1983) Intellectual Issues in the history of Artificial Intelligence. Machlup F és Mansfield U (Szerk): *The Study of Information: Interdisciplinary Messages*.Wiley and Sons: New York.

Nguyen A, Yosinski J, Clune J (2015) Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *Computer Vision and Pattern Recognition* (CVPR '15), IEEE, 2015.

Nyholm S (2020) *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.

OECD (2019) *Recommendation of the Council on Artificial Intelligence:* 1–12. [accessed 18 / 08 / 2020]. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Paksi D és Héder M (2020) *Guide to personal knowledge*. Vernon Press.

Paksi D (2017) Medium emergence—Part one—The personalist theory of emergence. *Appraisal* 11(2): 13–22.

Papenmeier A, Englebienne G és Seifert C (2019) *How model accuracy and explanation fidelity influence user trust*. arXiv preprint arXiv:1907.12652.

Pasquale F (2015) *The black box society.* Harvard University Press.

Penrose R (1994) *Shadows of the mind* (Vol. 4). Oxford University Press.

Penrose R és Mermin ND (1990) The emperor's new mind: Concerning computers, minds, and the laws of physics. American *Journal of Physics* 58(12): 1214-1216.

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

Polányi M (1965) The Structure of Consciousness. *Brain* LXXXVIII: 799-810

Polányi M (1976) Life's irreducible structure. *Science* 160:1308-1312.

Posner RA (1974) *Theories of economic regulation* (No. w0041). National Bureau of Economic Research.

Putnam H (1964) Robots: Machines or artificially created life? *Journal of Philosophy* 61(21): 668-691.

Quine WV (1951) Main trends in recent philosophy: Two dogmas of empiricism. *The philosophical review:* 20-43.

Rab Á (2024) A mesterséges intelligencia és az automatizált szolgáltatások percepciója a magyar társadalomban 2023 végén. Rab Á, Majó-Petri Z, Koltai A (szerk) *Felelős AI: Az AI irányításának gyakorlati megközelítése*: Budapest, Magyarország: Gondolat Kiadó, 297-306.

Rawls J (1971) *A theory of justice*. Cambridge, MA: Belknap Press.

Rawls J (1999) *A Theory of Justice* (átdolgozott kiadás). Cambridge, MA: Harvard University Press.

Regan T (1983) *The case for animal rights*. Berkeley: The University of California Press

Robbins SA (2019) Misdirected Principle with a Catch: Explicability for AI. *Minds & Machines* 29, 495–514.

Ropohl G (1998) Technological Enlightenment as a Continuation of Modern Thinking. *Research in Philosophy and Technology* 17: 239-248.

Rorty R (2007) Dewey and Posner on pragmatism and moral progress. *The University of Chicago Law Review* 74(3): 915-927.

Sandberg A (2013) Feasibility of whole brain emulation. Vincent C. Müller (szerk.), *Theory and Philosophy of Artificial Intelligence*: 251-64, Berlin: Springer.

Sandberg A és Bostrom N (2008) *Whole Brain Emulation: A Roadmap, Technical Report* #2008-3, Future of Humanity Institute, Oxford University.

Sanders AF (1991) Tacit knowing—Between modernism and postmodernism. *Tradition and Discovery* 18(2): 15–21.

Sauer H (2019) *„The Enemy of the Good. Towards a Theory of Moral Progress" (PROGRESS)*. ERC-fundend Starting Grant project https://progress.sites.uu.nl/

Saxena NA, Huang K, DeFilippis E, Radanovic G, Parkes DC, Liu Y (2019) How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society:* 99-106.

Schmidt P Biessmann F és Teubner T (2020) Transparency and trust in artificial intelligence systems. *Journal of Decision Systems:* 29(4), 260-278.

Searle JR (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3): 417–424.

Simon HA (1996) *The sciences of the artificial*. MIT Press.

Singer P (1975) *Animal liberation.* New York: Random House.

Singh IL, Molloy R és Parasuraman R (1993) Automation induced “complacency”: development of the complacency-potential rating scale. *Int J Aviat Psychol* 3:111e22.

Smith JK (2022) *Robot theology: Old questions through new media.* Wipf and Stock Publishers.

Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146, 102551.

Skitka L és társai (1999) Does automation bias decision-making? *International Journal of Human-Computer Studies* 991–1006.

Steedman I (1997) Jevons's Theory of political economy and the ‘marginalist revolution’. *Journal of the History of Economic Thought* 4(1): 43-64.

Stuart SAJ és Dobbyn C (2002) A Kantian prescription for artificial conscious experience *Leonardo* 35.4 407-411. https://doi.org/10.1162/002409402760181213.

Stump D (2006) Rethinking modernity as the construction of technological systems. Veak TJ. *Democratizing technology: Andrew Feenberg's critical theory of technology.* SUNY press.

Swartout WR (1977) A digitalis therapy advisor with explanations. *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, MIT A1 Laboratory, 819-826.

Swartout WR (1985) Explaining and justifying expert consulting programs. Reggia JA, Tuhrim S. (szerk) Computer-Assisted Medical Decision Making. *Computers and Medicine:* 254-271. New York, NY: Springer New York.

Taddeo M és Florid L (2018) How AI can be a force for good. *Science* 361 (6404): 751–752.

The European Parliament and the Council (2023) *Artificial Intelligence Act*. /Official Journal of the European Union, L series,/ 2024/1689. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

Toosi A, Bottino AG, Saboury B, Siegel E és Rahmim A. (2021). A brief history of AI: how to prevent another winter (a critical review). *PET clinics* 16(4), 449-469.

Turing AM (1950) Computing machinery and intelligence. *Mind* 49 (236): 433-460.

Umeda Y, Ishii M, Yoshioka M, Shimomura Y és Tomiyama T (1996) Supporting conceptual design based on the function-behavior-state modeler. *AI EDAM* 10(4): 275-288.

Vadász P (2021) Elkerülhető, hogy a robotok diszkrimináljanak bennünket? Török B, Ződi Zsolt (szerk) *A mesterséges intelligencia szabályozási kihívásai : Tanulmányok a mesterséges intelligencia és a jog határterületeiről.* Budapest, Magyarország : Ludovika Egyetemi Kiadó, 89-107.

Van Lent M, Fisher W, Mancuso M (2004) An explainable artificial intelligence system for small-unit tactical behavior. *Proceedings of the National Conference on Artificial Intelligence*. San Jose, CA: AAAI Press, 900–907. ISBN 0262511835.

Varzi A (2019) Mereology. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (szerk.), https://plato.stanford.edu/archives/spr2019/entries/mereology/.

Vica C, Voinea C és Uszkai R (2021) The emperor is naked: Moral diplomacies and the ethics of AI. *Információs Társadalom* 21(2): 83-96: https://dx.doi.org/10.22503/inftars.XXI.2021.2.6

Waterman DA (1986) *A guide to expert systems*. Reading, PA: Addison-Wesley.

Wetzel L (2018) Types and Tokens, *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.) https://plato.stanford.edu/archives/fall2018/entries/types-tokens/

Weiner JL (1980) BLAH, a system which explains its reasoning, *Artificial Intelligence* 15 (1–2): 19-48. https://doi.org/10.1016/0004-3702(80)90021-1

Weizenbaum J (1966) ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1): 36-45.

Wiener N (1950) *The Human Use of Human Beings*. Cambridge, Massachusetts: Riverside Press.

Whitby B (2008) Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. I*nteracting with Computers* 20(3), 326–333.

Wiener N (1950) *The Human Use of Human Beings*. Cambridge, Massachusetts: Riverside Press.

Williams B (1962) The Idea of Equality. Laslett P és Runciman R (Szerk.) *Philosophy, Politics and Society*, Blackwell, Oxford, 112–117.

Winfield AF és társai (2021) IEEE P7001: A proposed standard on transparency. *Frontiers in Robotics and AI* 8, 665 – 729.

Winograd T (1972) SHRDLU-Procedures as a representation for data in a computer program for understanding natural language. *Cognitive Psychology* 3.

Winograd T és Flores F (1986) *Understanding computers and cognition: A new foundation for design.* Intellect books, Bristol.

Wysocki O, Davies JK, Vigo M, Armstrong AC, Landers D, Lee R és Freitas A (2023) Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence* 316: 103839

Z Karvalics L (2007) Információs társadalom – mi az? - Information Society - what is it exactly?: Egy kifejezés jelentése, története és fogalomkörnyezete. Pintér R (szerk) *Az információs társadalom : Az elmélettől a politikai gyakorlatig.* Budapest, Magyarország: Gondolat Kiadó 29-46.

Z Karvalics L (2015) Mesterséges intelligencia – a diskurzusok újratervezésének kora. *Információs Társadalom* 15(4) 7-41.

Z Karvalics L (2024) Mesterséges intelligencia, technológia és társadalom. Néhány mesterkontextus. Rab Á, Majó-Petri Z, Koltai A (szerk) *Felelős AI: Az AI irányításának gyakorlati megközelítése*: Budapest, Magyarország: Gondolat Kiadó, 227-254.

Ződi Zs (2024) A mérnöki és jogi gondolkodásmód ellentéte a mesterséges intelligenciáról szóló jogszabályvitában. Rab Á, Majó-Petri Z, Koltai A (szerk) *Felelős AI: Az AI irányításának gyakorlati megközelítése*: Budapest, Magyarország: Gondolat Kiadó, 13-36.