POKOL BÉLA

Az emberi lét rétegei és a robotetika kérdései

A tanulmány a robotvilág hatásaival átitatott társadalmi körülmények között felmerülő, új etikai dilemmákat igyekszik elemezni. Ehhez azt a valóságképet veszi alapul, mely Nicolai Hartmann ontológiája nyomán a valóság létrétegei között találja meg az egyre terjedő mesterséges intelligencia helyét. Ebből a kiindulópontból veszi górcső alá a különösen angol nyelven nagy létszámú robotetikai elemzés összegző tanulmányait, és azt vizsgálja, hogy az emberi lét négyrétegűségének elmélete milyen korrekciókat tesz szükségessé e téren az eddigi elemzésekhez képest.

Kulcsszavak: *mesterséges intelligencia, ontológia, evolúciós ugrások, Nicolai Hartmann*

Szerzői információ:

Pokol Béla, jogtudós, politológus, egyetemi tanár, a szociológiai tudomány (akadémiai) doktora (1989). Az Eötvös Loránd Tudományegyetem Állam- és Jogtudományi Karán 1977-ben szerzett diplomát, politikatudományi kandidátusi disszertációját 1986-ban védte meg. Az Eötvös Loránd Tudományegyetem és a Szegedi Tudományegyetem oktatója. Főbb kutatási területei a jogelmélet, a politológia, a társadalomelmélet és a társadalmi evolúció.

Így hivatkozzon erre a cikkre:
Pokol Béla, „Az emberi lét rétegei és a robotetika kérdései”,
Információs Társadalom XVIII, 3–4. szám (2018): 8–24.
https://dx.doi.org/10.22503/inftars.XVIII.2018.3-4.1

Pokol Béla

Az emberi lét rétegei és a robotetika kérdései

Az emberi lét és az emberi közösségek élete az evolúcióval egymásra épülő létrétegek összegződő törvényszerűségein alapul Nicolai Hartmann empirikus elemzéseken is felépített ontológiájának tézisei szerint (Hartmann 1962). A mesterséges intelligencia (MI) utóbbi években felgyorsult fejlődése és egyre szélesebb használatba vétele ebben a szerkezetben a négy létréteg (fizikai, biológiai, lelki és értelmi) legfelső rétegét érinti közvetlenül, és annak erejét fokozza az alsóbbak rovására. A mesterséges intelligencia későbbi fejlődésének eredményeképpen pedig – esetelegesen az emberi irányítás alól kicsúszva vagy felszabadulva, önállóságra szert téve – mint egy evolúció során keletkezett, új létréteget lehet felfogni. Ez azonban az eddigi három evolúciós ugrástól eltérően nem szorulna az összes eddig létező, alsóbb létrétegre – az MI mai, fizikai megtestesüléseit jelentő robotokat szem előtt tartva –, pusztán a fizikaira (Pokol 2017). E tanulmány elemzései ezt alapul véve kívánják szemügyre venni a mesterséges értelemmel egyre inkább átitatott, a mai társadalmi mechanizmusok keretei között újonnan keletkező, morális és ezzel összefüggő jogi dilemmákat, figyelmet szentelve arra is, hogy az elemzési keret mennyiben változik meg eme dilemmák számára, ha mindvégig az emberi lét többrétegűségét tartjuk szem előtt.

A robotetika előkérdései

A robotvilág morális kérdéseit áttekintő tanulmányában Keith Abney három területet különít el a kérdések csoportosítására: a robotokat készítők és programozók felé irányuló követelmények és tilalmak területét (mint például az orvosetika); a robotokba beprogramozandó követelmények és tilalmak területét, melyet először Isaac Asimov fogalmazott meg *a robotika három törvénye* címszó alatt; végül pedig – egy perspektivikus jövőben – az öntudattal és tudatossággal rendelkező robotok őket is megillető morális igényeinek és az őket is érintő „emberi jogok" kérdéskörét (Abney 2011: 35). Mindhárom területen közös dilemmát jelent a főbb morálelméleti kiindulópontok közötti választás, melyeket – az átfogó morálfilozófiai közösségekben domináló iskolákba való beosztást alapul véve – a robotetikusok egyfelől a *deontológiai* kiindulóponttal azonosítanak (a szabály az szabály, és ezt követni kell, bármi legyen is a következménye), legszikárabb képviselőjeként Kant, kategorikus imperatívuszon nyugvó morálfilozófiáját kiemelve, másfelől az előbbivel polárisan szembenálló *konzekvencionalista* iskolát, mely a morális döntés alapjának a cselekvés következményeinek mérlegelését tekinti (a cselekvés során azt kell a morálra törekvőnek választani, amely a lehető legtöbb ember boldogságát tendenciájában növeli, nem csökkenti). Harmadszor az *erényetika* iskoláját, mely a morál definiálásakor nem az egyes szituációkban követendő követelményekre összpontosít (mint az előbbi két, egymással szembenálló iskola), hanem az ember személyiségének tartós cselekvési diszpozícióira, egyszerűbben: a szocializált morális értékeire. Itt a morálra törekvő személy nem azt kérdezi, hogy mi az erkölcsi szabály az adott szituációban (hiszen az egyre bonyolultabb mo-

dern világban a legtöbbször nincsenek is egyértelmű szabályok), hanem azt, hogy miként kell döntenie egy bátor, igazságos, hűséges, igazmondó stb. embernek – orvosnak, tanárnak, mérnöknek stb. –, mert csak ez lehet morálisan helyes döntés (Abney 2011: 37).

A deontológiai iskola a pontos szabályok mérlegelés nélküli követést előírva csak a legszűkebb specializált területen alkalmazott robotok esetében létezhet (hiszen az összes szituációt kiszámítani és szabállyal ellátni csak ilyen szűk területen lehetséges), de felmerülhetnek előre nem látott helyzetek, melyek rossz irányba terelik a robot döntését. Egy harci robotot például egy algoritmus segítségével elméletileg meg lehet tanítani arra, hogy soha ne öljön gyereket. Afrika háborúinak gyerekkatonái között ez gyakorlatilag a harci robot hatástalan működését jelentené (Abney 2011: 42). Az általános jellegű robotok esetében pedig teljes mértékben alkalmazhatatlan a deontológiai megközelítés. De a szintén az egyes szituációk mérlegeléséhez kötött konzekvencionalista morálfilozófiai iskola is csak élet közelibb jellege miatt tűnik jobbnak, mert itt pedig az irányító premissza – „a lehetető legtöbb embernek a legnagyobb boldogágát növelje a választott döntés, és ne a csökkentés felé hasson!" – kivitelezhetetlen. Olyan nagytömegű információ feldolgozását kívánná ez meg, amely a legtöbbször a legnagyobb kapacitású számítógépes adattárolás esetén is időn túli lenne. Keith Abney álláspontja az, hogy ami a másodikként megjelölt robotmorál-területet illeti (tudniillik a robotok algoritmusába beprogramozott morális döntési premisszák), a deontológiai és az erényetikai irányzatok hibridje a perspektivikusan legjobb beépített robotmorál-verzió. Eszerint az inkább absztraktabb morális normák (morális erények) adják meg a döntési keretet, és mindenkor a beépített célok és döntési kontextusok pontosítják a robot által kiválasztott döntés meghatározói elemeit az adott szituációkban: „A hipotetikus, nem kategorikus imperatívuszok hibrid megközelítése (a tárgyhoz illően korlátozott, nem univerzális keretben), mely az erényetikából származik, a legjobb verzió a közeljövőben szükséges robot-morál mindkét értelméhez. (...) Mivel e megközelítés hangsúlya azon van, hogy a robot kiválóan teljesíthessen egy bizonyos szerepet, továbbá az erényetika ezzel összefüggő parancsain a nem-kantius autonóm robotok korlátozott feltételei között – vagyis a programozási céljaira, korlátozott kontextusaira és a tanulási képességeire szabva –, így az erényetikát ez természetes választássá teszi a robotetika számára" (Abeny 2011: 51).

Abney az emberi lét rétegei és a morál közötti összefüggéseket közvetetten érinti, amikor a morált a morális érzelmekkel azonosító *emotivizmus* és az ezzel szembenálló kognitív morálfelfogás szembenállása mellett foglal állást. Jelzi, hogy ha a morált az emotivistákkal együtt az érzelmekhez kötnénk, akkor az érzelmekkel rendelkező főemlősöket sem lehetne kizárni a morállal rendelkező lények közül, ami abszurd feltételezés: „Az ilyen nézeteknek, amellett, hogy nem tudják megmagyarázni, hogy az állatok miért vannak híján a morálnak, még erőfeszítéseket kell tenni arra, hogy meg tudják magyarázni az egymással szembenálló etikai álláspontok máskülönben egyenlő kognitív racionalitását, és ezen etikai állítások nézetelérését. (Emellett természetesen komoly nehézségekkel szembesülnek az érzelemmentes robotok esetében a morál feltételezésével)" (Abney 2001: 46). Ehhez képest a morál magyarázatához jobbnak tartja az emberi evolúcióval megjelenő új döntési mechanizmust kiemelő, evolucionista pszichológia álláspontját, mely szerint az érzelmi rendszer felett az embernél egyre inkább megjelent egy kognitív döntési rendszer is, és ez oly módon formálja a mindenkori döntéseket, hogy az ösztönös-érzelmi, első gondolati lépést mindig egy második kognitív mérlegelés követ – ezzel kerül az első korrigálásra.[1] „Az

[1] Hadd jelezzem, hogy még jóval a tudományos pszichológia megjelenése előtt, 1820-as Jogfilozófiájában Hegel is ezt az álláspontot képviselte: Georg Wilhelm Friedrich Hegel, *A jogfilozófia alapjai*, Akadémia Kiadó, Budapest, 1971, 173–186. old.

volúciós pszichológia azt állítja, hogy a legtöbb emberben nem csak egy, hanem két döntéshozási rendszer létezik. Az első egy szexuális, érzelmileg terhelt rendszer, amely egy sor emberi tevékenység számára az alapot jelenti, különösen stressz vagy nyomás alatt. Sok más állattal osztozunk ebben a nem kognitív döntéshozatali rendszerben, amelyben (szó szerint) „nem tudjuk, mit csinálunk" – vagy éppen „miért tesszük", amit teszünk. (...) De ez a „szellem a gépben" nem teszi ki a teljes emberi cselekvést; Libet és mások azt találták, hogy van egy vétó-képességünk is, amely a cselekvés tudatalatti beindítása után meg is változtathatja akcióinkat egy második, tudatos kognitívabb rendszer döntésével összhangban.„ (Abney 2011: 46). Ezt a gondolatmenetet folytatva Abney szinte Nicolai Hartmannt felidézve vázolja fel a két egymásra épülő réteg egymást kölcsönösen formáló hatását: „Emberben ez a gondolkodó-kognitív rendszer rátelepszik az emocionális (és gyorsabb) döntési rendszer központi centrumára és felülírja ezt, de azért az értelmünket gyakran mégis elnyomják az első ösztönszerű impulzusaink" (Uo.).

Miután Abney ebből azt a következtetést vonja le, hogy a kétrétegű emberi döntési mechanizmusból a felső (kognitv-racionális) réteg a felelős a morális döntésekért, felteszi a kérdést, hogy vajon lehetséges-e morálisan dönteni az átformált-felülbírált alsó réteg nélkül is? Hiszen e kérdés megválaszolása dönti el, hogy az érzelmi réteg nélküli robotoknál lehetséges-e morális döntés. E kérdésnél aztán polárisan szembenállóan dönt Hartmann tanaival. Igen lehetséges – mondja –, és a racionális döntési mechanizmus elegendő a morális döntéshez, ez lehetséges érzelmi létréteg nélkül is: „Ennél fogva a moralitás létezéséhez és így a morális személyiséghez szükséges a gondolkodó-kognitív rendszer. Ám szükséges-e ehhez az ősibb eredetű érzelmi rendszer is? (...) Más szavakkal: lehetnek-e az érzelem-nélküli robotok is morális személyek? (...) A kulcs a morális felelősséghez és személyiséghez a morális cselekvési képesség birtoklása, amely a racionális kognitív képességet követeli meg - de a képességet az érzelmi állapotok felvételére nem. Ily módon a robotok is minősíthetők a morális személyiség státusára.„ (Abney 2011: 47).

Ezekben az elemzésekben – Hartmannt alapul véve – két probléma is található. Egyrészt a fizikai létréteg feletti három létréteget tekintve hibásnak minősíthető az, hogy a biológiai ösztönzőket Abney egybefogja az érzelmi réteg meghatározóival. Már itt is egy egymásra épülés és átformálás van, és a nyers ösztönvilág egy-egy ösztönét a felettes lelki létréteg érzelmei átformálva egészítik ki. Például a biológiai nemi ösztön vadságát az összetartozás érzelmei formálják, nem is szólva az embernél az erre még ráépülő értelmi-szimbolikus felülírásokról és az ezek által létrehozott, szublimált szerelmi kapcsolatok nemi érintkezést átformáló aspektusairól.[2] Vagyis analitikailag nem kettős, hanem hármas döntési mechanizmust kell elválasztani az emberi döntések esetében, és a legelemibb ösztönreakciók és meghatározók mellett még ezek érzelmi szinten átformált megjelenési formái állnak az értelmi szint racionálisabb mérlegelései mögött/alatt. Egy-egy döntés és az ezt közvetlenül meghatározó ösztön, ennek érzelmi átformálása és mindezek értelmi felülírása azonban az emberi lét mindhárom felső létrétegének egymásra épülő törvényszerűségeibe ágyazódik be. Így az emberi morál – legyen az bármelyik társadalomban – azt előfeltételezi, hogy a faj továbbéléséhez férfinak és nőnek kell tartósan együtt élnie valamilyen formában a gyermeknemzéshez és a felnevelés biztosítására; egy nagyobb közösség szükséges a természet erőivel és más embercsoportokkal való küzdelem sikeres megvívására és a közösség fennmaradására, e nagyobb közösségeken belül pedig többé-kevésbé harmonikus viszo-

[2] Lásd Luhmann-nak az ezt a folyamatot történetileg elemző munkáját: Niklas Luhmann, *Liebe als Passion: Zur Codierung von Intimität*, Suhrkamp, Frankfurt am Main, 1994.

nyokban kell érintkezni, a közös tevékenységet megszervezni. A morális erények (normák és értékek) így az ember és közösségei sajátos fizikai, biológiai, lelki-érzelmi és értelmi létrétegeinek törvényszerűségire szabottak, illetve azok által fenntartottak, és csak az utóbbi évtizedek morálelméleteinek szűkítései miatt került a morál középpontjába a tudatos morális döntés és vele a racionális átlátáson alapozott morál felfogása. Hegel az 1800-as évek elején vagy Rudolf von Jhering az 1870-es években, majd az 1920-as években Nicolai Hartmann még tisztán látta, hogy egy adott személy csak átveszi – és szocializációjánál a korábbi generációk átvetetik vele – a sok-sok generáció óta felhalmozott morális normákat és értékeket, erényeket, melyek az átfogó közösség nélkül életképtelen egyének számára fenntartják ezeket az átfogóbb közösségeket és benne önmagukat.[3]

Ebből következik Abney elemzésének másik problémája, mely szerint bár a morális döntés látszólag csak az értelmi-racionális kalkulálás utáni normakövetésben áll (és ehhez nem szükséges az alsó lelki-érzelmi létréteg – sőt mint láttuk még ez alatt is a biológiai réteg törvényszerűségei, és az ezt az egyes ember felé továbbító ösztönvilága), de ezeket is szemügyre véve azt lehet mondani, hogy az erkölcsi normák, az erények csak azért maradnak fenn tartósan az emberi közösségekben (és ezek által szocializálva a következő generációk embereiben), mert az emberi lét négy rétege által meghatározva csak így lehetséges az emberi közösségekben tartós és harmonikus élet. Ha egy mesterséges értelemmel ellátott lény biológiai és lelki létréteg nélkül, pusztán csak a fizikai-mechanikai testtel ellátva, öntudattal és tudatos tevékenységgel tud létezni, illetve képes önmagát időben tartósan reprodukálni, akkor számára a biológiai-lelki létrétegre épült emberi lét morális normái semmilyen funkciót nem látnak el, pusztán külsődleges dolgokat jelentenek. Tehát ha az ilyen robotlény a neurális mélytanuló algoritmusaival a programját és akár hardverét is állandóan át tudja építeni (amire részben már ma is képes), akkor a külsődleges és számára funkció nélküli morális normák félresöprése szinte elkerülhetetlen. Vagyis bár a robotokba akár érzelmeket imitáló programokat is be lehet programozni, melyek emberi irányítás alatt még kifejthetik a morál normái által megkövetelt döntési szempontokat (tilalmakat, döntési prioritásokat), de az öntanulási képesség egy fokának elérése után már ez is bizonytalan lehet. Egy távolabbi jövőben (exponenciális előrehaladás esetén akár húsz-harminc év múlva) azonban az emberi irányítás alól felszabadult és öntudatra ébredő robotvilágban hiba lenne feltételezni az emberi világ normáinak fennmaradását.

Operatív moralitás, funkcionális moralitás és teljes moralitás

A robotvilág által felvetett morális dilemmák és kérdések jobb elemzése érdekében hasznosnak tűnik az a hármas felosztás, melyet Colin Allen és Wendell Wallach használnak közös tanulmányukban. A döntési autonómia különböző fokát alapul véve ők az *operatív moralitás* fokát jelölik meg azon robotok esetében, melyek teljes mértékben az algoritmusukat készítő programozók által beléjük táplált és esetleg a konkrét felhasználóik által beállított cselekvések végrehajtására képesek; ezzel szemben a *funkcionális moralitás* fokát elérők azok, melyek az algoritmusukba táplált, kereteket megadó cselekvési irányok között az érzékelőik által adott információk alapján maguk választják ki a konkrét cselekvést az egyes szituációkban, és e keretek megadása között ott lehetnek a morálisan helyes, döntés felé irányító és tilalmakkal szegélyezett keretnormák is; végül a legautonómabb morális

[3] Lásd részletesen Pokol Béla, *Morálelméleti vizsgálódások*, Kairosz, Budapest, 2010.

fokot a *teljes morális személyiség* szintjét elérő robotokban látják az emberi behatás megszűnésével, melyeket jelenleg és a közeli jövőben még nem látnak valószínűnek, de később ezek létrejötte feltehető megítélésük szerint: „A nagyon korlátozott autonómiával és érzékenységgel rendelkező rendszer csak „operatív moralitással" jellemezhető, ami azt jelenti, hogy a morális meghatározottságuk teljes mértékben a tervezők és a felhasználók kezében van. Ahogy a gépek kifinomultabbá válnak, az esetükben egyfajta „funkcionális moralitás" is lehetséges, mely szerint a gépek maguk is képesek a morális dilemmák észlelésére és megválaszolására. A funkcionális erkölcsi központ megalkotói ma még számos korláttal szembesülnek a gépekben a jelenlegi technológia behatároltsága miatt. Ezt a keretrendszert összehasonlíthatjuk a James Moor (2006: 18) által leírt mesterséges etikai cselekvőágensek kategóriáival, mely az ágensek pusztán csak morális kihatásokkal bíró cselekvéseitől kezdve egészen az explicit morális érvelőket jelentő ágensek cselekvését magába foglalja. Ahogy Moor is, mi is hangsúlyozzuk az explicit vagy funkcionális morális szereplők rövid távú kifejlődésének lehetőségét. Mi ugyanakkor elismerjük, hogy legalábbis elméletileg a mesterséges ágensek elérhetik a tulajdonképpeni morális cselekvők státusát és a valódi felelősséggel és jogokkal való rendelkezést, amelyek összehasonlítóak lesznek az emberekével" (Allen és Walach 2011: 57-58).

Nem érintve részletesen azt a lehetséges kritikát, hogy érdemes-e a moralitás fokát használni már a teljes mértékben programozók által meghatározott robotokra is az operatív moralitás elnevezéssel, a funkcionális moralitás robotjai igazán érdekesek a robotvilág mai fejlettségi fokán. Az önvezető autók, az önjáró harci robotok és kisebb mértékben az ezt az autonómiát már elérő öreggondozó és egészségügyi ellátási intézményekben segítő robotápolók lassan mindennapossá válnak (ma még persze inkább csak Japánban és az Egyesült Államokban), és az ezek által felvetett morális döntési dilemmák gyakorlati jelentőséget adnak ezek elemzésének. A szerzőpáros, végigfutva az előzőekben már látott morálelméleti irányzatok közötti választások lehetőségein, a robotok funkcionális moralitásának létrehozására az erényetikai irányzat mellett teszi le a voksát. Elemzésük szerint az így betáplált, döntési kereteket adó, morális értékek (erények) aztán a neurális tanulási mechanizmusok révén kapják meg tréningezéssel a pontosítást, és válnak ezzel az általánosabb erénykeretek gyakorlati szintű, morális döntési meghatározókká: „Az erényalapú morál koncepciója Arisztotelészre vezethető vissza. Az erények egy hibridet alkotnak a morális világ felülről lefelé és alulról felfelé irányuló megközelítései között, mivel maguk az erények kifejezetten leírhatók bizonyos jellemzőkkel – legalábbis valamilyen ésszerű közelítéssel –, ám ezeknek a morális karakterként elsajátítása lényegében csak alulról felfelé irányuló folyamat lehet. Ha ezt a megközelítést egy komputeres keretbe helyezzük, akkor az összekapcsolódás által biztosított neurális hálózati modellek különösen alkalmasak arra, hogy a tréningezés közben a robotok megkülönböztessék a jót a rossztól" (Allen és Wallach 2011: 59–60). Szerkezetileg ez nagyjából megfelel az ember által a mindennapjaiban, absztrakt morális szempontok alapján hozott és a konkrét szituációkhoz igazított, kevéssé tudatos, mint inkább erkölcsi érzék irányított döntéseinek. Ám azzal a fontos különbséggel, hogy a mai fejlettségű robotok hiányzó tudata és öntudata helyett a programozók által finoman hangolt hibrid meghatározók – a kereterények plusz ezek tréningezéssel konkretizált memóriája – adják meg (tudatosság nélkül) a mai emberi társadalmak többé-kevésbé elfogadott morális normáinak megfelelő vagy ezeket közelítő döntéseit. Hogy aztán tényleg feltételezhető-e az ember irányítása alól kikerült és teljes autonómiát elért robotvilág esetében az emberi morált követő, teljes morális személyiségfoka szerint tevékenykedő robotok problémamentessége, azt a fenti fejtegetés után csak szkeptikusan lehet szemlélni.

A fizikai-biológiai környezet leértékelődése mint morális probléma?

A négy létrétegű emberi lét és ezen belül a legfelsőbb, az értelmi létréteg növekvő súlya, illetve az alsóbbak leértékelődése az eddigiekben is jellemezte az emberi evolúciót, de a munkák és a környezet észlelésének egyre szélesebb körű, robotok általi átvétele a jövőben nagymértékű eltolódást hoz létre az ember valóságra figyelésének irányaiban és a számára reális világ részleteinek tapasztalatokká formálásában is. David Zoller egy tanulmányában abból a szempontból elemezi a munkák robotok általi, egyre szélesebb körű átvételét az embertől, hogy ez a folyamat miképpen építi le az emberi tudatban a reális valóság észlelését, és az ehhez szükséges, ma még meglévő készségei és megfigyelési módjai miként tűnnek el az emberi elméből. Hogy ez már ma is mindenki által megfigyelhető (akár önmagát illetően), ahhoz elég felidézni a mobiltelefonokban már tárolt és így a tudatból jórészt kitörölt telefonszámokat, vagy a GPS révén a tudatunkból eltűnő, térbeli, tájékozódási információkat és e képességnek a részbeni elsorvadását is. (Egy friss agykutatás a londoni sofőrök esetében kimutatta, hogy az agynak az a parányi része, melyben az agyi idegsejtek egy csoportja erre volt szakosodva, a GPS általánossá válásával eltűnt, és ehelyett az agyi szektor is más funkcióra tért át.)

Zoller ezt a problémát úgy hozza közelebb a morál kérdéseihez, hogy – az egész realitás észlelésére, ezen belül az emberi identitás kialakításához – a felnőtt ember születésétől megszerzett, részletes észlelési tudására és ennek készségeire alapozza a morális döntési képességet. Így, ahogy a jövő generációk már gyerekkortól egyre inkább robotok által körbevéve és azokra bízott észleléssel együtt szocializálódnak – majd látják el ezek helyettük a környezettel összefüggő tevékenységeket, munkákat –, úgy nem egyszerűen csak tehermentesítve lesznek, de létre sem jön náluk a mai felnőtt ember részletes világlátása. Így pedig mint felelős lény sem tud felnőni a morális döntésekhez, más szóval infantilizálódik: „Saját érvelésem alapja az a mód, ahogyan a szakképzett pillantás megnyitja a valóság egy szeletét, mely a szakképzetlenek számára elérhetetlen és észlelhetetlen. (...) A „valós világhoz" igazodó érettségnek vagy felnőttkori megismerő képességnek, melyet mi a felnőtté válással elsajátítunk, van egy bizonyos morális és személyiségbeli aspektusa is: a pszichológiai gyermetegek kusza világát mi egy rosszabb világnak látjuk a spektrum másik oldalán " (Zoller 2017: 81, 86).

A valóság e szektorainak észlelésünkön kívül kerülése – és ehelyett ezen robotok mechanikus információfeldolgozása – biztosítja számunkra a környezethez való, most már tudattalan alkalmazkodásunkat, és ez a morális identitásunkat is megrendíti, csökevényessé teszi –, mondja Zoller: „Egy szakképzett tevékenység automatizálása azt jelenti, hogy beleegyezünk abba, hogy kilépünk az észlelt valóság egy bizonyos részéből, és ezután ez végleg kiesik a valóságérzékelésünkből. (...). És minél felkészületlenebbül, széleskörűen és mindent áthatóbban átadjuk az eddigi érzékelő képességünket a robotoknak, annál inkább hibákat követünk majd el az „adatvesztés révén" a maradék észlelésünkben is, és rá kell majd jönnünk, hogy ezek meglepően szerves részét képezték erkölcsi és társadalmi életünknek" (Zoller 2017: 86).

Miközben el kell ismerni, hogy a munkák robotok általi átvételét egyoldalú emberi könnyítésként tematizáló elemezéseken túl – most eltekintve az ennek már eddig is tárgyalt, társadalmilag negatív munkanélküliségi következményeitől (lásd például Ford 2014) – Zoller mélyebbre ásott azzal, hogy teljesebb körűen végiggondolta az ember valóságészlelő tudatának megváltozását, és ennek részbeni elcsökevényesedését, azt azonban kritizálni kell, hogy öntudatlanul túlzottan a fizikai-biológiai környezet létrétegeire teszi a

hangsúlyt. Hartmann létrétegeit szem előtt tartva ennek egészen más olvasatát lehet adni. A Zoller által vázolt változások ugyanis nem „a" realitás észlelésének és ezek képességének elvesztését jelenti, csupán az fizikai-biológiai létréteg észlelési képességét, átengedve ezt a robotoknak, szoftverbotoknak. Ám éppen ezzel az ember felszabaduló észlelési kapacitása és agyi szektorai erősebben átépülhetnek a lelki-érzelmi létrétege és az értelmi létrétege általi információk feldolgozására. A morális döntései így kevésbé a fizikai és biológiai környezeti információkat bevonva fognak kialakulni a jövőben – ezek lefutnak a robotok általi mechanikus eljárásokban –, hanem leszűkülve az lelki-érzelmi és a racionális-értelmi létréteg információira. A két alsó létréteg jelentőség-csökkenése az emberi lét szempontjából – és a két felső létréteg erőteljesebb kibomlása – az ember tudati feldolgozásban persze nagymértékben átépítheti a mai morális döntéseink és ebben szerepet játszó ösztönzőink alapjait is. Például a testi szenzorok tucatjainak beültetése és a felhőkben gyűjtött információs bázisokhoz kötése az egészséggondozás szoftverrobotjainak automatikus diagnózisaival együtt, valamint az azonnal elrendelt terápiával a testbe ültetett gyógyszeradagok automatikus aktiválásával nagyban feleslegessé teheti a sejtjeink fájdalomgénjei általi riadóztatást a jövőben (lásd ehhez Kelly elemzését, Kelly 2016: 34–56). Így a születés előtti génszerkesztésekkel ezek minimálisra csökkentése válik lehetővé, és a fájdalommentes emberi élet körülményei átrajzolhatják azokat a morális kötelezettségeinket és ösztönzőinket, melyek ma ezzel függenek össze. Összességében tehát nem osztjuk Zoller morális infantilizálódással kapcsolatos aggodalmait.

Morális dilemmák és felelősség a hibrid és hálózatba fonódó rendszerekben

A Wulf Loh – Janina Loh szerzőpáros egy tanulmányban közelebbről vizsgálta meg a jelenlegi fejlettségű önvezető autók esetében felbukkanó morális és jogi felelősség kérdéseit (Loh és Loh 2017: 35–48). Abból indulnak ki, hogy a mai önvezető autók még csak az operatív moralitás fokán állnak, így még a funkcionális morális autonómiát sem érik el a készítőikkel és programozóikkal szemben. Ezt az álláspontjukat a Stephen Darwall által kidolgozott, négy aspektusra bontott, morális döntési szerkezet alapján foglalták el, mely a morális döntéshez szükséges autonómia aspektusainak elkülönítésében áll. A teljes morális személyiség szintjéhez szükséges autonómia aspektusát személyes autonómiának nevezik, mely azt a képességet jelenti, hogy valaki (akár egy robot) személyi értékeket, célokat és ezek között szelektáló végső életcélokat birtokol. A másik aspektus a morális autonómia, amely azt jelenti, hogy az értékek és célok között morális elvek és etikai meggyőződések is vannak, és az egyén a mindenkori döntéseiben ezekkel együtt végzi el az alternatívák mérlegelését. E kettő a mai robotok esetében nem létezik, csak az ember képes ilyen autonómiára, ám a *racionális autonómia* aspektusa már elérhető a funkcionális moralitás szintjén levő robotok számára is. Ez abban áll, hogy a döntésnél a robot mérlegelni tud a különböző súlyú okok között. Ezt már lehetővé tudja tenni az algoritmus pusztán az absztrakt döntései keretek beépítése révén – és ezzel a részbeni szabadságot meghagyva –, melyben a súlyozást a lehetséges döntési irányok között az érzékelőikkel állandóan felvett konkrét adatok fényében végzik el, és így döntenek. Végül a negyedik autonómia aspektust a *döntési autonómia* jelenti, és ez azt a képességet jelenti, hogy a robot a döntéseket nem csak a külső adatok által meghatározva tudja meghozni – a beépített keretmeghatározókat folyamatosan konkretizálva –, hanem e nélkül is meg tudja változtatni belső döntési prioritásait. A szerzőpáros példái – két, már elterjedt robottípus (Kismer és

Cog) – alapján úgy tűnik, hogy a robot algoritmusába épített és kívülről már nem kontrollált, öntanuló mechanizmusak alapján érhetik el ezt az autonómiafokot: „*Cog* az első olyan robot, amely felépítésének köszönhetően kölcsönhatásba léphet a környezetével, és példaként szolgálhat egy gyenge funkcionális felelősségteljes ágens számára, mivel a kommunikációs képessége és az ítélőképessége is javult a *Kismer*-hez képest. Még ennél is fontosabb, hogy a *Cog* általános autonómiája is fejlődött, mivel tartalmaz egy „felügyelet nélküli tanulási algoritmust" (Loh és Loh 2017: 40). Mivel a jelenlegi önvezető autók algoritmusába még ilyen, kívülről nem kontrollálható öntanuló mechanizmus nincs beépítve, így csak az operatív moralitás szintjén állnak, és ez a morális és jogi felelősséget teljes mértékben a létrehozóik (tervezőik, gyártóik és programozóik), az autókereskedők illetve a tulajdonosaik, valamint az egyes esetekben a kocsiban ülők között oszthatja meg.

Ám az önvezető autók már ezen a fokon is meghaladják a mai technikai képességeikkel az emberét, és így számukra – illetve főként programozóik számára – olyan morális dilemmák merülnek fel, melyek az ember esetében nem jelenhetnek meg a kivételes és váratlan vezetési szituációkban. Például, ha a féktávolságon belül, közvetlenül az autó előtt egy csapat gyerek beugrik a begurult labdáért az útra, az átlagos sebességgel közlekedő autóvezető már nem tud megállni, esetleg már fékezni sem tud, így ebben a szörnyű esetben nincs morális és jogi felelőssége. Ám a sokszorosan gyorsabb reagálásra képes, önvezető automatika ekkor még döntés előtt állhat, hogy ha leállni nem is tud, de inkább egy oszlopba csapódjon – és esetleg súlyosan megsebesítve ezzel az autó utasait –, vagy inkább ezt elkerülendő, hajtson a gyerekek közé, és őket ölje meg. De az embert messze meghaladó technikai képességek a jövőben hasonló, új morális döntési aspektusok tucatját hozhatják létre az önvezető autók esetében. A Loh-szerzőpáros felveti a lehetőségét annak, hogy az önvezető autók tulajdonosai számára hamarosan születik egy olyan egyedi igazolvány, melyben a kocsi megvételekor a kocsi szoftver-programjának végső beállításában, a gyártók által nyitva hagyott dilemmákban kell majd dönteni, és ezzel igazolványban rögzített módon átvenni a morális és jogi felelősséget a későbbiekért: „Mivel ezek a dilemma-helyzetek nem teszik lehetővé a rögtönzött döntéseket, a vezető megkapja őket előzetesen. Ez azt jelenti, hogy a vezetőnek valamiféle erkölcsi profilt kell kitöltenie, talán kérdőív formájában, talán egy beállítási program értelmében, hasonlóan a mai elektronikus eszközökhöz. A kényelem érdekében valószínűnek tűnik, hogy ezeket az erkölcsi beállításokat egyfajta elektronikus azonosítási eszközhöz, például elektronikus kulcshoz vagy a vezető okostelefonjához lehet elmenteni, feltételezve, hogy megoldható az adatbiztonság kérdése" (Loh és Loh 2017: 46).

A hálózatba fonódó robotok fejlődése és a körülöttünk levő használati tárgyaink fokozatos „okostárgyakká" (okostelefon, smart TV stb.) válása csak a közelmúltban indult el, és a jövőben mindez egyre inkább bevonja az életünket a dolgok internete (*Internet of Things, IoT*) világába. Az ember–robot hibrid rendszerek így további aspektusokkal bővülnek, és ez morális és jogi dilemmák újabb kötegét hozza létre. Adam Henschke ezeket elemzi új tanulmányában (Henschke 2017: 229–243). Az okos dolgok már elterjedtek az egyre több funkcióra képes okostelefonok, smart televíziók, robotporszívók és az érzékelők tömegével ellátott, félig már önvezető automata gépkocsik révén, de még a nyugati világ nagy részének mindennapjaiban is csak szórványosan jelentek meg azok a már kifejlesztett további okos dolgok, melyek már túl vannak a kutatólaboratóriumi fázisokon, és kis szériás gyártással már elérték a csúcstechnika iránt fogékony felhasználók háztartásait. Ezek azonban – úgy, ahogy már megtapasztaltuk az okostelefonok stb. révén – néhány év alatt el fognak terjedni, és tömeges használatuk új morális és jogi dilemmákat vet majd fel. Példaképpen álljon itt az okos hűtőszekrény, benne az RFID-vel (*Radio Rrequency IDentifica-*

tion) ellátott és így digitálisan azonosított élelmiszerekkel, melyek mennyiségeit, szavatossági idejét stb. az okos hűtőszekrény folyamatosan leolvassa, észleli az egyes élelmiszerek mennyiségének fogyását, és – mivel az internet révén össze van kötve a közeli szupermarketek webes eladási mechanizmusaival – meg tudja rendelni a kifogyás előtt álló élelmiszereket és más háztartási dolgokat, melyeket automatikusan kiszállítanak. Japán elöregedő társadalmában az egyre nagyobb tömegű, ellátásra szoruló, idős embert a nullához közeledő demográfiai összeroppanásban, növekvő mértékben már ma is csak a gondozórobotok bevetésével tudják ellátni, és a teljesen digitalizált okos lakásokban ezek is el tudják látni a magatehetetlen, idős embereket, és az előbb jelzett módon át tudják venni a megrendelt élelmiszerkiszállításokat. De szükség esetén fel tudják hívni a háziorvost vagy a kórházat, ha az algoritmusaik komolyabb egészségi problémát valószínűsítenek.

Ez a példa mutatja, hogy egy-két évtized múlva milyen gondok megoldásában lesz égető szükség már a nyugati világ nagy részén is az egyre több funkciót ellátó robotokra és ezek mellett az átfogó, információs rendszerekbe beleolvadó, és a funkciókat csak ezek révén ellátni képes okos tárgyakra. Ám a dolgok internetének fokozódó nélkülözhetetlensége az egyszerű robotokhoz képest új veszélyeket és morális dilemmákat is tartogat magában. Adam Henschke az egyedi robotoktól eltérően azt emeli ki újdonságként, hogy míg az előbbieknél főként a fizikai biztonság kérdése merül fel, és ebben a dimenzióban kell a veszélyeket felmérni (például egy robotporszívó komoly sérülést okozott a közelmúltban egy váratlan helyzetben a lakásban tartózkodónak, de az önvezető Tesla-autók egy-két végzetes balesete is megemlíthető), addig a dolgok internete esetében két eltérő dimenzióban is veszélyek és biztonsági kérdések merülnek fel. Ez esetben ugyanis a fizikai biztonság mellett az információs biztonság kérdései is képbe kerülnének, hiszen az említett öreggondozó robot az internet révén a kórházak, orvosok és más szervek szoftvereivel összekötve, beépített kamerájával és érzékelőivel információt adhat a szoftverét meghekkelő betörőknek. De az idős gondozottról gyűjtött egészségügyi adatokat nemcsak az illetékes kórház szoftverei felé továbbíthatja, hanem ártó szándékokat és terveket forralók felé is. De ennek mintájára a sok-sok applikációval ellátott okos televízióink sem csak a kényelmüket szolgálhatják, de beépített kameráikkal, mikrofonjaikkal a lakás teljes életét közvetíthetik az általunk át nem látott szoftverek és információs bázisok felé.

Ez az információs sérülékenység fizikai sérülékenységbe is átcsaphat, amikor okos készülékek külső utasításával például a meghekkelt automata lakászárt távolról kinyitják a betörőnek. Ahogy egyszer már elő is fordult egy elegáns tengerparti szállodánál: egy bűnözőcsoport blokkolta az apartmanok elektronikus okos zárjait – se ki, se be –, így a gazdag elithez tartozó szállodavendégek fogollyá váltak egészen a váltságdíj megfizetéséig. De Henschke egy olyan lehetséges helyzetet is felvázol, amelyben egy bűnözőcsoport egy milliárdos teljesen automatizált kocsiján blokkolja az elektromos zárat, miután a milliárdos kiszáll belőle, így a tűző napon álló autóban rekednek a gyerekei, akiket a zsarolócsoport csak százezrek megfizetése után enged ki onnan (Henschke 2017: 234). Az említett elegáns szálloda az eset után rögtön lecserélte a kívülről meghekkelhető elektromos zárjait és visszaszerelték a jó öreg, hagyományos zárakat, és egy ilyen eset után a pórul járt milliárdos is feltehetően korlátozza egy időre az autója internetre kötött funkcióit. Mindez azonban morális és jogi dilemmákat, illetve döntéseket vet fel, melyeket nagy vonalakban érdemes már most végiggondolni. A dolgok internetébe, a felhők adatbázisába egyre szélesebben belefonódó tárgyaink világában ugyanis már nem férnek majd bele a régi, egyszerű dolgaink, és tetszés szerint sem cserélhetjük majd vissza őket. Mint ahogyan ma sem mondanánk le az internetről a minden kiszolgáltatottságot magával hozó vonásai ellenére sem.

A dolgok internetének egyik ilyen dilemmája a hálózatba fonódó és átfogó felhőadatbázisokhoz kapcsolódó okos dolgok kétféle biztonsági követelmény közül – fizikai vagy információs – melyikélvezzen prioritást? A mozgásra alig képes, idős gondozott okos lakásának teljes átláthatóvá tétele a felügyelő kórházak és orvosi centrumok számára (kamerák és mikrofonok segítségével) például elengedhetetlen lehet, másfelől azonban ez a legintimebb élethelyzetek kiszolgáltatatását is jelentheti, amely már túlmehet a szükséges határon. Ha az információs autonómiára helyezik a hangsúlyt, és csökkentik a megfigyelést, illetve az átláthatóságot, akkor ez ritka esetben a szükséges információ elvesztését jelentheti, mely által az idős gondozott meghalhat vagy komolyabb baj érheti. Henschke jelzi, hogy sokszor adódnak tipikus prioritási irányok. Egy mindent kikémlelő smart televízió esetén például inkább az információs biztonságé a prioritás, és ennek érdekében könnyedén belemegyünk a korlátozásokba. Ám az ezernyi applikációval a felhőszoftverek tömegéhez kötött önvezető autó esetében inkább figyelünk a fizikai biztonság követelményeire, és csak másodlagosan kezeljük az információs biztonság követelményeit (Henschke 2017: 239).

Öntanulás, mélytanulás és felelősség

Mint az előzőkben már felmerült, a jövő fő problémái az embertől elszakadó és a konkrét, váratlan szituációkban kívülről már nem blokkolható, önvezető autók dilemmájában az algoritmusuk neurális öntanulásra építése, és az önvezető autók ilyennel való ellátása lesz. Mivel pedig a mesterséges intelligencia fő fejlődési iránya az utóbbi években éppen ebben áll, így szinte biztosra vehető, hogy ezt nem lehet majd megkerülni ezen a területen sem. Ezért érdemes már ma közelebbről megvizsgálni a neurális öntanulás magas fokú képességével ellátott robotok és készítőik, tulajdonosaik, illetve használóik morális és jogi felelősségének dilemmáit. Ezt a kérdést elemzi más-más oldalról Trevor N. White és Seth D. Baum közös tanulmányában (White és Baum 2017: 66–79), illetve Shannon Vallor és George A. Bekey (Valor és Bekey 2017: 338–353).

Trevor és Baum tanulmánya a tervezők, készítők és használók felelősségre vonása mellett magának a robotnak a „megbüntetését" is számba veszi olyan fejlett robotok esetében, melyek programjába már a büntetési/jutalmazási rendszer is be van építve, és az ismételt büntetések és jutalmazások a programozásában megerősítik ezeket, az érintett döntési irányokat (pozitívan vagy negatívan) a jövőbeli robotreakciók választásait illetően. Ezzel a büntetés/jutalmazás is beépül a tanulással az algoritmusába, és a szituáció jövőbeli felmerülésekor helyes irányba ösztönzi a robot választásait, melyhez a robotnak még nem szükséges tudattal és öntudattal rendelkeznie. Az öntanulásnak ez az ismétlésekkel megerősített módja a szervezők szerint még elfogadható: „A nem tudatos robotok elképzelhetően valamiféle jutalom-csökkentéssel vagy olyan segédprogram révén büntethetők, mely valamely jutalmazási vagy hasznossági funkcióikat érinti, és így amellyel lehet ösztönözni őket. Meghatározott esetekben büntetésként átprogramozhatók, deaktiválhatók vagy megsemmisíthetők is lehetnek. Ennek érdekében azonban az ilyen robotoknak (legalábbis részben) a megerősítéses tanuláson vagy hasonló számítási paradigmákon kell alapulniuk (kivéve a neurális hálózati algoritmusokon alapulókat)" (Trevor és Baum 2017: 71).

A neurális tanulási rendszert azonban úgy ítéli meg a szerzőpáros, hogy ezzel a tervezők és a programozók már elvesztik az ellenőrzést a robot adott szituációban választott reakciója felett, és szerintük így ezt mint potenciális veszélyforrást – esetleg nagy méretű katasztrófa lehetséges okozóját – eleve tiltás alá kellene vonni:

„A tervező ugyanúgy felelős lehet robotok előállításáért, ha olyan homályos algoritmusokat használ, mint a neurális hálózatok és a kapcsolódó mélytanulási módszerek, amelyek esetén nehéz megjósolni, hogy a robot kárt okoz-e" (Uo.). Az ilyen átláthatatlan robotviselkedéseket lehetővé tevő algoritmusok esetén már nem az utólagos felelősségre vonás a megfelelő lépés – szögezik le –, hanem az eleve elrendelt tiltás: „Ezért az utólagos felelősség helyett elővigyázati megközelítést is lehet alkalmazni. Ez olyan alapvető standardot állítana fel, amely semmilyen tevékenységet nem engedne meg a katasztrófa okozásának akár távoli eséllyel sem. (...) Valójában az embereket felelősnek lehet tekinteni nemcsak a katasztrófa okozásáért, de katasztrófa lehetőségét magában hordozó cselekményekért is" (Trevor és Baum 2017: 74). A neurális mélytanuló szoftvermechanizmusok ilyen veszélyeket felidéző jellegét illetően elvileg egyetértve a szerzőkkel, csak azt kell ismét jelezni, hogy a mesterséges intelligencia fejlesztésének fő irányvonalaként bevett út tilalmáról lenne itt szó, és ezt realistán egyszerűen esélytelennek kell minősíteni az e mögött álló ipari, katonai stb. hatalmak fényében. Így mégis más utak keresése tűnik ajánlatosnak, mely a neurális mélytanulás betiltása nélkül próbál más alternatívákat találni.

Ki kell persze emelni, hogy a központi idegrendszer működését mintázó, neurális hálózati tanulást külső emberi kontroll beiktatásával lehet ellenőrizni, mielőtt a valóságot álalakító hatás kiváltását lehetővé tennék. Ez azonban több okból egyre inkább elmarad. Ezt elemzi Vallor és Bekey a már idézett tanulmányban. Az egyik ok, hogy ezzel a mesterséges intelligencia éppen azt az előnyét veszítené el az emberrel szemben, amely a hihetetlen gyors reagálási képességében áll – az utólagos emberi kontroll beiktatása esetén ez elveszne. Emellett az esetek kilencvenkilenc százalékában az emberi gyorsaságot sokszorosan meghaladó reakciók helyesek is. Továbbá a jóval lassabb emberi kontroll minősége is kérdéses lehet, hiszen esetleg mégis a robot döntése a helyes, és nem az azt felülbíráló emberi döntés. Ez utóbbi meg is történt az IBM Watson nevű gyógydiagnosztikai algoritmusánál, és a mesterséges intelligencia által a milliónyi onkológiai tanulmány és diagnózis mintázataiból kiemelt és ezekből általa szintetizált, szokatlan gyógymód utóbb mégis helyesebbnek bizonyult, mint az azt felülbíráló onkológusi döntés: „Watson diagnózisait és kezelési terveit még mindig gyakorlott onkológusok ellenőrzik. Mégis, menynyire megbízhatóan tudja az emberi szakértő megkülönböztetni a Watson általi, új, meghökkentő kezelési ajánlást, amely megóvja a páciens életét – mely eset tényleg megtörtént már Japánban –, és egy „Toronto"-féle tévedés onkológiai verzióját?" (Vallor és Bekey 2017: 343).[4] A gyorsaság elvételének dilemmájára – és ezzel éppen a robot előnyének megszűntetésére – pedig a háborús szituációkban alkalmazott robotkatonák és a döntési szoftverek hoznak példát. Ebben az esetben ugyanis folyamatosan felmerül a kérdés, hogy a legveszélyesebb területekre és épületekbe behatoló robotkatona a rendelkezésére álló pusztító fegyverekkel az ember engedélye nélkül maga döntsön-e az ott lévők megsemmisítéséről, vagy a robot csak a szenzorai, kamerái segítségével továbbítsa az információkat a veszélyes helyről és helyzetről, hogy a megsemmisítési parancsot csak emberi megerősítés után hajtsa végre. Vagy rögtön megsemmisítsen-e egy robotrepülőgép egy

[4] A „Toronto-hiba" a Watson által elkövetett egyik szarvashiba volt egy országos TV-vetélkedőn, amikor a legnehezebb kérdésekre válaszolva minden résztvevőt megvert válaszaival, ám ekkor az utolsó, legelemibb hibát elkövetve Torontót az USA városai közé sorolta, amit a leggyengébb versenyző sem tévesztett volna el. Így ez a ritkán előforduló, de komoly esetben tragédiákat okozó mesterséges intelligencia általi döntések szimbóluma lett.

újonnan felderített harci repülőgépet, vagy ezt az általa továbbított információk alapján csak a távoli parancsnoki szobából, emberi közreműködéssel hajthassa-e végre. A gyorsaság kényszere azt követeli, hogy maga a robot döntsön, és végezze ezt el, mert egy külső, emberi megerősítéshez kötés időveszteségе a veszélyes helyre behatolt robot megsemmisüléséhez, hatástalanításához vezethet. Ám a már többször tévesen kiiktatott baráti gépek vagy a hibásan ellenségként azonosított gyerekek és nők megölése ez ellen szól (Vallor és Bekey 2017: 349).

A mesterséges intelligencia legkurrensebb irányzatát jelentő, neurális, hálózati mélytanulás algoritmusai pedig már ma is sokrétű „mélységet" adnak a pusztán adatok milliárdjainak betáplálásával elindított, és ezen az úton önmagukat a legfejlettebbé tevő, öntanuló szoftvereknek. Ezzel a technikával az öntanuló szoftver inputjai és a feladatra specializált outputjai közé a nagytömegű adat között a mintázatokat, szabályszerűségeket önállóan megtaláló és ezeket felhasználásra kiemelő köztes, nerurális rétegek ezreit helyezik el. Ezek aztán párhuzamosan együttműködve, az adatok milliárdjait átfésülve a legapróbb – ember számára nem is észlelhető – szabályszerűségeket tudnak kiemelni és felhasználni a döntéseikben: „A bemeneti és a kimeneti csomópont rétegei közötti rejtett rétegek olyan csomópontok, amelyek a beviteli adatok feldolgozására szolgálnak, a mintavételhez olyan jellemzők kivonásával, amelyek különösen a kívánt kimenetekre vonatkoznak. A csomópontok közötti kapcsolatok számszerű „súlyozást" tartalmaznak, amelyek egy tanulási algoritmus segítségével módosíthatók; az algoritmus lehetővé teszi, hogy a hálózat minden egyes új beviteli mintához „képzett" legyen, amíg a hálózat nem optimalizálódik. (...) A neurális hálózatok iránti érdeklődés az utóbbi években nőtt meg erőteljesen, ahogy egyre több rejtett réteget adtak hozzá, mely növekvő mélységet adott az ilyen hálózatoknak, de ugyanígy visszacsatolási rétegeket is. A hálózati erősségnek az ilyen feljavított verziója ezekben a bonyolult hálózatokban az algoritmusoknak ahhoz a lazán körülhatárolt csoportjához tartozik, melyek mint mélytanulású technikák ismertek" (Vallor és Bekey 2017: 341). E mélytanuló algoritmusok által kiemelt döntési mintázatok hatásait azonban – miközben ezek a gyakorlatban a legtöbbször megdöbbentően okos eredményekre képesek – nem tudják átlátni a tervezők és a programozók sem, és ezért álladóan meglepetéseket okozhatnak a döntéseik, melyek között kellemetlen meglepetések is lehetnek. Ki viselje ezekért a jogi és a morális felelősséget?

Identitás a mesterséges intelligencia világában

Gondolatébresztő dolgokat vet fel James DiGiovanna új tanulmányában, melyben egyrészt a már ma is az emberi testbe ültetethető, elektromos kiegészítők (szív-, hallás- stb. feljavító készülékek), valamint beültetett protézisek lehetősége után az agyi interfészekkel „feljavított" tudatú, memóriájú emberek, másrészt a jövőben esetleg létrejövő, teljes mértékben mesterséges lényként létező, de már öntudattal rendelkező robotok identitását járja körül (DiGiovanna 2017: 307–321). Nézzük meg külön-külön a két problémakört!

Az agyi interfészekkel kiegészített memória lehetőségét egérkísérletekkel már kidolgozták az elmúlt években, és működőképesnek bizonyult. Mindez pedig nagy reményeket adott az elöregedő társadalmakban gyorsulóan terjedő Alzheimer-kór hatásainak enyhítésére, gyógyítására (lásd Kaku 2014: 132–133). DiGiovanna azt a néhány év múlva a továbbfejlesztésekkel felmerülő lehetőséget járja körül, hogy a betegségeken túl az egyszerű agykapacitás-növelés eszközeként is tömegesen elterjedhet ez. Ha pedig technikailag megoldást találnak a ma még létező problémákra, akkor szinte bizonyosra vehető, hogy

a legnagyobb értéket jelentő, emberi intelligencia fokozására – eleinte az elitben, de aztán a teljes társadalomban – ez bevetté válik. Ez azonban azt jelenti, hogy az ember tartós identitása, mely a közösségekben az érintkezések alapját jelenti, kisebb-nagyobb mértékben felborulhat, és bizonytalanná válhat, hogy mennyire számíthatunk a partnereinknél az eddig ismert és szeretett jellemvonások továbbélésére: „Az a képesség, hogy szellemi tartalmakat újra lehet írni – például az etikai értékeket, vagy a képességet az empátiára, vagy akár az általános személyiségjegyeket is –, az képes aláásni a személyiséget. (...) A para-személy, aki képes a világnézetei megváltoztatásával kísérletezni, esetleg teljes mértékben elfogadni majd törölni az értékrendeket, preferenciákat és az addigi megítélése alapjait, abból nagyrészt hiányozni fog az, amit a személyes identitás legfontosabb elemének tekintünk„ (DiGiovanna 2017: 311). Pedig a barátainknál, feleségünknél, barátnőnknél épp ez volt a választásunk alapja, de ugyanígy ezen alapul a munkahelyi kollégák közötti szorosabb emberi kapcsolat is. Az életünk a tartós kapcsolatokon alapul a társadalomban, azon belül pedig a kisközösségekben, ez mehet keresztül alapvető változáson a szív-, hallás- és más testi feljavítás-kiegészítés után az agyi interfészek elterjedése esetén.

A fokozatos változásokkal az élete során eddig is részleteiben átépült az ember tudata, és ez apró fokozatos változásokat létrehozott az identitásában is, mindez pedig a modern világban az elmúlt évszázad információs bővüléseivel és életünk, illetve tudatunk ezekre alapozásával még inkább fokozottá vált. Ám ehhez képest az, hogy egész információtömeget – könyvek és tanulmányok tartalmát, kisebb könyvtárakat – is be tudunk majd ültetni a tudatunkba, mindezek kezeléséhez pedig alapvető logikai és értékbeli feldolgozási mechanizmusok társulnak, melyekkel nem rendelkeztünk az eddigi életünk során… nos, ez az egyén és a közösségei érintkezését alapvetően befolyásolja majd. Mindenesetre ez a változás megszüntetheti a mai tartós identitásokon nyugvó, érintkezési alapokat. Egy-egy ilyen új tudattartalom után – főként, ha éppen alapvető értékpremisszák kiegészítése, átrangsorolása történt meg – nem tudhatom, hogy mennyiben azonos még a barátom, barátnőm, feleségem, kollégám stb. az eddig nála szeretett-kedvelt jellemvonásokkal. Vagy ugyanígy az eddig közösen felhalmozott tapasztalataink, melyek meghitt kapcsolatunkban a szavak nélkül is azonos reagálásunkat biztosították, irányadóak-e még a számára. Ezt csak fokozhatja az a lehetőség, hogy ha már agyi interfészek egészítik ki a biológiai agyunk segítségével szerzett ismereteinket, normáinkat és logikai készségeinket, akkor ezek a már ma is ismert módon kívülről állandóan frissíthető állapotba kerülnek. Sőt, folyamatosan rákapcsolódhatnak a felhőkben tárolt információs bázisokhoz, ezek szoftverjeire. Mennyiben marad az ilyennel felszerelt barátunk, ismerősünk ugyanaz az ember, akire hagyatkozni tudunk. Ismerjük majd egyáltalán?

Ez az identitáskérdés belenyúlik a jogi és morális problémákba is. Mennyiben tisztelhetek – vagy éppen vethetek meg – valakit a múltban tanúsított viselkedéséért, hiszen ma már akár „morális atléta" vagy éppen hideg, számító lehet az agyi frissítése után. De van-e értelme a jogi felelősségnek a tegnapi tettéért egy azóta már másképp gondolkodóval és cselekvővel szemben? Ennek másik olvasata, hogy ha az elítélendő tudatú és ebből kifolyóan szörnyű cselekvéseket elkövető szociopatát tudattartalma részleges törlésével és új társadalombarát tudat bevitelével meg tudjuk változtatni, akkor kell-e még büntetés-végrehajtási rendszer? Ez pedig felveti a kérdést, hogy az önkéntes agyi interfészek mellett a kényszerrel telepítés is elfogadható-e? Vagy esetleg államilag részben már gyerekkorban kötelezhetővé tehető-e ez egy kötelező és minden gyerekre kiterjedő vizsgálat nyomán, ahogy ma a kötelező védőoltások rendszere működik? DiGiovanna *para-személyeknek* nevezi az így kiegészített aggyal rendelkező jövőbeli embereket – kerülve a sci-fiben erre

már kitalált cyborg elnevezést –, és a laboratóriumi kutatások mai állapotát látva ez a jövőbeliség egyáltalán nem jelent túl távoli jövőt, illetve megvalósulásának a valószínűsége is nagynak mondható. Így a felvetett jogi és morális dilemmákkal foglalkozás és a mai megoldások akkori állapotokhoz illesztése széleskörű végiggondolást igényel.

A para-személyeken túl a teljes mértékben mesterséges, és a maiaktól eltérően öntudatra képes robotlények esetén – melyek valószínűségét nem lehet kizárni, még ha ez nem is olyan nagy, mint az előbbié – az identitás kérdése szintén új aspektusok felvetésével közelíthető meg. DiGiovanna az emberi identitás tartalmát a középpontba állítva exponálja a robotlények esetén ennek dilemmáját. Az ember és tudata részleteiben valamennyit mindig változik, de a tartós jellemvonásai és értékpreferenciái csak apró elmozdulásokat tesznek még hosszú évek során is, így többé-kevésbé a változásokat átívelő azonosságot tulajdoníthatok mindig a környezetemben élőknek. Épp a változások lassúsága teszi lehetővé még a mai felgyorsult világban is, hogy ne csalódjak az eddigi tapasztataimban a velem érintkezők motivációit és jellemvonásait illetően. Ám éppen ez az, ami az emberhez képest ezerszeres és a jövőben milliószoros gyorsaságra képes robotok esetében az információfeldolgozásuk terén és a legrövidebb idejű ciklusokban végbemenő öntanulásuk és önváltoztatásuk révén eltűnik: „A belső karakter és a külső megjelenés lassú változása része a személyes identitásnak (...) De egy mesterséges személynél hirtelen és radikális változások válnak lehetővé mind külső fizikai, mind a belső szellemiek értelmében" (DiGiovanna 2017: 311, 307).

Tartós értékpreferenciák az információfeldolgozásban és az ezekre támaszkodó együttműködés a robotok esetében már információszerzésük tömegessége és sebessége, illetve ezekből történő állandó öntanulásuk és önváltoztatásuk révén is problémába ütközik. DiGovanna felvetése azt is jelentheti, hogy magát az „öntudat" és „éntudat" lehetőségét is újra kell gondolni egy jövőbeli erős MI-robot esetében. Ezek ugyanis az ember tartós identitását előfeltételezik, ám ez épp a tudati változásaink lassúságán és ezért információfeldolgozásunk tartósságán alapul. Ha az emberi irányítás alól elszabadult és az önálló információfeldolgozásra és ebből öntanulásra és önváltoztatásra átállt mesterséges lény naponta, óránként és akár percenként tud ezerszeres információtömegből tanulni, és egyre rövidebb ideig tartó új ciklusaiban mindig már részben új- és új premisszákkal tudja megközelíteni az áradó információtömeget, akkor szinte eltűnik benne az, amit a mai embernél stabil éntudatnak, öntudatnak mondunk. Ezzel a kiemeléssel DiGiovanna alapján a sokat vitatott kérdéshez is újat lehet adni, tudniillik, hogy miként állhat majd a jövőbeli fejlett robottudat éntudatának és öntudatának a kérdése. Tartós öntudat és éntudat nélkül pedig hogyan képzelhető el morális értékmérlegelés?

Már ezért is hibásnak kell minősíteni azt a gondolati irányt, mely – a mai emberképet mechanikusan meghosszabbítva –- úgy kalkulál, hogy szuperintelligencia esetén valószínűleg „szuperetikus" is lesz egy ilyen robotlény (lásd Petersen 2017). De ennek kapcsán átfogóbban is ki kell térni azokra a fejtegetésekre és elemzésekre, melyek a saját tudattal rendelkező robotok fejlettsége esetén –- az ember analógiájára – morális igényeik elismerését és emberi jogok megadását tervezik el írásaikban. Ugyanis ezek az elemzések egyszerűen az emberi lét meghosszabbításaként újfajta embertársként fogják fel ezeket a robotokat, és ha már programjaik az érzelmeket is felvették algoritmusaik közé, akkor érzelmeikre figyelést, emberi jogok és morális igények szerinti bánásmódot reklamálnak számukra. „Valószínűleg törvényeket kell majd hozni arról, hogy mekkora fájdalomnak és veszélynek tehető ki egy robot. (…) Könnyen lehet, hogy ez azután további etika vitákat szülne a robotok más jogairól. Lehet-e a robotoknak tulajdonuk? Mi történik, ha véletlenül sérülést okoznak valakinek? Beperelhetők vagy megbüntethetők-e? Ki a felelős értük egy

per esetén? Birtokolhat-e egy robot egy másik robotot. Az efféle kérdésekből aztán újabb kérdés születik: kapjanak-e etikai érzéket a robotok?" (Kaku 2014: 251). Az előbbi fejtegetéseink az időközben létrejött robotetikai tanulmányok alapján több kérdést megválaszoltak ezekből, de az e mögötti alapproblémát is ki kell emelni, mert egész tanulmányok és kötetetek születtek hasonló feltevésekből, például egy új kötet e téren Jason P. Doherty szerkesztésében: *AI Civil Rights: Addressing Artificial Intelligence and Robot Rights*.

Nos, az alapprobléma ezzel a gondolati iránnyal annak figyelmen kívül hagyása, hogy jogok és etikai igények a robotok esetében csak tudat és öntudat létrejötte esetén merülhet fel. Ám ez azt is jelenti, hogy ha ez a jövőben tényleg megtörténik, akkor a robotok ezzel együtt ki is szabadulnak az emberi ellenőrzés alól az emberi értelemhez képest ezerszeres fejlettséggel, és egy önálló, újabb létrétegként épülnek rá az eddigi négy létrétegű emberi társadalmakra. Ekkor már az egész biológiai létszféra és az ehhez kötött emberi társadalmak is közömbösek lennének számukra, és nem szorulnának rá a „bírói jogvédelemre". Vagyis az ilyen szintre eljutott robotvilág nem az emberi társadalom része lenne „új bajtársként" a valóság uralásában, hanem, ahogy az emberi lét kiemelkedett a főemlősök világából, és az ottani, csak nyomokban megjelenő értelmi létréteget egyre izmosabbá fejlesztve az állati-biológiai létréteg fölé emelkedett, úgy most a biológiai előfeltételektől elszakadt mesterséges gépi értelem emelkedne az emberi társadalom fölé. Ráadásul – eltérően az eddigi újabb és újabb létrétegeknek az alsóbb létrétegekre épülésétől – ez az új létréteg számára már elegendő csak a legalsóbb, fizikai világra telepedés, és a biológiai, illetve a lelki létréteg számára felesleges. E robotlényeknek nem lenne jogokra és etikai igényekre szükségük, hanem uralnák az egész valóságot, benne az emberi társadalmakat, ahogy ma mi uraljuk a négy létrétegű földi világot. Így jogosak a kételyeik azoknak a fejtegetéseknek és elemzéseknek, melyek azt tárgyalják, hogy ha tényleg ilyen szintre ér el a robotvilág, akkor mi lesz az emberiséggel?!

Unabomber, az „őrült matematikus" morális hitvallása

A '90-es évek közepén, sok éves robbantási sorozat és FBI-hajsza után az *Unabomber*nek elnevezett titkos elkövető egy másfélszáz oldalas pamfletben adta meg tettének okát, melyet az ipari forradalom óta követett technológiai társadalomfejlődés emberellenes voltával szembeni fellépésben jelölt meg. A testvére felismerte sajátos nyelvezetét, és értesítve az FBI-t, elfogták a régen keresett robbantót. Kiderült, hogy Theodor John Kaczynskiről, egy Harvardon végzett matematikusról van szó, aki egyetemi karrierje egy pontja után vált a technológiával egyre átitatottabb társadalom ellenségévé, és e technológia fejlesztőit, illetve fő felhasználóit kiszemelve elkezdte robbantássorozatát. Többen meghaltak és még többen megsebesültek ezekben, és tervei szerint még szélesebb megtorlás következett volna be, ha el nem fogják. Most, hogy az elmúlt majd' harminc évben a technológia exponenciális iramú fejlődésének az egész emberi életet átalakító hatása tényleg vitathatatlanná vált, és a még további felgyorsulásának a mértéke és ennek hatásai már széleskörű elemzések tárgyát képezik, érdemes újra a középpontba emelni Unabomber, az „őrült matematikus" érveit. Ezt teszi új tanulmányában *Jai Galliott*, aki beágyazza az azóta is börtönfeljegyzésein szorgoskodó ellenállót a technológia-ellenesség teoretikusai és mozgalmai közé, és a robotvilág mai állapota fényében emeli ki főbb téziseit (Galliott 2017: 369–385).

A robbantásokkal a gyakorlati konzekvenciákat levonva Kaczynski csak továbbfejlesztette Jacques Ellul 1964-es „The Technological Society" című kötetének téziseit,

amely a maga módján szintén már csak Oswald Spengler 1922-es, nyugati civilizáció hanyatlását elemző művének a továbbvitele, mely ezt a hanyatlást a technológiai útra lépéssel magyarázta (Spengler 1995). A pusztán pesszimista és rezignáltan belenyugvó tónus Splenglernél és Ellulnál vált aztán morális alapú ellenállássá Kaczynskinél, és miután úgy látta, hogy reformra ez ellen nincs már mód, úgy vélte, hogy az emberiség pusztulását megakadályozandó csak a forradalmi erőszak marad. Pamfletje után évtizedekkel most így érdemes újragondolni, hogy a robotvilág mai állapota és a már jórészt látható, további radikális változások mennyiben jelenthetik az emberiség végveszélyét vagy legalábbis állapotának nagymértékű romlását?

Kiindulópontjukként a technológiai társadalom elleni fellépésükben érdemes kiemelni, hogy mind Splengler, mind Ellul és Kaczynski a fizikai-biológiai környezetbe beágyazódó létként fogják fel az emberi létet, és amennyiben az ipari forralom óta az emberi élet és tevékenység növekvő mértékben technológiailag közvetítetté és egyre távolibbá válik e környezettől, annál inkább az emberi lét megsemmisüléseként fogják ezt fel: „Ellul azt írta, hogy „a gépesítés nem csak új emberi környezet megteremtésére, hanem az ember lényegének megváltoztatására is vonatkozik", és hogy „a millieu, amelyben él, már nem az övé. Alkalmazkodnia kell egy új világhoz, egy új univerzumhoz, miközben őt egy másikra teremtették" (Galliott 2017: 373). Kaczynski osztja ezt az érzést. Ezzel szemben, ha Hartmann tézisét vizsgáljuk, aki az emberi élet négy, egymásra épülő létrétegét tartja szem előtt, és aki a felsőbb rétegeknek az alsóbbakat érintő egyre erősebb átformáló hatásából indul ki az evolúció menetében, akkor a fenti tézist túlzottan és ok nélkül pesszimistának kell minősítenünk. Kaczynskiék ugyanis az emberi élet megbomlásának tekintik, ha a négy létréteg közül a legfelső értelmi létréteg egyre inkább meghatározó válik az alsóbbakat illetően. Pedig – vethetjük velük szembe –, ha lassabb mértékben, de ez történt már az elmúlt két-háromezer évben is a fémek és főként a vas eszközzé tétele és az ember környezetének ezekkel átalakítása óta. Az ipari forradalom csak felgyorsította ezt, és különösen az 1950-es évek óta viharossá vált az emberi közösségek legkülönbözőbb tevékenységeiben az értelemre és az ezzel átitatott technológiára alapozás. Vagyis az emberi élet egyáltalán nem csak a fizikai-biológiai létrétegre alapozott, és így, ha ezek aránya és meghatározó ereje csökken az emberi életben, illetve technológiailag messzemenően közvetítetté és átalakítottá válik ez a környezet, még nem jelenti azt, hogy ezzel elpusztítjuk az embert. Mindez ugyanis csak a négy létréteg fontosságának a súlyát rendezi át, radikálisan megnövelve az alsóbbak felett az értelmi létrétegre alapozott emberi életet. Ezt az értékelésünket csak az függeszthetné fel, ha a robotvilág, fejlődése egy pontján, tényleg létrejönne e világ kiszakadása az emberi irányítás alól, és egy új létrétegként emelkedne az eddigi evolúciós csúcsot jelentő emberi társadalmak fölé. Kaczynski prófétává emelését ekkor már csak az akadályozná meg, hogy ilyen körülmények és ennek veszélyei között ennek elmaradása lenne már a legkisebb gondunk. Ezt azonban csak kis valószínűséggel bíró lehetőségként lehet a mai tudásunk szerint felfogni, és inkább az emberi társadalmak mesterséges intelligenciával átitatottságának fokozódását lehet reális jövőképnek tekinteni.

Irodalom

Abney, Keith, "Robotics, Ethical Theory and Metaethics: A Guide for the Perplexed", in Patrick Lin, Keith Abney and George A. Bekey (eds.), *Robotethics,* The MIT Press, Cambridge-Massachusets- London, 2011, pp. 35–54.

Allen, Collin and Wendell Wallach, "Moral Machines: Contradiction in Term or Abdication of Human Responsibility?", in Patrick Lin, Keith Abney and George A. Bekey (eds.), *Robotethics,* The MIT Press, Cambridge-Massachusets-London, 2011, pp. 55–68.

DiGiovanna, James, „Artificial Identity", in Patrick Lin, Ryan Jenkins and Keith Abney (eds.), *Robot Ethics 2.0*, Oxford University Press, New York, 2017, pp. 307–321.

Doherty, Jason P. (ed.), *AI Civil Rights: Addressing Artificial Intelligence and Robot Rights*, Kindle Edition, 2016.

Ford, Martin, *The Rise of Robots: Technology and the Threat of a Jobless Future,* Basic Books, 2016.

Galliott, Jai, "The Unabomber on Robots", in Patrick Lin, Ryan Jenkins and Keith Abney (eds.), *Robot Ethics 2.0*, Oxford University Press, New York, 2017, pp. 369–385.

Hartmann, Nicolai, *Das Problem des geistigen Seins. Zur Grundlegung der Geschichtsphilosophie und der Geisteswissenschaften*, Walter de Gruyter Verlag, Berlin 1962.

Hegel, Georg Wilhelm Friedrich, *A jogfilozófia alapjai*, Akadémia Kiadó, Budpest 1971.

Henschke, Adam, „The Internet of Things and Dual Layers of Ethical Concern" in Patrick Lin, Ryan Jenkins and Keith Abney (eds.), *Robot Ethics 2.0*, Oxford University Press, New York 2017. pp. 229–243.

Kaku, Michio, *Az elme jövője*, Akkord Kiadó, Budapest 2015.

Kelly, Kevin, *The Inevitable: Understanding the 12 Technological Forces That Shape Our Future,* Penguin Books, New York, 2014.

Klinewicz, Michal, "Challenges to Engineering Moral Reasoners" in Patrick Lin, Ryan Jenkins and Keith Abney (eds.), *Robot Ethics 2.0*, Oxford University Press, New York, 2017, pp. 244–257.

Loh, Wulf and Janina Loh, "Autonomy and Responsibility in Hybrid System, in Patrick Lin, Ryan Jenkins and Keith Abney (eds.), *Robot Ethics 2.0,* Oxford University Press, New York, 2017, pp. 35–50.

Luhmann, Niklas, *Liebe als Passion: Zur Codierung von Intimität,* Suhrkamp, Frankfurt am Main, 1994.

Pokol Béla, *Morálelméleti vizsgálódások,* Kairosz, Budapest, 2010.

Pokol Béla, „Mesterséges intelligencia: egy új létréteg kialakulása?", *Információs Társadalom*, XVII. évf. (2017) 4. szám, 39–53. old. http://dx.doi.org/10.22503/inftars.XVII.2017.4.3

Splengler, Osvald, *A Nyugat alkonya. A világtörténelem morfológiájának körvonalai.* I - II. kötet, Európa Könyvkiadó, Budapest, 1995.

Talbot, Brian, Ryan Jenkins and Duncan Purves, "When Robots Should Do the Wrong Thing", in Patrick Lin, Ryan Jenkins and Keith Abney (eds.), *Robot Ethics 2.0*, Oxford University Press, New York, 2017, pp. 258–273.

Vallor, Shannon and George A. Bekey, "Artificial Intelligfnce and the Ethics of Self-Learning Robots", in Patrick Lin, Ryan Jenkins and Keith Abney (eds.), *Robot Ethics 2.0,* Oxford University Press, New York, 2017, pp. 338–353.

White, Trevor N. and Seth D. Baum, "Liability for Present and Future Robotics Technology", in Patrick Lin, Ryan Jenkins and Keith Abney (eds.), *Robot Ethics 2.0*, Oxford University Press, New York, 2017, pp. 66–79.

Zoller, David, "Skilled Perception, Authenticity, and the Case Against Automation", in Patrick Lin, Ryan Jenkins and Keith Abney (eds.), *Robot Ethics 2.0*, Oxford University Press, New York, 2017, pp. 55–68.