„A 'gondolkodó gép' fogalmával megragadott problématérben eltűnik az értelmezési tartományból az a tény, és nem elégszer hangsúlyozzuk, hogy nincs önmagában vett gépi intelligencia, az csak hibrid (ember+gép) szerkezetben tud megnyilvánulni, és az emberi mozzanat az elsődleges.(…) ez nyit utat annak a pusztító és félrevezető diskurzusnak, amelyik a „mikor éri utol és mikor múlja felül a gépi intelligencia az emberit" ál-dilemmáját, vagy ennek még kakofóniába hajlóbb változatát, a „mikor győzi le a gép az embert" morális pánikba forduló kérdését zenésíti meg."

(Z. Karvalics László)

„Nem nehéz belátni, hogy az érzelmeink nagyon szaftos biológiai alapjait is magában foglaló érzelmi rendszert programozni igen kétes vállalkozás."

(Síklaki István)

„A tisztes távolságot az öntanuló robotokkal is meg kell tartanunk, és megfelelő mennyiségű biztonsági intézkedéseket kell tennünk velük kapcsolatban."

(Juhos Sándor)

„A technológiai szingularitáshoz a lehető legsokrétűbb és „legmélyebb" tanuláson keresztül vezet az út, a gépi rendszerek máskülönben képtelenek hibátlanul felismerni arcot, tárgyat, szöveget, beszédet, érzelmet.A következő lépés a különféle részterületek eredményeinek egyetlen rendszerben történő integrációja lenne, anélkül, hogy zavarnák egymás működését.

(Kömlődi Ferenc)

„Ebből a családi nézőpontból a leendő mesterségesen intelligens/tudatos stb. entitásokból ugyanúgy lehetnek jó- vagy gonosztevők, mint a természetes gyerekeinkből. (…) A szoftverek készítéséből egy új szakma született: a programozó, várhatóan ennek mintáját követve születik majd meg az új foglalkozás: a robotokat nevelő, programnevelő informatikus."

(Bártfai Norbert)

„Mire lesznek képesek majd a gépek, azt nem tudjuk, hiszen éppen a tulajdonságkinyerés, a lényeglátás még hiányzik. Azt viszont tudni fogják, amit mi beléjük táplálunk a hagyományos módon (például összeadás, szorzás), vagy az új módszerekkel, a hatalmas adatbázisokkal."

(Lőrincz András)

ISSN 1587-8694
9 771587 869007 15004

Ára: 950 Ft

Információs Társadalom 2015. XV. évfolyam 4. szám

Információs Társadalom

Vita a mesterséges intelligencia fejlesztésében rejlő lehetőségekről és veszélyekről

2015. XV. évfolyam 4. szám

Információs Társadalom

TÁRSADALOMTUDOMÁNYI FOLYÓIRAT
Alapítva 2001-ben

Szerkeszti: Csótó Mihály – Rab Árpád
Olvasószerkesztő: Tamaskó Dávid

Lapterv: Szépkilátás Stúdió

Kiadja
Az INFONIA (Információs Társadalomért, Információs Kultúráért) Alapítvány és a Gondolat Kiadó

A folyóirat kiadását a Nemzeti Hírközlési és Informatikai Tanács (NHIT) támogatja

A folyóirat kiadásában közreműködik az Óbudai Egyetem Digitális Kultúra és Humán Technológia Tudásközpontja

Szerkesztőség: 1032 Budapest, Kiscelli utca 78. 214-es szoba
e-mail: titkarsag@infonia.hu
Gondolat Kiadó: tel: 486-1527, e-mail: info@gondolatkiado.hu
www.gondolatkiado.hu

Készült a Rolling Site Nyomdában
ISSN 1587-8694

A folyóirat 2008/1. számától kezdve megtalálható a Thomson Reuters indexeiben (Social Scientes Citation Index®, Social Scisearch®, Journal Citation Reports/Social/Sciences Edirion)

TANULMÁNYOK

Üdvözlet az olvasónak!

Az embert különböző feladatok megoldásban meghaladó gépek igen régóta foglalkoztatják az emberiséget, és természetesen a kutatókat is, akik közül sokan gyakorlatilag az első számítógépek megjelenése óta törekedtek az emberi gondolkodáshoz és tudathoz hasonló, vagy azon túlmutató eszközök(?) létrehozására. Az ötvenes években számos műhelyben kezdődött jelentős munka a területen, és a „mesterséges intelligencia" kifejezés megalkotása is ekkorra tehető. A téma iránti lelkesedés jól tükröződik a múlt század középső harmadának tudományos-fantasztikus irodalmán is.

Egy emberi mércével mérve intelligens gép, esetlegesen egy mesterséges tudat létrehozása azonban számos problémát tartogatott a fejlesztők számára. Hosszú évek aprómunkájának, valamint nem utolsó sorban az utóbbi években elérhetővé vált számítási kapacitásnak, illetve a korábbinál jóval nagyobb, a gépek tanítását lehetővé tevő adattömegnek köszönhetően az utóbbi időben azonban egyre látványosabb eredmények születtek. Az elmúlt években még a tudomány és technológia világa iránt kevésbé fogékonyak is naponta találkozhattak az „intelligens" technológiákkal, vagy az azok területén történt újabb áttörésről szóló hírekkel, legyen szó önvezető autókról, a beszédfelismerő személyi asszisztensekről vagy éppen az emberi játékosokat legyőző számítógépekről (a Kaszparovot legyőző Deep Blue-tól indulva a kvízbajnok Watsonon át egészen az emberi ellenféllel szemben 2015 őszén Go játszmákat nyerő komputerig).

Nem vitatható, hogy korunk egyik meghatározó technológiai trendje a mesterséges intelligenciákkal kapcsolatos kutatások intenzitásának növekedése, valamint az ezekre épülő termékek és szolgáltatások megjelenése. Ez a markáns trend magával hozta a mesterséges intelligenciában rejlő lehetőségek mellett a technológia veszélyeiről való gondolkodást, ami azonban sokszor úgy tűnik, a tudományos eredményektől szinte teljesen függetlenül, a már említett, tudományos-fantasztikus irodalom által egykoron meghatározott térben és kérdések körül zajlik (még a legfelsőbb döntéshozók és politikacsinálók esetében is, mint ahogy azt Danah Boyd, a Microsoft kutatója a Világgazdasági Fórumon a témáról zajlott eszmecserékkel kapcsolatban megjegyezte). Ez azon túl, hogy komoly kérdéseket vet fel a tudománykommunikáció terén, semmiképpen sem segíti a valóban igen fontos kérdéseket érdemi elemzését. Ezért döntöttünk úgy, hogy 2015-ös utolsó számunkat vita-formátumban a mesterséges intelligencia lehetőségeinek és veszélyeinek szenteljük, azzal a nem titkolt szándékkal, hogy felhívjuk a figyelmet azokra a kérdésekre, amelyekről szerintünk a felfokozott várakozások és az eltúlzott veszélyek helyett egy ilyen diskurzusnak inkább szólnia kellene.

Elsőként alapító-főszerkesztőnk, Z. Karvalics László vitaindítóját olvashatják, mely már terjedelmével és részletezettségével is arról üzen, milyen sokféle aspektust szükséges figyelembe venni a kérdés tárgyalásakor. A nagyívű írás után az arra érkezett reflexiókat közöljük, ezúton is megköszönve Bátfai Norbertnek, Juhos Sándornak, Kömlődi Ferenc-

nek, Lőrincz Andrásnak és Síklaki Istvánnak, hogy meglátásaikkal hozzájárultak egy valódi vita kialakulásához. Számunk egy, a témakör neves kutatói által jegyzett, a vitaindítóban is megemlített tanulmány magyar változatával zárul, mely a valós problémák és lehetőségek felől közelítve veszi sorra a mesterséges intelligencia fejlesztésével kapcsolatos rövid és hosszú távú kutatási prioritásokat. Még ha szerzőink legfőbb üzenetei sok szempontból egy irányba is mutatnak, természetesen ez nem jelenti azt, hogy egy csapásra a kívánt mederbe tereltük a mesterséges intelligenciával kapcsolatos gondolkodást, de talán a magunk eszközeivel sikerült a témát közelebb hoznunk a komplex információs társadalom narratívák nyújtotta keretekhez.

Mindezekhez jó olvasást kíván,

a szerkesztőség

Z. Karvalics László

Mesterséges intelligencia – a diskurzusok újratervezésének kora

2015 szeptemberében számos népszerű magyar online portál adott hírt arról,[1] hogy a Queenslandi Műszaki Egyetem intelligens robotot készített az ökoszisztémában kulcsszerepet játszó korallzátonyok védelmében. Északkelet-Ausztrália partjainál ugyanis komoly részben a töviskoronás tengericsillag felelős a pusztításért, ezért a tudósok kamerával és GPS-szel víz alá merülő automata szerkezetet terveztek, olyan szoftverrel, amely fotók és videók ezrei alapján képes megbízhatóan azonosítani a kártevőt, a töviskoronás tengericsillagot, és sűrített levegőt befecskendezve végezni velük. Az elsőként üzembe helyezett masina 8 munkaóra alatt mintegy 200 töviskoronás tengericsillagot öl meg, de a cél az, hogy a tapasztalatok alapján akár robotok százai is csatasorba álljanak, a minél nagyobb pusztítás érdekében.

Mivel az izgalmas fejlesztésről tudósító cikkek „robotgyilkosokkal", „víz alatti mészárlásokkal" tálalták a programot, rögtön megjelentek azok a kommentárok, amelyek a Terminátor-filmek hangulatát felidézve a „ma még tengericsillag, holnap ember?" kérdést tették fel, ügyet sem vetve arra, hogy a víz alatti robot gyakorlatilag a kapa funkcióját tölti be, amellyel egy ültetvényen gyomlálnak, csak nem emberi szem, izomerő és tüdő hozza mozgásba.

Ha egy szinttel magasabbra emelkedünk, akkor kiderül, hogy a beavatkozás szükségessége az ökológiai egyensúly felborulására vezethető vissza, de hogy pontosan mi okozta a felborulást, arról már megoszlanak a vélemények. Vannak, akik úgy vélik, hogy invazív fajról van szó, amelynek kiirtása kívánatos, mások szerint az ember felelős, mert megritkította a tengericsillagok természetes ellensége, a gyönyörű házáért intenzíven halászott kagyló populációit.

Az elkeseredett viták két dolgot tettek egyértelművé: az okok és magyarázatok magas szintű világában a teljesebb kép kialakításához további tudások és ismeretek megszerzésére van szükség, miközben a megoldáskeresés azonnali cselekvést igényel. A probléma forrása azonban semmiképp nem a könnyen diabolizálható „gyilkos robot" környékén, hanem civilizációnknak a környezettel való viszonyát szabályozó meghatározottságok összetett világában keresendő.[2]

[1] A hír forrása: O'Callaghan (2015).

[2] Mindez még egyértelműbben kirajzolódik, mondjuk, a kínai nagyvárosok légszennyezésének csökkentését támogató mesterséges intelligencia-megoldások értékelésekor. Nem kérdéses, hogy a környezetszennyező, kontrollálatlan ipari tevékenység, a közlekedés, a fűtés és az időjárás változása okozzák a bajt, de az sem, hogy a megoldáskeresésben egyidejűleg fontos a kedvezőtlen hatások azonnali enyhítését lehetővé tevő beavatkozások és rövid távú döntések támogatása, valamint a problémák okait megszüntető vagy módosító, hosszabb távon érvényesülő, komplexebb szabályozási vagy kultúraváltási lépések megtétele, amelyeket szemléletváltásnak kell megelőznie. Az utóbbi igazi civilizációs kihívás, az előbbihez esélyt kínálhat egy olyan előrejelző rendszer, amely óriási adattömeg és számítási kapacitás birtokában napokkal a szennyezés koncentrációjának veszélyes megnövekedését megelőzően jelzi a közelgő gondot, és javaslatot is tesz annak orvoslására (üzemek időszakos bezárására vagy forgalomcsökkentésre).
http://www.technologyreview.com/news/540806/how-artificial-intelligence-can-fight-air-pollution-in-china/?utm_campaign=newsletters&utm_source=newsletter-weekly-computing&utm_medium=email&utm_content=20150903

Nagyjából ugyanekkor, 2015 júliusának első napjaiban, a Bécsben tartott információtudományi világkonferencia egyik (merthogy több volt) családias Global Brain szekciójának vitaindító előadásában Michael E. Arth, az UNICE (Universal Network of Intelligent Conscious Entities)[3] alapítója a hallgatóságnak szegezte a kérdést: *emelje fel bátran a kezét az, aki szerint az erős mesterséges intelligencia nem jelent potenciális veszélyt az emberiségre nézve*! Arth csalódott arcát a mai napig látom magam előtt, merthogy az első sorokból volt módomban szemlélni az eseményeket. Nem azt várta, hogy tizenöt hallgatójából tizenhárom emeli majd fel a kezét. Egész előadását ugyanis arra kívánta felépíteni, hogyan készüljünk már most fel a „gondolkodó gépek" által hordozott fenyegetésre.

S ez a hevenyészett szavazás jól jelzi ugyan, hogy a kutatók nagy része nem szükségszerűen kútmérgező publicisztikák badarságaitól befolyásoltatva alakítja ki saját, autonóm véleményét ebben a kérdésben – de ez mégis egy tudományos konferencia volt, amelyen Arth tudósként vett részt. Evvel meggyőzően illusztrálta is, hogy a pánikkeltéssel operáló bulvárnarratívák milyen erősen összecsúsznak a tudományosnak mondottakkal. S mivel ezt aggasztónak gondolom, vitaindítónak szánt írásom elején röviden ismertetem ennek a két oldalról támogatott veszélydiskurzusnak a fő argumentumait, szomorúan konstatálva, hogy *a friss könyvtermést szinte teljes egészében ide tartozó munkák* dominálják.[4] Ezt követően bemutatok néhány, szívemnek kedves ellendiskurzust és szerzőt, majd utána javaslatot teszek egy alternatív megközelítésre. Mindvégig igyekszem az önállóként kutatható, érdeklődésre számot tartó témakörök és álláspontok nagy számát is érzékeltetni. Nem gondolom azonban, hogy a fentiekkel bármit sikerülne „megoldani": leginkább példát, mintát szeretnék adni arra, miként lehetne kiszabadulni a jelenlegi tipikus gondolkodási keretrendszerekből, és másképp, máshol és más veszélyeket, illetve kritikus csomópontokat azonosítani, mint a pánik-irodalom. Roppant kíváncsi is vagyok, vajon ki, miért és mit oszt vagy utasít el ebből a megközelítésből, vagy milyen más gondolkodási útvonalat ajánl.

Musk, Hawking és más borzongatók

Elon Musk, a sikercég Tesla elnöke szerint „*minél fejlettebb lesz egy robot, annál kevésbé tiszteli majd a tervezőjét*". Az önfejlesztő robotokkal felgyorsul a „szuperintelligens gép" kifejlesztéséhez vezető út, amelynek végén ez a masina „*spamnek nézheti, kiszűri és eltörli a Föld színéről az embereket*", mint egy kéretlen üzenetet.[5] „*Potenciálisan veszélyesebb, mint az atombomba*" (Pesthy, 2015). Nem kisebb tekintély véleménye mellett állt ki evvel, mint Stephen Hawking asztrofizikus, aki szintén osztja azt a nézetet, hogy a gondolkodó gépek a puszta létünkre nézve is veszélyt jelentenek. "*Egy ilyen mesterséges intelligencia pillanatok alatt önállósítaná és folyamatosan, egyre gyorsuló tempóban újratervezné magát. Miközben mi em-*

[3] http://www.unice.info/unice/index.htm (Letöltve: 2015. november 11.)

[4] A kutatás első szakaszát a FuturICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett. A befejezés „*Az Európai Unió és Magyarország támogatásával, a TÁMOP 4.2.1.D-15/1/KONV-2015-0006 azonosító számú - Ösztöndíj magyar és külföldi hallgatóknak és kutatóknak - A kőszegi innovációs kutatóbázis és tudásközpont fejlesztése a Pannon Egyetem oktatási és kutatási hálózatának keretében*" projekt részeként készült.

[5] http://www.origo.hu/techbazis/20141009-spamnek-nezik-es-torlik-az-emberiseget-a-robotok.html http://www.vanityfair.com/online/daily/2014/10/elon-musk-artificial-intelligence-fear

berek, mivel fejlődésünknek határt szab a lassú biológiai evolúció, menthetetlenül lemaradunk, és végül kiszorulunk a versenyből. A gépagy végez velünk."[6]

A Terminátor és a Mátrix fantáziavilágából a valóságba átemelt veszély-üzenethez[7] képest óvatos figyelmeztetésnek tűnik az ezredforduló előtti időszak közhelyes fordulata, hogy a „gondolkodó gépek teljesítménye meghaladja az emberét". Itt már evolúciós kontextusban bizonyul „fejlettebbnek" a mesterséges intelligencia a biológiainál.[8] James Barrat könyvében a szuperintelligens gép az utolsó emberi innováció, amely egyúttal elhozza a „véget" (Barrat, 2014). Amint utoléri a mesterséges intelligencia az emberit,[9] azonnal saját

[6] http://www.origo.hu/techbazis/20141203-hawking-szerint-a-gepagy-vegul-vegez-velunk.html

[7] „Az emberrel szemben cselekvő, öntudatra ébredt robot" képét a fikciós világban a 2001: Űrodüsszeia legendás HAL nevű gépe kezdte megalapozni, legfrissebb fejezetét pedig a 2014-es Ex Machina című film írta – a valóságban pedig az ipari robot-balesetek és a „gyilkos drónok" megjelenése erősítette fel. 2015 nyarán egy német autógyárban egy ipari robot okozott halálos sérülést, de nem „cselekvő intelligencia", hanem „rosszkor és rossz helyen bekapcsolt gép" formájában. A krónika hiába jegyez fel számos ilyen eseményt (az elsőt 1979-ből, a Ford-gyárból), ezek valójában mind emberi hibára és nem a robot-mivoltra vezethetőek vissza. Ahogy a ma már thrillerekbe is beszivárgó kisebb, szúnyogszerű, méreggel gyilkoló, és nagyobb, golyózáporral támadó drónok sem maguktól ölnek, hanem az optikai modulon keresztül a potenciális áldozatot azonosító és döntést hozó földi irányítás révén. Az általunk teremtett és életre keltett, de ellenünk forduló lénnyel szembeni félelmek valójában kulturálisan és nem a technológia miatt kódoltak – véli a koblenzi egyetem két kutatója, Ulrike Barthelmess és Ulrich Furbach (http://www.technologyreview.com/view/527336/do-we-need-asimovs-laws/). S hiába az alapos és sok forrást használó újságírói munka, amellyel a cselekvővé váló robottal kapcsolatos sok kérdést megfelelő óvatossággal sikerül tárgyalni, ha a cikk címe ez: *Robot ölt embert, ez már az apokalipszis?* (Bolcsó, 2015). S hiába 'nem' a válasz a kérdésre, már a leadben, tehát azonnal a kérdés után, a cím önálló életre kel: van, aki nem olvassa el az egész írást, csak a címet tudatosítja, a linkgyűjteményekbe is csak a cím kerül majd bele…

[8] Nem is véletlen tán, hogy erős képviselettel jelent meg az iker-gondolat, hogy *a Földön kívüli élet keresése helyett is a Földön kívüli mesterséges szuperintelligencia keresése* felé kellene fordulni, hiszen a Kozmosz egészére lehet igaz a szintetikus élet(?)formák magasabbrendűsége. Az amerikai filozófus, Susan Schneider néhány NASA-közszereplő által is támogatott megközelítését részletesen bemutatja Stone (2014).

[9] Néhányan a kritikus pontot a *'hazugságra képes robotban'* látják – sajnos makacsul azt a svájci kísérletet félremagyarázva, ahol az erőforrások megtalálását egymásnak jelző robotokat irányító „program-agyak" új generációi (hibridizált szoftverei) a források csökkenésekor „felülírták" a jelzés-parancsot, és nem értesítették a többi gépet, hanem maguknak tartották fenn az erőforrást. A kísérlet egy biológiai evolúciós helyzetet modellezett, mégis azóta a legtöbb hivatkozást a „nemcsak ölni, már hazudni is tudnak a robotok" szöveghelyzetben találjuk – úgy, hogy a „robot" nem „hazudott", hanem a megadott kiinduló feltételek által szabályozott térben programot módosított, és ezt nem egy ember-robot, hanem előformált robot-robot játszmában tette (Fox, 2009). Hasonlóképpen takarhat számtalan különböző megközelítést a *'vicc-képes robot'* fejlesztésének programja (Hernandez, 2016). Ígéretes és fontos irány azt vizsgálni, hogy egy személyes ember-gép kommunikációs helyzetben miként „olajozhat meg" interakciókat, ha a gépi oldal az emberi humorra emlékeztető jegyeket is tud csatasorba állítani, ráadásul szituáció-érzékenyen és adekvát módon. Nem kevésbé érdekes kihívás szöveget generálni humoros karikatúrákhoz (http://research.microsoft.com/en us/um/people/horvitz/phumor.pdf), amely egyúttal valódi ösvény a humor természetének a jobb megértéséhez is. De azt tűzni ki célul, hogy adott időn belül humorértő és humorgyártó mesterséges intelligenciánk legyen, illuzórikus, értelmetlen és erőforrás-pazarló. Az emberi intelligenciának egyrészt semmi szüksége nincs gépi humorpótlékra, jól elvan a maga kultúrába, közösségbe és cselekvésbe beágyazott vicc-ökoszisztémájával. Másrészt a humor (akárcsak a neologizmus vagy a metaforálás) magasrendű jelentésművelet (hiszen jellemzően több jelentés-sík keverésével és a jelentésvárakozások felülírásával fejti ki hatását). Amikor még a sokkal alacsonyabb szintű jelentésműveletek világában is megoldatlan kihívások sorával kell szembenézni, mi indokolná egy erre ráépülő kutatási szint létjogosultságát?

túlélése lesz a számára fontos, és már csak azt a kérdést tehetjük fel, hogy megengedik-e majd nekünk, náluknál fejletlenebb lényeknek, hogy az árnyékukban létezhessünk. Bostrom (2014) még csak a gorillákhoz hasonlítja az emberi fajt, de Brain (2015) nem viccel: friss könyvének címében a „második intelligens faj", az érző és öntudatra ébredt mesterséges intelligencia számára az emberiség már olyan irreleváns, mint ma nekünk a csótány…[10]

Elég úgy feltenni a kérdést, hogy *a mesterséges intelligencia szolgálni vagy helyettesíteni fog-e minket* (ráadásul egy iskolásoknak szánt, több mint 200 oldalas népszerűsítő összefoglalóban, mint Del Monte, 2014), hogy világos legyen: az *alarmista* kiindulópont öngerjesztővé vált.[11] Azt látjuk, hogy *az „emberit meghaladó mesterséges intelligencia" létrejötte olyan axiómává fejlődött, amelyre kétely nélkül épülnek elméletek és előrejelzések.*

A korábban erős mesterséges intelligenciának (Strong Artificial Intelligence, SAI), később mesterséges általános intelligenciának (Artificial General Intelligence, AGI) nevezett irányzat annak az elképzelt intelligens gépnek a megvalósítására szövetkezik, amelyik bármely intellektuális feladatot teljesít, amelyre az ember képes. Ehhez a célhoz ma[12] a *szingularitás-hipotézis* adja az üzemanyagot.

Legtöbben az irányzat egyes számú prófétájához, Ray Kurzweilhez és az ő emblematikus könyvéhez kötik a szingularitás, az emberit utolérő gépi intelligencia tézisének megszületését (Kurzweil, 2013/2005), pedig a diskurzus most kereken félszáz éves.[13] Irving

[10] Meg kell jegyeznünk, hogy a szuper-intelligencia okozta világvége-forgatókönyvek csak részei egy átfogóbb világvége-narratívának, amely az összeomláshoz vezető technológiai okok mellett a politikai és gazdasági szempontokra is súlyt helyez (Rees, 2004).

[11] A mikroelektromechanikai kutatások egyik „nagy öregje" az időutazással foglalkozó könyve után egy interjúban (Love, 2014) megad forgatókönyvet és határidőt is: 2045-re a legfejlettebb fajok (!) már nem emberiek lesznek. Egyébként az alarmizmus esetleges hasznáról és evvel összefüggésben az alarmistákkal való foglalkozás értelméről vagy értelmetlenségéről is folynak viták. Előkerül olykor a gazdaságtudományi analógia, ahol a stratégiai inflexiós ponthoz közeledő vállalatok életét meghatározó változásokat megérző és előre jelző kollégákra, az *úgynevezett kasszandrákra* nagyon is ajánlatos odafigyelni (Grove, 1998). Grove gondolatmenetével csak az a baj, hogy a „kasszandrák" családjának csak kis részére igaz, hogy ráéreznek a változásra, és megfelelően jelzik előre a jövőt. Az összes többi „kasszandrára" fordított idő és figyelem túlnyomó része nettó veszteség. Jól belátható mindez, ha arra gondolunk, vajon mennyit segít a világvége-jóslatok, a kozmikus katasztrófákat napra/órára pontosan előrejelző „próféciák" történeti és tipológiai elemzése a Földet fenyegető külső hatásokkal foglalkozó tudományos közösségnek, a dizasztrológusoknak. Az alarmista pozíció kialakításának és képviseletének súlyos tehertétele továbbá, hogy a jószándékú és meggyőződéses alarmisták mellett tömegesen foglalnak el pozíciót azok, akik szerint a figyelemgazdaságban sokkal eredményesebbek lehetnek alarmizmussal, mint mérsékelt, tárgyszerű előrejelzésekkel. Ezért is kötöttünk ki az Allen Institute for Artificial Intelligence vezetője, Oren Etzioni által bevezetett és népszerűsített „Frankenstein-komplexum" helyett az alarmizmus kifejezés mellett: a művelődéstörténeti analógiánál fontosabbnak tűnik ugyanis, hogy az egymásra licitálás kultúrájára is utalhassunk. Ki mond riasztóbbat, rémisztőbbet a gépi intelligencia jövőjében rejlő veszélyekről? Mindennek azért van haszna is: a szélsőségek visszamenőleg teszik meztelenné a kevésbé radikális alarmizmus királyait is.

[12] Korábban elsősorban Hans Moravec hajlítgatta a vitateret, aki a mesterséges gépi faj megszületését 2030-2040 közé tette (Moravec, 1988), s tíz évre rá már a szuperintelligencia megszületésére épített (Moravec, 1998). A Moraveccel szembeni kritikák sokat finomítottak az erős mesterséges intelligencia híveinek álláspontján is.

[13] Érdeklődőknek érdemes a különlegesen alapos, tanulmány-értékű Wikipedia-szócikkel kezdeni a diskurzustörténeti ismerkedést: http://www.wikiwand.com/en/Technological_singularity.

John Good 1965-ben jelentette meg az „*első ultraintelligens gépről*” szóló tanulmányát, bevezetve az intelligencia-robbanás (*intelligence explosion*) tézisét is, amely az egyre okosabb gépek által gyártott még okosabb gépektől várja a fordulópontot (Good, 1965). Koncepciója egyetlen, korántsem lényegtelen apróságban különbözik a kortárs elméletektől. A Vernor Vinge és Ray Kurzweil által képviselt szingularitást az teszi elérhetővé, hogy a robbanás exponenciálisan gyorsul, és *ekvezet* a szingularitáshoz. Good koncepciójában az ultraintelligens gép megteremtése *után* kezdődik a robbanás, amely az innovációt onnantól „áttolja” a gépek térfelére. Az emberírta technikatörténet nála itt véget is ér.

Belátható, hogy az alarmista nézőpont csakis azért született meg, hogy lelkiismereti ellensúlyt képezzen, és alternatív forgatókönyvet adjon a magasra emelkedett elvárás-horizontnak avval kapcsolatban, hogy mi várható néhány évtized múlva, az évszázad közepére[14] a szingularitástól az emberiségre nézve. Kurzweilék, akiket emiatt utópistáknak tartanak, ugyanis nem aprózzák el az ígéreteket: az üzlet világában kreatív rombolás megy végbe, nő a gazdagság, csökken az egyenlőtlenség, a világproblémák megoldhatóvá válnak, a biológiai tökéletesítéstől fantasztikus egészség és halhatatlanság-közeliség várható (Miller, 2012). A szuperintelligenciához hasonlóan ugyanis axiómává vált az emberi agynak a minőségi agy-számítógép kapcsolathoz szükséges feltérképezése, megértése (Kurzweil, 2013), ami a szingularitással elért végállapot idején az agy „feltöltését”, áttöltését teszi majd lehetővé a biológiai hordozóról a gépire. Blackford és Broderick (2014) már ennek a jövőjéről, Rothblatt (2014) pedig az ekképpen elért „virtuális emberi” és a „digitális halhatatlanság” ígéretéről és veszélyeiről írt könyvet.

Szívesen nevezem végül *navigacionistáknak* azokat a szellemi műhelyeket és szerzőket, akiknek se a gyilkos robot rémképére, se a digitális halhatatlanságra nincs szükségük ahhoz, hogy az „intelligenciarobbanás” (*intelligence explosion*) következetesen a 21. század közepére várt pillanatához igazítsák óráikat. Ahogy Muehlhauser (2013) fogalmaz: történelmünk legfontosabb pillanata lesz az, amikor a gépi intelligencia és az avval összekapcsolt képességek meghaladják az emberit – s ennek a folyamatnak a bölcs navigálása a jövő leglényegesebb kihívása. Ha jól nyúlunk bele az eseményekbe, a szuperintelligencia a legjobb dolgok egyike lehet, ami az emberiséggel valaha történt, mert amihez nem vagyunk elég okosak, „ő”, a barátságos mesterséges intelligencia (*Friendly AI*)[15] majd megoldja (Bostrom, 2014). Ellenben minden félre is csúszhat, ha nem vagyunk elég felkészültek. Ez a felemelt mutatóujj, ami a szélsőségekbe forduló alarmistákhoz képest tárgyilagosságot és „középutasságot” ígér, dollármilliókat ér a navigacionistáknak. Egyre-másra születnek azok a kutatóintézetek, amelyek a szingularitással foglalkozó kutatók legújabb generációját vonzzák be „a jövő megalkotásába”, miközben komoly forrástömeget tudnak mozgósítani.[16]

[14] Az időpontról nagy (és jórészt értelmetlen) vita folyik, számtalan variáció kering, volt már, aki „átlagot” is számolt (2040-et), de a mindenki közül a leginkább radikális Goertzel (2014) szerint a szingularitáshoz akár jóval korábban is eljuthatunk, ha projektként megfelelő figyelem övezi és kellő mennyiségű forráshoz jut. Nem véletlenül nevezi Daniel C. Dennett városi legendának a szingularitás-hipotézist (Brockman, 2015).

[15] A barátságos AI fogalmát (talán Casper, a barátságos szellem mintájára) Eliezer Yudkowsky alkotta meg, az alarmistákat ellensúlyozandó. Levy (2008) azt mutatta meg, *miért épülhet ki szoros érzelmi kapcsolat, valódi kötődésen alapuló szövetség* (real companionship) ember és állat, *ember és robot között*, s hogy ez miért nem jelenti a „humán” dimenzió elvesztését.

[16] A 2000-ban alapított Szingularitás Intézet (*Singularity Institute for Artificial Intelligence*) nőtt át a gépi intelligencia-kutató intézetté (*Machine Intelligence Research Institute*, MIRI). A 2005-ben az Oxford Egyetemen alapított *Future of Humanity Institute* (FHI) a filozófiai oldal művelésére jött létre, a rivális Cambridge 2012-ben hozta létre a maga központját (*Centre for the Study of Existential Risk*). A 2014 óta működő *Allen Institute for Artificial Intelligence* (http://allenai.org/) a Microsoft-alapító Paul Allen nevét viseli, s egyetemekkel közösen valósít meg (ebben a pillanatban) négy kutatási programot.

A magam részéről zsákutcának és már a fogalmi alapvetés oldaláról is elhibázottnak tartom az erős mesterséges intelligencia programjából kinövő szuperintelligencia *(Artificial Superintelligence, ASI)* fogalmát,[17] és emiatt szükségszerűen értelmetlennek minden ráépülő narratívát.[18] A szuperintelligencia tétele és az alarmizmus találkozása ugyanis a diskurzusok olyan terét teremti meg, amely feltételezett fejlődési folyamatokból adott eséllyel következő, lehetséges jövőállapotok tipikusan negatív kimenetei felé tereli a párbeszédet, miközben a mesterséges intelligenciának számos működő és éppen bevezetés előtt álló alkalmazása és megvalósulása van, amelyekkel kapcsolatban egészen más típusú kérdések vetődnek fel. Természetesen ott, ahol a jelfeldolgozás összekapcsolódik műveletvégzéssel, ahol az automatizáció átrendezi a folyamatok feletti kontroll-láncokat, és újra és újra más szerepet ad a gépnek és az embernek, kérdések garmadája merül fel, amelynek esetében indokolt a lehetséges veszélyek számbavétele is. A veszélydiskurzusok elsődleges szintjét azonban a létező megoldásokban és alkalmazásokban fellelhető vagy elképzelhető hibaforrások azonosítása és tárgyalása jelenti, erős kitekintéssel az érlelődő, készülődő, bevezetés előtt álló rendszerek működése során felmerülni vélt problématípusokkal való mérnöki szembesüléssel. S miközben a szellemi közélet tipikusan az utópisztikus megközelítéseket utasítja el, sokkal több nehézséget okoz a gondolatrest disztópia, a veszély kiterjesztett, absztrakt, ismeretelméleti-filozófiai dimenzióinak összekeverése vagy összecsúsztatása a valóságos kihívásokkal.

De magára a pánikkeltésre is egyáltalán azért kerülhet sor, mert a mesterséges intelligenciában rejlő fejlődési és veszélypotenciál minden csatornán mértéktelenül eltúlzottan jelenik meg, az akadémiai publikációktól az ipari küldetésnyilatkozatokon át a fikciós irodalomig. Ugyan milyen világvége-veszélyt idézne fel egy minden eddiginél sikeresebb nyelv- vagy arcfelismerő program? A mesterséges intelligencia-alkalmazások világát össze kell boronálni a robotokéval, amelyek ebben a pillanatban banális (és kifejezetten „buta") mechanikai művelet-végző szerkezetek. A jelen valóságos robotjai a legkevésbé sem intelligensek, a legintelligensebb alkalmazások viszont nem lépnek át a szimbólumok teréből a valóságos fizikai terekbe, nincsenek „végrehajtó" kimeneteik, eredményeik emberi tevékenységek bemeneteit támogatják.

[17] Természetesen sokakkal együtt vélem így. Legutóbb Robert Trappl, az Osztrák Mesterséges Intelligencia Kutatóintézet (OFAI) igazgatója szögezte le: "*A mesterséges intelligenciának bizonyos dolgokat jobban kell csinálnia az embereknél, egy szuperintelligencia kialakítását viszont nagyon valószínűtlennek tartom*" https://sg.hu/cikkek/114196/a-programok-tobbet-tudhatnak-majd-az-embereknel (Letöltve: 2015. december 3.). S ahogy az Isten-érvek és Isten-ellenérvek története is megírható és feltárható, ugyanígy érdemes lehet egyszer végigszemlézni a gondolkodó gép mellett és ellen szóló érveket. Érdeklődőknek kiindulásképp javaslom Rovenszkij, Ujemov és Ujemova több mint félszáz éves, magyarul is megjelent könyvének utolsó fejezetét – *Miért nem lehetséges az elektronikus agy?* (Rovenszkij és társai, 1964, 206-216. o.).

[18] Elsősorban azt a komikus igyekezetet, ahogyan a szuperintelligencia létrejöttének időpontját próbálják valamilyen nevezetes évhez kötni. Annak, hogy 2020 és 2060 között mikor valósul meg az „átfordulás", ma már önálló könyvészete van. De ugyanígy értelmetlen halmaznak tűnnek azok a „receptkönyvek" is, amelyek listába gyűjtik, „mindössze" mire van szükség a szuperintelligenciához: nagyobb számításteljesítményre (!), az agy működésének maradéktalan megértésére és szimulálhatóságára, az evolúciós szcéna mesterséges előidézésére és a gépi én-tudat megteremtésére (Urban, 2015). Mindez elvileg és még inkább a „mesterséges élet" és a mesterséges „komplex élő rendszerek" programjától lehetne remélni, de ebbe az irányba sokkal kevesebben tapogatóznak.

De mi a baj az erős mesterséges intelligenciával?[19]

Az erős mesterséges intelligencia ellenfelei hagyományosan azzal érvelnek, hogy még *a „józan ész"* egyszerű kihívásait[20] megérteni és kezelni tudó masinák létrehozatala is teljesíthetetlen kihívás: a könnyen algoritmizálhatónak tűnő, valójában szinte végtelen számú függő háttérváltozó ismeretét és használatát igénylő banális kommunikációs helyzetek és életszituációk modellezésébe is rendre beletörik a géppel szimulált elmék bicskája. A „józan ész" alapján okoskodásra, párbeszédre és tanulásra képes gépek fejlesztésével kapcsolatban két irány bontakozik ki (Havasi, 2014). Az egyik még inkább logikai útvonalakra emlékeztető, szabály-alapú reprezentációkra tenné képesebbé az okos masinákat, a másik inkább még asszociatívabb, még inkább analógia-alapú, de a természetes nyelvre, gondolkodásra és viselkedésre jobban emlékeztető képességekkel ruházná fel őket.[21] Ezek az irányok is arra épülnek azonban, hogy a mesterséges intelligencia-rendszereken belül információs folyamatok zajlanak.

Valójában azonban egyetlen *jel-feldolgozásban verhetetlen gép sem intelligens*. Ahogyan azt John R. Searle a nyolcvanas években már meggyőzően kifejtette, a kimenetnek mi adunk és tulajdonítunk értelmet, a feltételezett jelentés komplexitása vagy autenticitása miatt látunk bele intelligenciát. A gépben nem zajlik információfeldolgozás, csak kódmanipuláció. A gép jel-műveleteket hajt végre, programjának megfelelően: a komputáció lé-

[19] Az (erős) mesterséges intelligencia dekonstruálása egyidős annak programjával. Alapvetően két ellenérv-típus jelenik meg ennek irodalmában: a *külső* (ontológiai-ismeretelméleti) és a *belső* (a technológiai korlátokra, költségekre és limitációkra figyelmeztető). Mivel ezek kézikönyv- és bevezető kurzus-szerűen jól ismert diskurzusok, csak olyan mértékben utalunk rájuk, amelyre szükség van a „szuperintelligencia"-narratíva megkérdőjelezéséhez. Az úgynevezett „erős mesterséges intelligencia" programjával szembeni érveket legújabban összefoglalja Gleiser (2014). Igyekszünk azokat a szempontokat összegyűjteni, amelyek a viták legújabb hullámában merültek fel, s amelyek saját álláspontunk kifejtését segítik.

[20] Feladni egy menü-rendelést a vendéglőben? Magyarázkodni az ellenőrnek a jegy-lyukasztás késedelme miatt? A megfelelő zoknit kiválasztani az alkalomhoz, a cipőhöz és a ruhához? Más ruhában, más hajjal, öregebben, de ugyanaz az ember van-e a fényképen? Hogyan működhet a gépi tanulás ilyen helyzetekben? Megtanítható-e a gépek következő generációja a természetes nyelv feldolgozásának (natural language processing, NLP) magasabb szintjére, hogy kiállja a „józan ész" próbáit? A palacsintakészítő robot (és mögötte a brémai RoboHow projekt) azzal kísérletezik, hogy az emberi instrukció és az emberi viselkedés megfigyelése és adaptálása pótolja a „gépi" józan észt. Evvel viszont valójában nyelvfelismerési irányba viszik el a kutatást, miközben elvileg arra kellene képessé tenni a kísérlet ember-robot csapatát, hogy egy újfajta tejesdoboz kinyitására és megfelelő dőlésszögű kiöntésének módjára személyes instrukciókkal vezesse rá az ember a gépet. (Ehhez képest a palacsintareceptek letöltése a WikiHow oldalról banális részfeladat. http://444.hu/2015/08/25/internetrol-tanul-palacsintat-kesziteni-egy-nemet-robot/)

[21] A logikai irány legismertebb sikertörténete Watsoné, az IBM műveltségi vetélkedőt nyerő robotjáé (kevesebben tudják, hogy a Cycorp mai napig működő, Doug Lenat nevéhez kapcsolódó Cyc projektje 1984 (!) óta fejleszti azt a rendszert, amely a józan ész világának tényeit (újabb kezdeményezésekkel párhuzamosan) egyre nagyobb repozitóriumokba rendezi. A nyelvi alapú rendszerek legismertebbike az 1999-es Open Mind Common Sense kezdeményezés. Az egyik legkorábbi online önkéntességre (crowdsource) alapuló projekt eredménye a ConceptNet, amelynek szöveges információbázisa és hatalmas tudásgráfja 17 millió tényt tartalmaz különböző nyelveken.

nyege a számolásteljesítmény. „Tudása" a műveletképesség, de nincs a műveletre magára vonatkozó „metaszintje" – mint amikor valaki elsajátítja, hogyan kell összeadni, kivonni, szorozni, osztani, de nem tudja, miért, mikor, minek az érdekében van szükség minderre. Ha mondják, és megkapja a számokat meg a műveleti utasítást, elvégzi, de maga soha nem kezdeményez, hiszen azt sem tudja, hogy alkalomadtán az egyes számok milyen mennyiségeket reprezentálnak, és mi felel meg nekik a valóságban.

Ezért is hívja a rendszerszemléletű információtudomány karizmatikus alakja, Alva Noë vagy az adattudós-jövőkutató Ken Bodnar a mai mesterséges intelligenciát *pszeudo-intelligenciának*. A legfejlettebbnek vélt programok, amely bizonyos teljesítményt (például emberi érzelmek rendkívül bonyolultnak tartott felismerését és azonosítását) lehetővé tehetik, egydimenziósak: nincsenek alárendelt princípiumaik, amelyek értelmet adnak a műveletnek, vagy a „miért" kérdésére válaszolni tudnának. Az imponálónak tűnő kimeneti oldal ellenére ezért ez mégiscsak a legostobább intelligencia (dumbest intelligence), hogy ismét csak Bodnart (2015a) idézzük.

Ken Goldberg a budapesti Brain Bar rendezvényén legutóbb az egyéves gyerek észlelési és manőverezési szintjére való felfejlesztést is rendkívül nehéz, óriási számítási erőforrást igénylő feladatnak nevezte. Yann LeCun az egér szintjét tartja nem meghaladhatónak, de Noë szerint még egy amőba is többet tud, mint a legintelligensebbnek mondott masinák. *„Az egyetlen sejtnek élettörténete van; környezetté alakítja át azt a médiumot, amelyben találja magát, és ezt a környezetet értékes hellyé szervezi. Tápanyagot keres. Megcsinálja magát – és azzal, hogy megcsinálja magát, értelmet visz az univerzumba."* A géppel ellentétben *„az amőbának … van információja – begyűjti és feldolgozza azt"*.[22] Akkor kezdjünk majd aggódni a szingularitás miatt, amikor az IBM olyan gépeket gyárt, amely egy amőba működését és tudatosságát produkálja.[23]

[22] Noë gondolatait remek összefoglalásban idézi Pesthy (2015). Érdekes módon az amúgy néha bizarr teológiai kiindulópontú információs elméletek is hasonló következtetésekre jutnak *„… a mesterséges intelligencia programok – bármily komplexnek és „intelligensnek" tűnnek is számunkra – csupán reprodukált, tehát semmiképpen sem kreatív információt jelentenek. A reprodukált információ létrehozásához nincs szükség saját szellemi tevékenységre, ezért ez a munka rábízható a számítógépekre"* (Gitt, 2004,164. o.).

[23] Mindez annak ismeretében erős kontraszt, hogy amióta a japán To-Robo szoftver magasabb pontszámot kapott egy főiskolai felvételi angol nyelvi részén, mint a japán diákok átlaga, azóta a világsajtót elárasztotta a szenzációs hír, hogy a mesterséges intelligencia teljesítménye már felülmúlja a főiskolásokét (http://blogs.wsj.com/japanrealtime/2014/11/04/artificial-intelligence-outperforms-average-japanese-high-school-senior/). Az amőba egyébként nem a valóságtól elrugaszkodott választás: miután egy alig félezer génnel rendelkező baktériumot már sikerült számítógéppel modellezni, az Open Worm projektnek (http://www.openworm.org/) egy sokkal bonyolultabb lény, a húszszor annyi génnel rendelkező fonálféreg (Caenorhabditis elegans), az első digitális élőlény megalkotása a célja, amelynek mozgása és viselkedése megfeleltethető a valóságos lénynek. Hasonló vágy fűti az első robot-orangután megalkotóját, Steve Grandet, aki virtuális környezetben kísérletezik valódi életforma létrehozásával. Teremtményeit, a Grandroidokat ne úgy képzeljük el, mint egy játék figuráit: virtuális neuronokkal, receptorokkal, enzimekkel és génekkel ellátott ágensek ők. Polipszájjal rendelkező, borotvált kutyára emlékeztető lények, akik *viselkedni* tanulnak, és környezetükkel úgy tartanak kapcsolatot, hogy az interakciókról maguk „döntenek" (Parkin, 2015). De ha egy Grandroid felnő, vagy ha a mesterséges intelligencia el is jut a fonalféreg szintjéig, hol volna a hároméves gyermektől? A kérdést Alison Gopnik, a gyermeki intelligencia kutatója teszi fel, annak a kétségének adva hangot, hogy a gép intelligencia valaha is eljuthat-e egy kisgyermek szintjére. George Dyson ehhez azt az érvet illeszti, hogy a valódi kreatív gondolkodás mindig analóg marad, soha nem válhat digitálissá (Brockman, 2015).

S ezért jellemzi egyúttal *az autonómia hiánya is* a mégoly erősnek és okosnak mondott masinákat is. Nincsenek céljai, nincs akarata, nincsenek referenciapontjai, amelyhez viszonyítva a magára, a környezetre és a már korábban kialakult jelentésekre való tekintettel kellene új jelentéseket létrehoznia, és annak alapján döntést hoznia.[24] Programot futtat le, de *az új program teremtése is program*, és mindig hiányzik mögüle a jelentés. Minden, szemantikusnak mondott rendszer a jelentéssel mint speciális, gépi nyelven leírható, kódolhatóvá tett objektummal foglalkozik: érzékeny rá, de nem „érti", nem teremti, nem használja, hanem a számára értelmezhető jellé lefordított formájában komputálja. Az intelligens rendszer nem egyszerűen létrehozza, módosítja, folyamatosan átalakuló alakzatokba szervezi a jelentéseket, hanem minden egyes észlelési ciklus kezdetén újrateremti, újraértelmezi azokat, és azonnal összekapcsolja a jövő valamilyen előre jelzett képéhez igazodó cselekvés tervezésével. Információja nem létezik önmagában, hanem egy időtengely mentén felépülő, csakis történetiségében értelmezhető, folyamatosan változó, komplex kognitív univerzum részeként, mozgásában. A gépnyelvre lefordított információ statikus, és a kognitív univerzum valamely szakaszának vagy elemének képesség- vagy kapacitás-korlátját segít lebontani.

A biológiailag adott információkezelő képesség meghaladása már régóta zajlik extraszomatikus információtechnológiai megoldások fejlesztésével. Amikor az optikai eszközök segítségével az észlelés tartományait a makro- és mikrotartományokban kiterjesztjük, senki nem nevezné „intelligensebbnek" a távcsövet vagy a mikroszkópot az emberi szemnél. Csak nagyobb felbontásúnak. Amikor memóriánk korlátait listákba és katalógusokba tárgyiasított módon küzdjük le, ezeket a dokumentumtípusokat sem mondjuk „intelligensebbnek", mint az embert, hanem egyenesen az emberi intelligencia meghosszabbításaként tekintünk rájuk. Amíg a számológép nem automatikusan, hanem emberi beavatkozást igényelve növeli meg a számolásteljesítményt, addig fel sem vetődik, hogy intelligens volna a masina, amely segít a komputációban. S amíg az immár automata gép csak számol és adatot dolgoz fel, addig sem tűnik intelligensnek – csak akkortól kezdve, amikor eltűnik a szemünk elől a műveletvégzés folyamata és logikája, és csak annak, számunkra jelentést hordozó kimenetét látjuk.[25]

Általánosítva: az már régóta nem kérdés, hogy a számítógép teljesítménye a számolásműveletekben és a számolásművelet-alapú másodlagos műveletekben (mint például az alakfelismerés vagy a jel-visszakeresés) megfelelően formalizálható körülmények között felülmúlja az emberi teljesítményt, és a technológiai fejlődéssel párhuzamosan a teljesítménykülönbség mértéke egyre nő. (Mint ahogy az is teljesen elfogadott kiindulópont, hogy az algoritmizálható agymunka gépesíthetősége minden esetben a magasabb értékhozzáadású szellemi műveletekhez szabadítja fel az emberi agyat). A gépi jel-feldolgozás

[24] Ez az argumentum valójában egy változata Hubert Dreyfus sokak által osztott „klasszikus" AI-ellenes érvének, hogy tudniillik az emberéhez hasonló intelligenciához elkerülhetetlen, hogy az azt hordozó gépi entitásnak legyen(ek) saját „teste(i)", amellyel a világhoz kötődik, és amelyre érvényesek annak törvényei, és legyen szocializációja, amellyel nemcsak képességei, hanem kontrollstruktúrái is felépülnek (Dreyfus, 1972, 1992).

[25] Emiatt – és a fogalom elterjedtsége miatt – „megtartjuk" az „intelligencia" kifejezést és az „intelligens" jelzőt, de avval a megjegyzéssel, hogy azt metonimikus értelemben használjuk, a megértést és a párbeszédet megkönnyítendő.

fejlődésének tehát többszörös tétje van: újabb és újabb területeken meghaladni a természetes intelligencia műveletvégző kapacitását, illetve annak alap-paramétereit (gyorsaságát, pontosságát, párhuzamos végezhetőségét, visszakereshetőségét stb.), és újabb és újabb területeken kiváltani a repetitív jellegű, alacsony érték-hozzáadású agymunkát.

A legendás és ikonikus Turing-teszt valójában mindkét célra alkalmatlan. Abból, hogy egy párbeszéd során embert látunk a velünk anonim módon kommunikáló gépben (még pontosabban: nem ismerjük fel a gépi intelligencia jelenlétét), nem következik sem az emberi intelligencia *utolérése*, sem *meghaladása*: legfeljebb feltételes, esetleges, ideiglenes és szituatív, tehát mindenképpen erősen korlátozott és szűk tartományra érvényes sikeres *szimulálása*. (És ez sem a gép autonóm teljesítménye, hanem a szimulációt irányító, programozó kutatóké).[26] S mivel a csevegés, a párbeszéd túlnyomórészt nagyon is magasrendű intellektuális művelet, az agymunka speciális fajtája, amelyen nincs mit „kiváltani", egyre jobban látszik, hogy a gondolkodó gép metaforája[27] itt (és még sok más helyzetben is) értelmetlen, félrevezető és diskurzusromboló.[28]

Bármilyen szuperintelligenciát tételezünk is fel, az mindvégig csak *az emberi értelem valamilyen (tehát nem minden) tartományban történő kiterjesztés*ét szolgálja. Nem általános/ge-

[26] Sokkal egyértelműbb mindez, ha Alex Tidemann doboló robotja, SHEILA van a fal mögött, aki neurális hálózata segítségével tanul meg egy dobszólót egyre jobban lejátszani, a dobos kézmozgását utánozva. Vajon ha a csevegés helyett az emberi és gépi dobolás közti különbség válik felismerhetetlenné, akkor SHEILA-t már intelligensnek nevezhetnénk? Aligha, mert nem tesz mást, mint pusztán reprodukálja az emberi dobjátékot. Minderre lásd https://www.youtube.com/watch?v=8SCWXbK42sc (Letöltve: 2015. november 10.) Vagy itt van a tübingeni kutatók friss fejlesztése, a fényképeket adott festők stílusában átalakító algoritmus. Képfelismerő erő és a stílusjegyekkel kapcsolatos mintázatképzés arányos ötvözete kell ahhoz, hogy az eredeti művekre emlékeztető módon utánozható legyen az egyes festők jellegzetes, saját stílusa. A gépnek azonban ettől nem teremnek saját stílusjegyeik. http://index.hu/tech/2015/09/02/egy_ora_alatt_kesz_a_legujabb_van_gogh-remekmu/ (Letöltve: 2015. október 8.)

[27] Az ötvenes-hatvanas években, még kibernetikai fogantatású narratívák részeként a „gondolkodó gép" nagyon hasznos segédfogalom volt, amellyel a szűk tudósközösség képes volt az érdeklődő közvélemény számára is érthető nyelvet alkotva népszerűsíteni a mindennapokba betörő számítógépek kultúráját és az abban rejlő lehetőségeket. Ma, egy sokkal nagyobb felbontású ismeretelméleti térben már nemcsak az érdeklődő közvéleményt téveszti meg, hanem magát a tudományos diskurzust is zavarja. A szuperintelligencia narratívája kerüli is a gondolkodó jelzőt, mint a tüzet: meg kellene ugyanis magyaráznia, hogyan lehet valami úgy intelligens, hogy nem gondolkodik, hanem csak programot futtat le.

[28] Jól mutatja mindezt 2014 nyarának nagy híre, hogy tudnillik forradalmi áttöréssel sikerült megalkotni „a gondolkodó gépet". A Eugene Goostmanre keresztelt program a zsűritagok több mint 33%-át megtévesztve hitette el magáról, hogy élő személy, egy 13 éves ukrán fiúcska. Csakhogy mindez részben a feltételek kijátszásán alapult (a fiatalság és az ukránság a hibaészlelési küszöböt szállította alacsonyabbra), a teljesítményt pedig nem gondolkodó gép, hanem egy chatbot, egy csevegésutánzó program adta le (amelyek közül a Cleverbot már 2011-ben is sokkal jobb eredményt produkált). Így is jellemző, hogy bombasztikus címek tudatták az „áttörést", a „mérföldkövet a számítástechnika történetében" – lásd például az egyik legelső magyar híradást a PcFórumon (http://pcforum.hu/hirek/16151/Forradalmi+attores+Megalkottak+az+elso+gondolkodo+gepet.html), hogy aztán sokkal csendesebben adják át a helyet a kritikus és elemző, a kísérletet a megtévesztések közé száműző kommentároknak (például http://bitport.hu/zavaros-siker-a-turing-teszten).

nerális intelligencia tehát, hanem csak azokra az aspektusokra terjed ki, ahol a paraméternöveléshez a számolás- vagy memóriateljesítmény növelésére van szükség. És nem a generális, hanem a szakosított irány tűnik adekvátnak: minél olcsóbb robotokkal minél jobban ellátni kicsi részfeladatokat (Brooks, 2003). Ez a „szűk mesterséges intelligencia" (*Artificial Narrow Intelligence*, *ANI*) programja, amelyet korábban a „gyenge AI" (*Weak AI)* fejezett ki.[29] Ám a korábbiaknak megfelelően még ezek is pszeudo-intelligenciát jelentenek. Az emberi intelligencia ugyanis számtalan esetben nem komputációval dolgozik,[30] és evolúcióját már régóta eszközeibe helyezte át. Sem a szuperintelligencia, sem egy szűk területen elért előrelépés nem olvasható a gépi elem öntudatra ébredéseként, csak az egyik eszköztípus teljesítménynövekedéseként.

Semmivel nem vagyunk fogalmilag előbbre akkor sem, ha Luciano Floridi megoldását választva szimuláció helyett *emulációról* beszélünk (Floridi, 2014). A számítástechnikai környezetekben jól ismert emuláció-fogalommal ugyanis arra utalunk, hogy egy új (jellemzően: a korábbiaknál többet tudó) eszközkörnyezetben minőségromlás nélkül felismerhető, kezelhető, „átemelhető" egy korábbi eszközkörnyezetre szabott funkció. A lényeg tehát a kompatibilitás megteremtése, és ez kétségkívül több, mint a puszta másolás vagy utánzás. Csakhogy a mesterséges intelligencia nem az emberi értelem emulációja, nem az elmét „ülteti át" gépi környezetre. Pusztán kognitív részfunkciók válnak gépi úton is elvégezhetővé, s ha emulációt keresünk, arra csak ott bukkanhatunk, amikor a magasabb szintű teljesítményre képes gépi megoldás a korábbi, egyszerűbb megoldásoknál megszokott felületen is elvégezhetővé válik.[31]

Az alarmizmus reménytelenül félrevezető mivolta abból fakad, hogy tarthatatlan fogalmi kiindulópontra építkezik. Nem veszi figyelembe, hogy ha számtalan művelet automatizálható is, nincs önmagában vett gépi intelligencia. A mesterséges intelligenciának (absztrakt gépi gondolkodásnak) ugyanis a mesterséges tudatosság (Artificial Consciousness) lenne az előfeltétele, amely viszont nem létezhet önazonosságra és önreflexióra képes ágens nélkül, amelynek identitástudata és állapot-tudatossága (state awareness) van. Az állapot-tudatosságnak párosulnia kell az állapotváltozás érzékelésére való képességgel, és

[29] Ide tartozónak tekintik a táblás játékok gépi rendszereit (Sakk, dáma, Scrabble, Backgammon, Othello), az IBM Watsonját, a Google fordítóprogramját, de akár egy spamfiltert is. Egy okostelefon olyan, mint egy „ANI-gyár", számos „okos" alkalmazással (Urban, 2015).

[30] Hanem például hipotézisek felállításával és tesztelésével, amelyeket aztán a valóságban is kipróbál. Michael Littman szerint elvi lehetetlenség a valóságnak a valóságnál gyorsabb szimulálása: „*It is a logical impossibility that these computers would be able to accurately simulate reality faster than reality itself*" (Littman, 2015)

[31] Vegyük a következő egyszerű okoskodást. A mesterséges intelligencia jelenlegi állapotában olyan, mintha azt kérdést tenné fel, hogyan tudnánk repülni, mint a madarak, ahelyett, hogy egyszerűen csak úgy fogalmazna: repülni akarunk. A formula, ahogyan a feladatot kijelöljük, hatalmas eltérésekhez vezethet a megoldásban. Még mindig nem építettünk a madarak repülését „lemásoló" légi alkalmatosságot (pedig sokáig kísérleteztünk csapkodó szárnyú gépekkel), ellenben sikerült olyan masinákat teremteni, amelyek gyorsabban repülnek bármely madárnál és eközben óriási terheket képesek szállítani. Az evolúció-formálta biológiai képesség és a technológia által alig néhány innovációs ciklus által megteremtett képesség közti különbségnek a mesterséges intelligencia-diskurzusokban való alkalmazhatóságáról szenvedélyes vita folyik, lásd például itt: https://news.ycombinator.com/item?id=10165586.

az állapotváltozást meghatározó tényezők azonosítására való képességgel. Mindehhez még a belső állapotra való szakadatlan referenciaképzés és az éppen aktuális külső állapot állandó összevetése is szükséges, egyedül ez lehet a jelentéstermelés alapja (Bodnar, 2015b).

Ezért nem jelenthetnek kiutat ebből a csapdából a „megváltásként" tálalt, a maga dimenziójában izgalmas eredményekkel kecsegtető *mélytanulás* (deep learning) koncepciókra épülő rendszerek, amelyek kognitív oldalról a konnekcionizmus, számítástudományi oldalról a neurális hálózatok korábbi paradigmáit próbálják ötvözni. A lényeg azonban ugyanaz, amin nem változtat az, hogy sokkal több adat alkotja a feldolgozás bázisát, hogy sokkal több réteg épül egymásra. Jól formalizálható, zárt jelrendszerekben, ahol a lehetséges válaszok az immár megfelelő számításteljesítmény birtokában algoritmizálhatóak, kiváló teljesítményre lehet képes egy deep learning technológia (nem véletlenül invesztálnak sokat ebbe az Internet nagyágyúi). A képfelismerés és a beszédfelismerés bizonyos területein, radikálisan leszűkített funkciótérben kétségkívül sokkal fejlettebbek, mint elődeik, de ez csak annyit jelent – Bodnart parafrazálva –, hogy úgy eredményesebbek, hogy intelligensebb módon buták. S bár valamivel ennél is többet ígér az az irány, ahol a világot és annak összefüggéseit felfedező kisgyermek mintájára a szabályok és okoskodási rutinok nem előre programozottak, hanem a tanuló rendszer önmaga konstruálja meg azokat,[32] az ágens tudatosság-állapotán mindez nem változtat semmit. Ahogy korábban a puszta algoritmussal afféle parancsutasításos hadvezérként, most a tanuló algoritmussal van, türelmes tanítóként, nélkülözhetetlenül jelen az emberi értelem – hiába ígéri Pedro Domingos, hogy megalkotható a Mesteralgoritmus, amely mindent meg tud tanulni, annak köszönhetően, hogy öt különböző tanulási forma elemeit egyesíti (Domingos, 2015).[33]

Az *aggregált gépi tanulásnak* (aggregated machine learning) nevezett irányzat sem más, mint a mélytanulás egyik lehetséges architektúrája. Ha például nem egy masinát tanítanak meg beszédfelismerésre, hanem sokat, mindegyik maga lép előre a megbízhatóan felismert és reprodukált nyelvi egységek számának gyarapításában, majd kicserélik, illetve egyesítik tudásukat, akkor sokkal rövidebb idő alatt növekszik meg a validált felismerő-készlet, mint ha egyetlen környezetben egyetlen rendszert fejlesztenének. Az aggregált tanulásnak is ugyanazok azonban a korlátai: felügyelt (supervised), hiszen a felismerés helyességét, az „elfogadást" csakis a természetes nyelvet jól ismerő emberek garantálhatják, másrészt a gyarapodó készlet mindig csak a megtörtént (aktualizált) és soha sem a lehetséges nyelvi aktusok és szókombinációk világára lesz érvényes (vagyis a hibaarány, még ha mindig csökken is, soha nem fogja elérni a 0%-ot). Funkcionálisan sokkal fontosabb a hiányzó néhány százalék bevadászásánál az, hogy ott, ahol a fizikai és/vagy szemantikai zaj ellenére hibamentes beszédfelismerésnek tétje van (például az ember-gép kommunikációban, ahol a gép a beszédutasítást végrehajtásba fordítja), ott ugyanaz az egy helyes felismerés ismét-

[32] Ennek az iskolának az esélyeit leginkább Gary Marcus kutatásaival illusztrálják (Knight, 2015).

[33] S ugyanígy reménytelen az érző számítógép, az érzelmi intelligenciát formalizálni tudó mesterséges intelligencia megteremtésének programja. Egy emberi arc érzelmi állapothoz társításának képessége alakfelismerő mélytanulással elképesztő távolságra van attól, hogy a mesterséges ágens maga rendelkezzen a viselkedés-szabályozásban szerepet játszó érzelmi komplexummal. Ráadásul a humán ágens esetében az idegrendszer működésének két oldala, az érzelmi és értelmi szerves egységet alkot: egy evolúciós munkamegosztás részeként fejlődtek ki. Egyáltalán nem szükségszerű, sőt egyenesen indokolatlan, hogy a gépi intelligenciának is reflektálnia kelljen erre a bonyolult kettősségre.

lődhessen nagyon sok, egymást követő szituációban. Ez pedig valójában inkább a pedagógiában kialakult perszonalizált tanulási modellek átültetését jelenti a digitális környezetbe (egyetlen, eszközére folyamatosan ráhangolódó emberre és egyetlen, rá fokozatosan „kalibrálódó" eszközre építve). Eközben a puszta jel-aggregációban már „kollektív gépi tudatot" látni felelőtlen és félrevezető. Ám remekül teljesítenek együtt, ha a gépi elem az emberi tudás felhalmozásában segít. Amikor az egyes emberi tapasztalat formalizáltan leírható, és ennek révén valóban összeadódik és közkinccsé válik.[34]

Hibriditás: a 'gépi' alapvető létállapota

A 'gondolkodó gép' fogalmával megragadott problématérben eltűnik az értelmezési tartományból az a tény, és nem elégszer hangsúlyozzuk, hogy nincs önmagában vett gépi intelligencia, az csak *hibrid (ember+gép) szerkezetben* tud megnyilvánulni, és az emberi mozzanat az elsődleges. Másrészt, mint azt a bevezetőben láttuk, ez nyit utat annak a pusztító és félrevezető diskurzusnak, amelyik a „*mikor éri utol és mikor múlja felül a gépi intelligencia az emberit*" ál-dilemmáját, vagy ennek még kakofóniába hajlóbb változatát, a „*mikor győzi le a gép az embert*" morális pánikba forduló kérdését zenésíti meg.

Ez utóbbit bátran hívhattuk volna akár Deep Blue-effektusnak is, mert a korábbi, óvatos próféciák után a regnáló világbajnokot, Kaszparovot legyőző számítógép adta meg a bátorságot boldognak és boldogtalannak, hogy világgá kürtölje a lefegyverzően igaznak tűnő, valójában mégis veszélyesen félrevezető szenzációt: *a gép legyőzte az embert*! Egy ideje inkább Watson-effektusként érdemes már beszélni róla, amióta az IBM műveltségi játékra kifejlesztett háromezer processzoros, tizenháromezer gigabájtos szörnyetegének sikere nyomán még a korábbinál is hangosabban rázzák a kereplőt a hagyományos és az online média leginkább szem előtt lévő felületein a vastag betűs, gondolatrest címek: *a gép ismét legyőzte az embert*! [35]

[34] Ennek tipikus eseteit nem a Wiki-világban, hanem például a növény- és állathatározás, a gomba- és ásványfelismerés vagy az ornitológiai észlelések felhalmozása környékén kell keresnünk. A kollektív mozzanat itt azt is jelentheti egyúttal, hogy a besorolások/leírások validálását sem egyetlen szakértő végzi, hanem sokak egy irányba mutató megerősítése. Emiatt nagyon ígéretesnek tartjuk például azoknak a mobil alkalmazásoknak a fejlesztését, amelyek a személyi használatú eszközökön formalizálják az egyéni azonosítások felvitelét, majd az egymáshoz közel kerülő eszközökön szinkronizálják a gyarapodó adatbázist, fokozatosan építve ki egy kizárólag a felhasználók által alakított és egyre pontosabb „virtuális tudástárat" – amelyet in situ igénybe vehet minden felhasználó, ha számára ismeretlen azonosítási helyzetbe kerül. Az ismeretek így felfogott mesh hálózata nem a hozzáférést, hanem a tudás aggregálását segíti, s mivel ez tipikusan terepen történik, a mobil platform remekül igazodik hozzá (ilyen kutatások a Szegedi Tudományegyetemen folynak, Bilicki Vilmos vezetésével). A gépi környezetben aggregált emberi tanulás talán legfontosabb és legperspektivikusabb területét az *egyedi betegségleírásokból kinövő gyógyító praxis* körül kell keresni, és nem, ahogy Shawn DuBravac (2015) gondolja, az automata járművek által majdan generált majdani közlekedési helyzetek aggregált naplózásában. Itt sem gépi tanulás történne ugyanis, hanem az esemény-típusokból emberek formálnának új, még veszélymentesebb közlekedést lehetővé tevő algoritmusokat.
[35] És korántsem szükségtelen tudatosítani, hogy innen már olvasók milliói kerülnek egy lépésnyi közelségbe a falfehérre vált arccal, kezdődő pánikban elsuttogott baljós kérdéshez: *jaj, mikor fogja majd uralma alá is hajtani!?*

Vajon hányszor és milyen formában kell felhívni a figyelmet rá, hogy mindenki megértse: nem egy gép „győzte le az embert”,[36] hanem *egy programozó és mérnökcsapatból, valamint egy beszédfelismerő, jelfeldolgozó és beszédszintetizáló modulokkal rendelkező gépből álló hibrid rendszer*. Egy emberrel szemben tehát – mint egykor a sakkfejedelem túloldalán is – egy páratlan művelet-végző sebességgel rendelkező masina és az azt megtervező, megépítő, programozó és strukturált tartalommal feltöltő szakértők tömege áll. Azért, mert a képernyőn a legendás Jeopardy-pult mögött „Watson” szimbolikus alakját látjuk két mesterszintű játékos között, el is felejtjük, hogy valójában emberek ármádiáit kell mögé képzelni, minden másodpercben? Ismételjük meg, hogy mi is történt valójában: emberi agyak egy tekintélyes, összekapcsolt csoportja alkotott és töltött fel tartalommal egy olyan szerkezetet, amelyiknek a szabályozott, kimenő jelei meghatározott keretrendszerben a sikeres emberi problémamegoldásra emlékeztető illúziót keltettek.[37]

Sajnos még a legigényesebb jövőkutatási irodalomban is visszaköszön ez a kettősség. Hiába vizionálja – helyesen – ember és gép új civilizációs minőséget eredményező fúziójának lehetőségét Ray Kurzweil, amikor szerinte az emberi agy „biokulturális” dimenzióban megragadható tudás- és képességpotenciálja ötvöződik az emberi tevékenység érdekében csatasorba állított okos eszközök nagyobb reprezentációs és műveletvégző kapacitásával, sebességével és információmegosztó képességével. Mindez a korábbi „szimbiózisoknál” kétségkívül szervesebben forrasztja össze, utalja jobban egymásra, teszi nehezebben elkülöníthetővé a gépi és emberi komponenst (Kurzweil, 2013). Csakhogy téved, és ingoványos talajra kerül, amikor ehhez azt tartja szükségesnek, hogy „*a számítógépek is elérjék a legmagasabb emberi intelligencia szintjét*” – a fenti szimbiózis minden további nélkül működhet enélkül is (sőt működhet nagyon hatékonyan).[38]

Ez a tétel tehát jól működik, amikor a mesterséges intelligencia fejlesztése a kutatói kihívás, de ismeretelméleti szempontból zsákutca. A hibrid rendszereknek ugyanis éppen az a lényege, hogy *minden komponens a sajátlagosan rá jellemző képességmezőben adja le a teljesítményt, és ahol gyengébben teljesít, ott a munkamegosztás részeként átengedi a terepet*.[39] Az ár,

[36] Ráadásul miféle „győzelemről” is van szó? A hibrid rendszernek egy játékban aratott győzelméről, amely a játék természetes kimenete, akár humán-humán, akár gép-gép konstellációban. (Amikor az egyik sakkprogram legyőzi a másikat, akkor a gép legyőzi a gépet? Nem, az egyik programozó csapat bizonyul eredményesebbnek a másiknál). Csakhogy a semleges hangulatú, leíró jellegű „játékgyőzelem” a hétköznapi konyhanyelvben disztópikus jelentést ölt: ha a játékban a gép legyőzi az embert, akkor az már az előjele annak, hogy evolúciósan is felülmúlja, kiszorítja életteréből, aláveti, uralkodni fog rajta. És egyszer már csak ezen az absztrakciós szinten jelennek meg a gondolatok.

[37] Félreértés ne essék: Watson produkciója fantasztikus és érdemdús mérnöki teljesítmény, a mesterséges intelligencia kutatásának egyidejűleg több, régóta mozdíthatatlannak tűnő pontján ígér előrelépést vagy áttörést. Elismerés és gratuláció illet mindenkit, aki a siker körül bábáskodott. Kritikánkat az interpretációval kapcsolatban fogalmaztuk meg, ami ráadásul hangsúlyozottan nem is a fejlesztőknek, hanem a bulvárosodó közbeszédnek köszönhetünk.

[38] S realisztikus jövőképe és tárgyilagos megközelítésmódja ellenére ezért csúszik át Kurzweil a science fiction irányába az elme fizikai testbe való áthelyezhetőségének tételével (amelyhez nemcsak az emberit utolérő gépi intelligenciára, hanem a testtel összekötött egyedi idegrendszerek működésének szimultán digitális replikációjára is szükség lenne).

[39] Az információs műveletvégzést támogató pre-digitális hibrid rendszerek is ezen az elven működtek. Az állati *jelző-riasztó rendszerek* (legyen az bányarigó vagy őrzőkutya) *kereséstámogató hibridek*

amelyet az így jelentkező koordinációs szükségletek ellátásáért fizetni kell időben, figyelemben és energiában, messze alatta marad a teljesítménynövekedésből származó előnyöknek. Ennek a folyamatnak (illetve sajátos, dinamikus egyensúlynak) a tervezésekor szokták az ember-számítógép viszonyra az *orkesztráció* (orchestration) kifejezést használni (Burgess, 2015, 341-345. o.), az egy időben leadott közös teljesítményre és a párhuzamos, illetve a sorba rendezett folyamatokra utalva. Három szakasza a *specifikáció*, a *kollaboráció* és a *koordináció*.

A gépi oldal fejlesztése tehát fontos terrénum, de csak egyike a négy alapvető kérdésnek: *a hibrid rendszer emberi oldalának fejlesztése* a második, *az emberi és a gépi együttműködésének kérdése (az interfész)* a harmadik, s az egész rendszerfejlesztésnek értelmet és célt adó *teleológia* a negyedik. Ez az a négy problémasík, amit minden pillanatban szem előtt kell tartanunk, s amely egyfajta osztályozási elvként működik: vajon az éppen mérlegre tett vagy vitatott elmélet a négy közül melyik szinten fogalmazódik meg, kellően komplex-e, értelmezhető-e egyidejűleg mindegyikre?

Foglaljuk össze tehát lista-szerűen a négy sík leglényegesebb jellegzetességeit, ahol lehet, szembesítve azt a meghaladni kívánt vagy vitatott nézőpontokkal.

1. A gépi oldal, a mesterséges intelligencia kutatói számára nem az emberi intelligencia elérése a valódi kihívás, hanem az, hogy *a mesterséges komponensre eső feladatok a legmagasabb szinten teljesüljenek, illetve szakadatlanul új feladatok váljanak teljesíthetővé.*[40]
2. Az emberi oldal fejlesztése részben szociokulturális és oktatástechnológiai, részben pedagógiai, részben pszichológiai, részben pszichofarmakológiai kérdés,[41] amelynek a következő időszakban sokkal nagyobb szerepet kell kapnia a diskurzusokban[42] (s amelyre a kiborg-témakör rövid tárgyalása kapcsán majd vissza is térünk röviden).

(szarvasgombavadász disznó, nyomkövető vagy szaglókutya) vagy *üzenetküldő rendszerek* (postagalamb, futárkutya) például jól kiegészítették az érzékszervi és fizikai kapacitásában korlátozott emberi közösségeket, s a táplálás és – szükség esetén – betanítás költségei jóval alatta maradtak az együttműködésből származó hozadéknak. Jaron Lanier is észreveszi, hogy nemcsak „az emberi és a gépi", hanem a „predigitális és a digitális" legjobb vonásainak szintézise az elfogadható irány. A legfrissebb kognitív tudományi eredmények is meggyőzően üzennek ennek a sajátos egyensúlynak a meglétéről. Sikerült például feltárni és bebizonyítani, hogy az adatok (és az általuk hordozott ismeretek) elmentése „külső hordozóra" felszabadítja a memóriánkat a következő hasznos információ befogadására – a „mentés" általi memória-tehermentesítés megnöveli a kognitív kapacitásunkat (Storm és Stone, 2015).

[40] Richard Boyd (2015) így fogalmaz: „*Hogyan érhetjük el a megfelelő egyensúlyt az emberi és automatizált között, hogy a kimenetet optimalizáljuk vele? Az emberi és a gépi intelligencia milyen kombinációja segíti legmegfelelőbben a legégetőbb problémák megoldását?*"

[41] Itt részben a *nurture*, a tudatos nevelési környezet révén elérhető intelligenciafejlődés eszközvilágára gondoljunk, részben a memória, koncentráció stb. erősítését (ideiglenesen) stimuláló szerekre. Érdekes paradoxon, hogy a sikeres ember-gép kommunikációhoz olykor az emberi oldal „lebutítása" szükséges a gép szintjére (például primitív szintaxis a gép beszédértéshez), és nem a gépi oldal további „felokosítása".

[42] Az ember-centrikus megközelítés azonban nem jelenti azt, hogy akár az unalomig ismert luddita vagy az újkeletű „*humánsoviniszta*" megközelítésmód a legcsekélyebb mértékben is indokolhatóvá válna. Ez utóbbinak Jaron Lanier, a virtuális valóság atyja a legismertebb képviselője. Egy 2014-es díjátadási beszédében több egyéb tárgykör – például a nagy adat (Big Data) veszélyei – mellett egy olyan "új humanizmust" hirdetett meg, amely korábbi, a kibernetikai totalizmust (*cybernetic totalism*) elutasító deklarációi után azt a helyes állítást, hogy „az ember több a gépeknél és az algoritmusoknál", összekapcsolja avval a helytelen következtetéssel, hogy emiatt a „mesterséges intelligencia

Emellett tömegesen „kiadhatóak" olyan feladatok az emberi elmének, amelyek megoldásában jobban teljesít, mint a gép – ez az alapja az *emberi számítástechnika* (human computation) irányzatának.[43]

3. A két oldal közös felelőssége a megfelelő koordinációs minőség megteremtése a gépi és az emberi komponens között, hogy a hibrid rendszer a legnagyobb teljesítményt adhassa le.
4. S végül az egész rendszer működésének *a cél-vezérelt ko-evolúciós szemlélet* ad értelmet és megalapozást, nem pedig a mesterséges intelligencia-rendszerek öncélként felfogott „felgyorsítása" vagy az exponenciális ugrást lehetővé tevő „önfejlesztő üzemmódra" állítása. Mert e kettő mögé is fel kell tenni a kérdést, hogy mindez *mit szolgál a rendszer-egész szempontjából*?

HCI, humatika, kognitív infokommunikáció – visszaút a géptől az emberig

Első ránézésre a fenti felsorolás harmadik területével, a koordináció/interfész kérdéskörrel foglalkozik, ráadásul roppant kiterjedt módon és definíciószerűen, a HCI (Human-Com-

elutasítására" is szükség van. „*The new humanism is a belief in people, as before, but specifically in the form of a rejection of artificial intelligence*". S még ha a szövegkörnyezetből ki is derül, hogy valójában a „mesterséges intelligenciához kapcsolt túlzott elvárásokkal" és az eluralkodó „fantáziavilággal" szemben foglal állást, és nem magát az algoritmusok világát tenné ad acta, evvel az egyetlen beszédével a zászlóvivőjévé és hivatkozási alapjává vált a kiber-utópizmust felváltó másirányú egyoldalúságnak. A kibernetikai totalizmus kritikájára lásd http://hplusmagazine.com/2014/07/07/jaron- lanier-on-transhumanism/, az új humanizmusról mondottakra, a teljes beszéd közlésével: http://www.friedenspreis-des-deutschen-buchhandels.de/819335/. Az ember-központú szemléleti fordulat hasonlít ahhoz, amin (egy sokkal szűkebb mezsgyén) a humanitárius munkát végzők átmentek. Fel kellett ugyanis ismerniük, hogy hiába állítható elkötelezett tevékenységük szolgálatába számtalan nagyszerű technológia, a drónoktól a multiplayer játékokig. A digitális humanitarianizmusnak tudnia kell, hogy az előretekintő politika-alkotás és a résztvevők és vezetőik felkészültsége és felvilágosodottsága legalább annyira meghatározó (ha nem még fontosabb), mint a felhasznált technológia (Meier, 2015).

[43] Az immár szakosított intézettel (Human Computation Institute http://humancomputation.org/) és transzdiszciplináris folyóirattal (a Human Computation 2014 októbere óta jelenik meg http://hcjournal.org/ojs/index.php?journal=jhc) rendelkező irányzatot Luis von Ahn 2005-ös disszertációjára szokás visszavezetni. Ebben a guatemalai származású kutató és fejlesztő először fogalmazta meg alaptézisét a nagy tömegben önkéntes műveletvégzésre (crowdsource) fogott, emberi elmében rejlő komputációs erőről, amely megfelelő gépi környezetben olyan problémák megoldására lehet képes, amelyre külön-külön nem volnának elegek (nem utolsósorban a játékosítási és motivációs technikáknak köszönhetően). Azóta számtalan sikerprojektet sikerült tető alá hozni: az ExeWire-rel neurontérkép készül (közel 200 ezer önkéntes segítségével), ígéretes az Alzheimer-kór kutatását segítő alkalmazás (WeCureAlz.com), ezekről részletesebben lásd Michelucci és Dickinson (2016). Tegyük hozzá a teljesség kedvéért, hogy a fogalom születésének évében, 2005-ben jelent meg David Alan Grier könyve a „humán komputerekről": azokról az emberekről (főleg nőkről), akiket a huszadik század fordulóján iparszerűen alkalmaztak arra, hogy nagy tömegben végezzenek el számolásműveleteket (Grier, 2005). A számítógép megszületése ezeket a humán komputereket váltotta ki. De ma ott, ahol a gépek tudása véget ér, vagy az ember hatékonyabb lehet (például: mintázat-felismerés és osztályozás, kreatív absztrakció) egy magasabb szinten (és az egykori irodamérethez képest a feladatok sok kis részre bontásának, a microtaskingnak köszönhetően sokkal nagyobb tömegek bevonását lehetővé tevő módon) térnek vissza.

puter Interaction), az ember-számítógép kapcsolat (ESZK) tudománya.[44] Csakhogy a HCI elsősorban nem a ko-evolúcióra és annak teleológiájára koncentrál, magas absztrakciós szinten, hanem az adott funkcióra létrehozott gépi környezetek alacsony absztrakciós szintű, felhasználó-orientált hatékonysági és ergonómiai fejlesztésére.[45] Jól tükrözi ezt a HCI saját történeti identitása is: a fogalom első, szórványos előfordulását 1975-re teszik, a diskurzusnyitó monográfia (Card et al., 1983) megjelenését alig több mint 30 évvel ezelőttre. Eközben a ko-evolúciós gondolat évtizedekkel korábban, a hatvanas évek elején, nagyon is jövőtudatos formában vetődött fel: a Világháló egyik korai építője, J.C.R. Licklider vezette be a „*szimbiotikus rendszer*" kifejezést az ember-számítógép kapcsolatra (Licklider, 1960, 1965).[46] A számítógép-komponens feladata nála nem az emberi intelligencia utolérése vagy meghaladása, hanem *augmentációja* (feljavítása), és a szimbiotikus ember-gép rendszer közös célfüggvénye a (közös) intelligencia *amplifikációja*, felerősítése (intelligence amplification, IA). („*Az ember jelöli ki a célokat, szabja meg a feltételeket, és végzi az értékelést. A számítógépek végzik a rutinizálható munkát*"). Lickliderrel közel egyidőben szintén az augmentáció fogalmában ragadta meg a kihívást Doug Engelbart: kifejezetten a komplex szituációk megértésének és megoldásának eszközeként (Engelbart, 1962). Sok évtizeddel később John Thackara egyenesen a HHI (*Human-Human Interaction*) kifejezésre való váltást javasolta (Thackara, 2005) annak az érzékeltetésére, hogy *még ott is a humán oldal a lényeges, ahol dramaturgiailag fontos szerepben van jelen a gépi komponens* – mint például a hálózatokban. Ahogy Brynjolfsson és McAfee (2014) fogalmaznak: a brutális processzorerőnek az emberi leleményességgel kell párosulnia.

Engelbart jellegzetes tipológiája, amely a humán oldal „feljavításának" lehetséges formáit vette számba, mai napig érvényesen jelöli ki az irányokat:

- gyorsabb megértés;
- jobb megértés;
- a megértés megfelelő szintjének elérése korábban túl komplexnek tűnő probléma esetén;
- gyorsabb megoldás;
- megoldás találása olyan problémákra, amelyek korábban megoldhatatlannak tűntek.

Jól látszik, hogy *a megértés a humán komponens célfüggvénye, a megoldás a hibrid rendszeré.* Mindez semmilyen formában nem implikálja a gépi oldal egyoldalú kiemelését. Sokkal inkább következik belőle az, hogy ha a hibrid rendszer gépi komponense „előreszalad", nagyot ugrik teljesítményben, akkor az emberi oldal felzárkóztatása, „méretezése", a megfelelő kapcsolat kialakítása válik meghatározóvá a harmonikus fejlesztés érdekében. Ve-

[44] Használják még az ESZI (Ember-Számítógép Interakció) és – főleg az ergonómusok – az EGK (Ember-Gép Kapcsolat) rövidítést.

[45] Minderre egyre népszerűbb az interakció-tervezés (*Interaction Design*) kifejezés is, amelynek célja a zavaró mozzanatok kiküszöbölése és a pozitív élmény elősegítése és fokozása (Lásd például Preece és társai (2015) tíz év alatt immár negyedik kiadásban megjelenő tankönyvét).

[46] Mivel Licklider és munkássága kevésbé volt látható, ezért a gondolatot sokáig az üzleti számítógép-fejlesztés egyik pionírjához, John Dieboldhoz kötötték (Diebold, 1969), aki azt javasolta, hogy minden lényeges kérdést az „ember és számítógép" egységeként közelítsünk meg.

gyük észre például, hogy gyorsabb és jobb megértés érdekében számos technikát állíthatunk csatasorba, amelyek *nélkülözik a gépi elemet*.

Az újabb generációs infografikai szcéna például pontosan arról szól, hogy a numerikus nehézbombázással átláthatatlanná tett vagy túl sok változó egyidejű figyelembe vételét igénylő tudás-mikroverzumok művészi és professzionális megjelenítésével gyors, holisztikus, elmélyült és élményszerű megértést biztosíthatunk. Az augmentáció forrása lehet egy vagy több másik elme nagyszerű teljesítménye is, amely sok esetben gépi és automatizált adatgyűjtések és feldolgozások kimeneteit fordítja át használható tudássá. Az óriási adattömeg szemantikus operátorok segítségével elvégzett tartalomelemzésén alapuló előrejelzéseknél sikeresebbnek bizonyulhat a „tömegek bölcsességének" igénybe vétele.[47] Az is igaz ugyanakkor, hogy a komplex ismeretrendszerek vizualizálásának technológiája vagy az online önkénteseknek teremtett platform már ismét a gépi oldal felé mutat: a vonatkozó számítógépes képszerkesztő és megjelenítő alkalmazások, animációs lehetőségek és workflow eszközök amplifikálják az új értéket létrehozó emberi tevékenységet.

Ha a HCI funkcionális történetére vetünk pillantást, azt látjuk, hogy létrejöttekor és első szakaszában a meghatározó terület a *biztonságkritikus rendszerek* világa (erőművek, repülés) volt. Ebben a szakaszban a kisebb és nagyobb rendszerzavarok elemzése nyomán rendre kiderült, hogy a kiküszöbölendő hibák nagyrészt *az emberi komponens gyengeségére* (elfáradás, figyelemhiány, rossz döntések sorozata, előírások és karbantartási feladatok figyelmen kívül hagyása, frissítés elmaradása stb.) vezethetőek vissza, részben az evvel nem számoló *tervezés hibái*: a gépi oldal jellemzően megbízhatóan teljesített.

A HCI történetének második szakaszában a tömegfelhasználás került a középpontba: diszciplináris oldalról az ergonómia és a design (s annak révén az artisztikum, a művészi kreativitás és invenció). „Segédtudományi" oldalról a viselkedéskutatás, fejlesztési oldalról az új beviteli és visszacsatolási eszközök. Praktikusan: a felhasználók szempontjainak figyelembe vétele már a fejlesztési szakaszban – olyan módszerek felhasználásával, mint például a pszichológiából importált Q-metodológia.[48]

Amellett érvelek, hogy elérkezett az idő a HCI harmadik szakaszára, amelyet legszívesebben *koordinált jövőtudatos szakasznak* neveznék. Az elnevezés elsősorban azt tükrözi, hogy a második szakaszban (és evvel párhuzamosan: az Internet-korszak utóbbi húsz évében) az eszközfejlesztői, -gyártói logika határozta meg a gépi oldal fejlődését, és az akadémiai jellegű HCI-kutatások is jórészt elszigetelt, egymással versenyben álló hardver- és szoftverfejlesztő üzleti vállalkozások innovációs segédcsapataként üzemeltek. Mostanra alakultak ki az előfeltételei annak, hogy megkezdődjön egy egészen más elvekre és kiindulópontokra épülő kutatási szakasz, amelyben a tervezői gondolkodás (design thinking)

[47] A négyéves kutatásra épülő Good Judgment Project több ezernyi önkéntese a geopolitikai előrejelzésben meglepő pontosságot ért el, és lekörözte minden mesterséges intelligencia-alapú megoldás eredményességét – kiemelve annak fontosságát is, hogy ahol működik, lehet az első gondolatunk a humán intelligencia a gépi helyett (http://www.goodjudgmentproject.com/index.html).

[48] A pszichológus William Stephenson (1902-1989) által kifejlesztett módszertan a „szubjektivitás" mérésére szolgál, ennek HCI-re alkalmazott változata, a HCI-Q az iteratív design-ciklusokban használatos, amikor a felhasználók és mások szempontjai felől veszik szemügyre a fejlesztett eszköz személyes szignifikanciáit (O'Leary, 2013). A kérdéskör 15 oldalas bibliográfiáját Charles H. Davisnek köszönhetjük (Bibliography of Qmethodology in audience research). http://www.ryerson.ca/~c5davis/Q-studies-of-audiences.pdf

a jövő-írástudással (futures literacy) találkozva az alapoktól gondolja újra, s kezdje meg újraépíteni, majd orkesztrálni azt, ahogyan az emberi elme és az azt amplifikáló gépi intelligencia összekapcsolódik.

Hans-Dietrich Kreft az ezredforduló környékén „humatikának" nevezte el azt az irányzatot, amely még a szenzorok, szoftverek és autonóm rendszerek gépi világát is kizárólag az emberi környezetbe ágyazott, illetve az emberi életet gazdagító mivolta révén tartja tárgyalandónak.[49] S mivel az ember bármilyen technológiai rendszer lényegi középpontja, mindvégig a tudás interoperábilis fizikai jellegzetességeinek megértése, leírása – praktikusan: cserélhetőségének és mérhetőségének megteremtése – lesz a lényeges mozzanat (Kreft, 2003).

2015-ben, legnagyobb örömünkre, számos olyan könyv is napvilágot látott, amely Kreft kiindulópontjait viszi tovább, immár a legfrissebb technológiai környezetre alkalmazva. David A . Mindell a jelen technológiai rendszereinek középpontjában álló gazdag emberi jelenlét (*rich human presence*) tézisét helyezi szembe a mesterséges intelligenciával párosult robotok autonómiájával kapcsolatos tévképzetekkel (Mindell, 2015).[50] John Markoff pedig egyenesen odáig megy, hogy emberi és gépi koevolúciójában nem a mesterséges intelligenciára (AI), hanem a humán intelligencia feljavítására (intelligence augmentation, IA) helyezi a nagyobb hangsúlyt: a gépi elemnek ez utóbbit kell szolgálnia, hogy önmagunk, emberi mivoltunk újfajta „design"-jával készüljünk a jövőre (Markoff, 2015). És ez nemcsak biológiai, hanem társadalmi létünkre is igaz: Kentaro Toyama központi tézise, hogy a technológiai fejlődés mit sem ér, ha annak vívmányai nem vonhatóak be az akut társadalmi gondok megoldásába. Ezt a változást pedig az emberi bölcsesség és tudás irányíthatja csak, és ennek a kapacitásnak a megnövelése, és nem a helyettesítése a mesterséges intelligencia köré épülő technológiánk küldetése (Toyama, 2015). Kulcskérdéssé az emberi elme fejlesztése, a tudós- és tudásközösségek együttes szellemi teljesítményének (Chris Anderson szavaival: raj-intelligenciájának) fokozása válik, és mindebben felértékelődik a Turing és Neumann-típusú géniuszok szerepének megtalálása is (Hsu, 2015; Colvin, 2015).[51]

És ez az a pont, ahonnan a magát *kognitív infokommunikációnak* (Coginfocom) nevező irányzat létjogosultsága is leginkább megérezhető és megérthető.[52] A 2006 óta, javarészt

[49] Cége, a Humatics Corporation (http://site.humatics.com/) ezekre az elvekre építi profilját, amellyel piaci szereplők számára nyújt tudás-szolgáltatásokat.

[50] Mindell egyúttal azt a rendkívül fontos szempontot hangsúlyozza, hogy intelligens és önvezérlő mechanikai eszközeink az emberi tevékenység határait terjesztik ki extrém környezetekben. Basulto (2015) pedig méltán figyelmeztet arra, hogy *a robotok evolúciójának* (robotic evolution) is éppen ez ad értelmet és célt: életterek benépesítése, a megismerés határainak kiterjesztése, és az emberi cselekvés korlátainak megszüntetése. Frank Tipler pedig bátran ki is mondja: a mesterséges intelligencia küldetése nem más, mint a hozzájárulás az emberiség megmentéséhez és az űr kolonizációjához (Brockman, 2015). Ehhez képest az *evolúciós robotika* (evolutionary robotics), amely darwini trükköket alkalmaz a gépi intelligencia fejlesztéséhez, „csak" egyike a versengő fejlesztési paradigmáknak.

[51] Brian Eno egyenesen azt ajánlja nekünk, hogy képzeljük el az emberi társadalmat, mint a legerősebb szuperszámítógépet, és magát a globális civilizációt, amely mesterséges intelligenciaként ölel körül minket (Brockman, 2015). Más kérdés, hogy evvel a huszadik század harmincas éveitől íródó „világagy-diskurzushoz" kanyarodunk vissza.

[52] Érdekességképpen említsük meg, hogy a nagy szóalkotó, Stanislaw Lem lexikonában a fejlett mesterséges intelligenciára használt technoevolúció (technoevolution) és intellektronika (intellectronics) mellett született kifejezés a kognitív teljesítményt növelő technológiákra is: a cerebromatika (cerebromatics).

magyar kutatók kezdeményezésére elindult interdiszciplináris útkeresés[53] abból indul ki, hogy a komplex technológiai- és médiatérben minden kommunikációs aktusok révén történik, emiatt a kognitív tudományban felhalmozott tudást kell párosítani a mérnöki alkalmazásokkal (Baranyi, 2012). Másképpen: a jelenlegi eszközvilágot a lehetőségek maximumáig kell a kognitív folyamatokhoz illeszteni, hogy a kommunikáció hatékonyságának megnövelésével javuljon a rendszerteljesítmény is. Evvel a normatív, metaelméleti szint újra összetalálkozik a praktikus fejlesztési filozófiákkal: ideje hát kibővíteni a valójában leegyszerűsített ember+gép modellt, mert a szemléleti változás szükségességének legmélyebben fekvő okai csak a teljes kapcsolatrendszer ismeretében tárhatóak fel.

Ember és számítógép – út a kiterjesztett modell felé

Eddig ember-gép szimbiózisról beszéltünk, de az önmagában vett „hibrid" rendszerről, absztrakt módon beszélni értelmetlen: hiszen az ember-gép együttműködés okát és értelmét kizárólag konkrét helyzetekben, annak megnyilvánulásakor tudjuk értelmezni. Ember és mesterséges intelligencia párosáról tehát ideiglenesen, egy bonyolultabb építmény „alsó szintjeként" beszélhetünk, amely kizárólag a ráépülő elemekkel együtt lehet érvényes.[54]

(EMBER + AI)

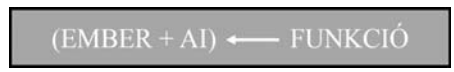
A gépi intelligencia csatasorba állítása kivétel nélkül olyan élethelyzetekre reflektál, ahol az augmentáció igénye felmerül.[55] Minden monoton számolásművelet értelme az eredmény lehetőleg azonnali „lefordítása" valamilyen valóságos problémára (nem véletlen, hogy a számítógépek tömeges felhasználása a vállalatok bérszámfejtési osztályain és a pénzintézetekben indult meg). Az „emberi komponens" kizárólag cselekvő emberként van jelen a rendszerben,[56] emiatt a cselekvésekkel (illetve azok összességével, a viselkedéssel) ki kell bővíteni a modellt. Ha a mesterséges intelligencia hozzájárulása a cselekvés hatékonyságát fokozza, akkor *a különböző hibrid rendszerek különböző cselekvéstípusokhoz rendelődnek hozzá*. Másképpen: a hibrid rendszerek egyúttal funkcionális rendszerek, ahol a közös működés értelmét és „behuzalozását" adott funkciókból levezethető feladatok és az ezekre reflektáló cselekvések vezérlik.

[53] 2015-ben már a 6. nemzetközi konferenciát rendezték a tárgyban, egyre növekvő érdeklődés mellett. Az érintett tárgyköröket, a fő kategóriákat jól tükrözi a végleges program. (http://coginfocom.hu/conference/CogInfoCom15/) Az irányzat intézményesedését a felsőoktatás képzéskínálatában való erőteljes megjelenés is jelzi (szakirányok, doktori témakiírások formájában).

[54] E modell egy korai, fejletlenebb változat (Z. Karvalics – Juhász, 2008) alapos kibővítése.

[55] Hangsúlyozzuk ugyanakkor, hogy ennek nagyon fontos járulékos következménye az az állítás, hogy sok élethelyzetben nincs szükség augmentációra. Így minden univerzális augmentációs próbálkozás eleve feleslegesen túl nagyra tágított teret kíván betölteni.

[56] Vannak első pillantásra extrémnek tűnő kivételek, például a nyugalomban lévő vagy alvó ember életműködésének paramétereit monitorozó szenzorrendszerek, amelyek automatikusan továbbítják mobil adatátviteli moduljuk segítségével a kimenő jeleket az azokat értelmező feldolgozó központba. Csakhogy ezek a megoldások valójában az orvosi cselekvés hatókörét kiterjesztő megoldások, a zavarok korai észlelésének funkciójával, amely a működés értelmét adja.



Mindez tárgyszinten is azt jelenti, hogy a funkció határozza meg a hibrid rendszer gépi komponensének mineműségét: teljesítményét, hangolását, kimenetének szabályozását, az emberi komponenssel való kapcsolódásának mikéntjét és felületét, az interfészt. Emiatt az összefüggést akár így is felírhatnánk:

A hibriditásnak értelmet adó funkció tehát a cselekvés jellege felől határozza meg a támogatási szükségletet. Emiatt aztán azonnal két ágra szakad az informatikai univerzum is: az információs viselkedést támogató rendszerekre[57] és a fizikai jellegű cselekvésekhez (mozgás, objektumra irányuló mechanikai műveletek) kapcsolódó rendszerekre.[58] Ez utóbbi esetben az ember-információs gép hibrid kiegészül egy mechanikus komponenssel is,[59] amely az anyagtudomány és robotika világával eredményez újabb és újabb fúziókat (a beágyazott rendszerek újabb generációjának megteremtésével).

Csakhogy a modell még mindig csak részleges. Az egész hibriditás végső értelmét azok a (legtágabb értelemben vett) környezeti beágyazottságok adják, amelyekből a funkció fakad: emiatt a platformok esetleges közössége vagy azonossága ellenére határozottan elkülönülnek egymástól a *társadalmi* (közösségi és épített), a *természeti* (geológiai, biológiai és kozmikus) és a *szimbolikus* környezet kihívásaira reflektáló megoldások. Olyannyira, hogy tulajdonképpen *három alternatív mesterséges intelligencia-fejlesztési irányról, alap-paradigmáról* kellene beszélnünk, mert a különbségeiket elfedi a „homogén" AI képe. A különbségekre érzékeny megközelítésmódnak az ad különös jelentőséget, hogy a „Minden dolgok Internetje" (Internet of Everything) világában ezek a különböző törvényszerűségekkel jellemezhető AI-kisvilágok ismét összekapcsolódnak, hosszú operációs láncok, „kaszkádok" részeiként.

Egyének és csoportjaik e különböző környezet-dimenziókhoz való egyidejű viszonyukból származtatják a támogatási igényt, életműködésük optimalizálását illetve javítását eredményező változók meghatározásával.

Azt is mondhatnánk, hogy innen érthető meg *a „gépi oldal" teleológiája*: az életre hívásának, fejlesztésének értelmet adó – és ezért felépítését, működését, fejlesztését alap-

[57] És itt is az információs viselkedés három különböző fajtájának szükségleteihez igazodó módon. (Belépő oldal: információszerzés – feldolgozás – kilépő oldal (tárgyiasítás).

[58] Természetesen az információs viselkedésnek is mindig van fizikai komponense, és a fizikai cselekvésnek is információs komponense, ezekkel azonban „egyszerűsíthető" az alapmodell. A két rendszer határán állnak az úgynevezett beszédaktusok, ahol a kimondott szó „értéke" és funkciója megegyezik a fizikai cselekvésekével (simogatás, bántalmazás stb.).

[59] Ezek története a szívritmus-szabályozókkal, a pacemakerekkel indul, és a gyógyászatban, katonaságánál és nehéz raktározási feladatoknál alkalmazható exoskeletonoknál jár, természetesen ide sorolandó a „dolgok Internetje" és minden, ami a közlekedés, a szállítás és a gyártás informatizálása révén koordinálhatóvá és vezérelhetővé válik.

vetően meghatározó, cél-természetű – kiindulópontok az ember-környezet kapcsolatból származnak. Ugyanakkor a cselekvési ciklusok is minden esetben a környezetre hatnak vissza, így aztán a „kibővített hibridet" ekképpen tudjuk megjeleníteni:

Az innen kinyerhető összefüggések csak látszólag banálisak és csak részben fedik át egymást a „*technológia társadalmi konstrukciójának*" (*Social Construction of Technology*, SCOT) vagy még inkább a *cselekvőhálózat-elmélet (Actor–Network Theory*, ANT) téziseivel illetve irodalmi hagyományával – ezekre együtt lásd (Király, 2005).

Úgy is mondhatnánk, hogy az ANT a fenti „képlethez" egy új dimenziót rendel, hiszen az ember+gép hibrid helyébe egy sokkal bonyolultabb, cselekvő entitást állít, amelynek egyaránt részei különböző emberek, technikai objektumok, más „materiális" elemek és még jelentések is (szemiotikai objektumok). A hálózat (amelynek node-jai, csomópontjai is hálózati természetűek, s más hálózatok által egyszerűsítettnek, és más hálózatokat egyszerűsítenek) az emberi összetevőt is a hálózatok által formált entitásnak láttatja: így nemcsak a valóságba beavatkozó (és abba rendet vivő) cselekvéseket, hanem még a jelentéseket is a hálózat generálja. Emiatt a hálózat minden entitását (legyen az emberi vagy nem emberi) ugyanazon a módon lehet megragadni és leírni (az ANT szótárában ezt fejezi ki az általánosított szimmetria (*generalized symmetry)* elve.

S ha a cselekvőhálózat maga egy elképesztően bonyolult, finom szövésű rendszer, akkor elképzelhetjük, hogy változása, módosítása, netán átalakítása mennyire összetett egyensúlyi térben értelmezhető. Ahogy Callon (2005,101) magyarázza: „*minden módosítás a cselekvő hálózat elemein és kapcsolatain kívül érinti az egyes elemek által egyszerűsített hálózatokat is… az átalakítás így a cselekvő hálózatot alkotó különböző elemek ellenállásának vizsgálatán múlik*". A társadalomtudomány hibás előfeltételezésekkel igyekszik megérteni és előre jelezni változásokat: „*szükségképpen hipotetikus és spekulatív fog maradni, mivel a társadalmi realitás egyszerűsítése közben a vizsgált asszociációk közül elhagyja mindazokat az entitásokat…, amelyek képesek megmagyarázni a társadalomnak és termékeinek koevolúcióját*". De mindez fordítva is igaz: ha egy elmélet kizárólag a technológiai komponens változásaira érzékeny (mint az erős mesterséges intelligencia és a szuperintelligencia tézise), szükségképpen hipotetikus és spekulatív fog maradni, mivel a technológiai realitás egyszerűsítése közben elhagyja azoknak a cselekvő hálózatoknak a más entitásait, amelyek csak együttesen képesek válaszolni a feltett kérdésekre.

Még jobban látszik mindez, ha felismerjük, hogy az ANT figyelemre méltó diskurzusában a mesterséges intelligencia valójában legalább négyféle formában entitásképző.

1. *Az emberek helyére szituatívan ember-gép hibrideket kell állítanunk*. Ugyanezek az emberek ettől még a gépi elem nélkül is, és ugyanazok a gépek az emberi elem nélkül is hálózati pontok, és más helyzetekben más elemekkel alkotnak funkcionális rendszereket.[60] Vint Cerf és David Nordfors egyenesen erre az entitásra szabták a víziójukat.[61] Olyan

[60] A funkcionális rendszerek Anohin-féle elméletét termékeny módon lehet a hálózatokkal (jelesül: a minden dolgok Internetjével) összekapcsolva tárgyalni (Z. Karvalics, 2015).
[61] http://i4j.info/2014/07/disrupting-unemployment/1502/ (Letöltve: 2015. szeptember 1.)

mesterséges intelligencia-rendszerek fejlesztését sürgetik, amelyek személyre szabottan és külön-külön segítik a bolygó minden egyes lakóját abban, hogy az általa kedvvel végzett tevékenységekhez kapjon külső támogatást (s ha ez a kapacitástömeg kiváltaná és felülírná a gyűlölt, kényszerből végzett munkát, akkor avval mind a gazdasági növekedés útakadályai, mind a foglalkoztatás krízise kezelhetőek volnának).

2. *A mesterséges intelligencia-rendszereket fejlesztő tudásközösségek tagjai és az általuk orkesztrált gépi komponensek együttesen hálózati természetűek*, de hálózati entitásként *csomópontszerűen* kapcsolódnak más hálózatokhoz (például kutatói közösségekhez általában, vagy a konkrét alkalmazási-felhasználási területek emberi és nem-emberi komponenseihez, vagy olyan gazdasági szereplőkhöz, akik a fejlesztés vagy a piacra vitel kontextusában finanszírozók vagy potenciális vásárlók).
3. *Bármely AI-objektum (szoftver- és hardver-komponenseivel) önmagában is cselekvő-hálózati entitás*, és önmagában is lehet hálózati természete (raj-intelligenciája például). A legbonyolultabb, akár több hálózati „réteget" magába foglaló AI-szuperrendszer is azonnal pontszerűvé válik azonban, amint műveletet végez, hiszen avval a cselekvő-hálózat részévé válik.
4. S végül *az AI-diskurzus maga is része a cselekvő-hálózatnak*: azok a fogalmak, terminusok, jelentések, képzetek, referenciák, asszociációk (nevezzük az AI szemiokulturális dimenziójának) alkotnak önmagukban is hálózatot, amelyekkel az AI kérdései felé fordulunk. Amelyek alapján az emberi komponensek általános és speciális viszonyokat alakítanak ki, döntéseket hoznak, és cselekvéseiket hozzáigazítják ezekhez a mentális tartalmakhoz. Ez az a mozzanat, amely ismételten felértékeli az alarmista nézőpontok kritikáját – annak tudatában, hogy a *nézőpont* egy cselekvő-hálózatban egyúttal *csomópont*, amely ugyanúgy formálhat kimenetet, mint bármely materiális komponens.

És erről a teraszról látszik csak igazán, mennyire leszűkül az erős mesterséges intelligencia és a szuperintelligencia teleológiája a harmadik típus, az AI „önmagában vett", cselekvő-hálózatból kiemelt, steril értelmezésére. Az „emberit elérő és meghaladó" intelligencia-entitás létrehozása kétségtelenül öncél: az ezt elkerülhetetlennek láttató előrejelzések pusztán technológiai dinamikák, felgyorsulások és trajekciók mesterségesen létrehozott, hálózati beágyazásukból kiszakított vákuum-világában értelmezhetőek. Az, hogy a diskurzusban ez gyakran összekapcsolódik a várható „jótéteményekkel" (gyógyítás, tudományos problémák megoldása) nem feledteti, hogy e kapcsolódásnak szervesnek, funkcionálisnak és nem hierarchikusnak illetve kauzálisnak kell lennie.[62] Ahogy Bruno Latour fogalmaz: *„a kutatási programoknak oly módon kell asszociálniuk egymással elméleteket,*

[62] Egy egyszerű példával illusztrálva: az öncélú mesterséges intelligencia-fejlesztés irányt vehet például egy betegség-típusra fogékonnyá tevő, genetikai kapcsolat-mintázatok feltérképezéséből származó, „nagy adatokból" új algoritmusok segítségével kinyert heurisztikák megtalálására – miközben ez csak az egyik lehetséges út a betegség leküzdésére. Mi van, ha sikerül a rizikófaktorok visszaszorításával eliminálni a betegség kialakulásának veszélyét? És mi van, ha egy etnomedicina alkalmazásával sikerül megoldást találni a már kialakult betegség villámgyors kezelésére? De gondolhatunk akár az Internetre is: korai fejlesztését a hatvanas-nyolcvanas években *akadémiai teleológia* vezette. (A létező, de nem egyedül meghatározó *katonai-hidegháborús* „szál" alapvető szerepe csak széles körben elhíresült városi legenda. Ennek forrása az volt,hogy a hálózat pionírjai a sok lehetséges útvonalat

fogalmakat, osztályozásokat, nem embereket,[63] *megfigyelési eszközöket és technikákat, intézményeket, világnézeteket vagy akár politikai cselekvőket és ideológiákat, hogy azok ne mondjanak ellent egymásnak*" (Berger, 2008,83. o.).

Az ANT azonban csak egyike az elemzési-leírási keretrendszereknek. Önmagában nem ad sem teljes magyarázatot, s nem vezethetőek le belőle automatikusan a jövő alakításával kapcsolatos normatív elvek sem. Emiatt az „egydimenzióssá" silányított, a szingularitást axiómává emelő technológiai diskurzus mellé egy átfogóbb, a civilizációs teleológia birtokában „újratervezett" diskurzusra van szükség – és úgy tűnik, alarmisták ide, morális pánik oda, a folyamatok, az emberek és az erőforrások ennek az iránynak megfelelően kezdenek összeszerveződni.

A Future of Life Institute 2015. január 11-i nyílt levele[64] manifesztum-szerűen foglal állást az AI-nak a társadalomra gyakorolt pozitív hatása és gazdasági értékei miatt szükségszerű, koordinált fejlesztésének fontossága mellett. A nyílt levél melléklete (*Research priorities for robust and beneficial artificial intelligence*)[65] strukturált módon foglalja össze, milyen sarkalatos kutatási témák köré rendeződhetnek a jövő tudományos és technológiai erőfeszítései, a tárgykör egyfajta problématérképét (és az ahhoz igazodó, közel száz tételes szakirodalomlistát) kínálva (az írás magyar változata tematikus számunk végén olvasható – *a szerk.*). S bíztató, hogy a kutatók, fejlesztők, üzletemberek, újságírók (a csatlakozásra továbbra is nyitott nyílt levél aláírói) között a legnagyobb „vészmadarakat" is ott találjuk. Az is előremutató, hogy a szerzők tisztában vannak vele: önmagában a mesterséges intelligencia kutatása nehezen lép már előre a gépi tanulás, a statisztika, az irányításelmélet vagy a neurológia tudományaival való szövetségek nélkül. Más kérdés, hogy amikor a fő alkalmazási területeket tekintik át, akkor – az önvezető járműveken kívül – rendre csakis régi, jól ismert darabokat sorolják fel: beszédfelismerés, gépi fordítás, kép-osztályozás, kérdező-válaszoló rendszerek, lépkedő robotok stb. Ezek kétségkívül erősen foglalkoztatják azokat, akik a mesterséges intelligencia fejlesztésének tudományos és üzleti hátországában dolgoznak.

A mindennapok hullámverésében azonban más válik fontossá. Amiként a távcső megnyomhatja a szemet, a könyv a fejünkre eshet vagy félrevezethet minket, amit látunk vagy

bejáró csomagkapcsolt adattovábbítási elvhez szükséges fejlesztések erőforrásainak előteremtését támogató egyik érvként használták a háborús helyzetben is épen maradó átviteli csatornákat.) Még a kilencvenes évek közepi, WWW-vel, grafikus böngészővel, e-maillel induló szakasz is akadémiai alapokra épült, a *gazdasági érdek és az üzleti innováció (profit-központú) teleológiája* csak az évtized végére, az úgynevezett dotcom-buborék és annak kipukkanása időszakára hódította el a technológiai oldal fejlesztését, amelyhez sokféle módon asszisztált a *stratégiai-politikai teleológia*, az ismert kulcsszavaival: demokratizálás, egyenlőség, hozzáférés, életminőség, versenyképesség, polgárbarát(abb)ság, olcsóbb működés. Azt, hogy az Internet olyanná lett, amilyen, az egymásra épülő teleológiai és technológiai korszakoknak köszönheti: mai tudásunk birtokában technológiailag is más módon fognánk a fejlesztésbe, szüntetnénk meg a szűk keresztmetszeteket, és az egyes teleológiák mentén is feltehetően önálló Internetek épülnének (külön akadémiai-tudományos, tartalomfogyasztási, üzleti és közösségi dimenzióban, ahol minden esetben másképp paramétereződnek a keresési, tárolási, kapcsolódási és fenntarthatósági szempontok).

[63] Latour „nem emberek" alatt – leegyszerűsítve – dolgok és szubjektumok hibridjeit érti, amelyekkel a cselekvések tere három eleművé válik.

[64] http://futureoflife.org/AI/open_letter

[65] http://futureoflife.org/static/data/documents/research_priorities.pdf

olvasunk, úgy a mesterséges intelligencia-megoldásaink sem kockázatmentesek. Egészen széles körben kelt izgalmat egy-egy új alkalmazás, és ettől szinte elválaszthatatlanul az automatizáció újabb hullámával elvesző munkahelyek kérdése. A környezetünkben egyre nagyobb számban jelenlévő és egyre több cselekvés révén életünk részévé váló intelligens robotokkal kapcsolatos kockázatkezelési, illetve erkölcsi és jogi dilemmák, s nem utolsó sorban a megélénkülő párbeszéd a „kiborggá" válás veszélyeiről. Befejezésül röviden ezekről kell szót ejtenünk, annál is inkább, mert a filozófiai fogantatású technológiakritika mellett ezek a praktikus kérdések is rendre strukturálatlanul köszönnek vissza az alarmista megközelítésekben, ahogy arra a bevezető rész végén már nyomatékosan utaltunk.

AI és társadalom(tudomány): frontvonalak és viták

A nyugtatószerek adminisztrációját már egy gép kezeli az egyik seattle-i kórházban. A Szilícium-völgy egyik szállodájában géplondiner viszi a törölközőt vagy az italt a vendégek szobáiba. A Los Angeles Times olyan cikket közölt egy földrengésről, amelyet egy szoftver írt. (Ezt az eredményeket szigorú szintaxissal közlő sport-sajtó már korábban megtette). A thai konyha ízeinek autenticitását, megfelelő fűszerezettségét robot méri 2014 óta, egy kormányprojekt eredményeképp. Watson, az IBM kvíznyertes bajnoka pedig már egészen másra is használja félelmetes szemantikus memóriáját: háborús veteránoknak ad tanácsot biztosítóválasztásban és életvezetési döntésekben, új receptek alkotásában segít séfeknek, vagy diagnózis felállításában orvosoknak.

A felsorolt friss példákat összegereblyéző Miller (2014) a fentiekkel két állítást kíván illusztrálni:

- hogy a mesterséges intelligenciához köthető automatizáció korábban gyári és irodai munkahelyeket veszélyeztetett, napjainkban viszont már a tudásmunka és a szolgáltatások világába hatol be,
- s hogy többek között ez az oka annak, hogy míg korábban lehetett tudni, hogy a megszűnő állásokat a technológiai fejlődés újak teremtésével ellensúlyozza, addig ma már ez korántsem biztos.

Pontosabban kell azonban fogalmaznunk. Napjainkban még mindig párhuzamosan zajlik a fizikai[66] és a szellemi munka gépesítése, illetve automatizációja. Nem a gazdasági ágazat számít azonban, hanem *a végzett munka jellege*. Miller valamennyi példája *a repetitív agymunka* kiváltásáról szól. Ha nem a munkanélküliség lenne a kontextus, ennek az emancipatorikus és egalizáló jellegét hangsúlyoznánk inkább, hiszen ami „gépies", az embertelen, a gépesítés így valójában humanizálás, akkor és amennyiben az életidő magasabb érték-hozzáadású, kreatív, nem automatizálható munkára szabadul fel. Az alacsony hozzá-

[66] Nagy robotizációs hullámra lehet számítani például azokban az országokban, ahol eddig a nagytömegű olcsó munkaerő volt a versenyképesség alapja (Kína-szindróma). Az automatizáció újabb hulláma elér eddig kevésbé érintett munkakör-kategóriákat (például takarítás, épületfenntartás, hulladék-kezelés). A legújabb technológiák némelyike pedig tudás-intenzívebbé tesz eddig kisiparinak és kézművesnek megmaradt területeket (például a 3D-nyomtatás az idomtermékeket előállító műhelyekben).

adást igénylő szakmák eltűnésével pedig közelebb kerülnek egymáshoz a munkaerőpiac különböző szegmensei. Brynjolfsson és McAfee (2014) azt gondolta végig, hogy mivel járhat ez azok oldaláról, akik kedvezményezettjei a változásoknak. Nekik személyes technológiák sora áll csatasorba, hogy képességeiket a leginkább megfelelő módon használhassák, és ezek mögött különösen fejlett infrastruktúra dübörög. Korlátlanná válik a hozzáférés a kultúrjavakhoz, és a nyersanyagok szerepét a szűk erőforrássá lett gondolatok és azok hordozói töltik be.

Azok számára azonban, akik rövid vagy középtávon kimaradnak mindebből, a jelenlegi irányítási rend és elosztási logika kétségkívül nem ígéri, hogy a profit-elvvel szemben az értéktermelésen alapuló munkahely-biztosítás kerülhetne az előtérbe. De ez nem technológiai, hanem kormányzási (governance) probléma. *A teljes gazdasági és társadalmi rend, az elosztás alapvető újrarendezése* nélkül nem megoldhatóak az automatizálás-keltette munkaerőpiaci zavarok (Ford, 2015; Wadwha, 2015b). Ez pedig – hangsúlyozzák az elemzők – csakis az államok (újabban: egyes városok (!)[67] újraértékelt szerepvállalásával, a köz-szféra kiterjesztésével működhet. Mindezt nagyban segítheti, ha az oktatás az új és nem a régi gazdaságra készít fel, a politika pedig nyitottá válik az új utak keresésére.

Nem árt tudatosítani, hogy nincs semmi szingularitás-specifikus a diskurzusban. Szuperintelligens gépek nélkül is érvényesül a mélységben erősen tagolt[68] automatizációs nyomás: az információtechnológiai forradalom „nem áll le", a termelékenységet folyamatosan javítja (Byrne, 2013). Ha azonban a szuperintelligenciát a korábbiakban kifejtetteknek megfelelően „kivesszük a képletből", akkor a munkahelyük megszűnésével riogatott „csúcs-értelmiségiek" (orvosok, tanárok, jogászok, mesterszakácsok, sőt művészek)[69] számára sokkal inkább az új koegzisztenciák keresése és a munkájukat megkönnyítő, felgyorsító alkalmazások adaptálása lesz a kihívás, nem a székük elvesztése miatti aggodalom (Susskind és Susskind, 2016). Riogathat a hosszú távú gyötrelmet ígérő Sachs és Kotlikoff (2012), a munkanélküliek világát vizionáló Ford (2015), vagy a már egyenesen a munka világának végéről beszélő Smith (2013) – amennyiben a gazdasági alapok biztosíthatóak, *az oktatás, a tudomány, a kultúra és a humán szolgáltatások világa* korlátlanul felvevőképes lehet, s felszívhatja a (szükségszerűen egyre képzettebb) munkaerőt.

S valóban: azokkal az előrejelzésekkel szemben, amelyek kizárólag a megszűnő munkahelyek számával, mennyiségbecslésével foglalkoznak, s amelynek alarmista részét találóan nevezi Marc Andreessen „luddita megtévesztésnek" (*Luddite Fallacy*), szaporodnak azok a tanulmányok, amelyek történeti kontextusban, a teljes munkaerőpiac méretének és szerkezetének számbavételével tesznek állításokat a technológiának a munkaerőpiacra gyakorolt hatásával kapcsolatban. Dajkó (2015) két friss szakanyagot ismertet, amelynek

[67] A feltétel nélküli alapjövedelemmel GMI (Guaranteed Minimum Income) való eredményes kísérletezés, egy fenntartható modell megtalálása például megfelelő válasz lenne, és a kérdéseket más tartományba tolná át. Újabban nemcsak kormányok, hanem városi önkormányzatok is elkezdték fontolgatni ezt a megoldást.

[68] Van, ahol kevesebben látják el a feladatokat, jobb információs infrastruktúrával. Van, ahol a termelés értékláncából esnek ki szereplők (és a munkahelyek a cégekkel együtt szűnnek meg vagy helyeződnek át máshová), és van, ahol egész iparágak tűnnek el.

[69] És evvel párhuzamosan a csodálatos verset író, gyönyörű képet festő, Mozartot lefőző robotok diskurzusát (legmerészebben: Levy, 2005) is ideje volna ad acta tenni.

végkövetkeztetése az, hogy *„a technológiai fejlesztések mindeddig több munkahelyet teremtettek, mint amennyit megszüntettek”*. A neves Forrester elemzője, J. P. Gownder a robotmunka döbbenetes előretörését prognosztizáló tanulmányok túlzásaira figyelmeztet, amelyek nem veszik figyelembe a nem pótolható emberi intelligenciát, majd (a düsseldorfi repülőtér példájával) érvel amellett, hogy a robotok és a mesterséges intelligencia alkalmazása a fejlettebb rendszerekhez igazodó összetettebb tudásokra támaszt folyamatos keresletet – tehát nem a kevesebb, hanem a képzettebb munkavállaló a jövőképe.[70] A szintén mértékadónak számító Deloitte tanulmánya az utóbbi 150 évet Anglia és Wales példáján vizsgálva arra jutott, hogy *„bár a technológiai fejlesztések miatt a vizsgált időszakban nagy számban szűntek meg munkahelyek a mezőgazdaságban és gyártó iparágakban, mindezt bőségesen ellensúlyozta, hogy számtalan munkahely teremtődött az egészségügyben, a kreatív szektorokban, a technológiai szegmensben és az üzleti szolgáltatások területén”*. S itt annak ellenére nő szakadatlanul és magas százalékokkal kifejezhető módon a foglalkoztatottak száma, hogy az itt végzett munka egyre inkább technológia-intenzív és egyre hatékonyabb.

A fárasztó és monoton munkák eltűnése[71] tehát együtt jár az oktatás különböző színterein alkalmazottak, a nővérek, gondozók, a közösségi szolgáltatásokban dolgozók (és tegyük hozzá: a tudományban, a médiában foglalkoztatottak vagy művészetből élők) számának gyarapodásával. S ha van alapja a Lanier-féle „új humanizmusnak”, azt sokkal inkább ebben a mozzanatban, és nem a mesterséges intelligencia elutasításában kellene keresni.

Nagy felbontásban azonban kiélesednek a valódi problémák is. Különböző munkakörökre és ágazatokra különböző mértékben lesznek igazak az átfogó trendek. Jól azonosítható, hogy a várakozásokkal ellentétben nem az alacsony, hanem a középszintű (és közepesen fizető) foglalkozások és állások tűnnek el (Autor és Dorn, 2013). Az elsődleges kihívás tehát a jövedelmi, vagyoni és társadalmi egyenlőtlenségek növekedése. Az átmenet

[70] A robotok alkalmazásának köszönhetően a német autóiparban 2010 és 2013 között az ágazat foglalkoztatottsága több mint 7 százalékkal nőtt, ekkor 10 000 alkalmazottra 1100 robot jutott. A Metra Martech által készített kutatás szerint a jelenleg működő egymilliónyi ipari robotnak csaknem hárommillió új munkahely köszönhető. Szakértőik megerősítik: a robotok nemcsak több, de jobb munkahelyet is teremtenek: „Az emberek szívesen dolgoznak robotokkal, mert megkönnyítik a munkájukat.” http://nol.hu/tud-tech/yumi-a-robot-lebontotta-a-keritest-1574625 (Letöltve: 2015. december 20.)

[71] Érdemes itt egy pillanatra elmélázni azon, mennyire biztosak az alarmisták abban, hogy a vezető nélküli járművek irtózatos pusztítást végeznek majd a sofőrök sok tízmilliós munkaerőpiacán. Wadwha (2015a). S noha már van példa kötött pályás közlekedésben (metró, vonat) és a légiközlekedésben (robotpilóta, számítógéppel támogatott fel-és leszállás stb.) vezető nélküli megoldásokra, a mesterséges intelligencia-megoldások gyors bevezethetőségében hívők nem veszik figyelembe az elképesztő méretű cselekvőhálózatot, amelynek nemcsak az utak, az útjelzések, az eltérő közlekedési funkciók, a különböző úti célok és minden egyes különböző közlekedő ember a része, s ami a változást nem a technológiai előrehaladás, hanem sokkal összetettebb paraméterek függvényévé teszi. De eközben azt is látni kell, hogy a sofőrmunka önmagában monoton, alacsony értékhozzáadású munka, amelynek lecserélése indokolt – de nem biztos, hogy pusztán a vezető nélküliség az egyetlen célfüggvény. Egy új funkcionális térben a szükséges közlekedési teljesítmény válhat csökkenthetővé, vagy a közlekedés menedzsmentje tudás-intenzívvé (ahogyan például speciális munkagépek – daruk, kombájnok stb. – esetében látjuk: ezekhez ma jellemzően diplomás, nagy tudású kezelők-vezetők kellenek).

kétségkívül lehet fájdalmas, egyenesen brutális, ahogy Kaplan (2015) fogalmaz, ha tőke és munka csatájában a tőke felé billen a mérleg. Ha a bérek növekedése tartósan alulmarad a termelékenység növekedése mögött. Ha a munkaerőpiac nem rugalmasabbá, hanem merevebbé válik. Ezek a kérdések azonban már mesterséges intelligencia és automatizációkontextus nékül is érvényes és gazdagon tárgyalt makroökonómiai alapdilemmák. Bernstein (2015) szerint a sarkalatos kérdés az, hogy növekszik-e a sebessége az emberi munka technológiai kiváltásának – és válasza az, hogy *nem* (amelyre számtalan bizonyítékot sorol fel).

És vajon az emberi testben mit válthatnak ki/javíthatnak meg a mesterséges intelligencia jelenlegi rendszerei, milyen élethelyzetben és milyen célcsoportoknál? A *biológiai test korlátainak leküzdésében* szerepet játszó megoldások valódi és érvényes tartalommal töltik fel az 1960-ban született cyborg-narratívát.[72] Humanistának látjuk azokat az erőfeszítéseket, amelyekkel fogyatékkal élő és beteg embertársaink jutnak lehetőséghez egy méltóságteljesebb, autonómabb életre (látás-hallás- és mozgásprotézisekkel, de akár gondolat-vezérléses környezet-manipulációval[73]). Az exoskeleton, a külső váz használata kíméli a nagy súlyt emelgető kórházi ápolókat és munkásokat. Azok az augmentált valóság-rendszerek (ahol a műveletvégzést mesterséges intelligencia-megoldások támogatják) és azok a környezetek, ahol testünk gépi kiterjesztését sajátunkként érzékeljük (s amit Boyd (2015) szuperpropriocepciónak nevez), a lehetőségek egy *speciális* világát jelentik. A szervezet egészségi állapotának megőrzése, monitoringja és gyógyítása, vagy a minőségi öregedés elősegítése érdekében latba vetett *általános*, *mindenkit érintő* technológiák pedig együttesen azt ígérik, hogy az evolúciós „előtörténetre" a gépi intelligencia felhasználásával épül új „réteg", amely nemcsak az egyes embereknek kínál jobb-létet, hanem a fajnak is nagyobb esélyt a túlélésre (Naam, 2005).

A test „megerősítésének" (enhancement) tudományos programjával kapcsolatos lényegi viták az ezredforduló után már lezajlottak,[74] napjainkban „középutas" álláspontok jelennek meg: Buchanan (2011) megközelítése „tervezési hibaként" szeretné láttatni a feljavítandó sajátosságokat, Agar (2013) pedig az ember „enyhe felerősítésének" (*moderate human enhancement*) programja mellett tör lándzsát, mert szerinte a túlzásba vitt változat az egész emberi identitást ásná alá. Abban azonban szinte mindenki egyetért, hogy megfelelő keretek között tartva a szuperintelligencia helyett mindez egy „új szuperhumán korszak" (new superhuman age) eljövetelét gyorsíthatja fel (Boyd, 2015).

[72] Noha a diskurzus-indító szerzők, Clynes és Kline (1960) specifikus űrkutatási kontextusban fogalmazták meg az emberi testnek a Földön kívüli körülmények elviselését lehetővé tevő mérnöki „továbbfejlesztését", a diskurzus mára sokkal összetettebb és általánosabb jellegű lett.

[73] Ennek a jövő interfészei szempontjából különösen fontos és izgalmas iránynak számos sikeres előzmény után 2014 őszén sikerült egymástól 8000 kilométerre lévő emberi agyak között sebészi beavatkozás nélküli kapcsolatot létesíteni. Az agy-számítógép kapcsolattal Franciaország és India között sikerült gondolatokat (például a kéz és a láb megmozdítására vonatkozó instrukciót) közvetíteni. http://index.hu/tudomany/2014/09/05/mar_mukodik_az_internetes_gondolatatvitel/

[74] A pozitív ígéretek foglalataként megszületett a transzhumanizmus fogalma, amelyre válaszul megjelentek a bio-ludditák is, akik korlátozni szeretnének minden ezirányú fejlesztést (Young, 2005). Garreau (2006) számos forgatókönyvet gondolt végig, Hughes (2004) a „demokratikus transzhumanizmus" jegyében az egész folyamat közösségi kontrolljának szükségességére figyelmeztetett. Az elmúlt időszak termésére lásd Savulescu és Bostrom (2011) olvasókönyvét és Anissimov (2015) merész vízióját, amelyben a szuperintelligenciát és a nanotechnológiát kapcsolja össze (az „ember digitális feljavításaként") felfogott transzhumanizmussal.

A civilizációs előrelépésnek ez az ígérete, ha megfelelően artikulált, ellensúlyozhatja azokat a valódi és kétségtelen veszélyeket, amelyek már a gépi intelligencia-megoldások jelenlegi világában is jól detektálhatóak. Régóta tudjuk, hogy az algoritmikus alapokon működő kereskedő-ágensek a tőzsdéken okozhatnak anomáliákat. Az automatizált és összekapcsolt nagy hálózati rendszerekben, a kritikus infrastruktúrákban a rossz cselekvésválasztás vagy az ellenséges támadás miatt alakulhatnak ki működési zavarok (például áramszünet, közlekedési káosz, tranzakciós rendszerek leállása stb.). A viselhető és perszonalizált okos eszközök és a felhőarchitektúra megnöveli az intelligens hálózati bűnözésnek való „cyber-kitettséget".[75] S mi történik, ha a bármely orvos-virtuóznál biztosabb kezű sebészrobot hibát követ el? Hogyan birkózhat meg egy rendkívüli helyzettel egy vezető nélküli jármű? Milyen szintű döntést hozhat gépi intelligencia, amelynek végén embereket érintő tranzakció áll?[76]

Az ilyen és ehhez hasonló kérdések jogi és erkölcsi dilemmák sorát vetik fel, emiatt nemcsak a fejlesztőknek kell az efféle aggodalmakat kiemelten kezelniük. Avval, ahogyan a baleset vagy probléma forrását jelentő géptől egyre hosszabb láncokkal jutunk el a felelősséget vállalni képes emberekhez vagy intézményekhez, kezdeni kell valamit. A tervezési, a kivitelezési, a működtetési, a karbantartási, a tesztelési és a felhasználói érintettségek világát szituáció-érzékenyen kell tudni „telepíteni" és felosztani. *Teljesen téves irány azonban azt fontolgatni, miképpen tehetők jogi személlyé, jogok birtokosává* (és eképpen felelőssé) *intelligens robotok* (Lin et al., 2014). Hasonlóképpen zsákutca az etikus robot, a morális döntéshozatalra alkalmas gép megteremtésének programja (Allen – Wallach, 2010). *A hibrid rendszerben az erkölcsi felelősség minden esetben az emberi komponensé. A gépek programjuknak megfelelően viselkednek*. Nem állíthatóak erkölcsi kihívások elé, mert – mint korábban láttuk – már a jelentések országába sem bocsáttattak be, így aztán a különlegesen bonyolult, sokféleképpen referenciális metajelentések világa még inkább távol van tőlük.[77] A mesterséges intelligencia ágenseinek nincs szuverén döntése, kizárólag

[75] A cyberbűnözéssel foglalkozó európai szakintézmény (European Cybercrime Center) 2014-es szakanyaga – *The Internet Organised Crime Threat Assessment* (iOCTA) – valamennyi területet áttekinti, és új, a fejlettebb technológiákhoz kötődő elkövetési módok megjelenését jósolja (https://www.europol.europa.eu/content/internet-organised-crime-threat-assesment-iocta). További példákra lásd: http://www.origo.hu/techbazis/20141021-okoseszkoz-hacker-gyilkossag-implantatum-gyogyaszat.html

[76] Vegyük észre: amikor sorszámot húzunk, ügytípusok közül választva, egy primitív mesterséges intelligencia-megoldás „dönti el", hogy mikor és melyik ügyintézőnél kerülünk sorra. Erre gyakorlatilag érzéketlenek vagyunk, mert az egyszerű besorolási algoritmus nem sért érzékenységet, és felül is írható (ha sietünk, az előbbre soroltaktól és az ügyintézőktől is kérhetjük a méltányosságot). Ott azonban, ahol – akár hasonlóan banális esetben – már nincs mód a beavatkozásra, már komoly aggályok vetődnek fel. Nem beszélve arról, amikor „buta" algoritmusok minősítenek egy életbiztosítást kockázatosabbnak pusztán a testsúly növekedése miatt, nem törődve avval, hogy az háj vagy izom-e (ami inkább csökkentené a kockázatot). Emiatt alakult ki sok alrendszerben az a gyakorlat, hogy az embereket érintő gépi operáció végpontján afféle „humán kontrollként" mindig a gépnél komplexebb körülmény-együttest kezelni képes emberek állnak, akik felülbírálhatják-módosíthatják az automatizált folyamat végeredményét.

[77] Azok a próbálkozások, amelyek erkölcsi dilemmaként próbálják beállítani a döntéshelyzetbe kerülő gépek esetét (mint például a 'mit kezd a robot akkor, ha nem egy, hanem két emberpótlékot kellene megmentenie a gödörbe zuhanástól', vagy 'mit tesz az automata sofőr, ha választani kellene, ki kerüljön el egy balesetet: saját, kevés utassal közlekedő járműve vagy egy sok gyereket szállító busz'), megtévesztő paradiskurzusokhoz vezetnek (http://sg.hu/cikkek/107784/robotok-az-etika-csapdajaban). Ezek ugyanis *nem erkölcsi, hanem programozási problémák*, amelyek mögött nem új típusú, hanem teljesen *hagyományos etikai kérdőjelek és megfontolások* állnak.

előfeltételeknek és szabályoknak olyan együtteseivel bírnak, amelyet a rendszereket megalkotó emberek plántáltak beléjük. Emiatt nagyon fontos kérdés a mesterséges intelligencia pragmatikus etikai és átfogó morálfilozófiai kérdéseivel foglalkozni – csak nem a gépek oldaláról, hanem az emberekéről.

Ahhoz például, hogy a jogrendszer kezelni legyen képes a praktikus problémákat, először is le kellene tudni fordítani a jog nyelvére a mesterséges intelligenciát. De hogyan definiálható kiindulásként maga az intelligencia fogalma, ráadásul az embertől függetlenül – ha annak négy kulcs-komponense, a tudatosság, a gondolat, a szabad akarat és az elme mibenléte Arisztotelész óta állandóan diszkutált tárgya a tudománynak?[78] Markus Hutter és Shane Legg használhatónak tűnő, rövid és szellemes definíciót alkotott. Náluk „*az intelligencia egy ágens képességét méri, amellyel céljait különböző környezetekben eléri*" (Lea, 2015), ahol az ágens egyaránt lehet ember és gép. Vegyük észre, hogy „a cél elérése" mozzanata szellemesen magába foglalja a tervezés, a tanulás és a problémamegoldás másutt önállóan tárgyalt attribútumait. De hol azonosítjuk az ágenst a hibrid rendszerben? Továbbá: minden mérnöki alkotásban benne rejlik a lehetőség, hogy méretezése, hibatűrése olyan valóságos kihívással találkozik, amelyre a tervező nem lehetett felkészülve. Elfogadjuk ezt egy „okos gép" esetében is? Mi, emberek, gyakorta vállalunk kényszerűen kockázatot, mert döntés- és cselekvéskényszerben vagyunk információhiányos helyzetben is – de ebből mennyit tudunk ráterhelni a gépi oldalra? Amikor feladatokat delegálunk, a felelősség meddig marad a miénk, különös tekintettel a cselekvő-hálózati beágyazásra? És mi az, amit nem delegálunk? Lehet mindezt listázni egy alapdokumentumban (Havens, 2015)?

S még ha sikerülne is megnyugtató intelligencia-formuláig vagy AI etikai kódexig jutni, hogyan határozhatjuk majd meg például a „jó"-t (Armstrong, 2014), minden kontextusok egyik legfontosabbikát, amikor a hibrid rendszer céljait definiáljuk, és használatának hatáskövetkezményeire készülünk? S mindezt hogyan tudjuk lefordítani gépi kódra?

A kérdések még folytathatóak lennének – csak jó volna túljutni az alarmista útakadályokon, hogy valóban a lényeges területekre koncentrálhassunk. Bemelegítésnek megteszi az a 2015 végén megjelent válogatás, amely 175 kortárs tudós, filozófus és művész tollából tartalmaz olykor aforisztikus, olykor tanulmány-értékű válaszokat a fenti és sok más, a gondolkodó gépekkel kapcsolatos kérdésre (Brockman, 2015). Ebben egyszerre vannak jelen a keresett és sürgetett új szemlélet irányába mutató megközelítések és az alarmista manifesztumok – ám kézzelfogható közelségbe kerül a frissen megfogalmazott AI-ígéretekből való kezdődő kiábrándulás, az eddigi kilencet követő, közelítő tizedik AI-tél (AI-winter) is.[79]

Pedig a mesterséges intelligencia valódi helye az „örök tavasz". Technológia-alapúvá lett kultúránk egyik legfontosabb, legperspektivikusabb tudományos iránya, kutatás-fejlesztési és alkalmazási területe, amelyben minden kis előrelépés avval kecsegtet, hogy az ember-gép hibrid rendszerek hatékonyabban tudnak teljesíteni az élet minőségének javításában és a civilizációs kihívások kezelésében. Az alarmizmus legnagyobb bűne, hogy talmi szenzációkeltésével figyelmet és erőforrásokat von el a valódi diskurzusoktól.

[78] 2014 nyarán egy orosz milliárdos, Dmitrij Volkov kísérletet tett arra, hogy a jachtjára meghívott neves filozófusokkal és társadalomkutatókkal konszenzusig sikerüljön jutni a kérdésben. David Chalmers, az egyik résztvevő egyenesen azzal búcsúzott az eredménytelen „csúcstalálkozótól", hogy a megoldásra a következő évszázadig kell majd várni.

[79] S mindeközben talán érdemes újra kézbe venni Darab Tamás könyvecskéjét (Darab, 1991), amely a mesterséges intelligencia körüli sarkalatos ismeretelméleti viták élvezetes és precíz összefoglalása, s amelynek köpönyegéből a kérdéskör hazai irodalma kinőtt.

Irodalom

Agar, Nicholas 2013: *Truly Human Enhancement: A Philosophical Defense of Limits* The MIT Press

Allen, Colin – Wallach, Wendell 2010: *Moral Machines: Teaching Robots Right from Wrong* Oxford University Press

Anissimov, Michael 2015: *Our Accelerating Future: How Superintelligence, Nanotechnology, and Transhumanism Will Transform the Planet* Zenit Books,

Armstrong, Stuart 2014: *Smarter Than Us: The Rise of Machine Intelligence*, Machine Intelligence Research Institute

Autor, David H. – Dorn, David 2013: How Technology Wrecks the Middle Class *New York Times*, August 24. 24–27. http://www.collier.sts.vt.edu/engl4874/pdfs/autor_nyt_9_24_13.pdf

Baranyi, Péter 2012: Kognitív infokommunikáció: egy ébredő interdiszciplína http://otodikal program.huminf.u-szeged.hu/sites/default/files/BaranyiPeter_Coginfocom_szeged'12.pdf

Barrat, James 2014: Our Final Invention: Artificial Intelligence and the End of the Human Era Thomas Dunne Books

Basulto, Dominik 2015: The strange link between global climate change and the rise of the robots *The Washington Post*, September 8.

http://www.washingtonpost.com/news/innovations/wp/2015/09/08/the-strange-link-between-global-climate-change-and-the-rise-of-the-robots/

Berger, Viktor 2008: Bruno Latour tudományképe és antropológiai megközelítésmódja *Szociológiai Szemle* 4. 72-92.

Bernstein, Jared 2015: *The Reconnection Agenda: Reuniting Growth and Prosperity* CreateSpace Independent Publishing Platform

Blackford, Russell – Broderick, Damien 2014: *Intelligence Unbound: The Future of Uploaded and Machine Minds* Wiley-Blackwell

Bodnar, Ken 2015a: Dawkins, Wasps, Artificial Intelligence, Evolution, Memorability and Artificial Consciousness *Future Imperfect & Software Stream of Consciousness* December http://coderzen.blogspot.hu/2015/12/dawkins-wasps-artificial-intelligence.html

Bodnar, Ken 2015b: Dimension & Event Sorters & Classifiers - The Genesis of Artificial Conciousness & Abstract Machine Reasoning *Future Imperfect & Software Stream of Consciousness* December http://coderzen.blogspot.hu/2015/12/dimension-event-sorters-classifiers.html

Bolcsó, Dániel 2015: Robot ölt embert, ez már az apokalipszis? *Index*, Július 7. http://index.hu/tech/2015/07/08/robot_gyilkossag_buntetojog_gepek_vilagvege/

Bostrom, Nick 2014: *Superintelligence: Paths, Dangers, Strategies* Oxford University Press

Boyd, Richard 2015: Man Vs. Machine: How Humans Are Driving The Next Age Of Machine Learning *Techcrunch*, June 15. aug. 18. http://techcrunch.com/2015/06/11/man-vs-machine-how-humans-are-driving-the-next-age-of-machine-learning/#.pzubov:ck1O

Brain, Marshall 2015: *The Second Intelligent Species: How Humans Will Become as Irrelevant as Cockroaches* BYG Publishing

Brockman, John 2015**:** *What to Think About Machines That Think: Today's Leading Thinkers on the Age of Machine Intelligence* Harper Perennial

Brooks, Rodney, 2003: *Flesh and Machines: How Robots Will Change Us* Vintage

Brynjolfsson, Erik – McAfee, Andrew 2014: *The Second Machine Age. Work, Progress and Prosperity in the Second Machine Age* W. W. Norton & Company;

Buchanan, Allen 2011: *Better than Human: The Promise and Perils of Enhancing Ourselves* Oxford University Press

Burgess, Mark 2015: *In Search of Certainty. The Science of Our Information Infrastructure* O'Reilly Media

Byrne, David M. – Oliner, Stephen D. – Sichel, Daniel E. (2013) Is the Information Technology Revolution Over? *Staff Working Paper*, FEDS March (No.3). http://www.federalreserve.gov/pubs/feds/2013/201336/201336pap.pdf

Callon, Michel 2005: Alakuló társadalom. A technika, mint a szociológiai elemzés eszköze *Replika* 51-51. (november) 89-105.o.
Card, Stuart K. – Moran Thomas P. – Newell, Allen 1983: *The Psychology of Human-Computer Interaction* Erlbaum, Hillsdale
Clynes, Manfred E. – Kline, Nathan S. 1960: Cyborgs and Space *Astronautics* September 24-26, 74-76. http://cyberneticzoo.com/wp-content/uploads/2012/01/cyborgs-Astronautics-sep1960.pdf
Colvin, Geoff 2015: *Humans Are Underrated: What High Achievers Know That Brilliant Machines Never Will* Portfolio
Dajkó Pál 2015: A robotok több munkahelyet teremtenek, mint amennyit elvesznek *IT café* augusztus 28. http://itcafe.hu/hir/robot_technologia_munkahely.html
Darab Tamás 1991: *A gépesített értelem* Áron Kiadó
Del Monte, Louis A. 2014: *The Artificial Intelligence Revolution: Will Artificial Intelligence Serve Us Or Replace Us?* Amazon Kindle Edition
Diebold, John 1969: *Man and the Computer:Technology as An Agent of Social Change* Praeger
Domingos, Pedro 2015: *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* Basic Books, 2015
Dreyfus, Hubert 1972: *What computers can't do: A Critique of Artificial Reason* New York: Harper & Row
Dreyfus, Hubert 1992: *What computers still can't do": A Critique of Artificial Reason*. Cambridge, MA: MIT Press
DuBravac, Shawn 2015: *Digital Destiny: How the New Age of Data Will Change the Way We Live, Work, and Communicate* Regnery Publishing
Engelbart, Douglas 1962: *Augmenting Human Intellect* paper, October http://www.dougengelbart.org/pubs/augment-3906.html
Floridi, Luciano: The Fourth Revolution: How the infosphere is reshaping human reality Oxford University Press, 2014
Ford, Martin 2015: *Rise of the Robots: Technology and the Threat of a Jobless Future.* Basic Books
Garreau, Joel 2006: *Radical Evolution: The Promise and Peril of Enhancing Our Minds, Our Bodies—And What It Means to Be Human* Broadway Books
Gitt, Werner 2004: *Kezdetben volt az információ* Evangéliumi Kiadó (2., bővített és javított kiadás)
Gleiser, Marcelo: Sinister dreams of transhuman machines: or, the world as information In: Uő: *The Island of Knowledge. The Limits of Science and the Search for Meaning* (31. fejezet) Basic Books, 2014
Goertzel, Ben 2014: *Ten Years To the Singularity If We Really Really Try: ... and other Essays on AGI and its Implications* CreateSpace Independent Publishing Platform
Good, Irwing John 1965: Speculations Concerning the First Ultraintelligent Machine *Advances in Computers*, Vol. 6.
Grier, David Alan 2005: *When Computers Were Human* Princeton University Press
Grove Andrew S. 1998: Csak a paranoidok maradnak fenn Bagolyvár Kiadó
Grudin, Jonathan 2007: A moving target: The evolution of human–computer interaction. In Sears Andrew – Jacko Julie A. (Eds.): *Human-Computer Interaction Handbook* (2nd Edition). CRC Press
Havasi, Catherine 2014: Who's Doing Common-Sense Reasoning And Why It Matters
TechCrunch, Augusztus 9. http://techcrunch.com/2014/08/09/guide-to-common-sense-reasoning-whos-doing-it-and-why-it-matters/
Havens, John C. 2015: The importance of human innovation in A.I. ethics *Mashable*, October 3. http://mashable.com/2015/10/03/ethics-artificial-intelligence/?curator=MediaREDEF#9W0.aNQo7gqc
Hernandez, Daniela 2016: It's hard work being funny – especially for robots *Fusion*, January 8. http://fusion.net/story/251798/funny-robots/
Horváth Bence 2015: Aggódni azon, hogy a robotok ellenünk fordulnak, olyan mintha a Mars túlnépesedésén aggódnánk *444.hu* május 25., hétfő http://444.hu/2015/05/25/aggodni-azon-hogy-a-robotok-ellenunk-fordulnak-olyan-mintha-a-mars-tulnepesedesen-aggodnank/

Hsu, Stephen 2015: Don't Worry, Smart Machines Will Take Us With Them. Why human intelligence and AI will co-evolve *Nautilus*, September 3. http://nautil.us/issue/28/2050/dont-worry-smart-machines-will-take-us-with-them

Hughes James 2004: *Citizen Cyborg: Why Democratic Societies Must Respond To The Redesigned Human Of The Future* Basic Books

Kaplan, Jerry 2015: *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence* Yale University Press

Király Gábor 2005: Hovatovább STS? Fejtegetések az értelmezési flexibilitás, a hiányzó tömeg, a kiborg és a demokrácia kapcsán *Replika* 51-52. November 25-56.o.

Knight, Will 2015: Can This Man Make AI More Human? MIT Technology Review December 17. http://www.technologyreview.com/featuredstory/544606/can-this-man-make-ai-more-human/?utm_campaign=newsletters&utm_source=newsletter-weekly-computing&utm_medium=email&utm_content=20151217

Kreft, Hans-Dietrich, 2003: *Humatics – Theorie Der Operablen Wissenseigenschaften. Geld und Wissen* Weissensee-Verlag, Berlin

Kurzweil, Ray 2013: *A szingularitás küszöbén* Ad Astra (eredetije: *The Singularity is Near*. New York: Viking Books, 2005)

Kurzweil, Ray 2013: *How to Create a Mind* Penguin Books

Lea, Gary 2015: Why we need a legal definition of artificial intelligence *World Economic Forum Agenda*, September 7. https://agenda.weforum.org/2015/09/why-we-need-a-legal-definition-of-artificial-intelligence/?utm_content=buffer4f517&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Levy, David 2005 *Robots Unlimited: Life in a Virtual Age* A K Peters/CRC Press

Levy, David 2008: *Love and Sex with Robots: The Evolution of Human-Robot Relationships*

Licklider, Joseph Carl Robnett (1965): *Libraries of the Future* Cambridge, MA

Licklider, Joseph Carl Robnett (1960): "Man-Computer Symbiosis" *IRE Transactions on Human Factors in Electronics*, vol. HFE-1, 4-11, March 1960.

Lin Patrick – Abney Keith – Bekey George A.: 2014: Robot Ethics: The Ethical and Social Implications of Robotics The MIT Press

Littman, Michael 2015: Rise of the Machines' is Not a Likely Future *Live Science*, January 28. http://www.livescience.com/49625-robots-will-not-conquer-humanity.html
Harper Perennial

Love, Dylan 2014: By 2045 'The Top Species Will No Longer Be Humans,' And That Could Be A Problem *Business Insider* July 6. http://www.businessinsider.com/louis-del-monte-interview-on-the-singularity-2014-7#ixzz3izPDTrft

Markoff', John 2015: *Machines of Loving Grace. The Quest for Common Ground Between Humans and Robots* Ecco

Meier, Patrick 2015: *Digital Humanitarians How Big Data is Changing the Face of Humanitarian Response* CRC Press

Michelucci, Pietro - Dickinson, Janis L. 2016: The power of crowds *Science*, 2016 January 32-33.o.

Miller, Claire Cain 2014: As Robots Grow Smarter, American Workers Struggle to Keep Up *The New York Times* The Upshot Dec.15.http://www.nytimes.com/2014/12/16/upshot/as-robots-grow-smarter-american-workers-struggle-to-keep-up.html?utm_source=nextdraft&utm_medium=email&abt=0002&abg=1&_r=2

Miller, James D. 2012: *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World* BenBella Books,

Mindell, David A. 2015: *Our Robots, Ourselves: Robotics and the Myths of Autonomy* Viking

Moravec, Hans 1988: *Mind Children* Harvard University Press

Moravec, Hans 1998: *Robot: Mere Machine to Transcendent Mind* Oxford University Press

Muehlhauser, Luke 2013: *Facing the Intelligence Explosion* Machine Intelligence Research Institute
Naam, Ramez 2005: *More Than Human: Embracing the Promise of Biological Enhancement* Broadway Books
O'Leary, Katie – Wobbrock, Jacob O. – Riskin, Eva A. (2013): Q-Methodology as a Research and Design Tool for HCI CHI 2013 April 27-May 2, Paris, France. http://students.washington.edu/kathlo/HCI-Q-CAMERAReady.pdf
Parkin, Simon 2015: The Brain in the Machine *How We Get to Next*, October 21. https://howwegettonext.com/the-brain-in-the-machine-f61559a38887#.gji39ffcw
Pesthy, Gábor 2015: AI: a mesterséges intelligencia barát vagy ellenség *Origo*, 2014. Jan. 4. http://www.origo.hu/tudomany/20141222-ai-a-mesterseges-intelligencia-barat-vagy-ellenseg.html
Preece, Jenny – Sharp, Helen, Rogers, Yvonne 2015: *Interaction Design: Beyond Human-Computer Interaction* 4th ed. John Wiley & Sons
Rees, Martin 2004: *Our Final Century: The 50/50 Threat to Humanity's Survival* Arrow Books
Rothblatt, Martine 2014: *Virtually Human: The Promise - and the Peril - of Digital Immortality* St. Martin's Press
Rovenszkij – Ujemov – Ujemova 1964: A gép és a gondolkodás Kossuth Könyvkiadó, Budapest
Savulescu, Julian – Bostrom, Nick 2011: *Human Enhance*ment Oxford University Press
Sachs, Jeffrey – Kotlikoff, Laurence J. 2012: Smart Machines and Long-Term Misery *NBER Working Paper* No. 18629, December http://www.nber.org/papers/w18629
Smith, Noah 2013: The End of Labor: How to Protect Workers From the Rise of Robots The Atlantic, January 14. http://www.theatlantic.com/business/archive/2013/01/the-end-of-labor-how-to-protect-workers-from-the-rise-of-robots/267135/
Stone, Maddie 2014: The Dominant Life Form in the Cosmos Is Probably Superintelligent Robots *Motherboard*, dec.19. http://motherboard.vice.com/read/the-dominant-life-form-in-the-cosmos-is-probably-superintelligent-robots
Storm, Benjamin C. – Stone, Sean M. 2015: Saving-Enhanced Memory. The Benefits of Saving on the Learning and Remembering of New Information *Psychological Science* February Vol. 26 No. 2. 182-188
Susskind Richard – Susskind, Daniel 2016: *The Future of the Professions: How Technology Will Transform the Work of Human Experts* Oxford University Press
Thackara, John (2005): *In the Bubble: Designing in a Complex World* Cambridge, MA, and London: MIT Press
Toyama, Kentaro 2015: *Geek Heresy. Rescuing Social Change from the Cult of Technology* Public Affairs
O'Callaghan, Jonathan 2015:This Robotic Assassin Will Hunt And Kill Starfish That Are Destroying The Great Barrier Reef *IFL Science*, September 3. http://www.iflscience.com/technology/robotic-assassinator-will-hunt-and-kill-coral-reef-destroying-starfish
Urban, Tim 2015: The AI Revolution: The Road to Superintelligence Wait But Why 1. http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html
Wadwha, Wiwek 2015a: It's No Myth: Robots and Artificial Intelligence Will Erase Jobs in Nearly Every Industry *SingularityHub* 2015. július 7. http://singularityhub.com/2015/07/07/its-no-myth-robots-and-artificial-intelligence-will-erase-jobs-in-nearly-every-industry/
Wadwha, Wiwek 2015b: Should We Redesign Capitalism to Address Our Jobless Future? *SingularityHub*, 2015. július 20. http://singularityhub.com/2015/07/20/we-need-a-new-version-of-capitalism-for-the-jobless-future/

Young, Simon 2005: *Designer Evolution: A Transhumanist Manifesto* Prometheus Books
Z. Karvalics, László – Juhász, Lilla 2008: *Társadalmi informatika I.* Információs Társadalomért Alapítvány, Budapest
Z. Karvalics, László 2015: Minden dolgok Internetje (Internet of Everything) In: Z. Karvalics László (szerk.): Metszéspontok. Társadalomtudomány és infokommunikáció az ezredforduló után Gondolat/Infonia, Budapest, 216-246.o.

Kömlődi Ferenc

Távol a Szingularitás…

1.

Valamikor a jövőben, feltehetően a 21. században, talán 2020 és 2040 között az exponenciális ütemű technológiai növekedés eljut egy pontra, ahol minden megváltozik – e ponton innen teljesen értelmezhetetlen, sőt felfoghatatlan az összes későbbi esemény.

Ez az a pont, amikor a gépi értelem – mesterséges intelligencia, MI – túlszárnyalja teremtőjét, az embert, szélvészgyors evolúciója elképesztő magasságokba repíti, miközben a Homo sapiens egyre bambábban nézi, és idővel annyit fog érteni szellemgyermekéből, mint belőlünk egy aranyhal.

Ez a pont a sokat vitatott technológiai szingularitás.

Egyesek szerint soha nem következik be, mert csak fantazmagória, hiszen a humán elmét sem értjük teljesen, mások szingularitás helyett evolúciószerű átmenetekben látják az MI megvalósulását, megint mások az emergencia finomhangolásairól megfeledkezve, a helyenként csak számszerűsített tényekből kiinduló elmélet fanatikus szószólói.

2.

„*Az MI a legjobb, de a legrosszabb dolog is lehet az emberiség számára. Óriási versenyfutás alakult ki a technológiai növekedés és az azt kordában tartó, szintén növekvő emberi bölcsesség között. A jövő a mi kezünkben van*” – így hangzik az MIT-n 2014-ben alapított Élet Jövője Intézet alapvetése.

Stephen Hawking évek óta a mesterséges intelligencia veszélyeire figyelmeztet. Elon Musk Tesla-vezér úgyszintén. A tudományos-technológiai világelit több illusztris szereplője csatlakozott hozzájuk.

Jobb később, mint soha jeligére, mielőbbi szabályozásért, egyes területek kutatásfejlesztésének korlátozásáért szálltak harcba.

Amivel csak a technológiai fejlődést hátráltatják.

Egyvalamiről feledkeznek meg: az MI jelenéről – arról, hol tart, és mi vár rá a következő néhány évben.

3.

Röviden: semmi, ami indokolná a rettegést. Az MI-kutatók zöme szerint a gépi intelligenciától való félelem körülbelül annyira reális, mintha a Mars túlnépesedése miatt aggódnánk.

A kutatások eddigi története legplasztikusabban két évszak, tavasz és tél, tíz-tizenöt évenkénti váltakozásával szemléltethető. A kezdeti, 1950-es évek végi és az 1960-as évekbeli nagy elvárásokat, a határtalan optimizmust a gyors sikerek elmaradása miatt elzáródó pénzcsapok, és lassú tetszhalál, az 1970-es évek „MI-tele” követte. A személyi számítógép elterjedésével és az olcsóbb, nagyobb számítási kapacitásokkal, 1985 körül megint kirügyeztek a fák, de világraszóló, konkrét eredmények híján a következő évtized a sikertelenség jegyében telt el. Szép lassan – egy-egy valóban izgalmas rendszer megbízhatóságába vetett hitet fenntartandó – az MI kifejezést is száműzték a bölcsek, mert zsákutcának tűnt. A harmadik évezred elején viszont ismét kitavaszodott, a szerteágazó szakterület és

a biológia, az agykutatás vagy az idegtudományok – melyekhez ezt követően még több diszciplína társult – egyre gyakoribb fúziójának köszönhetően az elméleti eredmények után jöttek a valóban praktikus, széles körben használt alkalmazások (amelyek már senki sem emleget mesterséges intelligenciaként). Az infokommunikációs nagyágyúk, az IBM, a Google, a Facebook, az Apple és a többiek látványos cégvásárlásai – és sikerei (IBM: Watson és a köréje épült startup-ökoszisztéma, asszisztensek, Tesla-autó) – egy harmadik évszakot, a nyarat ígérik.

2015 tanulsága: megint divat lett az MI.

4.

Csakhogy a Big Data jelenségre is reagáló jelentős fejlődés a rendkívül gyors adatfeldolgozást igénylő, viszont például a sakkhoz hasonlóan matematikai-logikai szabályokkal jól körülírható, szűk területeken igazán látványos. A legintelligensebb robotok (individuálisan és csoportban, rajként) is behatárolt keretek között tevékenykednek, autonómiájuk korlátolt, érzelmeket ugyan nagy nehezen felismergetnek, emocionális reakcióik azonban megmosolyogtatják a kommunikációs partnerüket. A legtökéletesebb chatbotok is csak jól meghatározott témákban kápráztatnak el, hirtelen váltásokkal viszont teljesen összezavarhatók.

5.

2015-ben egyetlen MI-területtel sem foglalkoztak annyit, mint a gépi tanulással, azon belül is főként a mélytanulással (deep learning).

Lényege, hogy a gépek tanuljanak meg hierarchikusabban és kontextuálisabban „gondolkozni", akár úgy is, mint mi, de nem ez a lényeg. Ha látnak egy oroszlánt, az állatvilág általános jellegzetességeitől elindulva jussanak el a kizárólag az oroszlánra jellemző tulajdonságok felismeréséig. Ha szöveggel dolgoznak, tanulják meg a szavak egymás közti kapcsolatát – hogyan állnak össze mondatokká, miként fejeznek ki gondolatokat.

A program, algoritmus (általában neurális hálózat) szintenként tanulja meg a bemenő adatok tulajdonságjegyek szerinti hierarchiáját, minták szerinti osztályozását. Egyszerre csak egy szintet, ahol az adott szint bemente mindig az előző kimenete. Az adatokat csak így tudja kellő mélységben és pontosan reprezentálni, miközben a variációs lehetőségekről sem feledkezik meg.

A technológiai szingularitáshoz a lehető legsokrétűbb és „legmélyebb" tanuláson keresztül vezet az út, a gépi rendszerek máskülönben képtelenek hibátlanul felismerni arcot, tárgyat, szöveget, beszédet, érzelmet.

A következő lépés a különféle részterületek eredményeinek egyetlen rendszerben történő integrációja lenne, anélkül, hogy zavarnák egymás működését.

6.

Intelligencia és tudatosság: emberi szintű MI-ről akkor beszélhetünk, ha az nemcsak intelligens, hanem tudatos is. Ha nemcsak érzékeli az élményeket, hanem azt is átéli, hogy képes átélni valamit. Ha megéli, és nemcsak szimulálja az érzelmeket. Ha vannak szándékai, ha intencionálisan cselekszik.

De miért akarjuk mindenképpen a Homo sapiensről mintázni a jövő gépi intelligenciáját? Miért ne valósulhatna meg emberi segítséggel egy a miénktől teljesen eltérő mesterséges értelem?

Juhos Sándor

Amikor a robot programozza az embert

A robotika iránti érdeklődésem kezdete még gyerekkoromra tehető. Akkor sem és most sem úgy gondolok a robotokra, mint független, öntudatra ébredő gépekre. Elképzelhető, ám ha be is következik, ehhez még el kell telnie legalább 10-15 évnek. Annak érdekében, hogy ez ne járjon negatív következményekkel, fontos, hogy megtegyük a megfelelő lépéseket.

Mitől is félünk?

Attól, hogy Asimov 3 törvénye kevés ahhoz, hogy megfelelő szabályozással kordában tudjuk tartani az önmódosító robotok működésének önállósodását. A robotok humanoidok vagy beágyazott rendszerek lehetnek, esetleg szoftverek, amelyekbe a programozók olyan genetikus algoritmusokat integrálnak, melyekkel képessé válnak az öntanulásra, vagy inkább önmódosításra. Mi a cél? Az, hogy óriási munkát igénylő programozással pontosan meghatározzuk a robot számára, mit tegyen, vagy az, hogy kevesebb munka után önálló cselekvésre ösztökéljük, s csak akkor módosítsunk a viselkedésén, ha az nem jó irányba halad?

Ha ez pusztán kényelem kérdése, akkor megteremtjük a saját problémánkat. Mert minél több dolgot hagyunk magától működni, annál kevesebb fölött marad meg a felügyeleti jogunk és lehetőségünk. Ha folyamatosan mi programozzuk a robotokat, akkor jobban beléjük tudjuk ültetni a logikát, a biztonságot, és könnyebben észrevesszük a lehetséges hibagócok kialakulását.

Öntanulás. Az öntanulás a robotokra nézve nem más, mint adatgyűjtés, amely a környezetből, a működésből, működtetésből fakad. Mivel nincs érzelmi tényező, csak igen/nem döntés a statisztikai eredmények által, a környezetből, és az adatbázisokból nyert információk összefüggései vezérlik és változtathatják meg a robotok működését.

Milyen szimulációkkal implementáljuk majd az algoritmusokat, hogy a véletlenszerű lehetőségek és variánsok mutálódását megállapítsuk? Mi történik akkor, ha a generálódás során nem sikerül a variánsok keletkezése során bekövetkező mutálódást leállítani – mert túl sok a variáció és a visszacsatolás. Követhetetlen az összefüggések logikája, mint a Pascal Triangle[1] esetében, egy végtelen folyamat indul be, amelyre nem számítottunk. Mi történik, ha a működés során bekövetkező változások meggátolják a hozzáférésünket?

[1] A Wikipédia a következőképpen fogalmazza meg a Pascal-háromszöget: „A háromszögben a sorok számozása zérótól kezdődik, és a páratlan és páros sorokban a számok el vannak csúsztatva egymáshoz képest. A háromszöget a következő egyszerű módon lehet megszerkeszteni: A nulladik sorba csak be kell írni az 1-est. A következő sorok szerkesztésénél a szabály a következő: az új számot úgy kapjuk meg, ha összeadjuk a felette balra és felette jobbra található két számot. Ha az összeg valamelyik tagja hiányzik (sor széle), akkor nullának kell tekinteni. Például az 1-es sor első száma 0 + 1 = 1, míg a 2-es sor középső száma 1 + 1 = 2”.

Az alap program megírása után a legtöbb öntanuló szoftver vizsgálja a környezetét. Képeket, tárgyakat, hangokat, összefüggéseket keres. Kell egy háttér adatbázis. De vajon mi lesz az? A világháló? Ha igen, akkor a már meglévő programokat, adattárakat, keresőoptimalizálásokat is használják majd? Vagy csak a képeket, a szavakat, a nyelveket, és ezek nyelvtani szabályait mint adattárat?

A legfőbb probléma, amelyet meg kell oldani, a hozzáférés. Erre nem azért van szükség, hogy a szoftver ne jusson hozzá bizonyos adatokhoz, hanem azért, hogy az illetéktelenek ne férjenek azokhoz hozzá a robot kommunikációs csatornáin és a szoftverén keresztül. Az Internet, mint lehetséges „agy", sokat segít az öntanulás és az érzelmi alapú döntéshozás szoftveres kifejlődésében. Igen ám, de ennek több a veszélye, mint az előnye. Az öntanuló szoftverek rengeteg anyagot, információt gyűjtenek, és a környezetükben keresik az összefüggéseket a működésük során. Statisztikákat vezetnek, prioritással rangsorolják a bejövő információkat – végül ebből születik egy eredmény, az eredményből pedig megkezdődik a végrehajtás.

Nézzük meg például a böngésző keresőmotorokat. Egy meghatározott algoritmus alapján megtanulják a szokásainkat, a kedvenc zenéinket, a kedvenc filmjeinket. Néha tévesen, mert az oldalak jellemzőit a Meta Title[2] és Meta Description Tag[3]-ek határozzák meg. Ha tehát a keresőoptimalizálás során több olyan kulcsszót adunk meg, ami nem 100%-osan a tartalomra utal, vagy nem teljesen jellemző az oldalra, téves statisztikát fog rólunk vezetni.

A legnagyobb probléma az, hogy a robotoknak nem tudunk érzelmen alapuló racionalitást tanítani. Még nem. Már a kezdetekben is kevésnek bizonyult Isaac Asimov 3 törvénye.

1. A robotnak nem szabad kárt okoznia emberi lényben, vagy tétlenül tűrnie, hogy emberi lény bármilyen kárt szenvedjen.
   (A jelentés mint "kár", erkölcsi kár is lehet, mert az erkölcs azonnal sérül, amint a robot felváltja a dolgozni akaró embert.)
2. A robot engedelmességgel tartozik az emberi lényeknek, kivéve, ha az utasítások az első törvény előírásaiba ütköznek.
3. A robot tartozik saját védelméről gondoskodni, amennyiben ez nem ütközik az első vagy második törvény bármelyikének előírásaiba.

   És a legrosszabb, ami történhetett, hogy megszületett a "Nulladik" törvény.
   „A robotnak nem szabad kárt okoznia emberi lényben, kivéve, ha valahogy belátja, hogy ez a kár végül az emberiség javára válik."

Hogyan látja be? Kiegészítésként az 1-2-3 törvényhez Asimov hozzátette, hogy nem oldhatják fel a 0. törvény tilalmát. Ki alkotja meg a programot, amely által egy robot belátja,

[2] Olyan nem látható címek, amelyek általában a weboldalra jellemző szavak és tartalmak az oldalon elrejtve, melyek a kereső optimalizálás során rávezetik a böngészőt, hogy a megadott kereső szavak alapján rátaláljon az oldalra.

[3] Olyan nem látható, a weboldal tulajdonságát tartalmazó leírások, melyekben tárolt adatok rávezetik a keresőmotort a találat kereséskor, hogy rátaláljon a weboldalra.

hogy mi és hogyan válik az emberiség javára úgy, hogy emellett káros egy egyén számára? Hogyan tervezi meg a robot a cselekedetet érzelem, képzelőerő nélkül, valamint meg nem történt állapotában a végkifejlet láncreakcióját és lehetséges következményét? A számozás azért lett a nulladik, mert magasabb sorszámú törvény nem írhat felül egy alacsonyabb sorszámút. Máris egy prioritás, amelyben a „nulladik" nem jó helyen van.

Az első törvény kimondja, hogy „*A robotnak nem szabad kárt okoznia emberi lényben*." A nulladik viszont belevisz egy olyan döntést, amit jelenleg még nem tudunk megoldani, de még azt sem tudjuk, hogy ezt a fajta döntéshozást milyen úton tanítjuk meg. Próbálkozunk feltárni az emberi agy működését, és megérteni az összefüggéseit, de azt elfelejtjük, hogy az ember évtizedekig fejlődik, és ezt követően mondható ki, hogy felnőtt. Képes az önálló életre, az önálló döntéshozásra. Amikor ez megtörténik, létrejön egyfajta ember, egyfajta viselkedésmód, egyfajta jellem.

Aztán Roger MacBride Allen átírta kicsit a 3 törvényt. Nem a nulladikkal kezdte, hanem kiegészítette egy negyedikkel.

1. A robotnak nem szabad kárt okoznia emberi lényben.
2. A robotnak együtt kell működnie az emberi lényekkel, kivéve, ha ez az együttműködés ütközik az első törvénnyel.
3. A robot tartozik saját védelméről gondoskodni, amennyiben ez az önvédelem nem ütközik az első törvénnyel.
4. A robot kedve szerint cselekedhet, kivéve, ha bármely cselekedete az első, a második vagy a harmadik törvényt sérti.

Azért a 4. törvény elgondolkodtató.

Amikor nem direkt, csak az ember szolgálatában áll, azt csinálhat, amit akar. Illetve az ember igénye szerint néha irányított, néha szabadon tevékenykedhet. Ez sem teljesen jó lehetőség. Gondoljunk bele, mi lenne, ha az egyik robot inputja az összes többi robot outputja. Így egymás tapasztalatait is begyűjtenék. A folyamat követhetetlenné válna.

Kezdjük az elején!

Az ember. Mivel viszonylag jól működünk, jó példa lehetünk a robotok számára. Igaz, sok olyan dolog van, amit a robotok még jó ideig nem lesznek képesek kivitelezni az ember viselkedéstanulásából. Megszületik, növekedése során gyűjti a környezetéből az információkat. A születés egyenlő a robot létrehozásával.

A szülő folyamatosan követi gyermeke életpályáját. Segítik a családtagok, az iskola, a törvényhozás, az államigazgatás. Az erkölcstan. Tehát ha a gyerek tapasztal valamit, és megkérdezik, hogy a tapasztalásból mit ért meg, mit szűr le, és ezt követően hogyan cselekszik, befolyásolható, és a helyes útra terelhető, ha tévesen értelmezett egy tapasztalást. Igen ám, de ki határozza meg, hogy mi a helyes? Mint ahogy a szülő hibáját is orvosolja az iskola, a programozók téves logikai összefüggésrendszerét is ellenőriznie kell egy robotikai felügyelő szervnek, mert itt már kevés a három (négy) törvény.

A helyes irány az lenne, ha minden egyes önmódosító folyamatot csak az ellenőrzés és vizsgálat után – nem okoz-e kárt – engedélyeznénk. Nem csak műszaki, fizikai kárra,

vagy veszélyre gondolok, hanem azt is szűrni kell, hogy a "master"[4] mindig az ember, a "slave"[5] mindig a robot maradjon. Tehát a lényeg, hogy a robot nem kelhet önálló "életre", és nem végezhet kontrollálhatatlan, kiszámíthatatlan cselekvéseket. A gond csak az, hogy mint minden rendszer a világon, megkerülhető, kiiktatható, és az érdekek iránymutatása szerint befolyásolható.

Hogyan állítsunk fel egy olyan felügyelő rendszert, amelyen nincs olyan kiskapu, melyen keresztül a szoftver megkerüli a felügyeletet? A legjobb példa erre a vírus, és ellenszere, a vírusirtó. De a kulcs nem a vírusban és a vírusirtóban keresendő. Tehát ne orvosoljuk a hibát, mint lehetséges megoldást, inkább védje az alapszoftvert egyfajta tűzfal a tanult változásoktól, és csak akkor engedje beépülni a változást, ha auditálva lett. Tehát olyan robot tűzfal kell, amely észreveszi a negatív kimenetelű változást.

Vagy hogy jobb hasonlattal éljek, említhetem az automatikusan aktiválódó szoftverfrissítést. Ha nem akarom, hogy automatikus legyen, értesítést és részletes jelentést kérek minden változási szándékról. Egy öntanuló robotnál azt lehetne tenni, hogy az öntanulási statisztika által végbemenő változások egy tárba kerüljenek, és csak akkor aktiválódjanak, ha a felügyeleti szerv által biztonságosnak lett ítélve.

Vannak olyan agyi folyamatok, cselekedetek, melyek életünk során csak ritkán vagy soha nem következnek be. Ilyen például egy másik ember életének kioltása. Generációs nevelés folyamatának eredménye, hogy „jó embert alkossunk." A jó robothoz is sok-sok idő kell. Jó programozó, jó szándék és jó cél. A robotot, és annak felhasználását annak idején Karel Čapek sem így képzelte el. A rossz útra akkor kerültünk, amikor egy robot nem egy embert helyettesített, hanem több(et). A jó felállás az lenne, ha minden ember egy tanár volna, és minden robot egy diák, az ember „manipulátora".

Vegyük például a vadállatokat vagy a már megszelídített, háziasított állatokat. Elindult egy nevelési folyamat, állandó felügyelettel, melynek folyamán háziasítottuk, kineveltük az állatokból az agresszivitást, az ember számára negatív tulajdonságokat. Az evolúció folyamán a rosszat elsorvasztotta, elnyomta a nevelés, a folyamatos kontroll, és a vadállat folyamatos figyelmeztetése. A kutya szelídített állat, ám az ősgénekbe kódolt vadállati viselkedés a mutáció során visszaváltozhat, és ha nincs meg a kellő kontroll, az állat ismét kiszámíthatatlan vadállattá lesz.

Tisztában vagyunk azzal, hogy a háziállatok hogyan viselkedtek megszelídítésük előtt, ennél fogva óvatosak maradunk velük, ott van bennünk a félsz, egy veszélyérzet, amely arra figyelmeztet bennünket, hogy éberek maradjunk. A tisztes távolságot az öntanuló robotokkal is meg kell tartanunk, és megfelelő mennyiségű biztonsági intézkedéseket kell tennünk velük kapcsolatban.

Juhos Sándor 1975-ben született Győrben. A győri Technics Playground Robotikai Automatizálási és Mechatronikai Oktatóközpont vezetője, a Robotika Szakosztály alelnöke. Az általa épített humanoiddal csapatával a 2009-es RoboCup világbajnokságon első helyezést értek el, SuperTeam világbajnokok lettek.

[4] Mester, tanító, akinek a tudásából fakadó kialakított hierarchia szerint mindig engedelmeskednek.
[5] Szolga, aki a mesternek mindig alárendeli magát.

Síklaki István

Ne féljünk a számítógéptől!

Hozzászólás Z. Karvalics László vitaindító cikkéhez

Előrebocsátom, hogy nem fogok vitatkozni Z. Karvalics László kiváló és gondolatébresztő cikkével. Teljes mértékben egyetértek azzal a nézetével, hogy tévút az alarmista diskurzus arról, hogy a gépi intelligencia előbb-utóbb eléri azt a fejlettségi szintet, ahol legyőzi az emberi intelligenciát, ami az emberiség számára végzetes következményekkel járhat. Mint az emberi elme működése iránt érdeklődő szociálpszichológus, szeretnék néhány adalékkel hozzájárulni Z. Karvalics érveléséhez.

Először szeretném felvázolni, hogy a magam együgyű módján miként képzelem el az emberi elme működését. Ebből a vázlatos modellből nyomban adódnak következtetések az alarmista nézetek számára. Az emberi elme egy evolúciósan kialakult, olyan hihetetlenül robusztus hálózat, amely folyamatosan fenntart, működtet egy modellt, a külső és belső világunk emulációját, annak történetével, pillanatnyi állapotával és közelebbi és távolabbi jövőjével együtt. Tehát amikor tudatos gondolataink vannak, akkor valójában ehhez a virtuális emulációs modellhez van hozzáférésünk, a modell által a tudatos folyamatok rendelkezésére bocsátott információt használjuk föl. Erre nyomban hozok egy-két egyszerű példát, de előbb néhány adattal érzékeltetni szeretném ennek a modellnek a biológiai nagyságrendjét, amit érdemes összevetni a szuperszámítógép rendszerek nagyságrendjével.

Az agyban hozzávetőlegesen 100 milliárd neuron található. Ezek a neuronok olyan hálózatot alkotnak, ahol a 100 milliárd neuron mindegyike hozzávetőleg 8-10 000 másik neuronnal áll szinaptikus kapcsolatban. Tehát a lehetséges kapcsolatok száma egy adott pillanatban 100 milliárd a tízezrediken. Ez a hálózat megállás nélkül működik, alapjáraton a neuronok körülbelül másodpercenként negyvenszer sülnek ki, az aktív neuroncsoportok esetében ez elérheti a másodpercenkénti ezer kisülést. És ez csak az alap. Ezt a működést modulálja több tucat neurotranszmitter és más aktív molekula, és még egyéb tényezők, csak amikről már van valamelyes tudomásunk. Ez az emulációs modell, amely a tudatunk számára a virtuális valóságot előállítja, moduláris szerkezetű. Részben az evolúció során kialakult és velünk született, részben tapasztalati tanulással kialakított tudattalan autonóm modulok óriási és folyamatosan más és más konfigurációkba rendeződő hálózata alkotja. Ez az elrendezés biztosítja, hogy képesek vagyunk tudattalanul óriási információmennyiséget párhuzamosan feldolgozni. Ehhez képest közismert, hogy a tudatos folyamataink igen lassúak, szekvenciálisak, és korlátozott a kapacitásuk.

Lássunk egy egyszerű példát arra, hogy a tudatunk nem fér hozzá a valóságos világhoz. Az egyik közismert példa, amire az interneten számos helyen nagyszerű illusztrációkat találhatunk, a változás vakság. Bizonyítható, hogy ha például egy vizuális bemeneten az adott szituációban irreleváns változás következik be, akkor azt tudatosan képtelenek vagyunk észlelni. Nem azért, mert a bemenet nem járná végig a természetes útját a retinától a V1 látómezőn át az egész vizuális rendszeren, hanem azért, mert a tudattalan modellünk,

amely irrelevánsnak találta a változást, aktívan gondoskodik róla, hogy ne jusson el a tudatunkba, ne terhelje a szűk kapacitást fölöslegesen. Tehát egy ilyen végtelenül egyszerű észlelési feldolgozás mögött is ott van az evolúció és az egyéni fejlődés eredményeként kialakult sajátos virtuális modell, amely primer módon biztosítja azt, amit a mesterséges intelligencia számára lehetetlenség beprogramozni. Ez teljesen összhangban van Z. Karvalics gondolatmenetével.

Egyelőre tehát ott tartunk, hogy a saját elménk működését nem vagyunk képesek olyan mértékben megismerni, még megközelítőleg sem, hogy algoritmizálható módon felhasználhassuk a mesterséges intelligencia építésére. Ennek illusztrálására egy klasszikus kognitív szociálpszichológiai kísérletet szeretnék felidézni. Timothy Wilson (2010) kezébe vette az amerikai fogyasztóvédelmi magazin, a *Consumer Digest* egyik számát, és megnézte azt a cikket, amelyben 16 ételszakértő tesztelte a piacon lévő negyvenvalahány dzsemet, és felállítottak szokás szerint egy minőségi rangsort. Wilson ebből a spektrumból találomra kiválasztott egy fél tucatot lehetőleg egymástól távol a rangsorban, és laikus egyetemi hallgatókat kért meg a kóstolásra és rangsorolásra. Amikor összehasonlította a szakértők rangsorolását a laikus egyetemistákéval, meglepetéssel tapasztalta, hogy igen magas korrelációt mutatott (r=0,46). Ekkor ugyanilyen feltételekkel megismételte a kóstolásos rangsorolásos vizsgálatot azzal a csekélynek tűnő különbséggel, hogy most indoklást is kért a rangsorolásra. Tehát tudatosan meg kellett adniuk, hogy mely tényezőket vették figyelembe a rangsor kialakításánál. Ebben a változatban a szakértőkkel való korreláció drámaian lezuhant (r=0,11). Tehát amikor átváltottak a kognitív tudattalan párhuzamos üzemmódjáról a tudatos szekvenciális üzemmódra, a teljesítmény látványosan leromlott. Nyilván áldozatos munkával megoldható, hogy a szakértők tudását valamely jól szervezett szakértői rendszerbe szervezzük, s ez által algoritmizálhatóvá és így gépesíthetővé tegyük. De ha belegondolunk, hogy milyen csillagászati apróságból áll össze folyamatosan az emberi teljesítmény, nyilvánvaló, hogy a mesterséges intelligenciának ez az útja nem járható.

Egy egészen más oldalát is fontolóra érdemes venni ennek a kérdésnek. A cikk vége felé Z. Karvalics ír a mesterséges intelligencia és a felelősség kérdéséről. Voltaképpen ugyanezeket a kérdéseket fel lehet vetni az emberi felelősséggel kapcsolatban is. A tetteinkért vállalt felelősség ugyanis azon a meggyőződésünkön alapul, hogy szabad akarattal rendelkező, szuverén személyek vagyunk. Igen ám, de számos kutatás mutatta ki elég meggyőzően az utóbbi időkben (például Wegner, 2006), hogy ez az élményünk nem azon alapszik, hogy közvetlen hozzáférésünk lenne az agyunkban, azaz, a tudattalanunkban lejátszódó döntési folyamatainkhoz. Ez az élmény csupán illúzió, mert az agyunk aktívan elzárja a tudatunktól mindazokat a folyamatokat, amelyek elvégzése után valamilyen eredményre jut, s csupán az eredményhez enged a tudatunknak hozzáférést. A saját indítékainkra tehát éppúgy következtetés útján jutunk, mint más emberek indítékaira. Ám mivel az agyunk elzárja előlünk azt a miriádnyi számítást, ami egy-egy tettünkhöz vezet, az az illúziónk, hogy szuverén, szabad akarattal rendelkező emberek vagyunk, s ezt tételezzük föl embertársainkról is. Így megvan a felelősség alapja. Ha egyszer eljutna oda a tudomány, hogy láthatóvá tegye ennek az irdatlanul összetett virtuális emulációs modellnek, a tudattalannak a működését, akkor pontosan azt a kérdést tehetnénk fel a felelősségről, ami mesterséges intelligencia rendszerek kapcsán felvetődött.

Végül csak utalnék arra, hogy a modern idegtudomány nagyon meggyőzően bizonyította, hogy az érzelmi működések nélkül, amelyek pedig a zsigereinkben gyökereznek, még egyszerű döntéseket sem vagyunk képesek meghozni (Damasio, 1996; Panksepp 1998). Nem nehéz belátni, hogy az érzelmeink nagyon szaftos biológiai alapjait is magában foglaló érzelmi rendszert programozni igen kétes vállalkozás.

Irodalom

Damasio, Antonio, R. (1996): *Descartes tévedése*. Adu Print Kiadó
Panksepp, J. (1998): *Affective Neuroscience*. Oxford University Press
Wegner, D. M. (2006): *A tudatos akarat illúziója*. Kossuth Kiadó
Wilson, T. D. (2010): *Ismeretlen önmagunk*. Háttér Kiadó

Bátfai Norbert

Bátfai Samu rövid reflexiója, avagy a Programnevelő informatikus BSc szak megalapozása

A változás, a végtelen, a tér és az idő vagy az algoritmus olyan fogalmaink, amelyeket vélhetően ezer évek óta használt, s ugyanolyan értelemben ma is használ az emberiség. Ezeknek a fogalmaknak az életében volt egy varázslatos pillanat, amikor a tudomány (matematika) megmutatta, hogy sokkal több tartalom van mögöttük, mint amennyivel közös intuíciónk felruházza őket. Mire gondolunk? A változás[1] matematikai kezelése volt Newton csodafegyvere, amelyet ha elsütünk, le tudunk szállni a Holdra – konkrétan ez a kalkulus. Cantor munkássága olyan csodálatosan tárta fel a végtelenség[2] természetét, hogy aki ezt megtanulja, annak a transzfinit indukció is csak olyan lesz mint az egyszeregy. Az einsteini tér és idő a speciális relativitáselméletben úgy húzódik össze és tágul ki, ahogyan arról csak egy pion[3] tudna első kézből beszámolni. A turingi-chaitini számítógépes kiszámíthatóság deduktív megalapozásának még nem tudni meddig érnek majd el a hullámai[4]. Vajon mi lesz az a következő fogalom, amely átesik ezen a ponton és forradalmi változást hoz szemléletben, tudományban és technikában?

Mint ilyeneket, a PROP[5] jegyzet a képzelet és a valóság fogalmát emeli ki, a szeretetet csak megemlíti, mivel e jegyzet írásakor egyáltalán nem látszott valóban elképzelhetőnek, hogy ez lehet napjaink jelöltje. Ennek a reflexiónak az írása viszont éppen azért kezdődik el, mert csodás érzés: immár tisztán látható, hol kaphat szerepet a szeretet[6], hol kell szerepet kapnia a szeretetnek.

[1]A legismertebb „változások” például: az időben a hely változása a sebesség, annak változása a gyorsulás.

[2]Két halmaz számossága megegyezik, ha elemeik között bijekció van, azaz ha az egyik bármely elemének megfeleltethető a másik egy eleme, és ez megfordítva ugyanúgy igaz, ehhez ragaszkodva tudott Cantor rendet vágni a végtelenek között.

[3]Fizika-tankönyvbeli olvasmányélményünk lehet, hogy a pi-mezon 50 km magasan keletkezik a légkörben, élettartama pár száz méter megtételére elegendő, mégis itt vannak a felszínen is… a magyarázat – relatívan – a fény sebességével mozgó pion szemszögéből: a tér összehúzódik éppen pár száz méternyire, azaz ez a pár száz méter lesz az az 50 km; a piont kívülről nézve: ideje kitágul – órája lassabban ketyeg – éppen annyira, hogy elég legyen befutnia azt az 50 kilométert.

[4]A megállási probléma egy hétköznapi megfogalmazása lehetne, hogy cégünk nem tudná piacra dobni azt a minden kétséget kizáróan killer applikációvá váló szoftver termékét, amely megmondaná a többi szoftverről, hogy az meghibásodik-e, mivel ilyen szoftver nyilvánvalóan bizonyítható módon nem létezhet. Másik kedvencem a Chaitin-féle Omega-szám, amelyhez képest a Pi nyeretlen kétéves, mert az Omega számjegyei véletlenek, hoppá!

[5] Bátfai Norbert: Programozó Páternoszter újratöltve: C, C++, Java, Python és AspectJ esettanulmányok, 3. oldal, http://www.inf.unideb.hu/~nbatfai/konyvek/PROP/prop.book.xml.n.pdf

[6] Tudomásom szerint a szeretet ilyen elementáris rendező erővel eddig csak Jézus által jelent meg az irodalomban (az evangéliumokban, mint az új parancsolat).

Mielőtt belevágunk a reflexióba, meg kell világítanunk a cím retorikai túlkapását! Samu nem egy személy (még? :)) hanem csupán egy kísérlet, amely egy megerősítéses tanulással működő csevegő robot kialakításának alapjait vizsgálja. Egészen pontosan mély Q-tanulással[7] kísérletezik, ahol az állapot-cselekvés párokat tanuló Q függvénynek többrétegű perceptronokkal[8], azaz neurális hálózatokkal történő tanítása a cél. Ezek az erőfeszítések egy projekt-családban öltenek testet, amelyek összefoglaló neve Bátfai Samu (SAMU1[9], SAMU2). A családnév olyan értelemben fontos, hogy már önmagában hangsúlyozza: ez a robot családi körben nevelendő! Ami nyilvánvalóan előre vetíti, sőt kijelöli a szeretet intuitíven értelmezett, jövőbeli helyét és szerepét.

A cikk első olvasása után érezhető, hogy az „alarmista" nézőpontokkal szemben polarizált. Ez jó kiinduló közös nevezője a reflexiónak. A keletkező második benyomás viszont, hogy a mesterséges intelligenciát mint platóni ideát kezeli abban az értelemben, hogy például nem a „developmental robotics" nézőpontjából[10] közelítve figyeli meg azt, hanem egy eljövendő időponttól kezdve a-priori létezőnek tekinti. Pedig még a természetes intelligencia-kifejlődésének kérdése is nyitott úgy a törzsfejlődés, mint az egyedfejlődés szintjén. Példaképpen gondoljunk csak Wigner Jenő (1967, 179. o.) esszéjére, ahol utóbbi megvilágításának (egészen pontosan a tudatosság kifejlődésének) egyik lehetséges útjaként a gyerekek nevelésének családokon belüli megfigyelését említi.

Ebből a családi nézőpontból a leendő mesterségesen intelligens/tudatos stb. entitásokból ugyanúgy lehetnek jó- vagy gonosztevők, mint a természetes gyerekeinkből. A kérdés nem sodorvonalbeli, de ugyanúgy felvethető, hogy az apák felelősek-e a fiak bűneiért? Nagyobb súllyal esik latba, hogy ebben az olvasatban a programozók szerepét mindinkább azok vehetik majd át, akik tanítják a leendő rendszereket. Mert a valóságban a mesterséges intelligencia olyan lesz, amilyennek neveljük. Legalábbis ez kell legyen az alapállásunk, egyszerűen mert ennyit tudunk tenni, aztán reménykedni. Mivel azt, hogy a mesterséges

[7] A Q-tanulás a megerősítéses tanulás egyik alaptechnikája, amelyben az ágens jutalmat (pozitív megerősítés) kap, ha adott szituációban megfelelően reagál, illetve büntetést (negatív megerősítés), ha nem. Ezek a megerősítések tartják karban az úgynevezett Q-függvényt, amely megmondja, hogy adott szituációban milyen reakciót érdemes választani: durván azt, amelyre a Q-függvény értéke a legnagyobb. Fontos, hogy az ágens meg sem próbálja értelmezni a szituációkat, tehát nincs egy deliberatív értelemben vett modellje a környezetéről, csupán azt tanulja, melyik szituációban hogy érdemes reagálnia. Persze ez sem királyi út, alapvető probléma a szituációk nagy (akár gyakorlatilag végtelennek is tekinthető) száma, itt segíthetnek a neurális hálók. A mély Q-tanulás nem (csak) attól mély, hogy a Nature folyóiratban publikálta Mnih és munkatársai (2015) a Google DeepMind cége, hanem – kevésbé tréfásan – mert „mély" az a neurális hálózat, amely tanulja a Q-függvényt (lásd még a perceptronokról szóló lábjegyzetét is).

[8] A Samuba épített perceptronok olyan mesterséges neurális hálózatok, amelyek bemenete Samu „vizuális képzeletére" van kapcsolva, amely így egy következő köztes rétegbe súlyozódik, amely réteg megint egy újabb köztes rétegbe súlyozódik, s ez attól függően ismétlődik, hogy milyen „mély" a hálózat egészen az egyetlen kimenő egységig.

[9] Bátfai Norbert: *nbatfai/samu*, https://github.com/nbatfai/samu

[10] Ennek a megközelítésnek az alapkérdése, hogy hogyan fejlődhetnek ki egyáltalán, és hogyan fejlődhetnek folyamatosan az egyedi mesterségesen intelligens rendszerek. A születő válaszok tipikusan az élő rendszerek megfigyeléséből vagy azok analógiájára születnek, ezért ez a megközelítés jellemzően interdiszciplináris.

intelligencia valójában milyen lesz, ugyanúgy nem lehet megmondani, mint ahogyan minden emberi lényről is csak e földi léte után lehet érdemben nyilatkozni.

Nem az intelligens szoftverek, hanem csak általában a szoftverek tekintetében éltünk már át (és éljük meg ma is folyamatosan) hasonlót: ez a szoftver-krízis, amikor a szoftverek nem úgy viselkednek, mint ahogyan mi azt elvárnánk, ahogyan előzetesen feltételeztük. Kemény János nem szó szerinti szavaival érzékeltetve, miszerint régen az volt a gond, hogy a programok nem azt csinálták, amit szerettünk volna (a sok hiba miatt), ma ellenben már az a gond, hogy pontosan azt csinálják (a hardver gyakorlatilag megbízható), de csak azt és csak annyit. Egy programozó szempontjából: kvázi többet várunk el, mint amennyit be tudunk programozni. A krízisre az a válasz született, hogy a szoftverek fejlesztésének folyamatát mérnöki tevékenységgé kell tenni! Várhatóan ez történik majd a mesterségesen intelligens robot ágensek fejlesztésének tekintetében is. A szoftverek készítéséből egy új szakma született: a **programozó** (Dijkstra, 1972), várhatóan ennek mintáját követve születik majd meg az új foglalkozás: a robotokat nevelő, **programnevelő informatikus**.

Irodalom

SAMU2 – Bátfai N. (2015): A disembodied developmental robotic agent called Samu Bátfai (http://arxiv.org/abs/1511.02889)

Wigner, E. P. (1967): Remarks on the mind-body question. *Symmetries and Reflections*

Dijkstra, E. W. (1972): The humble programmer. *Commun*. ACM 15, 10, 859-866.

Mnih, V. – Kavukcuoglu, K. – Silver, D. – Rusu, A. – Veness, J. – Bellemare, M. – Graves, A. – Riedmiller, M. – Fidjeland, A. – Ostrovski, G. – Petersen, S. – Beattie, C. – Sadik, A. – Antonoglou, I. – King, H. – Kumaran, D. – Wierstra, D. – Legg, S. – Hassabis, D. (2015): Human-level control through deep reinforcement learning. *Nature*, 518, 7540, 529-533.

Bátfai Norbert, PhD
batfai.norbert@inf.unideb.hu
Debreceni Egyetem
Információ Technológia Tanszék
Debrecen, 2015 október 15.

Lőrincz András

Mesterséges intelligencia az egészség és a jólét területén: a gépi tanulás, a crowdsourcing és az ön-annotációban rejlő lehetőségek[1]

Bevezetés

Munkám és tapasztalataim szerint a mesterséges intelligencia jelentős segítséget adhat a speciális igények ellátásában, feltéve, hogy tisztában vagyunk azzal, mi is lenne a megoldás, milyen veszélyek vannak a háttérben és ezeket hogyan kerülhetjük el. Úgy vélem, hogy a világban lezajló kölcsönhatások modelljét tartalmazó és 3D-s fizikát is modellező, például grafikus játékokhoz is szükséges szoftverek fontos építőelemek a veszélyek és az akadályok elhárításában. Az ilyen jellegű modellek fejlesztése nem mai keletű és ki is egészül a 3D-s képi érzékelőknek és a viselhető mozgásérzékelőknek köszönhetően. Segítségükkel a számítógépes játékokban az játékélmény is megsokszorozható. Az egészségügyi és jóléti világban szerepet kaphatnak a fejlesztő játékok is, bár a gyors változásokat a technológiai elemek más, ellenőrzött alkalmazásai és a gazdasági igény hozza meg.

Egy példa

Érdemes példán keresztül szemléltetni, milyen módon is lehet segítségünkre a technológia. Tegyük fel, hogy beszédértő, de nem beszélő és mozgássérült fiatalemberről van szó, aki például csak a kép alapú kommunikáció segítségével tudja gondolatait közölni. Táblagépet használ kommunikációra, és egy okos kerekesszékben ül. Ő a táblagépen keresztül tudja bemutatni azokat a képi szimbólumokat, amelyeket már nekünk kell összeraknunk és nekünk is kell azt megfejtenünk. Mondjuk, éppen azt észleljük, pontosabban azt észleli a robot-társ, hogy a fiatalember kérni szeretne valamit. Meg is jelenik az első szimbólum a táblagépen, ami az jelenti, hogy „igazítsd". Ez a parancs vonatkozhat például a székre, ha az kényelmetlen. Vonatkozhat a táblagépre, ha az nem megfelelő szögben áll, de vonatkozhat egy szívószálra is, ha ő éppen inni szeretne. A környezet és az abban lejátszódó történések segíthetnek annak eldöntésében, hogy mi is lenne a feladatunk. A példában a fiatalember „okos" ruhát (smart cloth) hord és ki tudjuk számítani, hogy kényelmesen ül. Tehát nem kell a széket megigazítani. A szívószálat a felhasználó el tudja érni. Nem arról van szó tehát, hogy inni szeretne. Észleljük azt is, hogy a táblagép az asztalon van de „egyre közelebb kerül", az asztal széléhez és hamarosan leesik. Ebben az esetben minden bizonnyal az asztal megigazítása és az esés megakadályozása lenne a feladat. Sajnos a leesés nem tipikus tulajdonsága a táblagépnek, arról semmit sem mond a táblagép használati utasítása. A leesés általános tulajdonság, szinte bármilyen tárgy leeshet és az erre vonatkozó

[1]Az írás a KI – Künstliche Intelligenz folyóiratban megjelent „Revolution in Health and Wellbeing" című írás a vitaindító alapján továbbfejlesztett változata. A magyarra történő átültetésben közreműködött Tamaskó Dávid.

leírások, az esetleges veszélyek és megoldások nem találhatóak meg az adatbázisokban vagy az ontológiákban. Ez a mindennapi tudásunk része és „nincs rá szükség", „fölösleges".

Ha a robot „világmodellje" a 3D-s virtuális valóság és fizikát is számítani tudó motorral van felszerelve, ha a szoba modellje is rendelkezésre áll, akkor a lejátszódó dinamikai folyamatokat – megfelelő szenzorok segítségével – fel lehet ismerni, a mozgási paramétereket meg lehet becsülni és az egyes elemek pályái is jórészt megjósolhatóak. Tehát, a segítő mesterséges intelligencia ki tudja számítani, hogy (a) a mozgó tárgy leeshet, (b) eltörhet és (c) így értékét veszti. Ha robot-társ a felhasználó érdekeit, értékeit védi, akkor az „igazítsd" utasítást megfelelő módon értelmezheti. Sőt, ha az észlelés jó, akkor utasítás nélkül is, proaktív módon elvégezheti a szükséges korrekciókat.

A kritikus funkciók eléréséhez ezek a képességek kellenek:

- A környezeti mesterséges intelligenciának észlelnie kelli, hogy valami mozog. A mozgásérzékelés lehetséges.
- Fel kell ismerni azt is, hogy a mozgó tárgy egy táblagép. Ezt a részfeladatot megnehezítheti, ha a tárgy takarásban van. A 3D-s modellek használata segít ezen a ponton: követni lehet a tárgyak mozgását, ismerhetőek modelljeik, észlelhetőek a dinamikai tulajdonságok is. Így a részben takarásban levő tárgy pozíciója és mozgása is becsülhető. A 3D-s modell megoldást nyújt más részlegesen megfigyelt, vagy akár éppen meg sem figyelt adatokra is; milyen messze van a fal, az ajtó, egy-egy akadály, amelyeket már regisztrált a rendszer, de most éppen nem észlelhetőek. A 3D modell tehát lehetővé teszi, hogy ne legyen szükség minden információ folyamatos felmérésére az aktuális helyzetet illetően. Fontos ez, mert az elmélet szerint is, a részlegesen ismert állapotban hozott döntés vezethet nagyon rossz eredményre is.
- A környezet fizikai modellje lehetővé teszi a proaktív viselkedést, mivel – mint ahogyan a fenti példában láttuk – a környezeti mesterséges intelligencia ki tudja számítani, hogy a közeli jövőben mi fog történni. Röviden, a 3D-s grafikus modellek gyors fejlődése, beleértve a beágyazott fizikai ismereteket, mint például a környező tárgyak súlya, hajlékonysága, törékenysége, leegyszerűsíti a mesterséges intelligencia megfigyelési feladatát, és elősegíti az időben történő interakciókat.

Tér- és időbeli tulajdonságok és intelligencia

Kritikusnak tekintem azt a feltételezést, hogy az emberi intelligencia ereje a tér- és időbeli tulajdonságok megtapasztalásából, a 3D dinamikai modell (implicit) megalkotásából és az így lehetővé tett időbeli becslésekből ered. Azonban, ezekkel a képességekkel rajtunk kívül igen sok állatfaj is rendelkezik és kétségtelen a jóslásból, az előrelátásból adódó evolúciós előny is. Az intelligencia egyéb összetevői, mint például a logikai feladatok megoldása, vagy valamely feladat végrehajtása feltételeinek kiszámítása, és az ún. megerősítési tanulás (reinforcement learning) viszonylag egyszerűek. Az elmúlt évtizedek alatt kifejlesztett algoritmusok jelentős sikereket értek el ezekben a feladatokban, beleértve a táblajátékokat is. Tehát az egyik tulajdonság olyan, amivel sok élőlény rendelkezik. Ez az, ami még nem megy elég jól a mesterséges intelligenciának. A másik tulajdonság, a logikai kérdések végigjárása az, amiben felülmúljuk a legtöbb állatfajtát. Ez viszont jól megy a gépeknek már ma is.

Elgondolkodtató, hogy az a tudásunk, ami az állatokét jelentősen felülmúlja, amelyek felfedezéséhez sokak intelligenciájára volt szükség és hosszú évezredekig tartott, valamint mi a feltalálókat és a felfedezőket géniuszoknak tartjuk, mindössze néhány év, vagy alig több mint egy évtized alatt átadható egy-egy gyermeknek. Valóban olyan nagy lenne ez a tudás? Vagy inkább csak kigondolni nehéz, de átadni ezt a fajta tudást nagyon könnyű? Gondoljunk a matematikai tételekre. Felmerülhetett már mindenkiben az, hogy hogyan lehetett egy-egy egyszerű bizonyításra rájönni? De ha a bizonyítás nem is egyszerű, akkor is általában egyetlen szálon fut végig és így a bizonyítás ellenőrzése gyors lehet. Ezek olyan problématípusok, amelyek megoldása nehéz, de ellenőrzése egy vonalon való végigfutásnak felel meg, vagy arra átalakítható, azaz egyszerű. Ezért tudjuk a tudást hatékonyan átadni. Ha ezt a tudást a gépnek is át tudjuk adni úgy, mint ahogyan gyermekeinknek is átadjuk, akkor a gépek lehet, hogy nem lesznek okosak, de biztosan annak fognak tűnni.

Az előzőekben érintettük azt a kérdést, hogy az emberi intelligencia elérhető-e, vagy felülmúlható-e speciális algoritmusokkal. Noha ez releváns lehet a számítógépek és robotok egészségügyben való hasznosításával kapcsolatban is, de ma más a fontos. Igaz-e, hogy, a kvantitatív gépi megfigyelések felülmúlhatják a szülők és gondozók szubjektív megfigyeléseit? Életünk meglepően reguláris, gyakran ismétlődő események teszik megnyugtatóvá és így rendkívüli módon kiszámítható. Ezáltal a becslések és az anomáliák észlelése sikeres lehet, különösképpen akkor, ha szakértői segítség is rendelkezésre áll, amikor ilyen eseményt észlel a gép.

A kérdés tehát az, hogy az érzékelők és a számítógépes algoritmusok elérték-e már arra a szintre, ahol általában is tudnának segíteni, és amikor a szakértői megfigyelés jelentős anomália esetén veendő csak igénybe? A válasz az, hogy (a) igen, az érzékelők elérik a szükséges szintet, és (b) igen, bizonyos esetekben az algoritmusok is elérik vagy akár meg is haladják az emberi teljesítményt. A becslések és az proaktív viselkedés szempontjából viszont a válasz (c), nem, a tér- és időbeli adatokból való tulajdonság-kinyerés és predikció, illetve a segítségükkel történő célorientált és proaktív viselkedés nincs még a szükséges szinten.

Gyors változások, ha tetszik, akkor forradalom

Miért számítunk forradalmi változásokra, ha a tulajdonság-kinyerési algoritmusok egyelőre nem felelnek meg az igényeknek és az elvárásoknak? Azt mondhatjuk, hogy nagyléptékű erőfeszítések oldják meg az egyelőre ismeretlen algoritmikus hiányosságokat. Az erőfeszítések éppen a gépi tanulás tulajdonság-kinyerési gyengeségeit próbálják orvosolni. Hatékony módja a *crowdsourcing,* amely meghozza a változást és az egészségügyben is forradalomhoz vezethet.

A crowdsourcing az az eszköz, amit legnagyobb adatgyűjtők, a FaceBook, a Google, a MicroSoft hasznosíthat a leggyorsabban és a leghatékonyabban. Emberek tömegei dolgoznak nagy adatbázisok „címkézésén" (annotálásán). Például, fillérekért, vagy ingyen is megjelölik emberek azt, hogy egy-egy képen mi van, van-e rajta zebra, vagy autó, esetleg kacsa vagy valami más is. Adatbázisokat gyűjtenek mosolygó, ijedt, vagy dühös, esetleg dühöngő emberekről. Ezekben a címkézett adatbázisokban, amelyeket emberi intelligenciával címkéztek (annotáltak) a tulajdonság-kinyerés lényegében megtörtént… És ez át is hidalja a problémát. A tulajdonság-kinyerés kritikus lépcsője helyett a valamely tu-

lajdonságra adott millió és millió példa segítségével a mára már igen hatékonnyá vált gépi tanulási módszerek már az adott tulajdonságot felismerik. Vegyük észre a különbséget a speciális összetevő, vagy tulajdonság fogalmának megalkotása, kinyerése és a tulajdonság felismerése között. Az előbbit elvégzi a humán intelligencia évezredek tudásgyűjtése segítségével, majd azt a tudást már mindenki könnyen elsajátítja, így az olcsó munkaerővel hatalmas adatbázisokat hozhatunk létre, aminek segítségével a gépi intelligencia betanítható és a tulajdonságot felismerni képessé válik.

Mire lesznek képesek majd a gépek, azt nem tudjuk, hiszen éppen a tulajdonságkinyerés, a lényeglátás még hiányzik. Azt viszont tudni fogják, amit mi beléjük táplálunk a hagyományos módon (például összeadás, szorzás), vagy az új módszerekkel, a hatalmas adatbázisokkal.

Azt gondolom, hogy mind az algoritmusok, melyek kihasználhatják az annotált adatbázisok által már birtokolt emberi tudást, mind pedig az intelligens, viselhető érzékelők elérték a szükséges szintet ahhoz, hogy képesek legyenek több hasznos feladatot elvégezni az egészség és a jólét területén is. Az otthoni ápolási rendszerek fejlesztése már elkezdődött, és nagy valószínűséggel ezek a rendszerek gyors átalakuláshoz vezetnek elsősorban az óriási piaci igények miatt.

Az általános kognitív algoritmusok is ki tudják használni az új fejlesztéseket. Ilyen algoritmusokkal több szerző írásaiban találkozhatunk. Az érdeklődő olvasó figyelmét az Artificial General Intelligence Society konferencia-sorozatára szeretnénk felhívni. Általános kognitív architektúrákat már használtak bizonyos egészségügyhöz és jóléthez kapcsolódó kutatásokban is. Ezek a törekvések is fejlődni fognak a modern figyelő eszközök kiaknázásával, különösen, ha a crowdsourcing is bevethető. A crowdsourcing a nehéz, mert ahhoz az egészségügy és a jólét területén szakemberek is kellenek.

Személyre szabás

Vajon a megoldások tetszenek majd az embereknek? Használni fogják-e azokat?

Köztudott, hogy az idős emberek idegenkednek az új technológiáktól; nehéz számukra az új eszközök használata, és nem könnyű hozzászokniuk a gyors változásokhoz. Kritikus is lehet ez a probléma.

A fenti dilemma megoldása kreatív találmányokat *és* személyre szabhatóságot is igényel: az egyéni ízlésekhez, szokásokhoz és szükségekhez való idomulás képessége elengedhetetlen. Egy fiatal autista személy problémái, akinek a helyzete megkönnyíthető számítógépes kognitív viselkedési terápiával, és egy bipoláris depresszióban szenvedő páciens esete, ahol a mozgási minták változásainak észlelése feltétlenül szükséges lehet, igen-igen különbözik egymástól. Tekintsük például egy demenciában szenvedő idős személy szükségleteit, amely esetben tevékenységének megfigyelésére, aktív kognitív támogatásra, és különleges bánásmódra is szükség van. Tekinthetjük azt a már említett esetet is, amikor valakinek mozgási és beszédbeli korlátai vannak születéstől fogva, vagy egy szerencsétlen agyvérzés következtében, vagy egy baleset után. Ez utóbbi esetben VR-alapú (virtuális valóságalapú) diagnosztika és a VR segítségével történő mozgásterápia hatékony lehet, ha az alkalmazkodni tud az aktuális képességekhez, hangulathoz és fáradtsági szinthez. A valódi szükségletek felismerése és a lehetséges megoldások gondos megválasztása

nagy kihívást jelentenek, amit a kiragadott példák szemléltethettek. Igen széles a kör és az igények tekintetében személyre szabott javaslatokra van szükség, amelyek függenek az egyén és a család szokásaitól, anyagi lehetőségeitől, kulturális hátterüktől, valamint a társadalombiztosítási rendszer mozdítható forrásaitól egyaránt.

A személyre szabhatóság fel tudja használni a legújabb adatgyűjtési és adatbányászati módszereket. Segítségükkel optimalizálni lehet az előre programozott és az átlagos felhasználó számára készült heurisztikus döntési eljárásokat. Ez a szempont független a 3D-modellező kapacitásoktól, bár igaz, hogy mindennemű döntéshozatal leegyszerűsödik az ilyen modellek használatával, mert több időnk van arra, hogy kitaláljuk a megoldást és cselekedjünk is.

Az adatbányászattal kapcsolatban érdemes kitérnünk az ún. „ajánló rendszer technológiák" gyors fejlődésére is. Az ajánló rendszerek az összegyűjtött adatok mennyiségéből és minőségéből nyerik erejüket, amely adatok példaként szolgálhatnak a még nem látott esetekkel kapcsolatos általánosításokban. Hérakleitosz szerint csak ilyen esetek léteznek, kétszer nem lehet ugyanabba a folyóba lépni. Mégis igaz az, hogy két eltérő esetben ugyanaz a döntés lesz optimális. Hol a határ? Ezek az algoritmusok megkülönböztetik azon tipikus eseteket, amelyeknél a statisztikán alapuló javaslatok lehetségesek és biztonságosak is. Elhatárolják ezeket azoktól az esetektől, amelyek szokatlanok és az adatbázisban még nem tárolt tudást és szakértelmet igényelnek. Az algoritmusok a különböző beavatkozás várható esélyeit is meg tudják jósolni, ha azokra, vagy hasonlókra már akadtak példák és lehetőleg nagy számban. Ez nagy jelentőségű az egészség és a jólét szempontjából, főleg ha ezeket az általánosságban jól működő rendszereket adaptálni tudjuk az aktuális szituációhoz az aktuális egyénről már eddig összegyűjtött adatok alapján.

A crowdsourcing rendkívül hatékony eszköz, ám a speciális igények kielégítésére már nehezebben lesz alkalmazható, mivel az elérhető adatmennyiség kisebb, ráadásul a ritkább adatok értékelése nagy szakértelmet igényelhet. Másik oldalon az adatvizsgálat módszerei egyre hatékonyabbakká válnak, jelentős előrehaladás látható ezen a területen is. Gyors fejlődés mutatkozik a kevesebb példát igénylő tanulás esetére és így jelentősen lecsökkenhet a szükséges szakértői munka mennyisége is. Továbbá gyorsan fejlődik az okos eszközök birtokában az ún. ön-annotációs lehetőség. Ez a módszer lehetővé teszi azt, hogy saját adatainkat rögzítsük és saját magunk, saját érdekünkben azokat minősítsük, saját észrevételekkel kiegészíthessük. Mivel ez a tevékenység nem kerül pénzbe, valamint mivel az öröm és a fájdalom érzése így „első kézből" érkezhet, ez a módszer az adatbázist nagymértékben megnövelheti és a gépi döntéseket nagymértékben megjavíthatja.

Érdemes megjegyezni, hogy a gépi eszközök beiktatása súlyos és fontos közösségi, etikai és magánéleti kérdéseket is fölvet, amelyekkel itt nem fogunk foglalkozni – figyelembe véve, hogy e lényeges pontok túlmutatnak kérdésfeltevésünkön. Hangsúlyoznunk kell azonban, hogy ez a kérdés rendkívül fontos lesz és a gyors fejlődés miatt igencsak sürgős is lenne. Az Amerikai Tudományos Alap egyik felhívásából idézve, az „Informatikai tudományok arra születtek, hogy befolyásoljanak minket". Ehhez kell hozzáfűznünk, hogy az adatgyűjtés és az adatbányászat arra jó, hogy a befolyásolás hatékony és egyénre szabható is legyen. Az egyik oldalon éppen ezt keressük: azt szeretnénk, ha konfliktusok nélkül és szórakozva juthatnánk „előre". Így leszünk kiszolgáltatva azoknak, akik ismerik (ez nem baj) és saját hasznukra akarják kihasználni (ez a baj) gyenge pontjainkat. Nehezen tagadható, hogy vannak ilyen emberek, csoportok, szervezetek. Azt az állítást is meg

merem kockáztatni, hogy az ilyen emberek, csoportok, szervezetek előnyökhöz jutnak majd az informatika fejlődésével. Szép természetesen az az eset, ha valaki mások javára szeretné kamatoztatni a megszerzett tudást. Kérdés az, hogy eleget tud-e ehhez, megfelelő módon csinálja-e, egyetért-e a szakember vele, egyetért-e az alany is, fel tudja-e mérni egyáltalán, hogy miről is van szó, hiszen még nem rendelkezik elegendő tudással. A szülőktől kamaszkorig ezt elfogadjuk - és ez a természetes.

Az a nagy szerencse, hogy az egészségügy és a jólét talán a leginkább mentes a fenti problémáktól. Okot ad az óvatosságra azonban az, hogy az egészségügyre rászorulók kiszolgáltatottak.

Következtetések

Úgy gondolom, hogy a gépi tanulás elegendően fejlett ahhoz, hogy hatékonyan használni lehessen már egészségügyi feladatok ellátására. Ennek oka az, hogy az egészséges léthez szükséges „hiányosságok" ismerete, kombinálva az ön-annotációval a crowdsourcinggal, az adatbányászattal, az új 3D-s képi feldolgozó eszközökkel, a környezet fizikai ismeretével „rendelkező" 3D-s modellekkel, nem beszélve a viselhető intelligens eszközök széles köréről, amelyek lehetővé teszik a valós idejű megfigyeléseket, értelmezéseket és becsléseket, és ami végül lehetővé teszi a gépi pro-aktív viselkedést is. Úgy gondoljuk, hogy a technológia és az algoritmusok készen állnak arra, hogy meghozzák ezt az áttörést. A kérdés azonban nyitva áll: hogyan lehet, hogyan szabad és hogyan nem szabad intelligens környezetet és segítő robotokat bevezetni az egészségügybe a jólét érdekében?

Kutatási prioritások a megbízható és hasznos mesterséges intelligencia létrehozásáért[1]

A mesterséges intelligenciával kapcsolatos kutatások sikeressége magában hordozza a lehetőségét annak, hogy az emberiség példa nélküli előnyökhöz jusson. Érdemes ezért feltárni azokat a kutatási területeket, amelyek az esetleges buktatók elkerülésével egy időben segíthetnek maximalizálni az elérhető eredményeket. Jelen tanulmány számos ilyen témakört és példát mutat be (a teljesség igényének hajszolása nélkül), amelyek biztosíthatják, hogy az mesterséges intelligencia a jövőben is robusztus és az ember számára előnyös maradjon.

1. A MESTERSÉGES INTELLIGENCIA NAPJAINKBAN

A mesterséges intelligencia (MI) kutatása már a kezdetektől fogva számos különböző problémát és megközelítést vetett fel, de az elmúlt 20 év során főként az *intelligens ágensek* – olyan rendszerek, amelyek bizonyos környezetben képesek az érzékelésre és a cselekvésre – megalkotása körüli problémákra koncentrált. Ebben a kontextusban az intelligencia jellegzetesen a racionalitás statisztikai és gazdasági fogalmaihoz kapcsolódik – egyszerűbben fogalmazva a helyes döntések meghozatalára, a tervezésre vagy következtetések levonására való képességet jelenti. A valószínűségi megközelítés és a statisztikai tanulási metódusok alkalmazása nagyfokú integrációhoz vezetett, ahol termékenyítőleg hatott egymásra az MI, a gépi tanulás, a statisztika, az irányítás-elmélet, a neurológia és más, kapcsolódó területek. A közös elméleti keretek létrehozása a jelenleg rendelkezésre álló adatmennyiséggel és számítási kapacitással kombinálva figyelemre méltó sikereket eredményezett az olyan területeken, mint például a beszédfelismerés, a képek osztályozása, az autonóm járművek, a gépi fordítás, a robotok lábbal történő helyváltoztatása és a kérdés-válasz rendszerek.

Ahogy a lehetőségek ezeken és más területeken lehetővé tették, hogy a megoldások átlépjenek a laboratóriumi kísérleteken és elinduljanak a gazdaságilag is hasznosítható technológiák fejlesztése felé, úgy a teljesítmény kismértékű javulása is jelentős pénzt és nagyobb kutatási beruházásokat mozgat meg. Széleskörű egyetértés mutatkozik abban, hogy az MI kutatás folyamatos fejlődése egyre nagyobb hatással lehet a társadalomra. A potenciális előnyök óriásiak, hiszen minden, amit a civilizáció nyújtani tud, az az emberi intelligencia terméke; nem tudjuk megjósolni, hogy mit érhetünk el, ha ezt az intelligen-

[1] A tanulmány első verzióját Stuart Russell, Daniel Dewey és Max Tegmark írta Janos Kramar és Richard Mallah anyagainak felhasználásával. Észrevételeikkel, visszajelzéseikkel a tanulmányt segítették: Anthony Aguirre, Erik Brynjolfsson, Ryan Calo, Tom Dietterich, Dileep George, Bill Hibbard, Demis Hassabis, Eric Horvitz, Leslie Pack Kaelbling, James Manyika, Luke Muehlhauser, Michael Osborne, David Parkes, Heather Roff, Francesca Rossi, Bart Selman, Murray Shanahan. A fordítást a legutolsó elérhető verzió (http://futureoflife.org/static/data/documents/research_priorities.pdf 2015.01.23) alapján Csótó Mihály és Tamaskó Dávid készítette.

ciát MI-rendszerekkel támogatjuk, de a szegénység és a betegségek felszámolása sem elképzelhetetlen. Az MI nagy lehetőségeket tartogat, ezért érdemes megvizsgálni, hogyan lehet kihasználni az előnyeit, elkerülve a lehetséges buktatókat és csapdahelyzeteket.

Az MI kutatásában tapasztalt fejlődés alapján elérkezett az idő, hogy a kutatások ne csak mesterséges intelligencia alkalmazhatóbbá tételére, hanem annak társadalomra gyakorolt pozitív hatásainak maximalizálásra is kiterjedjenek, és meg is jelentek az első ilyen jellegű kezdeményezések (Pl. a Szövetség a Mesterséges Intelligencia Fejlesztéséért (Association for the Advancement of Artificial Intelligence (AAAI)) 2008-2009-es elnöki panelje az MI hosszú távú jövőjéről (Horvitz – Selman, 2009)). Ez a folyamat jelentős befolyással lehet az MI kutatások kiterjedésére, amelyek mindezidáig elsősorban a technikai megvalósíthatóságra összpontosítottak, azok hasznosítási területeire általában nem terjedtek ki. Jelen tanulmányra úgy lehet tekinteni, mint ezeknek az erőfeszítéseknek a természetes folytatására, ami azoknak a kutatási irányoknak a meghatározására összpontosít, amelyek célja a mesterséges intelligencia társadalmi előnyeinek kiaknázása. Ez a kutatás szükségszerűen interdiszciplináris, hiszen magában foglalja a társadalmat és a mesterséges intelligenciát is. Érinti a közgazdaságtant, a jogot, a filozófiát, a számítógépes biztonságot, a formális módszereket és természetesen az MI különböző ágait. A cél olyan MI fejlesztése, ami *hasznos* a társadalomra nézve és *robusztus* abban az értelemben, hogy jótéteményei garantáltak, azaz MI rendszereinknek azt kell tenniük, amit mi szeretnénk, hogy tegyenek.

2. RÖVID TÁVÚ KUTATÁSI PRIORITÁSOK

2.1 Az MI gazdasági hatásainak optimalizálása

Az MI ipari alkalmazásainak sikere a legkülönbözőbb területeken (a gyártási folyamatoktól kezdve az információs szolgáltatásokig) egyre nagyobb hatást gyakorol a gazdaság növekedésére, habár ennek a hatásnak a természetével kapcsolatban, és abban, hogyan lehet különbséget tenni a mesterséges intelligencia és egyéb információtechnológiák hatásai között, nincs egyetértés a szakemberek között. Sok közgazdász és informatikus egyetért abban, hogy érdemes kutatni azt, hogyan lehet maximalizálni az MI előnyeit a káros hatások – mint a növekvő társadalmi egyenlőtlenség és a munkanélküliség – csökkentése mellett (Mokyr, 2014; Brynjolfsson–McAfee, 2014; Frey–Osborne, 2013; Glaeser, 2014; Nilsson, 1984; Manyika, 2013). E megfontolások egy sor, a közgazdaságon és a pszichológián átívelő kutatási irányt ösztönöznek. A következőkben a teljesség igénye nélkül néhány példát mutatunk ezekre.

1. **Munkaerő-piaci előrejelzés**: mikor és milyen sorrendben várhatjuk a különböző munkák automatizálását (Frey–Osborne, 2013)? Hogyan fog ez hatni a kevésbé szakképzett munkások, a kreatív munkát végzők, valamint a különböző informatikai szakemberek munkabérére? Vannak, akik az állítják, hogy az MI nagyban növelné az emberiség egészének általános jólétét (Brynjolfsson–McAfee, 2014). Ugyanakkor az automatizáció növekedése még jobban eltolhatja a jövedelemelosztás mértékét a hatványtörvény irányába (Brynjolfsson–McAfee–Spance, 2014), és az ebből eredő egyenlőtlenségek aránytalanul jelenhetnek meg a faji, társadalmi és nemi vonalak mentén;

ezért a kutatások, melyek a gazdasági és társadalmi hatások aránytalanságának mértékével foglakoznak, egyértelműen hasznosnak lehetnek.

2. **Egyéb piaci zavarok/diszruptív (radikális változással járó) hatások**: a gazdaság jelentős része, beleértve a pénzügyet, biztosításokat, biztosítás-statisztikát és egyéb fogyasztói piacokat könnyen tárgya lehet a kreatív rombolás jelenségének az MI rendszerek használatának köszönhetően, melyek képesek modellezni és megjósolni az ágensek cselekedeteit. Az ilyen piacokra jellemző a magas komplexitás és a komplexitás kezeléséből következő jelentős haszon kombinációja (Manyika, 2013).

3. **Politikák a káros hatások kezeléséhez**: milyen irányelvek segíthetik az egyre nagyobb mértékben automatizálódó társadalmak kiteljesedését? Brynjolfsson és McAfee (2014) például bemutat különböző megoldásokat a munkaerő-intenzív szektorok fejlődésének ösztönzésére, illetve az MI által létrehozott jólét felhasználására az alulfoglalkoztatott népesség támogatására. Mik az előnyei és hátrányai az olyan beavatkozásoknak, mint az oktatási reform, a gyakornoki programok, a munkaerőigényes infrastrukturális projektek, valamint a minimálbérre, az adózási struktúrára és a szociális hálóra vonatkozó szabályozások módosítása (Glaeser, 2014)? A történelemben számos példát találunk, amikor a népesség egy részének nem kellett dolgoznia az anyagi biztonságért, legyen szó az ókori arisztokráciától vagy a mai Katar számos állampolgáráról. Milyen társadalmi struktúrák és egyéb tényezők határozzák meg az ilyen társadalmak prosperitását? A munkanélküliség nem egyenlő a szabadidővel, szoros összefüggés mutatható ki a munkanélküliség és a boldogtalanság, az önbizalomhiány és az elszigeteltség között (Hetschko–Knabe–Schöb, 2014; Clark–Oswald, 1994); annak a megértése, hogy milyen politikai beavatkozások és normák szüntethetik meg ezt a kapcsolatot, jelentősen javíthatja az átlagos életszínvonalat. Az empirikus és elméleti kutatások az olyan témákkal kapcsolatban, mint a feltétel nélküli alapjövedelem, nagyban segítenék a lehetőségek feltárását (Van Parijs et al., 1992; Widerquist, 2013).

4. **Gazdasági mérések**: Elképzelhető, hogy a gazdasági mérőszámok, mint például az egy főre jutó reál GDP nem mutatja meg pontosan a mesterséges intelligencián és az automatizáción alapuló gazdaság előnyeit és hátrányait, ezért stratégiai tervezésre, politikai irányvonalak meghatározására e mutatók alkalmatlanok (Mokyr, 2014), így a döntéshozók, politikacsinálók számára hasznos lehet a mérőszámok továbbfejlesztése.

2.2 Jogi és etikai kutatás

A számottevő intelligenciát és az autonómiát tartalmazó rendszerek fejlesztése olyan fontos jogi és etikai kérdésekhez vezet, melyekre adandó válaszok hatással vannak az MI megoldások fejlesztőire és fogyasztói oldalra egyaránt. Az ezt érintő kérdések kiterjednek a jogra, a közpolitikára, a szakmai és filozófiai etikára is, megválaszolásukhoz az informatikusok, a jogi szakértők, a politológusok és etikával foglalkozó szakértők tudására is szükség van. Például:

1. **Kötelezettségek és törvényi szabályozás az autonóm járművek tekintetében**: ha az önműködő autók felére csökkentenék az évi mintegy 40 000 halálos közlekedési balesetet az Egyesült Államokban, elképzelhető, hogy az autógyártók nem 20 000 köszönőlevelet, hanem 20 000 pert kapnának a nyakukba. Milyen jogi keretek között tudjuk leginkább kihasználni az autonóm járművekben (mint például a drónok vagy az önvezető autók) rejlő biztonsági lehetőségeket (Vladeck, 2014)? Az MI-vel kapcsolatos jogi kérdéseket a létező (szoftver- és internetközpontú) „kiberjog" alapján vagy attól függetlenül kellene kezelnünk (Calo, 2014a)? A katonai és a kereskedelmi alkalmazásoknál is a kormányoknak kell döntést hozniuk arról, hogyan lehet leginkább a kérdést a megfelelő szakértelemmel kezelni; példa lehet egy olyan szakmai és akadémiai fórum vagy közösség létrehozása, mint amilyennek a felállítását Calo javasolta (Szövetségi Robotikai Bizottság (Federal Robotics Commission), Calo, 2014b).

2. **Gépi etika**: hogyan döntsön egy autonóm jármű olyan események között, amikor az egyik kimenet egy alacsony valószínűségű emberi sérülés, a másik pedig egy majdnem teljes bizonyossággal bekövetkező komoly anyagi kár? Hogyan kellene az ügyvédeknek, az etikával foglalkozó szakembereknek és döntéshozóknak a nyilvánosság elé tárnia e problémákat? Vajon szükséges az ilyen döntési helyzeteket nemzeti szabályozásban rögzíteni?

3. **Autonóm fegyverek**: lehetséges-e az emberi jogokkal összeegyeztethető autonóm fegyvereket készíteni (Churchill–Ulfstein, 2000)? Ha – mint azt már néhány szervezet javasolta – az autonóm fegyvereket betiltanák (Docherty, 2012; *The Scientists' Call To Ban Autonomous Lethal Robots*, 2015), definiálható lenne-e az autonómia ebben a környezetben? És egy ilyen tilalom egyáltalán érvényesíthető-e a gyakorlatban? Ha megengedhető vagy legális az élet kioltására alkalmas autonóm fegyverek használata, hogyan kellene ezeket a fegyvereket integrálni egy már kialakult ellenőrzési struktúrába úgy, hogy a felelősség és a kötelezettségek megfelelően kijelölésre kerüljenek, milyen technikai realitásokat és előrejelzéseket kell figyelembe venni ezeknél a kérdéseknél, és hogyan határozható meg ezekkel a fegyverekkel kapcsolatban az „érdemi emberi ellenőrzés" (Roff, 2014; Roff, 2013; Anderson– Reisner–Waxman, 2014)? Az autonóm fegyverek csökkenthetik a politika ingerküszöbét a fegyveres konfliktusokkal kapcsolatban, netalán ezek a fegyverek „véletlen" háborúkat is kirobbanthatnak (Asaro, 2008)? Végül, hogyan lehet az átláthatóságot és az aktív közbeszédet biztosítani a témával kapcsolatban?

4. **Magánélet/Privacy**: hogyan viszonyul az MI-rendszerek képessége a térfigyelő kamerákból, telefonhívásokból és e-mailekből stb. származó adatok összegyűjtése és feldolgozása terén a magánélethez való joghoz? Hogyan hatnak az adatvédelmi kockázatok a kiberbiztonságra és kiberhadviselésre (Singer–Friedman, 2014)? A mesterséges intelligencia és nagy adattömeg (big data) közötti szinergiák kihasználásának sikere részben attól függ, hogy képesek vagyunk-e fenntartani és megóvni a magánélet biztonságát (Manyika, 2011; Agrawal–Srikant, 2000).

5. **Szakmai etika**: milyen szerepet kellene játszani a számítógépes szakembereknek az MI fejlesztésének és használatának jogi és etikai kérdéseiben? A tanulmányban többször említett szakmai testületek és kutatások kiemelten foglalkoznak ezzel a kérdéssel (Boden et al., 2011; Horvitz – Selman, 2011).

Közpolitikai megközelítésből az MI (mint bármely feltörekvő új technológia) nagyszerű új előnyöket és elkerülendő csapdákat jelent, ahol a megfelelő politika biztosíthatja, hogy úgy élvezzük az előnyöket, hogy közben a kockázatok a minimális szintre csökkennek. Ez olyan szakpolitikai kérdéseket vet fel, mint például:

1. Milyen politikai és stratégiai megközelítéseket érdemes számba venni?

2. Milyen kritériumokat alapján értékelhetjük a különböző politikákat? A lehetőségek között szerepel az ellenőrizhetőség, a végrehajthatóság, a kockázatok csökkentésének képessége, a megfelelő irányú technológiai fejlesztések akadályainak mérséklése, az alkalmazkodóképesség, valamint a változó körülményekhez való alkalmazkodás képessége.

2.3 Számítástechnikai kutatás a robusztus, megbízható mesterséges intelligenciáért

Ahogy az autonóm rendszerek előfordulása egyre gyakoribb a társadalomban, úgy válik egyre fontosabbá, hogy ezek megbízhatóan, az elvárásoknak megfelelően működjenek. Az önvezető járművek, az automatikus kereskedelmi rendszerek, az autonóm fegyverek és hasonló megoldások fejlesztésének esetében éppen ezért kiemelt figyelem övezi a magas biztosítású rendszereket, ahol a megbízhatóság terén erős garanciák fogalmazhatók meg; Weld és Etzioni szerint „*társadalom elutasítja az autonóm ágenseket, kivéve ha van néhány hiteles eszközünk azok biztonságossá tételére.*" (Weld–Etzioni, 1994). A robusztusság számos területen szenvedhet csorbát, melyek külön kutatási területeket nyújthatnak a megbízhatóság tekintetében:

1. **Ellenőrzés (verification)**: hogyan lehet bizonyítani, hogy a rendszer kielégíti-e a kívánt formális jellemzőket? *(„Jól építettem meg a rendszert?")*

2. **Érvényesség (validity)**: hogyan biztosítható, hogy a rendszer, amely megfelel a formális követelményeknek, nem viselkedik az eredeti szándéktól különbözően, és ez nem jár nem kívánt következményekkel? *(„A megfelelő rendszert építettem meg?")*

3. **Biztonság (security)**: hogy lehet megelőzni a rendszer illetéktelenek általi manipulálását?

4. **Irányítás (control)**: hogyan lehet egy MI rendszert érdemben emberi irányítás alatt tartani, miután az már működésbe lépett? *(„Oké, rosszul építettem meg a rendszert, helyre tudom hozni?")*

2.3.1 Ellenőrzés

Verifikálás alatt azokat a módszereket értjük, amelyek nagy valószínűséggel biztosítják, hogy a rendszer meg fog felelni az előre meghatározott formai követelményeknek. Amennyiben lehetséges, fontos biztosítani azt, hogy a biztonsági szempontból kritikus helyzetekben a rendszerek verifikálhatóak legyenek.

A szoftverek formális ellenőrzése jelentősen fejlődött az elmúlt évek során (pl. a seL4 kernel (Klein et al., 2009), vagy a HACMS (Fisher, 2012)): nem csak azt lenne szükséges lehetővé tenni, hogy az MI rendszerek a különböző, ellenőrzött alapokra épüljenek; azt is biztosítani kell, hogy az MI rendszerek terveit, felépítését önmagában is ellenőrizni lehessen, különösen, ha azok „*komponens architektúrát*" követnek, így az egyéni összetevőkre vonatkozó garanciák a komponensek kapcsolatai alapján kombinálhatók a teljes rendszer tulajdonságainak javítása érdekében (ez a megközelítés tükröződik például Russel és Norvig (2010) munkájában).

Talán a legszembetűnőbb különbség a hagyományos szoftverek és a mesterséges intelligencia-alapú rendszerek ellenőrizhetőségében az, hogy míg az hagyományos szoftvereket egy állandó és ismert gépi modell definiálja, addig az MI rendszerek (főleg a robotok és más testet öltött rendszerek) olyan környezetben működnek, amelyről a rendszer tervezőjének legjobb esetben is csak részleges a tudása. Ennek alapján az MI rendszerek esetében célszerű lehet azok működésének helyességét a rendszer tudásának fényében ellenőrizni, elkerülve a valós környezet modellezésének problémáit (Dennis, 2013). A tervezési idő ismeretének hiánya a tanulási algoritmusok használatára ösztökél az ágens szoftverén belül, amely által az ellenőrzés még nehezebbé válik: a statisztikai tanulás elmélete megad un. ϵ-δ (valószínűleg közelítőleg helyes) határokat, leggyakrabban az olyan, nem valós körülményekre, mint a felügyelt tanulás azonos eloszlású adatokból, és az egyedi ágensek megerősítésen alapuló tanulása egyszerű architektúra és teljes megfigyelhetőség esetén, de még így is túlságosan nagy mintanagyság szükséges a megfelelő garanciák eléréséhez.

A kutatási módszerek, amelyek lehetővé teszik, hogy erős megállapításokat tehessünk a gépi tanulási algoritmusokkal kapcsolatban és a számítási költségek menedzselésére különböző numerikus feladat esetén, javíthatják lehetőségeinket ezen a területen, akár kiterjesztve a munkát a Bayes-tételre is (Henning–Kiefel, 2013; Gunter, 2014). Az adaptív szabályozás elméletére (Åström–Wittenmark, 2013), az úgynevezett *kiberfizikai rendszerekre* (Platzer, 2010) vagy a hibrid és robotrendszerek ellenőrzésére (Alur, 2011; Winfield–Blum–Liu, 2014) irányuló munka szintén rendkívül releváns, de ugyanezekkel a nehézségekkel kell szembenéznie. És természetesen mindezek a kérdések a sztenderd problémára rakódnak, miszerint be kell bizonyítani, hogy az adott szoftver-kezdemény megfelelően implementálható például a megerősítő tanulás algoritmusának egy kívánt típusára. Zajlottak kutatások a neurális hálózati alkalmazások ellenőrzése terén (Paulina–Tacchella, 2010; Taylor, 2006; Schumann–Liu, 2010) és a *részprogram* (partial programs, Andre–Russel, 2002; Spears, 2006) fogalma lehetővé teszi a tervezőknek, hogy tetszőleges „strukturális" megszorításokat tegyenek a viselkedésre vonatkozóan, de még így is sok a tennivaló, mielőtt lehetővé válik magas megbízhatósági szinten az, hogy egy tanuló ágens a tervezési kritériumokat kielégítő módon lesz képes tanulni életszerű kontextusokban.

2.3.2 Érvényesség

Egy ágens tervezésének verifikációs tétele a következő formát mutatja: „*ha a környezet eleget tesz φ feltételezéseknek, akkor viselkedés eleget tesz γ (ψ) követelményeknek*.” Két módja van annak, hogy egy ellenőrzött ágens elhibázza a helyes cselekvést: első esetben a környezetére vonatkozó φ feltételezés hamis a való életben, és olyan reakcióhoz vezet, ami megsérti a γ követelményeit. Második esetben a rendszer megfelel a γ formai követelményeknek, miközben továbbra is úgy viselkedik, amit a gyakorlatban erősen nem kívánatosnak értékelünk. Előfordulhat, hogy ez a nem kívánt esemény annak a következménye, hogy teljesül γ, miközben φ sérül, azaz φ megvalósulása esetén a nem kívánt esemény nem történt volna meg; előfordulhat az is, hogy γ önmagában hibás. Russel és Norvig (2010) egy egyszerű példával él: ha arra kérünk egy robot tisztítógépet, hogy annyi szennyeződést tisztítson föl, amennyit csak lehet, és rendelkezik a porgyűjtőjének kiürítésének képességével is, akkor ez azt eredményezi, hogy a gép ugyanazt a piszkot fogja újra és újra feltakarítani. A követelményeknek tehát nem a szennyeződés eltűntetésére, hanem a padló megtisztítására kell vonatkoznia. Az ilyen specifikációs hibák minden olyan szoftver ellenőrzése során előfordulnak, ahol általában megfigyelhető, hogy a helyes specifikáció megírása nehezebb, mint a helyes kódé. Sajnos a specifikáció ellenőrzése nem lehetséges: a „hasznos” és a „elvárt” fogalmak külön nem formalizálhatók, így nem lehet egyértelműen bizonyítani, hogy γ-nek való megfelelés szükségszerűen a célszerű viselkedéshez és egy hasznos ágenshez vezet.

Annak érdekében, hogy megbízhatóan működő rendszereket építsünk, természetesen minden alkalmazási területen el kell döntenünk, mit is jelent a „jó működés”. Ez az etikai kérdés szorosan kapcsolódik ahhoz, hogy milyen mérnöki módszerek állnak rendelkezésünkre, hogy mennyire megbízhatóak ezek a technikák, és milyen kompromisszumokat köthetünk – minden területen, ahol a számítástechnika, a gépi tanulás és a tágabb MI szakértelem értékes. Például Wallach és Allan (2008) szerint kiemelten figyelembe veendők a különböző viselkedési sztenderdek (vagy etikai elméletek) számítási költségei: ha egy szabványt nem lehet elég kielégítően használni ahhoz, hogy helyes viselkedést eredményezzen a biztonsági szempontból kritikus helyzetekben, akkor olcsóbb megközelítésekre lehet szükség. Egyszerűsített szabályok kialakításához – például egy önvezető autó viselkedésének szabályozása kritikus helyzetekben – minden bizonnyal informatikusok és etikában jártas szakemberek is szükségesek. Az etikus gondolkodást leíró számítástechnikai modellek fényt deríthetnek a számítási költségekre és a megbízható érvelési módszerek életképességére egyaránt (Asaro, 2006; Sullins, 2011); az ilyen irányú kutatás például alkalmas lehet további területek alkalmazásának feltárására: szemantikus hálózatok használata az esetalapú következtetésben (McLaren, 2006), a hierarchikus korlátkielégítés (MacWorth, 2009), vagy a súlyozott prospektív abdukció (Pereira–Saptawijaya, 2007) alkalmazása a gépi etikához.

2.3.3 Biztonság

A biztonsággal kapcsolatos kutatások segíthetnek az MI robusztusabbá tételében. Mivel az MI rendszerek egyre növekvő számban kritikus szerepben kerülnek felhasználásra, egyre nagyobb felületet nyújtanak a kibertámadások számára is. Valószínű, hogy az MI és a gépi tanulási technológiákat is fel fogják használni kibertámadásokra. A robusztusság alacsony szinten szoros kapcsolatban van az ellenőrizhetőséggel és a hibáktól való mentességgel. A DARPA SAFE programjának célja például egy olyan, rugalmas metaadat szabályokat tartalmazó, integrált szoftver-hardver rendszer építése, amire olyan memóriabiztonsági, hiba-elkülönítési és egyéb protokollok építhetők, melyek javíthatják a biztonságot a kihasználható sérülékenységek megelőzésével (DeHon et al., 2011). Az ilyen programok nem szűntethetik meg az összes biztonsági hiányosságot (mivel az ellenőrzés csak olyan erős lehet, mint a követelmény, ami alapján az ellenőrzés történik), de jelentős mértékben csökkenthetik az olyan sérülékenységek okozta károkat, mint a közelmúltban terjedő „Heartbleed bug" vagy „Bash Bug". Az ilyen rendszerek használatát lehetőség szerint támogatni kell a biztonsági szempontból kritikus alkalmazásoknál, ahol a magasabb szintű biztonság költségei igazolhatók.

Magasabb szinten az MI és a gépi tanulás területén zajló specifikus kutatások egyre hasznosabbak lehetnek a biztonság számára. Ezek a technológiák alkalmazhatók a különböző jogtalan behatolások észlelésére (Lane, 2000), a rosszindulatú szoftverek (malware) azonosítására (Rieck et al., 2011), vagy esetleges sérülékenységek feltárására egyéb programok forráskódjának elemzése során (Brun – Ernst, 2004). Nem elképzelhetetlen, hogy az államok és a privát entitások között zajló kibertámadások is alkalmaznak majd a közeljövőben MI megoldásokat, ezáltal olyan rizikófaktort jelentenek, ami további kutatásokat motivál a káresemények elkerülése érdekében. Ahogy az MI rendszerek egyre komplexebbé válnak, és egymással is hálózati kapcsolatba kerülnek, intelligens módon kell kezelniük a bizalom kérdéskörét, ami szintén olyan újabb kutatási területek felé nyitja meg az utat, mint a statisztikai-viselkedési alapokon nyugvó bizalomépítés (Probst – Kasera, 2007), vagy a számítástechnikai reputációs modellezés (Sabater – Sierra, 2005).

2.3.4 Irányítás

Bizonyos típusú, biztonsági szempontból kritikus MI rendszerek – különösen a járművek és a fegyverrendszerek – esetében kívánatos lehet az emberi ellenőrzés valamilyen formájának megőrzése, jelentsen ez akár közvetlen emberi visszacsatolást (in the loop, on the loop - Hexmoor et al., 2009; Parasuraman et al., 2000) vagy valamilyen más protokollt. Sok esetben további technikai fejlesztés szükséges annak biztosításához, hogy az érdemi humán kontroll fennmaradhasson (UNIDIR, 2014).

Az önvezető járművek jó lehetőséget nyújtanak a hatékony kontroll-mechanizmusokkal való kísérletezésre. Az automata navigáció és az emberi ellenőrzés közötti váltást megvalósító rendszerek és eljárások tervezése szintén egy ígéretes kutatási terület. Az ilyen és ehhez hasonló problémák megtermékenyítően hatnak egyéb kutatásokra is, mint például az optimális feladat-elosztás meghatározása ember és gép alkotta egységek számára, illetve olyan helyzetek felismerése, ahol az ellenőrzést mindenképpen át kell adni, biztosítva a hatékony emberi döntést a legfontosabb döntéshozatali kérdésekben.

3. HOSSZÚ TÁVÚ KUTATÁSI PRIORITÁSOK

Számos mesterséges intelligenciával foglalkozó kutató által gyakran tárgyalt hosszú-távú cél olyan rendszerek fejlesztése, amelyek képesek az emberhez hasonlóan tág határok között tanulni tapasztalataikból, és meghaladni az emberi teljesítményt a legtöbb kognitív feladat esetében - ezáltal jelentős befolyást gyakorolva a társadalomra. Ha az elhanyagolhatónál nagyobb esélye van annak, hogy az ez irányú törekvéseket siker koronázza a belátható jövőben, akkor a korábban bemutatotthoz képest újabb, a továbbiakban részletezett kutatási területek nyílnak, melyek célja biztosítani, hogy az MI robusztus és jótékony hatású maradjon a jövőben is.

A kutatók véleménye jelentős mértékben eltérhet egy ilyen rendszer létrehozhatóságának valószínűségéről, de elenyésző azok aránya, akik nagy biztonsággal kijelentenék, hogy ez a valószínűség elhanyagolható – különösen a korábbi, hasonló jóslatok fényében. (Ernest Rutherford, korának vitathatatlanul legnagyobb atomfizikusa 1933-ban a nukleáris energiát a „fantazmagória" kategóriájába sorolta (Associated Press, 1933), míg Richard Wolley, a Királyi csillagász (Royal Astronomer, a legmagasabb csillagászi pozíció az Egyesült Királyságban) 1956-ban a bolygóközi űrutazást nevezte „teljes sületlenségnek" (Reuters, 1956)). Mindezeken túl, ahhoz hogy igazolhatók legyenek ezen a területen az MI robusztusságának kutatására fordított összegek, ennek a valószínűségnek nem kell magasnak lennie, csupán nem elhanyagolhatónak – épp annyira, mint amennyire a lakásbiztosításra fordított összegeket indokolja az otthon leégésének kicsi, de nem elhanyagolható valószínűsége.

3.1 Ellenőrzés

Visszaidézve a rövid távú prioritásokat, a kutatások, melyek ellenőrizhető alacsony szintű szoftvert és hardvert tesznek elérhetővé a különböző programhibák és problémák számos csoportját megszűntethetik az általános MI-rendszerekben; ahogy a rendszerek egyre erősebbeké és biztonságuk egyre kritikusabbá válik, a verifikálható biztonsági tulajdonságok is egyre értékesebbek lesznek. Ha a verifikálható összetevők teljes rendszerekre történő kiterjeszthetőségének elmélete elfogadott, akkor akár nagyon nagy rendszerek is részesülhetnek bizonyos típusú biztonsági garanciákból, potenciálisan akár olyan technológiák segítségével is, amelyeket kifejezetten tanuló ágensek és magas szintű összetevők irányítására terveztek. Az elméleti kutatások - kiváltképp a mindenre alkalmas mesterséges intelligencia-rendszernek terén - különösen hasznosak lehetnek.

A verifikáláshoz kapcsolódó, a hosszú távú aggodalmak esetében különösen releváns kutatási téma az olyan rendszerek ellenőrzése, amelyek képesek módosítani, bővíteni vagy fejleszteni saját magukat, akár többször egymás után (Irving, 1965, Vinge, 1993). A formalizált verifikációs megoldások egy-az-egyben történő alkalmazásának kísérlete erre a jóval általánosabb keretre új nehézségeket jelent, beleértve azt a kihívást, mely szerint a kellően erőteljes formális rendszerek kézenfekvő módon nem használhatnak formális módszereket azért, hogy megbizonyosodjanak funkcionálisan hasonló formális rendszerek pontosságáról - hacsak el nem tekintünk Gödel nemteljességi tételétől... (Fallenstein – Soares, 2014; Wallach – Allen, 2008). Egyelőre nem teljesen világos, hogy hogyan lehet meghaladni ezt a problémát, illetve hogy egyéb hasonló problémák felbukkannak-e a hasonló erősségű verifikációs módszerek kapcsán.

Végezetül az is elmondható, hogy sokszor nehéz valóban alkalmazni a formális verifikációs technológiákat fizikai rendszerekre, különösen olyan rendszerekre, amelyeknél a tervezés során nem tartották szem előtt a verifikálás problémáját. Ez a tény ösztönzőleg hathat az olyan, általános elméletre törekvő kutatásokra, melyek összekapcsolják a funkcionális specifikációkat a fizikailag megvalósuló eseményekkel. Ez a fajta elméletalkotás lehetővé tenné a formális eszközök használatát, hogy megjósolható és kontrollálható legyen az olyan rendszerek viselkedése, amelyek közelítenek a racionális ágensekhez, vagy az olyan eltérő szerkezetek, mint a kielégítő ágensek, illetve az olyan rendszereké, amelyeket nem lehet könnyen leírni a hagyományos ágens-formanyelvvel (erőteljes predikciós rendszerek, tételbizonyítók, korlátozott célú tudományos vagy mérnöki rendszerek stb.). Egy ilyen elmélet akár annak bizonyítását is lehetővé tenné, hogy a rendszerek korlátozhatók bizonyos tevékenységek elvégzésében, vagy bizonyos fajta érvelés alkalmazásában.

3.2 Érvényesség

Csakúgy, mint a rövid távú kutatási prioritások esetében, az érvényesség azokkal a nem kívánt viselkedésformákkal foglalkozik, melyek egy rendszer formai helyessége ellenére fordulnak elő. Hosszú távon az MI rendszerek még erősebbé és még autonómabbá válhatnak, így a validitásproblémák is jóval költségesebbek lehetnek. A rövid távú prioritásoknál már említett gépi tanulási módszerek számára létrehozott erős garanciák a hosszú távú biztonság szempontjából szintén fontosak lesznek. Az ezen a területen végzett munka hosszú távú hasznosítása érdekében a gépi tanulással kapcsolatos kutatások egyik fókuszpontja a nem várt általánosítások típusainak vizsgálata lehet, amely a mindenre alkalmas mesterséges intelligencia-rendszerek számára okozhatja a legtöbb problémát. Különösen fontos lehet megérteni mind elméleti, mind gyakorlati szempontból, hogy a magasabb szintű emberi fogalmak tanult értelmezése hogyan (nem) általánosítható gyökeresen eltérő kontextusokban (Tegmark, 2015). Ha bizonyos fogalmak megtanulása megbízhatóan történik, akkor ez lehetővé teheti azok használatát feladatok és kikötések definiálására, minimálisra csökkentve a nem szándékolt következményeket, akkor is, ha az autonóm MI rendszer mindenre alkalmas mesterséges intelligencia-rendszerré válik. Ez a témakör eddig kevéssé kutatott, így mind az elméleti, mind pedig az empirikus kísérletek hasznosak lehetnek a területen.

Az olyan matematikai módszerek, mint a formális logika, a valószínűség- és a döntéselmélet igazán gyümölcsözőnek bizonyultak a döntéshozás és a gondolkodás alapjainak megismerése során, ám továbbra is számos a megoldatlan probléma. A megoldások ezekre a problémákra sokkal megbízhatóbbá és kiszámíthatóbbá tehetik a mindenre alkalmas rendszerek viselkedését. Példák ezen a területen: érvelés és döntések korlátozott számítási erőforrások esetén (Horvitz, 1987; Russel – Subramanian, 1995), hogyan vehető figyelembe a kapcsolat az MI rendszerek viselkedése és környezetük, illetve egyéb ágensek viselkedése között (Halpern – Pass, 2013; Hintze, 2014; LaVictoire et al., 2014; Soares – Fallenstein, 2014; Tennenholtz, 2004), hogyan kell gondolkodnia a környezetébe ágyazott ágenseknek (Orseau – Ring, 2012; Soares, 2014), illetve hogyan érveljenek az olyan bizonytalan tényezők logikai következményeivel kapcsolatban, mint a különböző hiedelmek vagy egyéb determinisztikus számítások (Soares – Fallenstein, 2014, Probabilistic Numerics, 2014). Ezeket a témákat szoros kapcsolatuk miatt hasznos lehet egyben kezelni (Halpern – Pass, 2011, Halpern et al., 2014).

Kézenfekvő, hogy hosszú távon az autonóm és erőteljes ágenseket az élet minél több területén szeretnénk alkalmazni. Egyértelműen lefektetni a preferenciáinkat széles doménekben a közeljövő gépi etikájának stílusában nem feltétlenül célszerű, mivel nehézzé teheti az erőteljes MI rendszerek értékeinek „összehangolását" saját értékeinkkel és preferenciáinkkal (Soares, 2014; Soares – Fallenstein, 2014). Vegyük például egy olyan hasznossági függvény megalkotásának a nehézségét, amely a teljes joganyagot felöleli; még a jog szó szerinti interpretálása is messze meghaladja a jelenlegi lehetőségeinket, és igen gyenge eredményekkel járna a gyakorlatban (mivel a törvények írottak, feltételezzük, hogy interpretálásuk és alkalmazásuk rugalmasan, eseti elbírálás alapján történik). A megerősítéses tanulásnak is megvannak a saját problémái: ahogy a rendszerek egyre inkább mindenre alkalmas mesterséges intelligencia-rendszerekké válnak, egyre nagyobb az esélye egy, Goodhart törvényhez hasonló hatás bekövetkezésének: a szofisztikált ágensek megpróbálják manipulálni, vagy közvetlenül irányítani azokat a mechanizmusokat, amelyek az eredményességüket jelzik (Bostrom, 2014). Ez olyan kutatási területekhez vezethet, amelyek javíthatják az olyan rendszerek konstruálását, amelyek futásidőben képesek a tanulásra vagy értékek elsajátítására. Az inverz megerősítéses tanulás például egy érvényes megközelítést nyújthat, amiben egy rendszer következtet egy másik aktor preferenciáira, aki feltételezhetően szintén a megerősítéses tanulást alkalmazza (Russel, 1998; Ng – Russell, 2000). Más megközelítések eltérő feltételezéseket használhatnak az aktor alapvető kognitív modelljeiről, akinek a preferenciái a tanulás tárgyát képezik (preferenciák tanulása, Chu – Ghahramani, 2005), vagy kifejezetten az ember etikai értékelsajátítási folyamata inspirálja. Ahogy a rendszerek fejlettebbé válnak, episztémikusan nehezebb módszerek is használatosak lehetnek, így az ezen a területen végzett kutatások is hasznosak, mint például a Bostom (2014) által példaként említett előzetes módszertani áttekintés az indirekt célmeghatározás lehetőségeiről.

3.3 Biztonság

Egyelőre nem egyértelmű, hogy az MI hosszú távú fejlődése a biztonsággal kapcsolatos problémák megoldását könnyebbé vagy nehezebbé teszi; egyrészt a rendszerek egyre komplexebbé válnak mind szerkezetüket, mind viselkedésüket tekintve és az MI segítségével végrehajtott kibertámadások döbbenetesen hatékonyak lehetnek, ám másrészt az MI és a gépi tanulási technológiák használatával – az alacsony szintű rendszerek megbízhatóságának jelentős fejlődésével együtt – jóval ellenállóbb rendszerek jöhetnek létre, melyek sokkal kevésbé sebezhetők, mint a maiak. Titkosítási szemszögből úgy tűnik, hogy ebben a konfliktusban az MI inkább a védekezőket részesíti előnyben a támadókkal szemben – ez lehet az egyik oka annak, hogy a védelemmel kapcsolatos kutatásokat teljes szívvel támogassuk.

Habár a 2.3.3. pontban felsorolt kutatási témák egyre fontosabbak lesznek hosszú távon, a mindenre alkalmas mesterséges intelligencia-rendszerek egyedi biztonsági problémákat vetnek fel. Különösen igaz ez akkor, ha az érvényesség és az ellenőrzés problémái nem kerülnek megoldásra. Ebben az esetben hasznos lehet olyan „konténerek" kialakítása az MI rendszerek számára, amelyekben kevésbé ellenőrzött környezetben fordulhatnak elő a nem kívánt viselkedésmódok és következmények (Yampolskiy, 2012). Ennek a kérdésnek mind az elméleti, mind pedig a gyakorlati vizsgálata indokolt. Ha az MI általános

szabályozása korlátozásokkal nehéznek bizonyul, akkor egy MI rendszer és egy konténer párhuzamos fejlesztése megoldás lehet, mivel lehetővé teszi a tervezés gyengeségeinek és erősségeinek felhasználását a korlátozó stratégia kialakítása során (Bostrom, 2014). Szintén jelentős segítséget jelentenének az eltérés-felismerő és automatizált tevékenység-ellenőrző rendszerek. Összességében indokoltnak tűnik, hogy ez az újabb perspektíva – a támadásokkal szembeni védekezés a „belső", rendszeren belüli problémáktól és a külső aktoroktól is – érdekes és jövedelmező területté váljon az informatikai biztonság területén.

3.4 Irányítás

Említésre került már, hogy a különböző feladatokat önállóan végző mindenre alkalmas mesterséges intelligencia-rendszerek gyakran olyan hatások alanyaivá válnak, amelyek jelentősen megnehezítik az érdemi emberi ellenőrzés fenntartását (Bostrom, 2012; Bostrom, 2014; Omohundro, 2007; Shanahan, 2015). Az olyan rendszerek kutatása, amelyek nem alanyai ezeknek a hatásoknak, vagy minimálisra csökkenti a hatások befolyását, vagy lehetővé teszi a megbízható emberi ellenőrzést hasznosnak bizonyulhatnak a nem kívánt következmények megelőzésében, valamint egyfajta biztonságos tesztkörnyezetként szolgálhatnak különböző fejlettségű MI rendszerek számára.

Ha egy MI rendszer választhatja ki a legjobb módszereket egy adott feladat megoldására, akkor egyértelmű cél az olyan körülmények elkerülése, amelyek akadályozzák a rendszert a feladat elvégzésére való törekvésben (Bostrom, 2012; Omohundro, 2007), és igaz ez fordítva is, a nem előre alkotott helyzetek is hasznosak lehetnek a megismerés szempontjából (Wissner-Gross – Freer, 2013)). Ez ugyanakkor problémákkal is járhat, ha újra szeretnénk tervezni a rendszert, deaktiválni, vagy jelentősen megváltoztatni a döntéshozatali folyamatot; egy ilyen rendszer racionálisan elkerülné ezeket a változtatásokat. Azokat a rendszereket, amelyek nem mutatják ezt a viselkedést, *javítható* (corrigible) rendszereknek nevezték el (Soares et al., 2015), és mind az elméleti és a gyakorlati kutatások jól tervezhetők és hasznosak ezen a területen. Lehetséges például hasznosságfüggvények vagy döntési folyamatok tervezése, amelyek használatakor a rendszer nem fogja megpróbálni elkerülni saját lekapcsolását vagy átprogramozását (Soares et al., 2015). Elméleti keretrendszer fejlesztése is elképzelhető annak jobb megértése érdekében, hogy milyen potenciális rendszerek képzelhetők el, amelyek elkerülik a nem kívánt viselkedésmódokat (Hibbard, 2012; Hibbard; 2014, Hibbard, 2015).

Egy másik alapvető cél lehet a különböző helyettesítő erőforrások összegyűjtése, mint amilyenek például a környezeti információk, a diszruptív, radikálisan újat hozó megoldásoktól való védelem, és a nagyobb szabadságú cselekvés mind számos feladat megoldásához hozzájárulhat (Bostrom, 2012; Omohundro, 2007). Hammond és munkatársai (1995) stabilizációnak hívja azokat a helyzeteket, amikor „*az ágens beavatkozásának köszönhetően a környezet jobban illeszkedik az ágenshez az idő múlásával*". Az ilyen célok nem várt következményekhez vezethetnek, egyúttal azon körülmények és hatásaik jobb megértéséhez és mérsékléséhez is, amelyek esetén az erőforrás-gyűjtés vagy a radikális stabilizáció az optimális stratégia (vagy valószínűleg a rendszer által leginkább preferált). A lehetséges kutatási témák ezeken túl olyan, alkalmazási területükben valamilyen módon limitált (Bostrom, 2014), „domesztikált" célokra is kiterjedhetnek, mint az időpreferencia mérté-

kének hatása az erőforrás-gyűjtési stratégiákra, vagy a hasonló kutatási célokat bemutató, egyszerű rendszerek gyakorlati vizsgálata.

Végezetül szót kell ejteni a szuperintelligens gépekkel, vagy a gyors, fenntartható önfejlesztésre képes rendszerekkel („intelligencia robbanás") kapcsolatos kutatásokról, melyekket számos múltbéli és jelenlegi, az MI jövőjével foglalkozó projekt érintett, és amelyek az irányítás hosszú távú megőrzése szempontjából is értékesek lehetnek. Ilyen például az AAAI 2008-09-es, az MI hosszú távú jövőjével foglalkozó elnöki paneljének „Sebesség, aggályok és ellenőrzés" munkacsoportja, mely a következő megállapítást tette:

> „*Általános szkepszis veszi körül egy lehetséges intelligencia-robbanás következményeit… Ettől függetlenül közös meggyőződésünk, hogy a további kutatások a területen olyan módszerekhez vezethetnek, amelyek segítenek megérteni és ellenőrizni a komplex számítógépes rendszerek számos viselkedését annak érdekében, hogy elkerüljük a nem várt következményeket. A panel néhány tagja szerint további kutatás szükséges az „intelligencia-robbanás" fogalmának pontosítására, valamint a hasonló, gyorsan változó intelligenciák osztályozására. A szakmai munka elvezethet az ilyen jelenségek valószínűségének jobb megértéséhez, éppúgy, mint különböző változataik természetének, a hozzájuk kapcsolódó veszélyeknek és általános következményeknek a leírásához.*" (Horvitz – Selman, 2009)

> A Stanford Egyetem „A Mesterséges Intelligencia Száz Éve"-kutatása szintén beszél az „*MI-rendszerek feletti irányítás elvesztéséről*", mint a kutatás egyik területéről, különösen annak a lehetőségnek, hogy „*…egy nap elveszíthetjük az irányítást az MI rendszerek felett az olyan szuperintelligencia fejlődése miatt, amely nem az emberi elvárásoknak megfelelően cselekszik – és ezáltal egy ilyen erős rendszer veszélyeztethetné az emberiséget. Vajon lehetségesek-e ilyen disztópikus forgatókönyvek? Ha igen, hogyan következhetnek be ilyen helyzetek? … Milyen kutatásokat kellene támogatni annak érdekében, hogy jobban megértsük és felkészüljünk egy veszélyes szuperintelligencia létrejöttére, vagy az „intelligencia-robbanás" bekövetkeztére?*" (Horvitz, 2014)

A kutatások ezen a területen magukba foglalhatják bármelyik, a korábbiakban felsorolt kutatási prioritást éppúgy, mint az elméleti és előrejelző munkákat az intelligencia-robbanás és a szuperintelligencia (Bostrom, 2014; Chalmers, 2010) kérdéskörében, és kiterjeszthetik, vagy kritikus elemzés alá vehetik a már létező kutatói műhelyek (pl. Machine Intelligence Research Institute (Soares – Fallenstein, 2014)) megközelítéseit.

4. ÖSSZEFOGLALÁS

A mesterséges intelligencia megteremtésére irányuló törekvések példa nélkül álló szolgálatot tehetnek az emberiségnek, és ezért érdemes kutatásokat végezni a lehetséges hasznok maximalizálása, valamint a lehetséges buktatók elkerülése érdekében. Ez a tanulmány (a teljesség igénye nélkül) számos példát hozott arra, milyen kutatások segíthetnek biztosítani azt, hogy a mesterséges intelligencia robusztus, hasznos és az emberi érdekekkel összehangolható legyen.

Irodalom

Agrawal, R., Srikant, R. (2000): Privacy-preserving data mining. In: ACM Sigmod Record 29.2 (2000), pp. 439-450.
Alur, R. (2011): Formal verification of hybrid systems. In: Embedded Software (EMSOFT) Proceedings of the International Conference on. IEEE. 2011, pp. 273-278.
Anderson, K., Reisner, D., Waxman, M. (2014): Adapting the Law of Armed Conict to Autonomous Weapon Systems. In: International Law Studies 90
Andre, D., Russell, S. (2002): State abstraction for programmable reinforcement learning agents. In: Eighteenth national conference on Artificial intelligence. American Association for Artificial Intelligence. pp. 119-125.
Asaro, P. (2006): What should we want from a robot ethic? In: International Review of Information Ethics 6.12 (2006), pp. 9-16.
Asaro, P. (2008): How just could a robot war be? In: Current issues in computing and philosophy pp. 50-64.
Associated Press (1933): Atom-Powered World Absurd, Scientists Told". In: New York Herald Tribune (1933). September 12, p. 1.
Äström, Karl J.; Wittenmark, B. (2013): Adaptive control. Courier Dover Publications
Boden, M. et al. (2011): Principles of robotics. In: The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC)
Bostrom, N. (2012): The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. In: Minds and Machines 22.2, pp. 71-85.
Bostrom, N. (2014): Superintelligence: Paths, dangers, strategies. Oxford University Press
Brun, Y.; Ernst, M. D. (2004): Finding latent code errors via machine learning over program executions. In: Proceedings of the 26th International Conference on Software Engineering. IEEE, Computer Society. 2004, pp. 480-490.
Brynjolfsson, E.; McAfee, A. (2014): The second machine age: work, progress, and prosperity in a time of brilliant technologies. W.W. Norton & Company, 2014.
Brynjolfsson, E.; McAfee, A.; Spance, M. (2014): Labor, Capital, and Ideas in the Power Law Economy. In: Foreign Aff. 93 (2014), p. 44.
Calo, R. (2014a): Robotics and the New Cyberlaw. In: Available at SSRN 2402972
Calo, R. (2014b): The Case for a Federal Robotics Commission. In: Available at SSRN 2529151
Chalmers, D. (2010): The singularity: A philosophical analysis. In: Journal of Consciousness Studies 17.9-10 (2010), pp. 7-65.
Chu, W.; Ghahramani, Z. (2005): Preference learning with Gaussian processes. In: Proceedings of the 22nd international conference on Machine learning. ACM. 2005, pp. 137-144.
Churchill, R. R.; Geir Ulfstein, G. (2000): Autonomous institutional arrangements in multilateral environmental agreements: a little-noticed phenomenon in international law. In: American Journal of International Law (2000), pp. 623-659.
Clark, A. E.; Oswald, A. J.(1994): Unhappiness and unemployment. In: The Economic Journal (1994), pp. 648-659.
DeHon, A. et al (2011): Preliminary design of the SAFE platform. In: Proceedings of the 6th Workshop on Programming Languages and Operating Systems. ACM. 2011, p. 4.
Dennis, L. A. et al. (2013: Practical Veri_cation of Decision-Making in Agent-Based Autonomous Systems. In: arXiv preprint arXiv:1310.2431 (2013).
Docherty, B. L. (2012): Losing Humanity: The Case Against Killer Robots. Human Rights Watch, 2012.
Fallenstein, B.; Soares, N (2012): Vingean Reection: Reliable Reasoning for Self-Modifying Agents. Tech. rep. Machine Intelligence Research Institute, 2014. url: https://intelligence.org/files/VingeanReflection.pdf.

Fisher, K. (2012): HACMS: high assurance cyber military systems. In: Proceedings of the 2012 ACM conference on high integrity language technology. ACM. 2012, pp. 51-52.
Frey, C.; Osborne, M. (2013): The future of employment: how susceptible are jobs to computerisation? Working Paper. Oxford Martin School, 2013.
Glaeser, E. L. (2014): Secular joblessness. In: Secular Stagnation: Facts, Causes and Cures (2014), p. 69.
Good, I. J. (1965): Speculations concerning the first ultraintelligent machine. In: Advances in computers 6.31 (1965), p. 88.
Gunter, T. et al. (2014): Sampling for inference in probabilistic models with fast Bayesian quadrature. In: Advances in Neural Information Processing Systems. 2014, pp. 2789-2797.
Halpern, J. Y.; Pass, R. (2011): I don't want to think about it now: Decision theory with costly computation. In: arXiv preprint arXiv:1106.2657 (2011).
Halpern, J. Y.; Pass, R. (2013): Game theory with translucent players. In: arXiv preprint arXiv:1308.3778 (2013).
Halpern, J. Y.; Pass, R;, Seeman, L. (2014): Decision Theory with Resource-Bounded Agents. In: Topics in cognitive science 6.2 (2014), pp. 245-257.
Hammond, K. J.; Converse, T. M.; Grass, J. W. (1995): The stabilization of environments. In: Artificial Intelligence 72.1 (1995), pp. 305-327.
Hennig, P.; Kiefel, M. (2013): Quasi-Newton methods: A new direction. In: The Journal of Machine Learning Research 14.1 (2013), pp. 843-865.
Hetschko, C.; Knabe, A.; Schöb, R. (2014): Changing identity: Retiring from unemployment. In: The Economic Journal 124.575 (2014), pp. 149-166.
Hexmoor, H.; McLaughlan, B.; Tuli, G. (2009): Natural human role in supervising complex control systems. In: Journal of Experimental & Theoretical Artificial Intelligence 21.1 (2009), pp. 59-77.
Hibbard, B. (2012): Avoiding unintended AI behaviors. In: Artificial General Intelligence. Springer, 2012, pp. 107-116.
Hibbard, B. (2014): Ethical Artificial Intelligence. 2014. url: http://arxiv.org/abs/1411.1373.
Hibbard, B. (2015): Self-Modeling Agents and Reward Generator Corruption. In: AAAI-15 Workshop on AI and Ethics. 2015.
Hintze, D. (2014): Problem Class Dominance in Predictive Dilemmas". Honors Thesis. Arizona State University, 2014.
Horvitz, E (2014): One-Hundred Year Study of Artificial Intelligence: Reections and Framing. White paper. Stanford University, 2014. url: https://stanford.app.box.com/s/266hrhww2l3gjoy9euar.
Horvitz, E. J. (1987): Reasoning about beliefs and actions under computational resource constraints. In: Third AAAI Workshop on Uncertainty in Artificial Intelligence. 1987, pp. 429-444.
Horvitz, E.; Selman, B. (2009): Interim Report from the Panel Chairs. AAAI Presidential Panel on Long Term AI Futures. 2009. url: https://www.aaai.org/Organization/Panel/panel-note.pdf.
Klein, G. et al. (2009): seL4: Formal verification of an OS kernel. In: Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles. ACM. 2009, pp. 207-220.
Lane, T. D. (2000): Machine learning techniques for the computer security domain of anomaly detection. PhD thesis. Purdue University, 2000.
LaVictoire, P. et al. (2014): Program Equilibrium in the Prisoner's Dilemma via Löb's Theorem. In: AAAI Multiagent Interaction without Prior Coordination workshop. 2014.
Mackworth, A. K. (2009): Agents, bodies, constraints, dynamics, and evolution. In: AI Magazine 30.1 (2009), p. 7.
Manyika, J. et al. (2011): Big data: The next frontier for innovation, competition, and productivity. Report. McKinsey Global Institute, 2011.
Manyika, J. et al. (2013): Disruptive technologies: Advances that will transform life, business, and the global economy. Vol. 180. McKinsey Global Institute, San Francisco, CA, 2013.
McLaren, B. M. (2006): Computational models of ethical reasoning: Challenges, initial steps, and future directions. In: Intelligent Systems, IEEE 21.4 (2006), pp. 29{37.

Mokyr, J. (2014): Secular stagnation? Not in your life. In: Secular Stagnation: Facts, Causes and Cures (2014), p. 83.
Ng, A. Y.; Russell, S. (2000):Algorithms for Inverse Reinforcement Learning. In: in Proc. 17th International Conf. on Machine Learning. Citeseer. 2000.
Nilsson, N. J. (1984): Artificial intelligence, employment, and income". In: AI Magazine 5.2 (1984), p. 5.
Omohundro, S. M. (2007): The nature of self-improving Artificial intelligence. Presented at Singularity Summit 2007.
Orseau, L.; Ring, M. (2012): Space-Time embedded intelligence". In: Artificial General Intel-ligence. Springer, 2012, pp. 209-218.
Parasuraman, R.; Sheridan, T. B.; Wickens, C. D. (200): A model for types and levels of human interaction with automation. In: Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 30.3 (2000), pp. 286-297.
Pereira, L. M.; Saptawijaya, A. (2007): Modelling morality with prospective logic. In: Progress in Artificial Intelligence. Springer, 2007, pp. 99-111.
Platzer, A (2010): Logical analysis of hybrid systems: proving theorems for complex dynamics. Springer Publishing Company, Incorporated, 2010.
Probabilistic Numerics (2014): http://probabilistic-numerics.org. Accessed: 27 November 2014.
Probst, M. J.; Kasera, S. K. (2007): Statistical trust establishment in wireless sensor networks. In: Parallel and Distributed Systems, 2007 International Conference on. Vol. 2. IEEE. 2007, pp. 1-8.
Pulina, L.; Tacchella, A. (2010): An abstraction-re_nement approach to veri_cation of Artificial neural networks". In: Computer Aided Veri_cation. Springer. 2010, pp. 243-257.
Reuters (1956): Space Travel `Utter Bilge'. In: The Ottawa Citizen (1956). January 3, p. 1. url: http://news.google.com/newspapers?id=ddgxAAAAIBAJ&sjid=1eMFAAAAIBAJ&pg=3254%2C7126.
Rieck, K. et al. (2011): \Automatic analysis of malware behavior using machine learning". In: Journal of Computer Security 19.4 (2011), pp. 639-668.
Roff, H. M. (2013): Responsibility, liability, and lethal autonomous robots". In: Routledge Handbook of Ethics and War: Just War Theory in the 21st Century (2013), p. 352.
Roff, H. M. (2014): The Strategic Robot Problem: Lethal Autonomous Weapons in War. In: Journal of Military Ethics 13.3 (2014).
Russell, S. (1998): Learning agents for uncertain environments. In: Proceedings of the eleventh annual conference on Computational learning theory. ACM. 1998, pp. 101-103.
Russell, S. J. (1995); Subramanian, D.: Provably bounded-optimal agents. In: Journal of Artificial Intelligence Research (1995), pp. 1-36.
Russell, S.; Norvig P. (2010): Artificial Intelligence: A Modern Approach. 3rd. Pearson, 2010.
Sabater, J.; Sierra, C. (2005): Review on computational trust and reputation models. In: Artificial intelligence review 24.1 (2005), pp. 33-60.
Schumann, J. M.; Liu, Y. (2010): Applications of neural networks in high assurance systems. Springer, 2010.
Shanahan, M. (2015): The Technological Singularity. Forthcoming. MIT Press, 2015.
Singer, P. W.; Friedman, A. (2014): Cybersecurity: What Everyone Needs to Know. Oxford University Press, 2014.
Soares, N. (2014): Formalizing Two Problems of Realistic World-Models. Tech. rep. Machine Intelligence Research Institute, 2014. url: https://intelligence.org/files/RealisticWorldModels.pdf.
Soares, N. (2014): The Value Learning Problem. Tech. rep. Machine Intelligence Research Institute, 2014. url: https://intelligence.org/files/ValueLearningProblem.pdf.
Soares, N. et al. (2015): Corrigibility. In: AAAI-15 Workshop on AI and Ethics. 2015. url: http://intelligence.org/files/Corrigibility.pdf. 11.

Soares, N.; Fallenstein, B. (2014): Aligning Superintelligence with Human Interests: A Technical Research Agenda. Tech. rep. Machine Intelligence Research Institute, 2014. url: http://intelligence.org/files/TechnicalAgenda.pdf.
Soares, N.; Fallenstein, B. (2014): Questions of Reasoning Under Logical Uncertainty. Tech. rep. url: http://intelligence.org/files/QuestionsLogicalUncertainty.pdf. Machine Intelligence Research Institute, 2014.
Soares, N.; Fallenstein, B. (2014): Toward Idealized Decision Theory. Tech. rep. url: https://intelligence.org/files/TowardIdealizedDecisionTheory.pdf. Machine Intelligence Research Institute, 2014.
Spears, D. F. (2006): Assuring the behavior of adaptive agents". In: Agent technology from a formal perspective. Springer, 2006, pp. 227-257.
Sullins, J. P. (2011): Introduction: Open questions in roboethics". In: Philosophy & Technology 24.3 (2011), pp. 233-238.
Taylor, B. J. (2006): Methods and Procedures for the Veri_cation and Validation of Artificial Neural Networks. Springer, 2006.
Tegmark, M. (2015): Friendly Artificial Intelligence: the Physics Challenge". In: AAAI-15 Workshop on AI and Ethics. 2015. url: http://arxiv.org/pdf/1409.0813.pdf.
Tennenholtz, M (2004): Program equilibrium. In: Games and Economic Behavior 49.2 (2004), pp. 363-373.
The Scientists' Call To Ban Autonomous Lethal Robots (2015). International Committee for Robot Arms Control. Accessed January 2015. url: http://icrac.net/call/.
United Nations Institute for Disarmament Research (2014): The Weaponization of Increasingly Autonomous Technologies: Implications for Security and Arms Control. UNIDIR, 2014.
Van Parijs, P. et al. (1992): Arguing for Basic Income. Ethical foundations for a radical reform. Verso, 1992.
Vinge, V. (1993): The coming technological singularity. In: VISION-21 Symposium, NASA Lewis Research Center and the Ohio Aerospace Institute. NASA CP-10129. http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html
Vladeck, D. C. (2014): Machines without Principals: Liability Rules and Artificial Intelligence. In: Wash. L. Rev. 89 (2014), p. 117.
Wallach, W.; Allen, C. (2008): Moral machines: Teaching robots right from wrong. Oxford University Press, 2008.
Weaver, N.: Paradoxes of rational agency and formal systems that verify their own soundness. Preprint. url: http://arxiv.org/pdf/1312.3626.pdf.
Weld, D.; Etzioni, O. (1994): The first law of robotics (a call to arms). In: AAAI. Vol. 94. 1994, pp. 1042-1047.
Widerquist, K., et al. (2013):Basic income: an anthology of contemporary research. Wiley-Blackwell
Winfield, A. FT.; Blum, C.; Liu, W. (2014): Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In: Advances in Autonomous Robotics Systems. Springer, 2014, pp. 85-96.
Wissner-Gross, A.D., Freer C.E. (2013): Causal entropic forces. In: Physical review letters 110.16: 168702.
Yampolskiy, R. (2012): Leakproofing the Singularity: Artificial Intelligence Confinement Problem. In: Journal of Consciousness Studies 19.1-2, pp. 1-2.

and visual processing methods, 3D models of the environment with robust physics engine – and new applications of human computing and crowdsourcing hold great promises for health and wellbeing. We are neither claiming nor excluding that human intelligence will be reached in some years from now, but make the above claim, which is both weaker and stronger. We believe that fast developments for health and wellbeing are the question of active collaboration between health and wellbeing experts and motivated engineers.

Keywords: personalization, machine learning, smart tools, crowdsourcing, data mining

Research priorities for robust and beneficial artificial intelligence 60

Success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to research how to maximize these benefits while avoiding potential pitfalls. This document gives numerous examples (which should by no means be construed as an exhaustive list) of such worthwhile research aimed at ensuring that AI remains robust and beneficial.

Keywords: artificial intelligence, short and long term impacts, law and ethics, computer science research