Gáspár László

# A MESTERSÉGES INTELLIGENCIA LEHETSÉGES „ETIKAI ASPEKTUSA" ÉS HATÁRAI

*„A tárgykörök elhatárolása, ezek speciális övezetekre osztása nem szakítja szét a tudományokat, hanem csak érintkezéseket teremt a határok között, aminek révén határterületek rajzolódnak ki."*
(Heidegger)

Az ember és a technika kölcsönhatását társadalmi krízisek jellemzik. A kettő kölcsönhatásából eredő konfliktusok jó része abból adódik, hogy „az új technikák új felelősségeket hoztak létre olyan időben, amikor még semmiféle törvény nem szabályozta ezeket a felelősségeket".[1] Mind a technikafilozófia, mind a kibernetika látómezején igen korán, a bioetikát, a környezeti etikát és más alkalmazott etikai irányzatokat megelőzve jelenik meg a „felelősség" kérdése. Nem feladatunk az alkalmazott etikai irányzatok fejlődését vázolni, de érdemes megjegyeznünk, hogy az alkalmazott etika természethez köthető ágai, mint a bioetika, vagy korábban a környezeti etika esetében kézenfekvő módon, a természettel kapcsolatos elméletek horizontján jelennek meg az etikai kérdések. E tény azt a kézenfekvő instinkciót közvetíti, hogy az ember nagyobb technikai arzenál birtokában nagyobb „felelősséggel" tartozik környezete iránt. Más-más megfogalmazásban, de minden ilyen etikai vonatkozású írás sorai mögött ott lappang az ember a természettel szemben érzett „rossz lelkiismeretének" a gondolata. A technikával kapcsolatban azonban sokkal inkább ellentétes kiindulási alapról beszélhetünk, ahol sokkal inkább a technika térhódításával szembeni védekező álláspont kiépítése húzódik meg az elgondolások mögött. Az ember a természet és a technika Szkhüllája és Kharübdisze között hánykódik. A technikával kapcsolatos etikai problémák sokrétűségéről és a hánykódás viszontagságairól árulkodik, hogy a másiktól, ti. a természettel kapcsolatos „felelősségetikától" eltérően a technika nem csak egyirányúan, az ember gépekhez való viszonya miatt problematikus



[1] Norbert Wiener: Az első és a második ipari forradalom. In: *Válogatott tanulmányok.* Gondolat, Bp., 1974. 327.

(technofília, technofóbia), hanem fordítva is. Ma már egyértelműen körvonalazódnak egy küszöbön álló fejlett gépi intelligencia emberhez való viszonyának problémái is. A téma számos kérdést vet fel és ellenérzéseket ébreszt mind a laikusok, mind a témához kapcsolódó szakterületek művelői között, sőt az ellenzők és a pártolók sorai között is, ezért érdemes előre rögzítenünk, hogy nem a technikai fejlődés jövőben lehetséges távlatai alapján kialakított fantasztikus feltételezésekből szeretnénk kiindulni, éppen ellenkezőleg, az ember és gép kétirányú viszonyának lehetséges határait szeretnénk megvonni.

Mivel a technikával kapcsolatos etikai terminusok nem rögzültek, a gép emberhez való lehetséges viszonyának jelölésére a nyugati szakirodalomban használatos „machine ethics"-nek a „gépi etika"-ként történő fordítását használjuk.

A technika-filozófus H. Jonas öt tétele, mely a technika általában vett etikai aspektusának okait sorolja fel és a technika környezetre gyakorolt hatásait tematizálja, akadálytalanul vonatkoztatható a gépi etikára is. A „technika hatásainak ambivalens jellege", az „alkalmazás kényszere", melynek során a technika az élet legapróbb, nanoléptékű pórusaiba is beszivárog, a „globális méretű" elterjedés, az „antropocentrikus szemlélet áttörése", a „metafizikai kérdések felvetése", megfeleltethető az MI fejlődési tendenciái során keletkező problémákkal is.



Gépekkel kapcsolatban mint említettük, etikáról kétféle értelemben beszélhetünk. Első formája, egyidős az emberiséggel és az ember eszközeihez való archaikus-mitikus viszonyán alapul, ez az ember gépekkel szembeni magatartását és a gépek alkalmazási területeinek jogosságát foglalja magában, ám ez nem választható el ellenkező irányú párjától, mivel éppen a technika alkalmazási módjainak kérdése vezet át a második formájához, mely e viszony fordítottját, a gépnek az emberhez való viszonyát helyezi kilátásba. A természet leigázása nem csak a természet önkorrekciója révén bosszulja meg magát, hanem a leigázás eszközeiben, a tudomány és technika médiumában is. A természeturalás ipari forradalom idején kibontakozó lendülete mechanikus gépekkel ment végbe. Az e korszakból öröklött „holt gép" szemlélete alapján a gépi etika értelmetlennek tűnik, hiszen eszközként értelmezi a gépet. A szemléletváltásra a posztindusztriális és információs társadalom 1950-es és 1960-as éveiben kerül sor, amikor az információelmélet és információtörténet dinamikus információs halmazokként

kezdte értelmezni a gépeket. Itt csak arra a nézőpontra utalnánk, miszerint a kétirányú „viszonyulás" nem választható el egymástól sem történetileg, sem tematikusan. A kölcsönviszony minden időben jelen volt a folyamatban, az új elem a gépek interaktív tulajdonságaival jelent meg, olyan interaktív gépi intelligencia megjelenésével, mely lehetővé tette a gépi viselkedés etikai kérdésének felmerülését. Az ember-eszköz, ember-gép, ember-interaktív gép relációban e „viszony" jól lemérhető átalakulásának lehetünk tanúi, éppen ezért, a kettő elválaszthatatlansága miatt jelen tanulmányban először az ember gépekkel szembeni viselkedésének dichotómiáit fogjuk érinteni, majd fordítottjának lehetőségét és feltehető határait tesszük vizsgálat tárgyává.

Hogyha hitelt adhatunk Heidegger vagy Marcuse technikai haladással kapcsolatos kritikai meglátásainak – és a szellemnek az intelligenciával szembeni Heidegger által megállapított háttérbeszorulásának, és a technika hatalmi apparátussal történő összefonódása láttán minden okunk megvan rá, hogy ezt tegyük – akkor éppen e „fejlődéssel" szembeni aggályok miatt tűnik ésszerűnek bizonyos etikai elvárások megfogalmazása az interaktív gépi viselkedéssel szemben.



Ha tehát a klasszikus ipari forradalmak által belénk sulykolt előítélettel közelítünk a gépi viselkedés etikai limitálásának kérdéséhez, akkor a „gépi etika" fogalma is puszta értelmetlenségnek fog tűnni, hiszen a gőzgépek visszacsatolási mechanizmusa semmiféle viselkedésbeli jellemzővel nem rendelkezik. A gépi etikával szembeni ellenvetések és ellenérzések magvát az a megállapítás alkothatja, miszerint a gép csak holt eszköz, mely programokat hajt végre, így autonóm döntések, öntudat és érzelmek híján működése nem értelmezhető etikailag, vagyis a legügyesebb programozás is az ember erkölcsi magatartásának „szimulációját" képes biztosítani csupán. Mind az ipari forradalom eszközszemlélete, mind a tradicionális etikai elméletekben központi szerepet játszó öntudat és autonóm cselekvés normája ignorálja a gépi etika létjogosultságát, s egy erőteljes mozdulattal határt von a gépi működés és az emberi viselkedés közé. Ehhez jön még a felvilágosodás racionális antropológiája, mely egyetlen intelligencia, az értelmes emberi lények egyenlőségének deklarálásán, következésképpen a nem értelmes és a nem emberi intelligencia automatikus kizárásán alapul.

Mindezek konklúziója összességében annak elismerése lesz, hogy a nyilvánvaló határok mentén elkülönülő emberi és gépi világ, kölcsönö-

sen értelmezhetetlen a másik szabályai szerint, amivel ráadásul mind az MI lelkes rajongói, mind ellenzői egyet kell, hogy értsenek. Viszont ez se nem igazolja, se nem cáfolja a gépi etika lehetőségét, minthogy a szétválasztásban foglalt probléma alapvetően nem etikai vagy technikai, hanem pusztán szemiotikai természetű, s annyit állít csupán, hogy az emberi erkölcs fogalmaival megragadhatatlan a gépi viselkedés, és viszont. Azonban éppen e szemiotikai törésvonalnak kell arra emlékeztetnie, hogy ott, ahol határ húzódik, az átjárás elvi lehetősége is adott valamilyen formában. Ember és gép érintkezését hajlamosak vagyunk kizárólag programozási kérdésnek tekinteni. Amíg nem létezik öntudattal bíró MI, addig ez alighanem tartalmaz is féligazságot, ám az igazság másik oldala az, hogy a gép emberrel szembeni viselkedését vezérlő programok is számos kérdést támasztanak. Elég, ha csak a jóváhagyás és a legitimáció kérdését vagy a beprogramozandó maximák körének kérdését említjük. Ki és milyen elveket tartatna be a gépekkel? A beprogramozott maximák azonban nem csak a legitimáció miatt lehetnek aggályosak, hanem éppen a két világ átjárhatóságának, a szemiotikai problémának a látványos megkerülése miatt is. Könnyen adódhat az a fonák helyzet, hogy a gépi intelligenciával szembeni humán ellenszenv és a „gépi etika" létjogosultságával szemben érzett kétely, valamint az intelligens gépek viselkedésében rejlő veszélyek felismerése vezet el együttesen oda, hogy éppen e fenntartásokból kiindulva ruházzák fel a gépeket az ember tiszteletére felszólító, legitimitását tekintve azonban kétes eredetű maximákkal, a gép számára olyan áthághatatlan parancsokkal, melyek minden gépi tevékenység „tízparancsolatának" számítanának. A gépek, pontosabban a gépi intelligencia iránt érzett humán ellenszenv a döntések „parancsolatszerű" korrekciós szabályainak betáplálása és korlátozása révén zárkózna el a gépi etika lehetőség-feltételétől. Ugyanakkor könnyű belátni, hogy a döntéseket vezérlő „parancsolatszerű" maximák beprogramozása mindenkor legitimációs kérdések megválaszolásával maradna adós, nem beszélve arról a kibernetikai problémáról, hogy a valódi szituációk egyedi és valószínűsíthető változó értékeit nem minden esetben lehetne behelyettesíteni ellentmondásmentesen az általános maximák szűk körén értelmezett szabályrendszerének képleteibe, vagyis a programozás nem tenné a gépet „jónak és rossznak tudójává", aminek következtében a szabályok valódi, egyedi szituációkban történő alkalma-

zása hibákkal járhatna, amit minduntalan korrekciós szabályoknak kellene újra meg újra felülírniuk.[2] Annak ellenére, hogy a rendszeres kibernetikai elmélet kidolgozója, N. Wiener már az 1960-as években rámutatott ennek az eljárásnak a gyakorlati veszélyeire, hogy ti. a gép számára a programozás révén adódó logikus megoldás nem feltétlenül jelent humán értelemben is jó megoldást, az MI modern elméleteiben lépten-nyomon felbukkan a korrekciós szemlélet. Létezik olyan megoldási javaslat, mely a gépi cselekvés korrekciós elvének a cselekvés-utilitarizmus Bentham-féle elméletét tenné. (Azt persze el kell ismernünk, hogy a haszonelvűség elmélete formalizálható valószínűségi komponensekkel, így értelmezhetővé válna a gépi problémakezelés számára is. De éppen az a kérdés, hogy miért pont az utilitarizmus szabályai szerint kellene cselekednie egy gépnek.)

Ha tehát feltételezzük, hogy a gépi és emberi világ határvonalán létezik „átjárás" vagy valamiféle érintkezés a kétféle világ között – márpedig az ember-gép kommunikáció a jövőben valószínűleg számottevőbb lesz a mostaninál is –, akkor annak módszerében az ember erkölcsi világa és a gép működése között végbemenő szintézisen kell alapulnia. Mivel a „gépi etikával" szemben felhozható legsúlyosabb (és persze jogos) ellenérv magját az öntudat és az autonómia hiánya, valamint a cselekedetek szimulált jellege alkotja, ezért a két világ érintkezésének a lehetőségét, vagyis szintézisét is az öntudat és az autonóm döntéshozatal támasztéka nélkül kell feltételeznünk. Hogy az etikai kérdésekben milyen módon játszik szerepet az autonómia és az öntudat és, hogy mit takar a cselekedetek gépi „szimulációjának" ténye, kimeríti jelen tanulmány kereteit. Az ember etikai döntéseit feltételező öntudattól és szuverén egyéniségtől elhatárolt etikának persze végzetes következményei lehetnének, de a gép esetében ez nem feltétlenül releváns következmény, hiszen működésük során sem hiányoljuk ezeket az esszenciális emberi tényezőket, pragmatista értelemben csak az életünket befolyásoló gépi működés outputjainak jótékony hatására számítunk.



*

[2] Vö: Susan A. J. Stuart – Chris Dobbyn: A Kantian Prescription for Artificial ConsciousExperience. Leonardo, Vol. 35, No. 4. pp. 407, 2002.

Az információs forradalmak kibontakozásában szabályszerűség fedezhető fel:[3] a paradigmaváltás-értékű információs forradalmak mindegyike magába olvasztotta az előzőek vívmányait, miközben egyre kevesebb időt hagytak a velük járó szemlélet elsajátítására. Amíg az írás ezer, a nyomtatás száz, a távközlési eszközök már évtizedes léptékben, addig a digitális eszközök még gyorsabban váltak életünk részévé. Miközben a céhes mesterségek esetében egy-egy eszköz működtetésének elsajátításához hosszú „tanulóévekre" volt szükség, addig mára voltaképpen egy használati utasítás elolvasásának időtartamára zsugorodott az elsajátítás ideje. A megtakarított betanulási időnek ára van. A mai gépek gombnyomással és érintéssel történő működtetése, a vezérlés kifinomulása és az interface digitalizálásnak köszönhető fizikai egyszerűsödése nem a gépek egyszerűsödésének, vagy szakértelmünk növekedésének, hanem a gép interaktív képességeinek, kvázi „tudásának" köszönhető, vagyis annak, hogy nagymértékben igényeinkhez igazítottuk. Kényelmünkért cserébe észrevétlenül saját önállóságunkról mondunk le olyan gépi intelligencia javára, mely köztudottan nem rendelkezik autonómiával és öntudattal, de a speciális feladatsorokat messzemenően az emberi képességek túlszárnyalásával hajtja végre.



Eszközeink és világképünk viszontformálták egymást, de a nyugati tudomány globális, vagy Heideggert idézve „planetáris" méretűvé válása, az információs forradalom tekintetében a paradigmaváltás erejével bíró reneszánszkorihoz hasonlítható csupán. Az interaktív gépek és döntéstámogató, valamint szakértői rendszerek korszakában a „holt gép" kifejezése egy csapásra idejétmúlt rögeszmeként lepleződött le, hiszen bizonyos képességekkel felruházva őket megtanítjuk nekik a

[3] Nem feladatunk a nyugati tradícióban meghonosodott technikai szemlélet világméretű exportálásának hatásmechanizmusát nyomon követni, mint ahogyan az sem, hogy e földrész „technikai sorsának" esetlegessége vagy szükségszerűsége felett töprengjünk. Számolnunk kell azonban a ténnyel, hogy bár az 1) írás feltalálása nem az európai kultúra sajátja, de annak 2) nyomtatás révén történő meggyökerezése, a 3) távközlés elterjedése, valamint a 4) digitális eszközök használata, tehát az információs forradalmak jó része a nyugat tudományos-technikai szemléletén belül bontakozott ki, s mint ilyen, ellentmondásainak is legfőbb hordozójává vált.

természetes nyelvi kommunikáció fortélyait, vagy azt, hogyan néz ki bolygónk felszíne, milyen az emberi test, s orvosi diagnosztika felállítását, űrutazások tervezését, műholdak és interkontinentális távközlési rendszerek vezérlését, atomerőművek fenntartását bízzuk rájuk. A folyamatot, melynek során a hatékonyságnövelés érdekében (legyen annak célja politikai, gazdasági, hadászati vagy társadalmi nyereségnövelés) emberi életvilágunkhoz „idomítjuk" a gépeket, s észrevétlenül felvértezzük őket a világunkról való „tudással", az emberi kultúra és életvilág intelligens gépek általi elsajátításának nevezhetjük. A gép az ipari forradalomban csak annyiban hasonlított emberi kultúránkra, amennyire használati szempontból kellett alkalmazkodnia hozzánk; ettől eltérően az információ feldolgozását végző gép nem csak összehasonlíthatatlanul több információt tárol kultúránkról, de azzal, hogy kognitív sémánkhoz, gondolkodásmódunkra hangoljuk őket, ezen információk kezelésében és alkalmazásában minőségileg is „fejlődik".

Ember és eszköz viszonyának kettőssége az információ-történetírás oldaláról nézve persze nem kelt különösebb megütközést. Az eszközhasználat elsajátításának hajnalán az *archaikus ember* még szent alázattal vette kézbe a technikai tudás birtokában kialakított eszközeit, mely „megragadás" konfliktusairól és jelentőségéről hűen árulkodnak a görög kulturhéroszok és a hozzájuk kapcsolódó mitikus alakok történetei. A ma már bensőségesnek tűnő hajdani viszonyt a szakrális szemlélet esetében a technika fétise, a hozzáértés helyét pedig renegát elutasítás vagy felületes funkcionális alkalmazás vette át. A termelő tevékenység éppen az eszközök révén a közvetlen kapcsolatok fellazulását és a specializált termelési módok kialakulását eredményezte. Először az emberi munkavégzés forgácsolódott fel szakmákra, majd a szakmák más-más gépekkel történő munkavégzést igényeltek, mígnem bizonyos munkavégzés (világméretű pénzügyi rendszerek fenntartása, vércukor- és szívritmus szabályozó) mára elképzelhetetlen ilyen technológiák nélkül. A technikai kompetencia felaprózódása az intelligens gépek korszakában minden addigi mértékét felülmúlja, azzal kiegészülve immár, hogy a technikai kompetencia visszafordíthatatlan felmorzsolódásával és specializációjával arányos mértékben növekszik az intelligens gépek önállósága és működési köre is. Az önállóság és a működés körének rohamos kiszélesedése még csak az első lépcsőfok. Habár öntudattal a legintelligensebb gépek sem bírnak, bizonyos terü-

leteken nem csak helyzetkiértékelő, de döntési lehetőséggel is rendelkeznek. Az indusztriális társadalomban először a fizikai munkavégzés zömét ruházták át a gépekre, ma a posztindusztriális és információs társadalomban a döntéshozatal majd a tervezés feladatát is megkapták. A puszta erőmunka áthárítását mindinkább az információátvitel és kezelés áthárítása váltotta fel, mely utóbbi a puszta mechanikai ismeretek helyett előbb számítástechnikai, majd a kommunikációval kapcsolatos kognitív folyamatok elsajátítását igényelte és igényli egyre nagyobb mértékben. Az archaikus és a modern közötti szemléletváltás az átmenet korában, az indusztriális korszakban ment végbe, amikor fokozatosan elveszítettük a feudális céhekre még jellemző bensőséges szemléletet, amikor nem csak megerősödött a munka éthosza, de egyben dogmatizálódott s megmerevedett a „holt eszköz" gondolata is.

Az átmenet évszázadaiból öröklött mechanikus szemléletmód határozza meg a gépekhez való mai viszonyunkat is. Emiatt gondoljuk úgy, hogy mindaddig, amíg az intelligens gépek „teszik a dolgukat", mint a GPS, a mobiltelefon stb, addig nem kérdezünk működésük módjára. Ha nagy a baj, akkor is nyugodtak maradunk mindaddig, amíg *hatalmunkban áll* kikapcsolni azokat, hiszen az indusztriális korszak mechanikus gépeihez való viszonyunkat is a ki- és bekapcsolás uralmi aktusa határozta meg évszázadokig.[4] Csakhogy a mai atomerőművek és műholdak leállítása nem csak technikailag nem lehetséges, de egyenesen a modern társadalmak regresszióját és összeomlását is eredményezné. A posztindusztriális korszakban a gépekre való rá-utaltság egyaránt tarthatatlanná teszi a mechanikus lendkerekek olajozott zakatolásához kötődő bensőséges kötődés és a „bármikor-kikapcsolás" uralmi aktusának szemléletét egyaránt. Minden partikuláris jelenség ellenére úgy tűnik, hogy a modern világ mindinkább két ál-alternatíva, a természeti állapotba való regresszió és a technikával való egyre dinamikusabb összefonódás iránya közül választhat.[5] Ráadásul a „bármikor-kikapcsolás" aktusa társadalmi és legitimációs vo-



[4] Vö: Michio Kaku: Robotok. In: Uő. *A lehetetlen fizikája.* Akkord Kiadó, Bp., 2010. 165.

[5] Vö: Hans Jonas: Miért tárgya a technika az etikának? Öt ok. In: *Legyenek-e a fáknak jogaik. Környezeti-etikai szöveggyűjtemény*. Molnár László (szerk.) Typotext, Bp., 1999.

natkozásaitól független, kibernetikai-rendszerelméleti akadályokba is ütközik, hiszen „ahhoz, hogy valóban kikapcsoljunk egy gépet, rendelkeznünk kell azzal az információval, vajon elérkezett-e a veszélyes helyzet".[6] Az ipari társadalom mechanikus gépei esetében a „veszélyes helyzeteket" viszonylag könnyen felmérhetőek, míg a mai összetett rendszerek esetében a veszély felismerését is elemző szoftverek munkájával kerül sor.

Míg a gép kezelésének nehézsége és a gép önállósága között fordított arányosság figyelhető meg, addig az ember szemléletmódja, habitusa, morális felelősségérzete fokozatosan kívül reked a technikai fejlődés tendenciáitól. Az ember még ma is mintegy „gépiesen" ragaszkodik az indusztriális korszakból öröklött beidegződéseihez és megpróbálja azokat alkalmazni az információs társadalom interaktív „evolúciójának" szerkezeteire is. Míg az emberi génszekvenciák az elmúlt tízezer évben nem változtak, és minden emberre irányuló javító szándék útját politikai balsikerek áldozatai szegélyezték (politikai reformok, eugenika, stb), addig a gép fél évszázad alatt átesett azon a döntő változáson, mely alapján kétségkívül nem tekinthető embernek, de több is lett puszta eszköznél, vagyis immár „félúton" van ember és gép között.



Ugyanakkor az út-metafora is egy alapvetően humán értékrendet tükröz, amennyiben az ember végcél-mivoltát tételezi. Ezen a ponton a mesterséges értelemmel szemben felhozható legsúlyosabb ellenérv látszik körvonalazódni, mivel az MI éppen az embert célként tételező humanitás eszméjével tűnik összeegyeztethetetlennek. Az ember: cél, emberre eszközként nem tekinthetünk. Az egyetlen és alapvetően emberi intelligencia paradigmája, az értelmes emberi lények egyenlősége és céljellege az értelem oszthatatlansági eszméjének jegyében fogant, mely eszme a nem emberi intelligencia eszméje révén a szemünk láttára változik oszthatatlansági dilemmává. Az MI a totalitarizmusoktól eltérően és ennek megfelelően azonban nem megkérdőjelezi az ember cél-mivoltát, sokkal inkább mintegy az ember „mellé tolakszik", s megosztani kívánja helyét. (Erre utal H. Jonas negyedik pontja, az „antropocentrikus szemlélet áttörése".)

[6] Norbert Wiener: Tanuló- és önreprodukáló gépek. In: *Válogatott tanulmányok*. Gondolat, Bp., 1974. 179.

A Marcuse-féle kijelentés, mely szerint a „technikai racionalitás" védelmezi az „uralom legitimitását", aligha megkérdőjelezhető kijelentés, de amikor Gilbert Simondonra hivatkozik, egyúttal arra a hagyományos, ipari társadalomból öröklött „gép-eszköz" szemléletre támaszkodik, mely szerint »A gép csupán eszköz; a cél a természet meghódítása… a gép szolga, mely arra szolgál, hogy további szolgákat gyártson«.[7] Csakhogy éppen az emberi történelem egésze azt bizonyítja, hogy a technika segítségével végrehajtott természeturalás a nagyobb energiapotenciál kiaknázását célozta meg, (mint ahogyan valószínűleg már a prehistorikus természeti állapotban is a hatékonyságnövelés érdekében hoztak létre pattintott szerszámokat). A gép azonban az információs társadalomban éppen az említett kognitív sajátosság megjelenésével kerül más megvilágításba. Ha elfogadjuk, hogy a történelem menete – mind a hegeli mind a marxi koncepció keretein belül – a szolga funkciójának átalakulásaként, a despotizmus felszámolásaként és minden ideológiától függetlenül, bizonyosfajta emancipációs folyamatok kibontakozásaként értelmezhető – márpedig éppen a technikai szemléletet képviselő Nyugat ezt az utat látszik követni, és Simondonnal és Marcuseval állítjuk, hogy a gép „szolga", s feltételezzük, hogy a mesterséges intelligencia valamilyen úton tovább fejlődik a jövőben, akkor nem tudjuk az elvi lehetőségét megkerülni annak, hogy ezen emancipációs folyamat aktív részeseként kezeljük a gépeket. Már az öntudattal nem rendelkező összetett gépi intelligencia esetében is elképzelhető oly mértékű önálló döntéshozatali képesség, ami bizonyos döntéshozatali „jogosítvánnyal" ruházza fel azt. Ilyen szakértői rendszerek jelenleg az embertől elszigetelt speciális munkaterületeken (atomerőművekben, részecskegyorsítóknál, űrrepülőkben, orvosi analíziseknél) működnek, a technikának a mindennapi életbe történő beszivárgása azonban alighanem a gép és az ember viszonyának elmélyülését eredményezheti. Fontos tény azonban, hogy az MI, legyen az bármilyen összetett és fejlett is, nem cáfolja azt a Marcuse-féle megállapítást, mely szerint „a technika az eldologiasodás fő hordozójává



[7] Herbert Marcuse: A negatívtól a pozitív gondolkodás felé: a technikai racionalitás és az uralom logikája. In: *Az egydimenziós ember*. Fordította Józsa Péter. Kossuth Könyvkiadó, Budapest, 1990. 181.

lett".[8] A társadalomtudományok mindezidáig többnyire figyelmen kívül hagyták azt a tényt, hogy a technika ugyanolyan hordozója és elszenvedője az eldologiasodásnak mint az ember, sőt elsősorban az ember az, aki mindinkább bevonja a technikát e folyamatba. A technika fejlődése semmiféle ellenérvet nem szolgáltat sem a Marcuse-féle, sem más technika-kritikával szemben, sőt, sokkal inkább aláhúzzák jelentőségüket, és a legitimáció súlyponteltolódására és a technikakritika történelmi korlátaira, valamint továbbgondolásának szükségességére figyelmeztetnek.

Minden kultúra esetében hasonlóak az eszközhasználat elsajátításának archaikus gyökerei, de nem úgy az átvett technika alkalmazási módja, melyben radikális eltérések mutatkoznak, elég, ha csak a lőpor keleti és nyugati alkalmazására gondolunk. Egy az iparosodás időszakából öröklött és a gyors technikai fejlődéssel járó előítéletünk nyilvánul meg a gépeknek főként távol-keleten divatossá vált, ember-formájú kialakításában. Mert bár „a tudomány a Nyugat övezeteiben és történeti korszakaiban egy másutt a Földön példa nélkül álló hatalommá fejlődik, s arra törekszik, hogy e hatalmát végül az egész fölgolyóra kiterjessze",[9] a kultúra melynek nem sajátja a technikai szemlélet, hanem külsődlegesen átvett forma csupán, naivitásánál fogva bátrabban is viszonyul az átvett technológiákhoz. (Míg nyugaton heves vitákat vált ki a géntechnológia alkalmazása, és egyben nagyban hátráltatja a gyógyításban való alkalmazás fejlesztését, addig távol-keleten számos sikereket értek el a kísérletek bizonyos legitimálása révén.) Az emberarcú gép illúziókeltése, alighanem a technika gyors átvételének és elhamarkodott exportálásának a számlájára írható. Az emberarcú gép ugyanis nem a gépi intelligencia mássága iránti érdeklődésről, hanem a hasonlóság kereséséről és a gép partikuláris céljaink alá való betagozásának, gleichschaltolásáról árulkodik, végső soron arról, hogy éppen azért akarunk egy alapvetően nem emberi dolgot ember alakúvá tenni, mert nem ismerjük fel, nem fogadjuk el a mássága mögött rejlő kogni-



[8] Herbert Marcuse: A negatívtól a pozitív gondolkodás felé: a technikai racionalitás és az uralom logikája. In: *Az egydimenziós ember.* Fordította Józsa Péter. Kossuth Könyvkiadó, Budapest, 1990. 190.

[9] Martin Heidegger: Tudomány és eszmélődés. In: *A későújkor józansága*. Göncöl, Bp., 1994. 49.

tív struktúra kvalitását. Amikor azt látjuk, hogy a távol-keleti autógyárak valósággal versenyeznek egymással a tökéletesebb humanoidok előállítása területén, lelepleződik a szemléletmód azon hiányossága, mely bár más formában, de a Nyugatot is jellemzi, hogy ti. a technikai fejlettség ellenére nem értünk meg morálisan egy magas szintű technikai rendszer megalkotására. A folyamat persze aligha áll meg Keleten, hiszen a humán szemlélet alapvetően empatikus társas relációk kialakítására törekszik, melynek nem hús-vér lények semmiképpen nem lehetnek alanyai. (Képesek vagyunk szeretni a macskánkat, egyes állattartók még a hüllőiket is, de rögtön lelepleződik humán szemléletünk, ha a ma már létező, hüllők fejlettségét meghaladó intelligens gépekre próbáljuk meg vonatkoztatni ezeket az érzéseket.) Ha számba vennénk az emberi és az intelligens gép kognitív sajátosságai közötti különbséget, teljesen értelmetlennek találnánk az emberi formát követő kialakítást: nem azért, mert a gép tényleg nem ember, hanem mert alapvetően nem ember, olyasvalami, ami számára értelmezhetetlen és teljességgel indifferens a „kinézet" kérdése. Az emberi magatartást ügyesen szimuláló japán robotok nem mások, mint az emberi pszichikum és morális érzék önmagán végrehajtott kísérletei. Az emberi arcot követő kialakításban a gyorsan átvett technológiával szemben érzett zavar kerül kifejezésre, nem csak a Kelet, hanem az egyes ember esetében is, melynek során a gép másságával szembeni ellenérzéseinket próbáljuk elaltatni.



Az ember gépekhez való felemás viszonyát történetileg az iparosodás idejétmúlt szemléletére vezethetjük vissza, elméleti értelemben pedig arra a szemléletmódra, melynek során a technikai eszközök létrehozatalához szükséges tudományos módszertől megtagadjuk azt a „bizalmat", amit a technikai szenzációknak mindenkor megelőlegezünk. A gazdasági és legitimációs érdekektől mélyebb, az emberi természetben gyökerező okai vannak, hogy „a tudomány feladatait magának a tudománynak a logikája helyett növekvő mértékben külső érdekek szabják meg".[10] Elfogadjuk a tudás fájáról letépett gyümölcsöket, de mélységesen elítéljük a mozdulatot, amivel a gyümölcs birtokába jutottunk, fetisizáljuk a technikai vívmányokat, de nem fogadjuk el az őket életre hívó tudomány módszereit és logikáját, emberi fiziog-

[10] Hans Jonas: A kutatás szabadsága és a közjó. In: *A későújkor józansága I.* Göncöl, Bp., 1994. 38.

nómiát kölcsönzünk az MI-nek, csak ne kelljen tudomást vennünk arról, hogy egy kibernetikai jelenség bújik meg a borítás alatt. Elfogadjuk a dolgot, ha az hasznos számunkra, de nem a dolog természetét, mely adott esetben merőben idegen a miénktől, de egyedi és sajátos módján ettől függetlenül még befolyást gyakorol világunkra. Ésszerű azt feltételezni, hogy ha a tudományhoz való viszonyunk archaikus mélységekben gyökerezik, valamint az iparosodás konvencióival terhelt, akkor félelmeink és fenntartásaink is e viszonyon alapulnak. Rossz lelkiismeretünk pedig éppen ezért nem megalapozatlan. Isten megteremtette az embert, az ember felvilágosodott és maga mögött hagyta istenét. Az ember pedig megteremtette az MI-t. Mi lehet a kissé irodalmias gondolatmenet folytatása? Vajon a technikai civilizáció lehetséges jövője egyben önbeteljesítő jóslatnak bizonyuló jövőalternatívát vetít előre?

A gép kognitív sajátosságainak vizsgálatai azt bizonyítják, hogy a gépi értelem kognitív sémája feletti kontroll elvesztése megalapozott félelem lehet. Az ember és a gép közötti kognitív különbségről jól árulkodnak az 1960-70-esévek MI-kutatásaiban uralkodó pozitivista irányzat eredményei. Az ún. tételbizonyító programok sikerei közé tartozott Pat Langley BACON nevű programja, mely többek között bizonyította Kepler harmadik törvényét, megalkotta a fénytörésmutató fogalmát, valamint Douglas B. Lenat AM nevű programja, mely számelméleti alapfogalmakat bizonyított. Az AM 115 alapszabályával és számok bizonyos halmazával többek között az egész számok fogalmát, a négy alapműveletet, és a prímszámok fogalmát definiálta. Számelméleti alapszabályaiban nem volt semmi újdonság, ám a matematikusokat is meglepte „észjárása": a prímszámok fogalmához úgy jutott, hogy észrevette, a három osztóval rendelkező számok négyzetszámok és négyzetgyöküknek mindig két osztója van. Lenat Lisp-nyelven írt EURISKO programját pedig az tette egyedivé, hogy olyan ún. metaheurisztikával rendelkezett, ami képes volt saját programmagját is felülírni.[11] Az intelligencia fogalmába tartozik az önfejlesztés képessége is, ami ha nem emberi programozó kognitív képességei szerint történik, hanem gépi „észjárás" alapján, akkor elvi lehetősége van annak,



[11] Vö: Mérő László: Észjárások. A racionális gondolkodás korlátai és a mesterséges intelligencia. Akadémiai Kiadó, Optimum Kiadó, Budapest, 1989. 80-85.

hogy a gépi intelligencia oly mértékben szakad el az ember gondolkodási sémájától, melyet csak nagy nehézségek árán tudnánk átlátni vagy megérteni. Talán éppen ezek a veszélyek teszik indokolttá az MI-vel szemben erkölcsi elvárásokat támasztani?

Az 1990-es években döntő fordulat következett be az MI-kutatásokban: az érdeklődés homlokterébe került az ember-gép kommunikáció, a természetes-nyelvfeldolgozás kérdése és a pszichológia, mely folyamat az MI-kutatások interdiszciplináris tudományként történő definiálásával végződött, ami aztán másfél évtized elteltével megnyitotta az utat a gépi viselkedés etikai vizsgálata előtt is.

Az 1990-es évek két MI-fejlesztője, Peter Norvig és Stuart J. Russel ennek szellemében revideálta az MI-t: az információszerzés, tanulás, memorizálás, kategorizálás és a tanultak felhasználása, vagyis az intelligencia programnyelvi átültetése és gépi értelmezése esetükben messze túlmutat a számítástechnika megszokott illetékességi körén, vagyis implicit módon elismerésre került, hogy az MI-t alkotó intelligens ágensek kognitív struktúrájának kialakítása a számítástechnikán túl nyelvészeti, pszichológiai, valamint filozófiai ismeretek nélkül immár lehetetlen. Tágabb értelemben a lélektan, filozófia, informatika, játék- és döntéselmélet, kibernetika, információelmélet, nyelvészet, matematika és etológia, valamint segédtudományaik konvergálnak az MI-fejlesztésekben. Talán nem véletlen, és van valami elgondolkodtató abban, hogy P. Norvig és St. J. Russel a kognitív folyamatokat lajstromba vevő ominózus művüket az Arisztotelész kategóriatana előtti tisztelgéssel kezdik, s az ókori filozófus *Metafizikájában* a tudatmodellezés egyik alapművét látják. Patetikus túlzás volna az MI-fejlesztéséhez hozzájáruló tudományterületek katedrálisának még hiányzó zárókövére „emelve tekintetünket" kijelenteni, hogy a nyugat technikai szemléletének projektuma valósul meg benne.



*

Mi lehet a gépi intelligencia viselkedését vizsgáló etikai aspektus reális kiindulási alapja?

Hogy a programozott viselkedés miért nem lehet kiindulási pont, könnyen belátható, hiszen a programozás a gépek működéséről nem szolgáltat megfelelő garancialevelet az ember számára; az adott prog-

ramnyelven megalkotott maximák egyrészt csak „szimulálni" tudnák az adott etikai normák szerinti döntést, másrészt pedig hatalmi kérdésként fogalmazódnának meg és kimondottan a programozást jóváhagyó döntéshozók állásfoglalását tükröznék, ami Marcuse azon megállapítását húzná alá csupán, miszerint „a gép közömbös társadalmi alkalmazásaival szemben",[12] s mint ilyen végső soron nem a technikai, hanem a legitimáció kérdését feszegetné, valamint bizonyos kibernetikai-rendszerszervezési problémákat támasztana. Ahhoz, hogy a nyugati szakirodalomban „machine ethics"-ként meghonosodott kifejezés létjogosultságát felismerjük, olyan etikai rendszer felé kell fordulnunk, mely a tradicionális etikai elméletektől (aranyközépút, kardinális erények, etc.) merőben eltérő módon közelíti meg a praxist.

Minden környezetével kommunikáló entitás lépései – legyen szó akár egy hüllőről, barátunkról, szomszédunkról, a munkatársunkról, egy mobilkészülékről, vagy a macskánkról – értelmezhetőek etológiailag, a tapasztalt lépések leírhatóak a játék- és döntéselmélet módszereivel, majd e leírások értelmezhetőek pozitív vagy negatív attitűddel bíró lépésekként. Ez alapján nevezhetjük emberi fogalmainkkal semlegesnek vagy ridegnek, barátságosnak és segítőkésznek, vagy ellenségesnek a cselekvőt és fordítva, a pozitív vagy negatív attitűdöt ugyanilyen módon értelmezhetjük a játékelméleti analízis módszerével. (Természetesen a „fordításra" nem egészen reverzíbilis értelemben kerül sor.)



Neumann J. és Oskar Morgenstern *Game Theory* című műve 1947-ben jelent meg. A racionális cselekvés elemzésének elméleti alapjait lefektető *Játékelmélet* iránti érdeklődést jól mutatja, hogy Merrill Flood és Melvin Dresher 1950-es munkájukban megalkotva az egyik leghíresebb játékelméleti alapjátszmát, az ún. fogoly-dilemmát már továbbfejlesztették. A fogolydilemma idővel méltán az egyik leghíresebb játékelméleti alapjátszma lett, ugyanis hűen reprezentálja mind az informális emberi kapcsolatokban, mind a makroökonómia és a közgazdaságtan szintjén kialakuló csapdahelyzeteket. A fogolydilemma érdekessége, hogy nem egy tisztán megoldható szituációra, sokkal inkább fel-

[12] Herbert Marcuse: A negatívtól a pozitív gondolkodás felé: a technikai racionalitás és az uralom logikája. In: *Az egydimenziós ember*. Fordította Józsa Péter. Kossuth Könyvkiadó, Budapest, 1990. 177.

oldhatatlan dilemmára épül.[13] A fogoly-dilemma erkölcsi problémája abban áll, hogy preferenciasorrendjéből szükségszerűen csapdahelyzet áll elő minden döntési helyzetben: nincs optimális megoldása.

A Michigan Egyetem politikatudósának, Robert Axelrodnak éppen a fogolydilemma csapda-jellege adta az ötletet arra, hogy az 1980-as évek elején pályázatot írjon ki többfordulós fogolydilemma lejátszására, melynek pikantériáját az adta, hogy a versenyt programoknak kellett lejátszaniuk. Az a program nyert, amelyik a fordulók során a nyereségoptimalizálás elve alapján a legtöbb pontot érte el. A versenyre többféle programot küldtek el, voltak engedékeny és szélsőségesen agresszív stratégiát választóak, s olyanok, melyek bonyolult módszerekkel próbálták felülmúlni ellenfelüket. A nyertes Anatol Rapaport, ukrán származású amerikai matematikus-pszichológus programja lett, mely a „tit for tat", (a magyar fordításban nem jól visszaadható) „szemet szemért" stratégiát követte.

Axelrod négy kijelentésben foglalta össze Rapaport nyertes prog-

[13] Két gyanúsított esetében a vádemelést a rendőrség csak beismerő vallomásra alapozhatja. Az egymástól elkülönített gyanúsítottak kihallgatása során mindketten ugyanazt az ajánlatot kapják: ha az egyik vall, a vallomást tevő szabadon távozhat, hallgató társa 10 évet kap, ha egyikük sem vall, akkor csupán kisebb, régebbi bűntényekért kaphatnak 6-6 hónapot, ha mindketten vallanak, 6-6 évet kapnak. A játékelmélet módszerének köszönhetően a szimpla erkölcsi dilemma a matematika formális eljárásával is elemezhetővé válik.

| | **A tagad (kooperál)** | **A vall (dezertál)** |
|---|---|---|
| **B tagad (kooperál)** | A: 6 hó (3 pont-R),<br>B: 6 hó (3 pont-R) | A: 0 év (5 pont-T),<br>B: 10 év (0 pont-S) |
| **B vall (dezertál)** | A: 10 év (0 pont-S),<br>B: 0 év (5 pont-T) | A: 6 év (1 pont-P),<br>B: 6 év (1pont-P) |

A kifizetési mátrix kanonizálásával a következő egyenlőtlenséget kapjuk: T>R>P>S, valamint 2R>T+S, mely egyenlőtlenségek a fogolydilemma kialakulásának feltételeit jelentik. A játékelméleti szakirodalomban használatos jelölések: S = Sucker's payoff (vágy kooperálásra) (k-d), P = Punishment for mutual defection (büntetés kölcsönös dezertálással) (d-d), R = Reward for mutual cooperation (jutalom kölcsönös kooperálással) (k-k), T = Temptation to defect (kísértés dezertálásra) (d-k). A legelőnyösebb tulajdonságnak a kölcsönös kooperálást tartó viselkedés ideális preferenciasorrendje az R > S > T > P egyenlőtlenség.

ramjának viselkedését: nice („kedves"), retaliating (megtorló), forgiveness (megbocsátó), non-envious (nem-irigy), vagyis a kezdő lépésben megelőlegezve a bizalmat alapvetően „barátságos", ha támadást tapasztal, csak akkor „megtorló", de ezt követően ha kooperációt érzékel „elnéző", ezért újabb lépés esetén ő is újra „jóindulatúan" tehát kooperatívan viselkedik.

Fontos hangsúlyoznunk, hogy az ember etikai cselekedeteinek és a gépi működés közötti szakadék áthidalására képes játékelméleti megközelítés nem új keletű módszer, és nem is mond többet, mint amit a játékelmélet fogalma takar. A játékelmélet voltaképpen olyan antieszencialista etikai konstrukción alapul, mely mindkét fél, mind az ember, mind a gép részéről kiiktatja a tettek motivációit, előzetes normatív sémáktól függetlenül csak a megnyilvánuló tettek stratégiai lehetőségeit mérlegeli. Bármennyire is kvalitatív „felszínesség" jellemzi a játékelméleti etika módszerét, hiszen e módszer matematikai formalizmuson alapul, egyúttal „korrektnek" is nevezhető abban az értelemben, hogy nem merészkedik olyan területre melynek feltárására nincsenek meg az eszközei, ami nem is volna szerencsés, hiszen „minél tágabb és mélyebb egy pszichológiai jelenség, annál megkérdőjelezhetőbbé válik matematikai formalizációja".[14] Az más kérdés, hogy ennek révén jó és rossz tettek között teszünk különbséget, emberi értelemben kvalifikáljuk a cselekedetek stratégiai vonatkozása mögött rejlő emberi motivációkat. De ami az ember-gép interakcióban rejlő etikai lehetőségek játékelméleti megközelítését illeti, fő „erényének" mégsem a normatív etikai ítéletektől való tartózkodás tekinthető, hanem az, hogy mindkét fél viselkedését képes közös platformra helyezni. A gépi és emberi interakció közös platformra emelése nem az emberi motivációk denaturalizálását célozza meg, hanem a – kanti műszóval élve – a *patologikusan afficiált* emberi tettek és érzelmek interakció során kívánatos semlegesítését és a gép lépéseinek stratégiai minősítését.



*

A kooperatív viselkedés társadalomszervezésünk és közösségi életünk

[14] Jáki Szaniszló: Az agy, az elme és a számítógépek. Kairosz Kiadó, Bp., 2011. 161.

alapjait képezi, ez az etológiai felismerés ma már széleskörű alkalmazást nyert a populációbiológiától kezdve a személyes emberi viszonyokon át a közgazdaságtanig. Axelrod versenyének végkifejlete azt bizonyítja, hogy a kooperatív viselkedés nem csak a bioszféra élőlényei vagy a társadalmi környezet emberi játékosai, hanem még programok esetében is ésszerűen alátámasztható. Axelrod versenyének tanulsága alapján arra a kérdésre tehát, hogy milyen attitűddel rendelkezhet a gépi értelem a jövőben, azt felelhetjük, hogy ha úgy implementálunk egy stratégiai ágenst, hogy a Rapaport-féle programhoz hasonlóan kooperációra törekszik, pontosabban ha a viselkedés vezérlőegysége alárendelődik az így kialakított ágensnek, döntéseit a gép ennek alapján hozza meg, akkor szükségszerűen pozitív attitűddel fog rendelkezni. Takarjon bármit is a fogalom: emberi értelemben „jóindulatú" lesz, mely a kooperációra törekvő lépésekben fog megnyilvánulni. Nem a programozók által elültetett és a legitimálók által jóváhagyott erkölcsi maximák és előítéletek, hanem a kooperatív döntések ágensei képezhetik a gépi etika programnyelvi implementációjának az alapját, ráadásul ez a megközelítés függetlenítené a gépi viselkedést az öntudat vagy az autonómia lehetséges vagy lehetetlen voltának problémájától is. A gép-ember interakció erkölcsi vonatkozásainak játékelméleti alapokra történő helyezése nem utolsó sorban azt a kibernetikai-rendszerelméleti problémát is képes áthidalni, mely szerint lehetetlen „a számítógépekbe programozni azt a fajta információt, amely jelentős mértékben túlmegy a tisztán kvantitatív kifejezések használatán".[15] Ez esetben nem általános törvények és rendelkezések rendszerét kellene betáplálni, és nem ezeknek a való életben felmerülő szituációkkal történő megfeleltetését és kvantifikálhatóságát kellene kivitelezni, hanem az éppen adódó egyedi szituációkat kellene csupán a stratégiai-játékelméleti ágens segítségével feldolgozni és értékelni.



A címben a gépi viselkedéssel kapcsolatos „etikai aspektusra" tett utalást úgy oldhatjuk fel tehát, hogy a játékelmélet alapján a gépi viselkedés – erre alkalmas technológia híján – egyelőre még nem alapulhat autonóm etikai döntésen, ezért helyénvalóbb a „gépi etiká"-t stratégiaként értelmeznünk. Ez alkotja az etikai aspektus címben jelzett „határát", mely határ az ember erkölcsi praxisa és a gépi stratégiai ágens

[15] Jáki Szaniszló: Az agy, az elme és a számítógépek. Kairosz, Bp., 2011. 279.

lépései között húzódik, és az imént jellemzett játékelméleti kooperativitás az a közös platform, ami formális átjárást biztosít a két világ, az etikailag cselekvő ember és a pusztán stratégiai alapon működő gép világa között.

Nem érdemes azonban lebecsülni sem az átjárás formális, sem a gépi praxis stratégiai jellegét. A stratégiai lépésekben kimerülő „gépi erkölcs" már a jelenkori kivitelezhetőség szintjén is megbízhatóbb partnerséget jelent, mint az erkölcsi-gépekként működő, konvencionális ítéletekből merítő embereké. A gépiesen működő konvencióembereket Immanuel Kant erkölcstanának terminusát kölcsönvéve „patologikusan afficiáltnak", azaz érzelmek és érzékcsalódások által befolyásoltnak nevezhetjük.

E gép eszméje mentes a legitimált maximáktól és az öntudat hiányában már-már aggasztóan tiszta. Ám éppen az emberi tényező támaszt egy megfontolandó ellenvetést még a gépi etika játékelméleti megközelítésével szemben is. Nem ígér semmi jót ugyanis a „játszma" ember és a stratégiai alapokon működő, pozitív attitűddel rendelkező kooperatív gép között. A futurisztikus optimizmust, mely szerint az ilyen elveken működő gép megvalósítható és számos rendszerszervezési és legitimációs problémát kiküszöböl, beárnyékolja az a racionálisan belátható tény, hogy ebben az esetben, a legbarátságosabb program is a második stratégia meglépésére kényszerülne, hiszen erkölcsi esendőségünk szükségszerűen az átmeneti megtorlás stratégiáját váltanák ki a Rapaport-féle program alapján működő „legbarátságosabb" programból is. Az okító szándékú megtorlás elviselése pedig, nem tartozik az emberi erények közé.



Nem kell előremennünk egy esetleg a távoli jövőben kifejlesztésre kerülő, önálló döntéshozatalra alkalmas technológia kifejlesztésének a gondolatához, hiszen már a játékelméleti alapon cselekvő és stratégiai ágensekkel működő gép azt bizonyítaná, hogy babonáink és mitologikus sémáink ellenére nem vagyunk oly mértékben „isteniek", hogy önálló értelemmel rendelkező lények technikai megalkotására egyben morálisan is alkalmasak legyünk. Nem tudásunk és technológiánk, hanem újfent tudománytalan módszerünk veszélyeztet e technológia életre-hívásának beláthatatlan balsikerével. A „sikert" ugyanis egy nem a memóriáját, hanem kognitív képességeit folyamatosan és beláthatatlan mértékben fejleszteni képes, a mi kognitív sémánktól merőben

eltérő észjárással működő entitás elfogadása volna. Kérdéses, hogy mennyiben volna siker egy ilyen értelem megalkotása? Mennyire volna sikernek nevezhető egy az embert felülmúló mesterséges „szuperentitás" megalkotása, amelynek egyetlen hibája az volna, hogy nem ember, s erénye az volna, hogy a legtisztább döntéshozatali képességgel rendelkezik? Önmagunk felülmúlása csak a legtisztább morális értelemben volna siker, mely oly mértékű morális érzékről tenne tanúbizonyságot, ami egyben szükségtelenné tenné a tiszta moralitással rendelkező lények megalkotását.



**Irodalomjegyzék**

ANDERSON, Michael – ANDERSON, Susan Leigh: Machine Ethics. Create an Ethical Agent. AI Magazine Vol. 28. No. 4. (2007)
HEIDEGGER, Martin: Tudomány és eszmélődés. In: *A későújkor józansága.* Göncöl, Bp., 1994.
JÁKI Szaniszló: Az agy, az elme és a számítógépek. Kairosz, Bp., 2011.
JONAS, Hans: A kutatás szabadsága és a közjó. In: *A későújkor józansága.* Göncöl, Bp., 1994.
KAKU, Michio: Robotok. In: Uő. *A lehetetlen fizikája.* Akkord Kiadó, Bp., 2010.
MARCUSE, Herbert: A negatívtól a pozitív gondolkodás felé: a technikai racionalitás és az uralom logikája. In: *Az egydimenziós ember.* Fordította Józsa Péter. Kossuth Könyvkiadó, Budapest, 1990.
MÉRŐ László: Észjárások. A racionális gondolkodás korlátai és a mesterséges intelligencia. Akadémiai Kiadó, Optimum Kiadó, Budapest, 1989.
NORVIG, Peter – Russel, Stuart J.: A mesterséges intelligencia modern

megközelítésben
STUART, Susan A. J. – DOBBYN, Chris: A Kantian Prescription for Artificial ConsciousExperience. Leonardo, Vol. 35, No. 4. pp. 407-411, (2002)
WIENER, Norbert: Válogatott tanulmányok. Gondolat, Bp., 1974.

**Tudományterület:** alkalmazott etika, technikafilozófia

**Tárgyszavak:** etika, machine ethics, mesterséges intelligencia, játékelmélet

**Névmutató:** Anatol Rapaport, Hans Jonas, Robert Axelrod, Norbert Wiener, Peter Norvig, Herbert Marcuse, Mérő László, Martin Heidegger