Horváth Zoltán:

Mesterséges intelligencia, mesterséges rugalmasság, mesterséges türelem

Balogh Zsuzsanna, Szalai Judit, Zvolenszky Zsófia (szerk.):
Artificial Intelligence[1]

Öt tanulmány jelent meg a *Magyar Filozófiai Szemle* 2019/4. angol nyelvű számában a mesterséges intelligenciáról (MI). Éppen elég ahhoz, hogy felmutassák a tárgykör legfontosabb, egyre szélesebb körben vitatott problémáit. Ahhoz is ugyanakkor, hogy demonstrálják a filozófusok többségének az MI-hez való szkeptikus vagy kritikus viszonyulását. Érveik szerint az MI-rendszerek nem lehetnek valódi szubjektumok, nem nyújthatnak igazi tudományos felfedezéseket, nincs értelme a moralitáshoz szükséges autonómiát tulajdonítani nekik. Óvakodni kell attól, hogy személyesnek véljük bármely kapcsolatunkat velük, etikai beállítást ruházzunk rájuk a közlekedésben és emberi felügyelet nélküli vezetést a politikában.

Ezt az egyöntetűnek mondható negatív hozzáállást nem osztom, és a szerzők sem győztek meg. Hozzászólásom számára alapvető Fabio Tollonnak az a megkülönböztetése, mely szerint a mesterséges rendszerek autonómiáját tekinthetjük mérnöki értelemben, vagyis a *tervezésük* szintjén, illetve a „mi *leírásaink*" szintjén (11, 18–20).[2] Javaslatom szerint MI-rendszerekről gondolkodva a tervezés szempontját, a célt mindig vegyük figyelembe. Ez megóvhat attól, hogy idegenként tekintsünk a gépekre vagy az algoritmusokra általában, eleve kockázatot és veszélyt látva bennük. Létrejöttük okaként saját gondolatainkat, céljainkat kell minden esetben feltételeznünk. Továbbá annak firtatása helyett, hogy algoritmusok „előidézhetnek-e" gondolatot, „létrejöhet-e" vagy „felmerülhet-e" tudatosság gépekben, ismerjük fel, hogy éppen a szimulált értelem az, ami biztosan valódi abban az értelemben, hogy létrehozása során a gondolkodást nem pusztán „tulajdonítjuk" a gépeknek, hanem *megvalósítjuk* általuk.

[1] Balogh Zsuzsanna, Szalai Judit, Zvolenszky Zsófia (szerk.): *Artificial Intelligence. Magyar Filozófiai Szemle*. 63 (2019) 4. A számban az alábbi tanulmányok szerepelnek: Fabio Tollon: Moral Agents or Mindless Machines? A Critical Appraisal of Agency in Artificial Systems, 9-24; Zsuzsanna Balogh: Intersubjectivity and Socially Assistive Robots, 25-46; Tomislav Bracanović: No Ethics Settings for Autonomous Vehicles, 47-60; Miklós Hoffmann: Science as a Human Vocation and the Limitations of AI-Based Scientific Discovery, 61-74; Zsolt Kapelner: Why not Rule by Algorithms?, 75-90.

[2] A szóban forgó kötetre hivatkozáskor csak az oldalszámot közlöm, a szerző kiléte mindig kiderül a szövegből.

Ennek az alapállásnak az ihletőjeként a descartes-i elmére és támogatójaként a Turing-gépre hivatkozom az első részben. Az MI jelen fejlődése a gépi tanulás előtérbe kerülésével ugyanakkor súlyos ellenérvet szolgáltat e racionális képpel szemben. Ez a második rész tárgya, amely egyúttal komoly kihívást jelenthet Hoffmann Miklós tézise számára is. A harmadik részben még tovább viszem az aktív „karteziánus számítógép" gondolatát egy radikális, metaszintű funkcionalizmusig. Eszerint maga a Descartes-féle elmefelfogás ugyanúgy mesterséges konstrukció a gondolkodás funkcióinak hatékonyabb gyakorlása céljából, mint az MI. Igen drasztikus következménnyel jár ez az asszisztens robotok gondozottjainak Balogh Zsuzsanna által nehezményezett „önámítására" vonatkozóan. A negyedik rész foglalkozik azzal a területtel, ahol ez a funkció az akarat (gyakorlati ész), így az erkölcsiséggel is. Itt Kant – akár gépesítésre is kínálkozni látszó – etikája adja a tárgyalás fogalmi keretét. Tomislav Bracanović is részben erre támaszkodik, így alkalmas lesz az ő autonóm járművekről szóló érveit is mérlegre tenni. Ugyanebben a keretben lesz jó az algoritmusok Kapelner Zsolt által fölvetett emberi közreműködés nélküli politikai vezetésének lehetőségével szembenézni, mert ez az előzőeknél is közvetlenebb és átfogóbb módon irányul az emberiség jövőjére.

Mesterséges *cogito*

Hoffmann Miklós az MI-nek a tudományos felfedezésben lehetséges szerepét kutatja. Azt állapítja meg, hogy az igazi felfedezés ezen a területen továbbra is az emberi értelem privilégiuma marad két olyan lényegi feltétel miatt, amelyek az MI számára elérhetetlenek. Ezek egyike a specializálás abban az értelemben, hogy a kutató tudós képes annak kiválasztására, ami tudománya szempontjából valóban érdekes lehet (65–67), a másik pedig a tudományos felfedezéshez elengedhetetlen lelkesedés, szenvedély, személyes elkötelezettség (61, 63, 67–71). Bevezetésképpen és záró megjegyzéseiben is összeveti ezt az MI számára reménytelennek tűnő két kihívást azokkal a sikerekkel, amelyeket viszont képes elérni egyes szellemi játékok terén. Kiemelt példája a sakk. A sakkot korábban magam is megpróbáltam példaként felhasználni egy már akkor is a fent említett, itt képviselt szemléletű írásban. Érdemes lesz annak felidézésével kezdenem.

1. Hagyományos sakkprogramok tanulsága

Hoffmann indoklása szerint a sakkhoz hasonló játékokban könnyű konstruálni olyan matematikai eszközt, célfüggvényt (vagy akár azok sokaságát), amely a szituációk értékelését és a választás optimalizálását vezérli, míg a kutatásban ilyen kvantifikáció nemigen lehetséges, a tudományos érdekesség, fontosság

aligha mérhető (72–73). Valóban, cikkem sakkprogramozásra vonatkozó első fele éppen ezt a konstrukciót próbálta feltárni (Horváth 2013, 52−63). Ilyenek létrehozása a magas színvonalú sakkjáték esetében korántsem bizonyul könnyűnek, de ami témánk szempontjából lényegesebb: matematikai eszközként *emberi felfedezések* sokaságát foglalják magukban.

Arról van szó, hogy a két fél egymásra következő lehetséges lépéseinek elágazásaiból álló „keresési fa" a maga teljes terjedelmében olyan hatalmas, amelynek a program csak egy töredékét tudja megvizsgálni. E részfára szűkítés miként-jében, az ígéretesnek vélhető lépések kiválasztásában szintúgy elkerülhetetlenül szerephez jut az emberi sakktudás, mint a részfa végeit („leveleit") jelentő sakkállások értékelésében, hiszen azok többnyire nem lesznek végállások a játék szabályai szerint (matt vagy patt például), hanem a program ismert sakkelveken alapuló számértéket rendel hozzájuk egy értékelő függvény révén. Mindezen tényezők emberi szakértők empirikus általánosításai, így tág lehetőséget nyújtanak a kísérletezéshez, például egyes súlyok beállítása és összegzése során. Szinte az összes sakkprogram ma is ezeket a lényegében már Claude Shannon 1950-es cikkében lefektetett elveket követve működik (Shannon 1950).

Nemkülönben a sakkvilágbajnokot végül legyőző, gyakran, így Hoffmann cikkének felütésében is említett Deep Blue, amely sakknagymesterek közreműködésével kialakított keresőeljárással és (több ezer paramétert összegző) értékelő függvénnyel bírt (Hsu 1999, Campbell – Hoane – Hsu 2002). A *puszta „brute force*" legendáját már az hitelteleníti, hogy a mai programok a Deep Blue-nál jóval (több száz Élő-ponttal)[3] erősebbek sokkal *kevesebb* sakkállást megvizsgálva másodpercenként akár egy laptopon vagy telefonon (az egykori óriásgép átlagos teljesítménye 200 millió állás/sec volt). Tehát nem pusztán a számítógépek fejlődése, hanem további kutatások és kísérletezések vezettek a lépéskeresés és az állásértékelés ilyen finomodásához, melynek során az *emberi heurisztikák* azonosítása és gépi megvalósítása kiemelkedő szerepet játszott. Úgy gondoltam, ezek vázlatos ismertetése után joggal vonom le azt a következtetést, hogy a keresést mindenképpen a *tudás vezérli*. Gondolataink, elveink „gépesítése" történik, s idevehetjük akár a Hoffmann által kiemelt invenciót és heurisztikát is. Ha talán még nem is mint a gépek produktumait, de legalábbis a programok eljárásainak általános, előbb köz-, majd gépnyelvi megfogalmazásaiban igenis megragadhatókat.

2. A „transzcendentális" Turing-gép

Ezt az alapgondolatot cikkem második felében megpróbáltam kiterjeszteni magára a számítógépre mint olyan testre, amelyet az ember saját *elméje képessége-*

[3] A Élő Árpád fizikaprofesszor által kidolgozott, ma már a sakkon kívül más sportokban is alkalmazott játékerőt mérő pontrendszer szerint.

inek ideális használatára alkotott. Legalábbis descartes-i eszmény szerint, ahol az értelem evidens belátásra jut és az akarata eszerint dönt: „A gépek tökéletes karteziánusok. Csakis azokkal a diszkrét és meghatározott információdarabkákkal tudnak bánni, melyeket Descartes »világos és elkülönített ideáknak« nevezett" (Dreyfus 1965, 66). A gép felépítése a program és az adatok olyan *funkcionális* (nem metafizikai) dualitására épül, ahol az adatok az „elme" számára transzparens, könnyen elérhető „emlékezet-" és „képzelettartalmak" gyanánt vehetők. A programot végrehajtó processzor működése mint kanti „öntudat" irányítja ezeket a „képességeket", és egyúttal a gép önmegfigyelésének és önuralásának *funkcióját* is ellátja. Egy beépített „kis számoló", amely egyáltalán nem fenyeget sem ontológiai problémával, sem végtelen regresszussal (Horváth 2013, 63skk.).

Az utóbb idézett, Turingra utaló kifejezés nyomán akkori merész hasonlatomat most megtoldom egy olyannal, ahol az elme descartes-i felfogásának gépi modelljéhez nem a számítógép szolgál mintaként, hanem az annak alapjául vehető absztrakt Turing-gép. A számítási eljárások általános leírására Turing kifejezetten *az ember számolási tevékenységét* mintául véve alkotta meg a számoló gép definícióját. A gép konfigurációi az ember „elmeállapotainak" felelnek meg, melyek mindegyikében az előtte levő „papíron" (szalagon) egyetlen szimbólumot „figyel" (*see*, *observe*) vagy „ír le". Egyszerre (egy „pillanatban") csakis a „közvetlenül felismerhetőnek" van „tudatában", habár tud „emlékezni" is a már „letapogatott" (*scanned*) szimbólumokra. Céljai érdekében tehát Turing a számoló ember (*computer*) hasonmását konstruálta meg (Turing 1936, 231–232, 250–253).

Ha továbbvisszük a hasonlatot, azt is mondhatjuk, hogy a Turing-gép egy elemi műveletekre bontott tudat modellje. Kezdve a gép *belső állapotának* fogalmával, de mindjárt hozzávéve az *idő* elemi szimulációját, az egymást követő diszkrét gép-elme állapotok posztulálását. Az adat felvételét a szalagról, a „külvilágból" az *észlelés*, a jobb-bal elmozdulást és a kiírást a *cselekvés* legelemibb formájaként értelmezhetjük. Különös jelentősége a belső-külső kettősségnek, hogy ez teszi lehetővé az algoritmuskészítés *függetlenségét* az adatoktól. A program futása során viszont már kölcsönösen függenek egymástól a belső és külső állapotok, annyira, hogy a következő állapotot megszabó átmenetfüggvény mondhatni a szigorú *ok-okozat* összefüggést képviseli. Az előbbi szolgálja azt a tervezői törekvést, hogy minél nagyobb autonómiát ruházhassunk a gépekre, az utóbbi révén támaszkodhatunk olyan nagy bizalommal rájuk. Mindkettő a morális gondolkodásnak is alapvető aspektusa: szabadon választhassam elveimet, céljaimat és eszközeimet, ugyanakkor szilárdan kövessem őket a végrehajtásban a cselekvés során.[4]

[4] A szabadság és a szükségszerűség e szimulált formái az „antinómiájuk" feloldását is jól megragad-

A fenti hasonlat kifejezetten csak a számolási feladatokat kitűző és végrehajtó embert veszi tekintetbe mintaként, de éppen az a kulcskérdés az MI témájában, hogy mennyire jogos ez a „csak". Ismeretesek az azzal szembeni ellenvetések, hogy puszta komputáció reprodukálni tudná az ember elmebeli állapotait vagy képességeit. Itt csupán ezek *indítékára* szeretném fölhívni a figyelmet. Közös jellemzőjük, hogy az algoritmusokat veszélynek vagy ellenségnek látják, ami ellen védekezni kell. Egyik típusuk a Gödel-tételre hivatkozva próbálja „legyőzni az algoritmust" (Penrose 1993, 83–85, 134–135), a másik úgy akarja a hatálya alól kimenekíteni az embert, hogy valamely, a puszta számolás számára elérhetetlen, intim elmeállapotra (esetleg tárgyra vonatkozásra) hivatkozik, legyen az kvália, megértés, intencionalitás. Mindkét változat idegennek tekinti az algoritmust és igyekszik azt belekényszeríteni egy korlátozott „leírásba". Egyik sem értékeli, vagy nem is látja, a számolás mögött az *algoritmusok alkotásának* – természetesen: emberi – szándékát és képességét. E munka gyakorlója, például az MI szakember számára viszont az állítólag sajátosan emberre jellemző nem-algoritmikus mentális minőségek értéke is azon múlik, mi azok *funkciója*, illetve van-e egyáltalán – már azon kívül, hogy kitüntetettségünkről biztosítanak minket.

Ha „megengedjük a gépesített gondolkodás lehetőségét" (Shannon 1950, 1–2), akkor világosan tudatában lehetünk, hogy annak szerzői mi magunk vagyunk, akik a gondolkodás képességét a biológiaitól eltérő testben is gyakoroljuk. Eközben Turinggal tartva nyugodtan lemondhatunk arról, hogy állást foglaljunk (vagy akár „udvariasan megegyezzünk") abban a kérdésben, „valóban" gondolkozik-e a gép valamilyen kimondottan emberi értelemben (Turing 1965, 139). Csak azt kell figyelembe vennünk, képes-e annyira vagy még inkább intelligensen cselekedni, mint mi emberek. Ezért attól a gondolattól sem kell viszszariadnunk, hogy *nálunk okosabb, tudatosabb és erényesebb gépeket* építhetünk, amelyek szabályokra, rendre, módszerre alapozottak és a ránk jellemző testi gyengeségektől mentesek. Sőt, mivel ezek azért gondolkodók, mert *mi* így gondoltuk ki és alkottuk meg őket, létünk értelmének és rendeltetésének kérdéséhez nyújtanak segítséget: általuk megtanulhatunk engedelmeskedni az „Ismerd meg önmagad!" régi parancsának (Simon 1980, 6268; Horváth 2013, 69–71).

Mesterséges tapasztalat, mesterséges érzés, mesterséges intuíció

1. Zérók tanultsága

A mesterséges intelligencia fejlesztése a legutóbbi időben olyan eredményeket

hatóvá teszi. A program *megalkotása* „kívülről", az adatokon „túli" világból határozza meg annak „immanens" *futását*.

produkál, amelyek kellemetlenül érintik mind Hoffmann érveit az MI-rendszerek lehetőségeinek korlátait illetően, mind az általam felvázolt képet a „tökéletes karteziánusként" gondolkodó gépről – noha nem ugyanazon okból. Maradjunk elsőre most is a játékoknál, kivált a sakknál. A Google DeepMind 2017. december 5-i publikációja AlphaZero nevű programjukról ad hírt, amely az eddigi sakkprogramok szofisztikált keresési technikái és a játék szakértőinek segítségével finomított („kézzel gyártott") értékelő függvényei helyett, „tiszta lapról" indulva, pusztán a játékszabályok ismeretére épít (erre utal a „zéró"). A tudását önmaga ellen játszva, *megerősítéses tanulással* kialakító mesterséges neurális háló (NN) 4 óra tanulás után elérte a legerősebb sakkprogram (Stockfish) játékerejét, majd a 9 órás betanulás végén kb. 100 Élő-ponttal bizonyult jobbnak.

Az eljárás a shannoni elvektől abban különbözik leginkább, hogy *statisztikai* mutatók játszanak benne vezető szerepet. A program az önmaga elleni játék során alakítja ki azokat a paramétereket, amelyek alapján tetszőleges álláshoz hozzá tudja rendelni az abból lehetséges lépések „valószínűségének" (jóságának) eloszlását és az állás értékét mint a játszma onnan várható kimenetelét. A keresés sem a hagyományos programoknál alkalmazott „alfa-béta" algoritmus (a játékelméletből ismert minimax eljárás javítása), hanem Monte-Carlo fakeresés (MCTS). Szimulált játszmákat generálnak, melyek lépéseinek kiválasztását az aktuális NN szerinti valószínűségek mellett az adott állás addigi kutatásának eredményei is befolyásolják. Egyfelől a lépésre kapott (nagyobb) érték az abból lejátszott játszmák végállásai +1, 0, vagy −1 (győzelem, döntetlen, vereség) értékének átlagaként – *exploitation*: a már jónak bizonyult választások kiaknázása. Másfelől a lépés addigi *kisebb* „látogatottsága" – *exploration*: a még kevéssé választott alternatívákban rejlő esetleges lehetőségek felkutatása. A két tényező közti balanszírozást vezérlő paraméterek a kísérletezés további tárgyai (Silver et. al. 2017).[5]

Hasonlóan a Deep Blue utáni fejlődéshez kisebb vállalkozások is próbálják követni az AlphaZero NN és MCTS technikáját a DeepMind közleményei nyomán. Legsikeresebb az LCZero, amely (a Google kapacitásai híján) önkéntes közreműködők segítségével tanítja be az NN-t, és ha nem is egy nap, de egy év alatt elérte a Stockfish szintjét. Jelenleg folyik a csata a „hagyományos" gépekkel, ám nemcsak eredményességben, hanem játékmódban, stílusban is: mélynek ható „stratégiai érzék", fokozatos pozíciós fölénybe kerülés a „hálós" oldalon,

[5] A történet a Hoffmann által szintén említett gó játékra írt programok fejlődésétől vezetett ide, amelyre az NN jó ideje alkalmasabbnak tűnt. Ebben a játékban ugyanis nagyobb jelentősége van az alakfelismerésnek és kisebb a változatszámításnak, lévén számításfájának mérete a sakkét is messze felülmúlja. Csak 2016-ben tudta a legerősebb gójátékost legyőzni egy részben emberi játszmákon, részben önmaga ellen játszva tanuló NN. Az AlphaZero az addigi legjobb góprogramokat is megverte teljesen „tudásmentesen", 24 órai tanulással.

pontos számolás, néha váratlan taktikai csapások az egzakt minimaxoson.[6]

Ám folyik az interpretációban is. Külön figyelmet érdemel a szerzők megítélése az eredményeikről. A játszmák során az AlphaZero ezerszer kevesebb állást vizsgált meg másodpercenként, mint a Stockfish (60–80 ezer vs. 60–70 millió), mert az NN sokkal *jobban szelektál* az ígéretes változatok között. Ezen szerintük *emberszerűbb* kiválasztási módon túlmenően jó néhány játszmában hosszú távú stratégiai előny ellenében áldozott tisztet, ami azt sugallja, hogy kifinomultabb, kontextusfüggő pozíciós értékelése van a korábbi sakkprogramok szabályalapú értékelő függvényeihez viszonyítva. Mi több, kijelentik, hogy a konvencionális emberi tudástól és normáktól nem korlátozott öntanuló program „érzést, intuíciót, belátást tesz magáévá", és „kreatív". Ennek köszönhetően felfedezte az ismert stratégiai eszméket és új, izgalmas ideákkal gazdagítja a több évszázados sakktudást (Silver et. al. 2017, Silver – Hubert – Schrittwieser – Hassabis 2018).

Korábbi cikkemben amellett kardoskodtam, hogy az ember a sakkprogramok által az önmaga számára is közvetlenül megfogalmazható sakktudást hozza létre a gépben, s ez döntően fontos a teljesítményeikhez. Most azt kell megállapítani, hogy a korábbiakat is túlszárnyaló öntanuló programoknál egyáltalán nem erről van szó. Azok az itt szóba jöhető *tapasztalatot*, vagyis sakkállások sokaságát hozzák létre és az abból leszűrhető *empirikus tudást* alkotják meg, számunkra nem követhető és átlátható módon. Úgy, ahogyan az emberi sakkozó tanul a játszmákból és játszmaelemzésekből, anélkül, hogy fejlődéséről önmagának pontosan számot tudna adni. Továbbá az NN adta értékelés és a statisztikai alapú keresés annak szimulációját képes nyújtani, amire a legsajátosabban *emberiként* szoktunk hivatkozni: egyfajta *mesterséges érzést és intuíciót*, és az AlphaZero ilyenek produkálása révén igazi *felfedezésekhez* jut a sakkjátékban. Az e képességekkel fölvértezett program néhány óra alatt felülmúlja „hagyományos" társainak fél, az emberi sakktörténetnek öt évszázados munkáját.

2. Mesterséges invenció

Karteziánus eszméim szempontjából e felfedezések *miként*je a kellemetlen: általános elveink tudatos gépi megvalósítása helyett „alulról", egyfajta mesterséges tudattalanból látszanak eredni. Hoffmann Miklós viszont *egyáltalában* vitatja el a valódi felfedezések lehetőségét az MI-rendszerektől. Az ehhez hiányolt képességek közül a *specializáció* tulajdonképpen a *szelekciót*, az adott tárgy szempontjából „érdekesnek" a kiválasztását jelenti (66). Első megítélésre nehéz ezt megtagadni akár a kifinomult célfüggvénnyel és lépéskiválasztó eljárásokkal működő sakkprogramoktól, akár öntanuló változataiktól. Igaz, „végtelenül sok érdektelen" lépést is kiválasztanak, de ebből nem következik, hogy a né-

[6] Itt élvezhető: https://tcec-chess.com/

hány ígéretes, további vizsgálatra érdemes idea észlelésének valószínűsége a gép részéről „a zéróhoz nagyon közel" volna (67). A jobb programoknál éppenséggel általában nagyobb az a valószínűség, mint bármely emberi sakkozónál.

Hoffmann persze aligha tart sakklépést tudományos kutatáshoz fogható felfedezésnek. Az ilyenekhez szükséges specifikáció tudományterületek, problémák, teorémák közötti szelekciót kíván, tekintettel arra, hogy – ha csak a matematikát nézzük is – az egyre hatalmasabb számú fölfedezett tételt senki sem képes átlátni, legfeljebb kisebb részterületeket (65–66). Ám éppen ezért: miért *nem* az MI-rendszerek azok, amelyeknek bővülő tudásától és kapacitásaitól a nagyobb tárgyterületek figyelembe vételének és összekapcsolásának képességét várhatnánk? Ráadásul éppen a matematikai fogalmak definíciói nagyon is alkalmasak a problémák megoldásainak vagy az állítások bizonyíthatóságának „mérésére" (72). Maga Hoffmann mutat egy szép tételt, megjegyezve, hogy formulázása és a pontos állítás igazolása „egyszerű mechanikus számítás" dolga (68). Szándéka persze itt az, hogy a tétel felfedezésére irányítsa a figyelmet, de a precíz formulázás és bizonyítás lehetősége már alapul szolgálhat a gép tanulása számára is. Mind a betanulást szolgáló adathalmaz generálása, mind a hipotézisek tesztelése könnyebb lehet, mint például a természetes nyelvek gépi fordítása esetében.[7]

Ezeknél még súlyosabb ellenvetéseket látok a tudományos felfedezéshez Hoffmann által (is) elengedhetetlennek vélt másik emberi tényezőt, az *entuziazmust* illetően. Idetartozik az elhivatottság, a szenvedély, a belső érdeklődés, a személyes elköteleződés. Ilyenek nélkül nem merül föl bennünk valamely ígéretes, de atipikus ismeret előrelátásának izgalma, az invenció megmagyarázhatatlan „gravitációja", amely azután valódi felfedezéshez vezethet. Minderre azonban az MI-rendszerek nem képesek (hacsak nem bírnak esetleg majd a jövőben mentális állapotokkal), mert külsőleg, a programozók és az adatok által vezéreltek, s így eleve csak tipikus mintákat követhetnek (67–69). Hadd maradjak még a sakknál. A nagymester „tipikus mintákat" pásztázva érzi, hogy „valaminek lennie kell az állásban", keresi, felsejlik benne egy „motívum" s egyszer csak fölfedezi az aktuális „adatok" által lehetővé tett egyedi, „atipikus" megoldást. Ehhez sokszor számára is szükséges az entuziazmus szóval összefoglalt mentális állapotok némelyike. A sakkprogram (akár hagyományos, akár öntanuló) tipikus minták – szakértők és saját programozói számára is sokszor atipikus – kombinációit igazán szenvtelenül kutatja, szelektálja és találja meg *ugyanezt* a lépést.

Általában sem igen látom elkerülhetőnek a fájdalmas következtetést: az általunk annyira nagyra becsült fenti mentális állapotok – ha a felfedezésben

[7] Ezen a módon megtanult már NN nagyon jól függvényeket integrálni és differenciálegyenleteket megoldani, ami – e műveletek jellege miatt – a mélyebb felfedezések felé tett lépésnek is vehető (Lample – Charton 2019).

betöltött szerepüket tekintjük – inkább az ember szegénységi bizonyítványát jelentik a gépekhez viszonyítva. Ami az ember számára olyan nagy dolog: az *elhivatottság*, hogy egy speciális területnek szentelje magát, kifejlessze és gyakorolja az ahhoz szükséges képességeket, belső érdeklődésre tegyen szert, ami azután sejtésekhez, „heurisztikus impulzusokhoz" és gyümölcsöző felfedezésekhez vezethet (61, 63–65, 67) – az a „hivatás" a gépeknél *eleve adott*. Azért *hívtuk* életre őket, hogy egy bizonyos tevékenységet (emberi mércével mérve) fáradhatatlanul és szó szerint teljes odaadással végezzenek. Talán félrevezethet minket, hogy nekik nem adózunk ezért megbecsüléssel vagy tisztelettel, mivel részükről, a fenomenális szféra híján, mindez nem jelent áldozatot. Ez azonban morális kérdés, önmagában (és majd alább is) megfontolandó ugyan, de nem érinti a felfedezések tudományos értékét.

Ám még ha elfogadjuk is a Hoffmann által feltárt mentális képességek elengedhetetlen szerepét a felfedezésben, akkor is szembe kell néznünk vele, hogy azok – egye kifinomultabb – *gépi szimulációi* is elláthatják ugyanezt a *funkciót*, melynek mintegy eszközeiként szolgálnak az ő tanulmányában is, és nem hivatkozik valamilyen metafizikai sajátosságukra. Ebben mindenesetre követem őt.

A tudat funkciója

Hoffmann csak érintőlegesen nevezi az MI fundamentális kérdésének, hogy bírhatnak-e ilyen rendszerek mentális állapotokkal, illetve miben is állna az, Balogh Zsuzsanna tanulmányában viszont ez döntő jelentőségű. Az általa tárgyalt időseket és testileg vagy értelmileg sérülteket gondozó robotok ugyanis érzelmeket, sőt szeretetet mutatnak azzal, hogy képesek ápoltjaik arckifejezését, tekintetét, hangtónusát, testmozdulatait detektálni és azokra adekvát válaszokat adni (36–38). Kulcskérdésnek és veszélyt hordozó ténynek tartja, hogy az asszisztens robotok nem valódi szubjektumok, holott csakis azok között lehetséges kölcsönös és fenomenológiai interszubjektivitás. A páciensek viszont gyakran tévesen így tekintenek a szociális robotokra, pedig azok csupán egyfajta „mintha" interszubjektivitásra alkalmasak (27–28, 37–43).

Az AlphaZero csak egy példáját, ezek a robotok már bizonyos szintézisét is látszanak nyújtani az „emergens" megközelítésnek, amely egyre több területen vezet gyakran az embert felülmúló teljesítményekhez. Szerzőnk témáját, az interszubjektivitás kérdését is érinti, hogy egyúttal hódit az a nézet is, mely szerint az NN „emberibb" modell a szimbólummanipuláció „merev" elvénél, amely az MI korai szakaszának vezéreszméje volt. Úgy látszik, éppen azért tartják

emberinek, mert az *agyat* tekinti mintának, nem pedig az *elmét* vagy a tudatot.[8] Éppolyan kellemetlen ez a Balogh által hivatkozott elmefilozófusok és fenomenológusok nézőpontjából, mint Descartes-éből, akire viszont én szeretnék támaszkodni, habár megint nem egészen ugyanaz a gondunk. Ő attól tarthat, a szimuláció becsapja az érintetteket, sőt talán el is halványul számukra a genuin szubjektum különbözősége. Ezért igyekszik fogalmi elkülönítései révén a sajátosan emberit megragadni és a gépitől elhatárolni. Számomra a gépi szimulációnak ez a *módja* és az ebben való hit rejti annak veszélyét, hogy az embert felmenti és eltántorítja valódi önmaga fel- és megismerésének feladatától. Ezért törekvésem továbbra is az *értelem gépben* való *működésének* felmutatására irányul, még az NN esetében is.

1. Karteziánus test, karteziánus agy és karteziánus elme

Témánk irodalmában gyakoriak a Descartes-ot érő vádak az MI ellenzői és támogatói részéről egyaránt, amelyeket az NN sikerei még támogatni is látszanak. Ezek egyike az *egyszerűséget* illeti, amellyel Descartes a gondolkodásról számot ad, holott azt nagyfokú bonyolultság, biológiailag „többszörös vázlat", de még mesterséges formáját is az algoritmusok sokasága jellemzi („*e pluribus unum*"). A leggyakoribb vád kétségkívül a „karteziánus *dualizmus*", ami különösen súlyos abban a formában, hogy az elme sajátos „anyagát" feltételezvén a tudatot misztériummá teszi. Vannak, akik bizonyos passzivitásként írják le a „karteziánus színház" és a „néző" én elképzelésében, vagy mint az elme pusztán érző „fenomenológiai" fogalmát (szemben a kauzális „pszichológiai" oldallal). Mások az MI híveit vádolják azzal, hogy (tudtukon kívül) az elme-test probléma új változatát keltik életre, amennyiben „valamiféle test nélküli »létezést«" tulajdonítanak az algoritmusnak, miáltal érthetetlen marad, hogyan hat a világra, illetve hogy a bármely megvalósulásától független program hogyan „reprodukálja és magyarázza a mentálist" (Dennett 1991, 33, 41–42, 101skk., 227, 261skk.; Dennett 1998, 222, 487skk.; Chalmers 1996, 11–13; Penrose 1993, 37, 457–458; Searle 1980, 423–424).

Meggyőződésem szerint ezek a kritikák nemcsak Descartes iránti ellenszenvről vagy értetlenségről tanúskodnak, hanem az MI alkotóinak szemléletétől való idegenkedésről is. Tömören így fogalmaznám meg ennek okát: a hivatkozott filozófusok *magyarázni* akarják az elmét, az MI létrehozói úgy gondolják, itt az ideje, hogy *megépítsük*, Descartes szerint mindig is annak van itt az ideje, hogy *használjuk*. Márpedig az elmét nem magyarázni, hanem funkciójának megfelelően használni, *rendezőt* jelent a színházban, nem nézőt.

[8] Mintha maga a „mesterséges intelligencia" kifejezés használata is leszűkülne, és egyre inkább kifejezetten az NN-t vagy legalábbis a tanuló rendszereket érdemesítené rá a köznyelv.

Még többről van azonban szó: magáról a „színház" *alapításáról*. Gondoljunk először arra, hogy Descartes az emberi test vagy a világ igazi természetének kifejtése helyett hipotéziseket és hasonlatokat adott egy ehhez a miénkhez pusztán hasonlító világ kialakulásának, illetve a testet csak utánzó gép működésének bemutatásához (Descartes 1992, 54, 57; AT XI. 120, 200–202). Ugyanígy nem követelmény, hogy a mesterséges NN az agy neuronjainak tényleges kapcsolatait adja vissza, az a fontos, hogy az emberi viselkedéshez megtévesztésig hasonló vagy azt valamilyen (végső soron emberi) mérce szerint fölülmúló teljesítményt produkáljon. A magam részéről nemcsak e modellalkotást látom descartes-inak, de megvalósítása során is találok az elme elsőbbségére utaló karteziánus mozzanatokat. Maguk az emergens jelenségek lehetnek mondjuk egy humanoid robot alsóbb, „szubszimbolikus" szintjei, ahol a hierarchikus struktúrán belül a mesterséges tapasztalatot és intuíciót mesterséges értelem vezérli. Továbbá miként Descartes a hétköznapi életből vett hasonlatokra vagy mesteremberek munkáira, úgy az MI kutatói szintén támaszkodnak emberi gondolkodásból merített heurisztikákra még statisztikai alapú módszereknél is. Például az *exploration* és *exploitation* egyensúlyáról való legjobb elgondolásokat a „félkarú (vagy n-karú) rabló" nevű szerencsejáték ihlette. A legfontosabb azonban, hogy ezekben a rendszerekben is *világosan megfogalmazható* emberi elvárások irányítják a megtartandónak a – végül is *mesterséges* – *kiválasztását* az adatok sokaságából. Az adathalmaz *közvetett uralása* eredményezi a szóban forgó teljesítményeket.

Még tovább menve tekinthetjük úgy már magát Descartes-nak a gondolkodásról elmélkedő filozófiáját is, hogy azt az elme önnön *irányító funkciója érdekében* fundálta ki. Az elme tevékenységét segíti a belső szabadságot teremtő elme-test dualitás elmélete.[9] Ez volna a „mesterséges" értelem legradikálisabb felfogása, mely szerint az értelem *eleve mesterséges* abban az értelemben, hogy az elméről, értelemről alkotott *fogalmainkkal célunk* van, azért olyanok, amilyenek, nem pedig, mert ilyenek „valójában", vagy így adódnak, tárulnak fel stb. Ezt az elvet követve próbáltuk fentebb az elme descartes-i felfogásának gépi modelljét látni a Turing-gépben, amelyben egy *mesterséges egyszerűség* és *mesterséges dualitás* segítségével végrehajtott program képviseli az elmeműveleteket. E mesterséges elménk nem kell másként „léteznie", mint valamely funkció szolgálatában *működni* – s eközben nem fenyeget semmiféle „szubsztanciadualizmus": eleve „testesült", még ha az emberi elme (és nem az agy) mintájára alkották is meg.

[9] Értelmezésem szerint tehát Descartes nem magyarázni akarja a tudatot (vagy annak megmagyarázhatatlan voltát), hanem *használni*. A *cogito* ugyan a „szilárd talajt" adja, de nem azért, hogy abban ássa el kincsek gyanánt a módszert (legyen az intuíció, belátás vagy empátia), hogy semmiféle magyarázat ne férjen hozzá. Ellenkezőleg, a gondolkodás munkáját valamely *célra* használja fel, végső soron arra, hogy elültesse és nevelgesse a tudománynak az emberi nemet szolgáló „fáját" (Descartes 1996, 16).

Fordítva pedig: ez a hasonlat arra indít, hogy önmagunkat *aktív*, gondolkodó, a gondolatait irányító és a gondolataival irányító lénynek tartsuk.

2. A funkció tudata

Olyan radikálisan „funkcionalista" válasz(tás) ez, amely a tudat Chalmers-féle „nehéz problémáját" könnyeden átlépi: a tudatnak nem problémája van, hanem feladata. Nem az a kérdés Descartes-nál sem, hogy milyen érzés, hanem hogy mit tegyünk vele (vö. Chalmers 1996, xi–xiii, 11–13). Úgy tűnik, azt a Chalmers által Dennettnek felrótt vonalat követem, miszerint annyit kell megmagyarázni a tudatból, amennyi szükséges, vagyis csak a tudat *funkcióit* (uo. 114, 190). Ám ha a vonal dennetti is, az irány épp ellentétes. A tudatot az itt képviselt szemlélet a legkevésbé sem „kimagyarázni" (*explaining away*) kívánja a funkcióiban feloldva, hanem megint csak használni: *legyünk tudatában* elménk funkcióinak. Igazán remek a Dennett által is propagált hasonlat, miszerint a tudat tapasztalata „felhasználói illúzió", de az ő szempontjából túlságosan is jól sikerült. A felhasználó ugyanis pontosan *tudatában van* az illúzió illúzió voltának is és annak is, hogy az illúziónak *funkciója* van, ennek megfelelően is használja, miközben létrejöttének részletei, informatikai és fizikai háttere nem érdekli. Nem is lehetne jobb példát adni rá, mint amit Dennett: az általa épp használt szövegszerkesztő keltette illúziót, hogy a billentyűzet és az egér „audiovizuális metaforáival" tudja vezérelni azt, ami a képernyőn, a gépben, no és ezáltal majd a filozófiai gondolkodás kultúrájában szándéka szerint megváltozik (Dennett 1991, 216–220, 311–312).[10]

Mesterséges rendszereink esetében nyilvánvaló is, hogy minden komponensükkel és állapotukkal együtt mindig valami cél eszközei, így és ezért alkotjuk meg őket. Eleve fals a „rendelkezhetnek-e mentális állapotokkal?" kérdés, ha nem irányul valamilyen célra. Nem csoda, ha az MI szakértőit nem foglalkoztatja. Ők megelégedhetnek a Balogh Zsuzsanna által leírt „mintha" szubjektivitással és „vékony" (*thin*) interszubjektivitással (38–39), ha produktumuk jól látja el a funkciót, amelyre szánták, és partnerük számára magas „érzelmi intelligenciájúnak" mutatkozik. Baloghnak az asszisztens robotok megtévesztett páciensei iránti aggodalmai éppen azért merülhetnek föl, mert a másik (ember) mentális állapotainak léte és mibenléte az ő számukra is elsősorban e *funkciók* miatt fontos. Ha *ezek érdekében* szükséges a robotokba érzelmeket, empátiát belelátniuk, akkor e „tévedésükkel" hozzájárulnak az MI jogos térnyeréséhez. Feltehetően

[10] Persze e *szándék* is ahhoz a tudathoz tartozik, melynek ki-, el- és szétmagyarázására Dennett vállalkozik. Nem ugyan tudományos elmélet révén, hanem a karteziánus színház metaforáját mint a gondolkodás eszközét olyan alkalmasabbakra cserélve, mint virtuális gép vagy szoftver (Dennett 1991, 454–455). A metaforáknak ebben a „háborújában" (uo.) próbálom én az utóbbiakat is a karteziánus oldalra állítani.

figyelmeztetés nélkül is tisztában vannak partnereik kilétével vagy mi létével, ám ők maguk azok, akik relativizálják ezt a különbséget azzal, hogy az (ál) érzések *hatása* és nem az ontológiai státusza érdekli őket. Megérthető a nagyobb bizalom egy „gyér" szubjektivitású, mondhatni „együgyű" társ iránt, ha megkapják tőle, amire szükségük van, és még csak nem is lehet azok részéről hamis érzéseket vagy hátsó szándékot gyanítani.

Filozófusok, metafizikusok, fenomenológusok viszont „valóságos" mentális állapotainkat féltik és próbálják őket egyre távolabbi vagy szűkebb, az MI számára elérhetetlennek hitt területekre menekíteni. Ennek egy lehetséges oka, amit a szintén így érző Balogh egy helyen megemlít, hogy talán valamiféle kötelességünk volna nem megtéveszteni önmagunkat, jóllehet ő maga nyitva hagyja e kérdést, nem volt célja morális problémákat tárgyalni (43, 19. lábjegyzet). Mi ellenben a most következő, erkölcsiséggel foglalkozó részben nem hagyhatjuk majd ki a szociális segítő robotjaira való hivatkozást sem.

Mesterséges moralitás

Tekinthetők-e maguk a társadalmi életben egyre nagyobb szerephez jutó MI-rendszerek felelősnek a viselkedésükért, vagy e felelősséget mindig inkább embereknek, intézményeknek kell tulajdonítanunk? Fabio Tollon az autonómia fogalma felől tárgyalja ezt az alapvető kérdést, amely kifejezetten morális problémák esetében válik élessé. Gyakran fordulnak az MI kutatói és filozófusai hagyományos erkölcsfilozófiai elméletekhez, elsősorban az utilitarista és a kanti etikához. Erre nyújt példát Tomislav Bracanović írása is az önvezető autók előzetes etikai beállításáról. Kapelner Zsolt tanulmánya, amely az MI-rendszerek önálló politikai vezetésének lehetőségét és (nem) kívánatos voltát taglalja, már az MI-vel kapcsolatos igazán végsőnek mondható kérdésekhez visz bennünket. Válaszaik megítélése során magam is a kanti morálfilozófiát hívom segítségül, annak tágabb gyakorlati fogalmi keretére is tekintettel.

1. Mesterséges gyakorlati ész

A mesterséges értelemnek már azért is egyik mintája lehet a Kant-féle gyakorlati ész, mert *ész*, azaz a következtetés képessége, tehát logikai tevékenységet végez. Mindig leírható egy szillogizmussal az, ahogyan eljut valamilyen felső tételből, „maximából" ***hipotetikus imperatívuszokon*** (akár ilyenek láncolatán) keresztül egy adott cselekvés akarásához. Ezek az imperatívuszok valójában elméleti tételek alkalmazásai, legyenek azok tudományos vagy hétköznapi „elméletek". Mindenesetre egy már adott célhoz adnak eszközöket: mit *kell* tenni a *cél eléréséhez*. A gyakorlati ész konklúziói tehát nem „van", hanem „kell" ítéletek

– éppen ezért *gyakorlati*. A célt végső soron *maximák* határozzák meg, amelyek mindig emberi szándékokból és érdekekből származó szubjektív szabályok (Kant 2004, 25–26, 33–34).

Ha csupán a hipotetikus imperatívuszok végrehajtását, sőt akár felfedezését tekintjük, akkor az eddig mondottak is tanúsítják, hogy az emberekkel szemben támasztott elvárásoknak egyre inkább megfelelnek MI-rendszerek is. A gépek jobbak nálunk (vagy hamarosan jobbak lesznek), mégpedig már nemcsak szakértői, hanem intellektuális értelemben is. De félő, hogy értelmünkön túl nem hivatkozhatunk kitüntetettségünk igazolásaként olyan sajátosságainkra sem, mint érzelem, intuíció, kreativitás vagy akár a történeti tapasztalat. Ráadásul a gépek az előírt szabályokat pontosan betartják, és nincs jelen náluk az „emberi gyengeség" számtalan formája, vágyak, indulatok, szenvedélyek. Vagy ha minket sokszor éppen ezek ösztönöznek nagyobb teljesítményre, felfedezésre, alkotásra, nos, a gépek ilyesfélék hiányában is mindig állhatatosan, kitartóan, teljes koncentrációval végzik munkájukat. Hiszen pontosan ezért alkottuk őket: hogy nálunk erősebbek és gyorsabbak legyenek, majd pedig ügyesebben és okosabban végezzenek el bizonyos feladatokat.

Mint a korábbiakból kiderült, idesorolhatónak tartom akár a tudományos felfedező, akár a szociálisan érzékeny tevékenységeket. Természetes, hogy idetartozik az autók önvezetése is, amellyel Tomislav Bracanović foglalkozik. Az ő cikke kimondottan etikai vonatkozású, mégis ebbe a moralitás témakörét csak felvezető szakaszba illik, mert a járművek autonómiája *tervezési* szintű. Ez Bracanović probléma-felvetéséből is jól látható: ha kell lennie „etikai beállításuknak" (*ethics setting*), akkor azt a közlekedő *személyek* egyénileg válasszák meg, vagy mindenki számára egységesen az *állam*? Az előzetes etikai beállítás jó példája lehet a kanti értelemben vett maximának is.[11] Az autonóm járművek (AV-k) számára közvetlenül a saját érdekünkben, a *mi* szolgálatunkra írunk elő szabályokat. Vagyis *emberi* célok és érdekek adják a „felső tételeket" a „ne ártsanak" nekünk minimumától a „saját boldogság általános elvéig" (Kant 2004, 28–29). Ám éppen azért, mert mindegyik szóban forgó választás végső soron emberi, Bracanović összes érve visszavezet és érvényes kell, hogy legyen az *emberi* járművezetők etikai „beállítására" is. Nézzük érveit ebből a szempontból!

[11] Habár elég durva és kezdetleges maximák az olyasféle „beállítások", mint hogy „egy (vagy akár több, de ismeretlen) embert föláldozok, ha ezzel a másik hármat (vagy magamat) megmentem". Úgy gondolom, az „élet-halál" szituációkra érvényes etikai beállítások is körültekintőbbek ennél az emberi vezetőknél is, próbálván tekintetbe venni például a közlekedők sérülésének *esélyeit* (mondjuk: autós vagy gyalogos a partner). Az efféle esélyek kalkulálása szintén olyan, ahol az MI támogatásától sokat várhatunk.

A személyes etikai beállítást gyorsan elintézi, leginkább azért, mert az önzésnek ad nyílt teret, amely elfogadhatatlan a kanti és az utilitarista etika számára egyaránt. Az igazán vizsgálódásra érdemes választás az egységesen államilag elrendelt etikai beállítás, amely részrehajlás nélkül osztja el a veszélyeket. Az ellenvetés itt kanti szempontból az volna, hogy az állami döntéshozatal felfüggeszti az egyének autonóm választását épp ezekben a morális tekintetben kritikus szituációkban, és hogy ezúttal nem egy egyén, hanem az állam kezelné puszta eszközként a „feláldozandó" személyeket (52). Az utilitaristát pedig e beállítások „mellékhatásai" kellene, hogy aggasszák, amelyek növelnék az emberekben a félelmet és a bizalmatlanságot. A közlekedésben különösen gyakori dilemmás helyzetekben ők nem láthatják át azt a nagyobb egész szerinti célszerűséget, amelyet az MI-rendszerek kezelnek. Egy kifinomultabb „2.0" változat pedig, amely figyelembe képes venni a résztvevők korát, egészségi állapotát és egyéb tényezőket, az adatkezelés átláthatatlanságához vezet és diszkrimináció, sőt totalitarizmus gyanúját kelti. Amellett az AV-k sikerességét mutató statisztikáknál az emberek inkább törődnek döntésük szabadságával „élet-halál" kérdésekben, valamint „bizonyos morális elvekhez és értékekhez való elkötelezettséggel" (56). Bracanović egyenesen ez utóbbiakkal indokolja, hogy miért azt tartja egyedül elfogadhatónak, ha az AV-ket *semmilyen* etikai beállítással nem látják el: e megoldás nem okoz „morális kárt".

Miért ne volnának érvényesek – önző vagy szigorú szabálykövető – *emberi* autóvezetőkre mindezen emberileg és emberi erkölcsfilozófiák alapján is méltányolható érvek? A válasz már az önzést támogató változat lehetséges fatális következményeinek említésekor felrémlik, de igazán világossá az állami rendelkezések precíz és kérlelhetetlen végrehajtása teszi. Az AV-k Bracanović okfejtéseiben potenciálisan *tökéletes végrehajtói* (ilyen vagy olyan *emberi*) morális elveknek – és *ez* a baj velük. Ezt látva az (autóvezető) ember számára is bármiféle etikai beállítás hiánya volna leginkább ajánlatos, de szerencsére a gyakorlatban etikai elvei sem okoznak súlyosabb „morális károkat", tudniillik azok követésének – és talán egyáltalán: komolyan vételének – *megbízhatatlansága* miatt. Nem így az MI-rendszerek, náluk az erkölcsiség többé nem játék, nem szabad megengedni.

Ha azonban az erkölcsfilozófiák intelmei, amelyekre érveiben Bracanović is támaszkodik, nem vezetnek ennyire negatív konklúzióhoz, és az emberek számára mégis ajánlható, sőt tőlük elvárható, hogy etikai beállításokkal, maximákkal a fejükben üljenek a volánhoz, akkor az érvelés is visszájára fordul. Hiszen ha így van, akkor nemcsak az azok szilárd követését segítő eszközöket kell egyenesen előírni, hanem magukat a beállításokat megválasztó MI alkalmazását is.[12] Ebben az esetben van létjogosultsága a következő szakasznak.

[12] Szerzőnk előveszi az autonóm fegyverek témáját is, mondván, ha ezeket elutasítjuk a szándé-

2. Mesterséges erény és mesterséges politika

Fabio Tollon szerint az autonómia fogalma az MI-rendszereknek csak az általunk való „leírásában" juthatna szerephez, ám inkább csak konfúziókhoz vezet, mert metafizikai terminusokat és vitákat hív elő. Mérnöki értelemben, a tervezés szintjén pedig egyáltalán nem lehet a gépek autonómiájáról beszélni, lévén itt valamennyire mindenképp függenek emberi irányítástól vagy ellenőrzéstől. Emiatt úgy tűnik, nincs értelme a mesterséges rendszereknek saját felelősséget tulajdonítani. (9, 11, 14–15, 18–19)

Természetesnek látszik, hogy ha a maximákat tőlünk kapják, a felelősség a miénk, még ha azt nyomon követni sokszor nehéz is vagy lehetetlen. Így aztán nem is *vonjuk* felelősségre őket, nem kérjük számon hibáikat, nem büntetjük őket, csak módosítjuk vagy átalakítjuk a rendszert. De talán *viselhetnek* felelősséget. Ennek elgondolására az egyik mód a Kapelner Zsolt által leírt Algoritmusok általi Vezetés, valamint belátása arról, hogy azt teljesen emberi felügyelet nélkülinek tarthatjuk, ha „elégségesen független" alkotóitól, ahogyan az emberi döntéshozók is függetlenek már szüleiktől, tanáraiktól (78–79, 83). Itt tehát egy elegendően távolra ható tervezési munka eredményét elfogadjuk autonómnak (nem firtatva a szabadság metafizikai kérdéseit), ugyanúgy, ahogyan általában emberként gondolkodunk önmagunkról.

A másik lehetőség is saját autonómiánkról veszi a mintát, szintén a tervezésre építve, de a lehető legközelebbre hozva azt. A kanti gyakorlati ész ugyanis a *maximáit megválasztó* ész, ennek közvetítésével hat a viselkedésre. Ezt gépektől is várhatjuk, és ha azt is feltehetjük, hogy a döntések legfelső maximája is nekik tulajdonítható, akkor MI-rendszerek kanti értelemben autonóm eszes lények lehetnek. Ehhez egyáltalán nem az szükséges, hogy saját céljaik legyenek, és ez nem is volna elégséges.[13] A kanti etika szerint először is azt jelentené, hogy a gép a tőlünk kapott vagy önállóan alkotott maximát mindig az *erkölcsi törvényben* megfogalmazott univerzalitás követelményével összevetve fogadja vagy vesse el a cselekvés szabályaként. Ennek az elvnek „tesztként" alkalmazása komoly problémákat vet fel akár gépek, akár emberek esetén. Azt viszont bizton állíthatjuk, hogy az előző szakaszban erről *egyáltalán nem* volt szó még az *embert illetően sem*. Az emberi boldogság, emberi érdek szolgálata mint legfőbb maxima még nem jelenti a morálisan jót. Kant szerint autonómiánk

kolatlanul feláldozott életek miatt, akkor nem fogadhatjuk el azt sem, hogy AV-ket szándékosan programozzunk emberéletek közti választásra (58). Ám ha arra jutunk, hogy az AV-k etikai beállítását el kell fogadni, akkor az érvelés itt is visszavág: egyre kevésbé lesz megengedhető, hogy *ne* MI-rendszerek kezeljék a fegyvereket is, amelyek sokkal biztosabban követik a morális szabályokat.

[13] Jellemző az a nézet, amely abban látná a gépek autonómiáját, ha „saját céljaik" volnának, amely nem „esik egybe" a mieinkkel. Az itt szóban forgó autonómia kritériuma azonban egy olyan törvény elfogadása, amely éppen azt követeli meg, hogy bármely célunk mindenki máséval összeférjen.

mércéje magasabban van, mi azonban az MI-rendszerek lehetséges autonómiájáról értekezve a magunkét (egy alacsonyabbra tett lécen üldögélve) többnyire evidensnek vesszük.

Holott az MI tőlünk kapott küldetése akár arra is szólhat, hogy ezt a mércét minálunk jobban megközelítse. A kanti erkölcsiségnek azt a további, az emberi természet számára igen kemény követelményét például, hogy ne pusztán a törvénynek megfelelően, hanem a *törvény kedvéért* cselekedjünk, a gépek a maguk módján nagyon könnyen teljesítik. Itt ugyanis az erkölcsiség mozgatórugójának kérdéséről van szó, hogy az nálunk, érzéki lényeknél ne legyen más, mint „az egyetlen a priori érzés", a morális törvény iránti tisztelet (Kant 2004, 89 skk.). A gép cselekvését nem befolyásolja érzéki mozgatórugó, eleve parancskövetőnek van megalkotva, folyton a neki megszabott szabályokra figyel. A számonkérhetőség, felelősségre vonhatóság, amely után a gépi etikában kutakodunk, csak speciálisan emberi ösztönző, avagy fék. Nem volna funkciója a gépnél, ami ha tudja, mi a kötelessége, biztosan meg is teszi.

Ha a gépek erkölcsi elvek felállításához, megfogalmazásához, ellenőrzéséhez, sőt követésükhöz is alkalmasabbak nálunk, akkor nemigen lehet kétséges, hogy *morálisan is jobbak* lehetnek. Ilyen eszközökként alkottuk és fejlesztjük tovább őket az emberi igények számára: engedelmes és önzetlen lényekként. Az MI sok tekintetben és egyre több területen már ma is az erény példaképének volna tekintendő. Kiváló példát nyújtanak rá a Balogh Zsuzsanna által leírt asszisztens, sőt az emberi ápolóknak a páciensekhez fűződő viszonyát is ellenőrző robotok, aminek egyébként ő maga még a gondolatát is sértőnek és embertelennek látja (38–39). Én ellenben azt mondom: csak remélhetjük, hogy az MI saját képére formál minket, hiszen ez a kép nem más, mint a *mi* – tőlünk telhető világossággal elgondolható és benne megvalósított – *legjobb énünk*.

Aligha lesz ezek után meglepő, ha bevallom, hogy Kapelner Zsolt „extrém" elképzelését a *mesterséges politikai vezetésről* nagy lelkesedéssel olvastam. Tanulmányának különleges érdeme az a gondosság, ahogyan a demokrácia alapértékei melletti érvelés kedvéért elgondolt Algoritmusok általi Vezetés (legyen itt AáV) főbb pontjait meghatározza. Kiemelkedően fontos közülük, hogy nem egy saját érdekekkel, vágyakkal, akarattal rendelkező mesterséges személyt feltételez, ezért is beszél algoritmusok, nem pedig mesterséges intelligencia általi vezetésről, amely inkább kelthetné egy koherens elme képzetét. Okkal hangsúlyozza ezt Kapelner többször is, mert erősebb érvet céloz meg, mint amilyet az egyszemélyi uralomtól való félelem MI-re kivetítése nyújthat. Az AáV-vel rokonszenveznie kellene mindenkinek, aki rendszerint támogatja az intézményi vagy egyéb „mechanizmusok", „automatizmusok" bevezetését célzó javaslatokat a sokszor nem megbízhatónak tartott személyi vezetéssel szemben. Éppen személytelen volta

miatt nincs okunk gyanakodni más, kevéssé átlátható módon érvényesülő elfogultságára sem, így az *egyenlőséget* sem fenyegeti (81–83).

Mi hát a baj a teljesen emberi felügyelet nélküli AáV-vel Kapelner Zsolt szerint? Az, hogy megengedhetetlenül korlátozná a *szabadságunkat*. Nem egymás közti egyéni viszonyainkban, hanem az *uralom*, tehát a törvényhozás és a politikai döntéshozatal fölötti ellenőrzés vonatkozásában, még egy személytelen és igazságos rendszerben is (84–85). Súlytalanná tenné, „elnémítaná" (*mute*) a polgárok akaratát társadalmi kérdésekben, akkor is, ha nem egy „idegen akarat" uralkodna helyettük vagy fölöttük (85–86). Nem volnánk szabad morális cselekvők, ha közös társadalmi életünk alapvető jellegéről nem dönthetnénk, ez pedig elfogadhatatlan. A szabadságot nem cserélhetjük el nagyobb gazdasági hatékonyságért vagy növekedésért, több innovációért vagy bármi másért, amit az AáV politikai döntéshozatala ígér (87). Mert természetesen ilyesmiket ígér, ez az alapja az egész elképzelésnek. Kapelner első feltevésként rögzíti, hogy az AáV emberi döntéshozóknál lényegesen jobb eredményeket produkál (77). A relativista ellenvetéssel szemben pedig a politikai kognitivizmus legalább olyan fokának elfogadására apellál, mely szerint megismerhető „objektív tény", hogy bizonyos politikai döntések jobbak másoknál, például a szabadságot és a prosperitást támogatók azoknál, amelyek a nélkülözést és a zsarnokságot mozdítják elő (80). Az AáV-ről tehát az előbbit feltételezi, de ismét hozzá kell tennünk, hogy szabadságon akkor itt csak az egyének kölcsönviszonyában felmerülő jogok garantálását értheti. *Uralom* alóli szabadságot a polgárok számára nem adhat szerinte az AáV.

De miért nem? A demokráciát keresztül-kasul áthatják mechanizmusok, automatizmusok, olyanok is, amelyek biztosítani hivatottak, hogy az emberek akarata igenis számítson (86), közülük is kiemelhetjük Kapelnerrel, mint alapvetőt, a közvetlen szavazás intézményét (76). Mármost ha valami, akkor az AáV nemcsak az ilyenek működésének, de kiterjesztésüknek és gazdagításuknak is biztosítéka lehet. A voks ugyan közvetlenül a miénk, de a közvetítéseket, melyek révén hozzájárul a politikai döntésekhez, a Kapelner által leírt algoritmusoknál jobb kezekbe alig adhatnánk (fokozatosan persze: 77). Bár szabad morális döntéshozó mivoltunk ugyanúgy *általános alapelv* lehet a politikában, mint minden emberi élet megóvása a közlekedésben, azt nemigen hihetjük, hogy bármiféle *tényleges* szavazat intelligensebb döntés volna, mint egy „1 vs. 3 halott gyalogos" típusú „etikai beállítás". Ezért ha valaki hisz abban, hogy a demokrácia garantálja politikai döntésképességünk meglétét vagy növekedését, akkor bizalommal fogadhatja az AáV lehetőségét, és vele (a szavazások eredményeiben is megnyilvánuló) döntéseink *összehangolt* érvényesítését. „Uralmában" nem kellene semmi mást látnunk, mint a törvényhozás és a politika fölötti *közös* ellenőrzésünk biztosítékát, legyen szó a társadalmi életünket alapvetően meghatározó döntéseinkről, azok súlyozásáról, adott esetben „elnémí-

tásáról". Nem valamely személynek vagy bármelyikünknek ítéletéhez, hanem egy általános morális törvényhez igazodva természetesen – azaz mesterségesen.

Befejezés?

A szerzőink gondosan kifejtett aggodalmainak rendre ellentmondó nézetemmel úgy tűnik, teljesen semmibe veszem az MI-vel kapcsolatos félelmeket. Könnyedén lemondok emberi mivoltunk kitüntetettségének olyan jegyeiről, mint a fenomenális szféra, a kreativitás, az autonómia, miközben az MI kezébe adnám át a jólétünkről való gondoskodásnak és a közösségi életünk vezetésének felelősségét minden szinten. Látszólagos árulásom csak erősíti a balsejtelmet, hogy az emberiségnek szembe kell néznie egy nála hatalmasabb és magasabb rendű létező megjelenésével az MI révén. Talán még saját váratlanul közeli végével is.

E félelmek legsötétebb fajtája szerint az MI egy bizonyos időpontban majd „átveszi az uralmat" az ember fölött és „leigáz" bennünket. Ilyesmit feltételezni a mesterséges lényekről, amikor azok mindenhol és mindenben az ember céljait és érdekeit szolgálják az őket fejlesztő szakemberek jótékony tevékenységének eredményeképpen, ha nem csupán rémálmaink kivetítése, akkor a rossz lelkiismeretünké. Képességeik és hatékonyságuk tudata azért táplálja ezt a szorongást, mert (még) *idegennek* tekintjük saját, önnön érdekünkben és céljainkra létrehozott, minket felülmúló teremtményeinket.

Reálisabb az az aggodalom, hogy MI-rendszerek morálisan rossz *emberi* szándékok vagy egymással ellenséges emberi törekvések eszközei lesznek. Ilyen értelemben veti fel Neumann János is a kérdést: „Túlélhetjük-e a technikát?", 1955-ben elsőként a nukleáris fegyverkezés lehetséges következményeire gondolva, de tárgyalva már az automatizálás jelentette lehetőségeket és veszélyeket is. Az MI mai stádiuma ez utóbbiak minőségi ugrását sejteti, ám egyben új megvilágításba helyezi az egyetlen „gyógymódot" is, melyet Neumann a fejlődés gyorsasága ellen kínál. A napi „megalkuvó intézkedések, az apró, korrekt döntések hosszú láncolata" és „intelligens végrehajtás[uk]" tipikusan olyanok, amelyekre az MI-rendszerek alkalmasabbak az embereknél sok tekintetben már ma is. Mi pedig, ahogy az itt tárgyalt tanulmányok is tanúsítják, képesek vagyunk arra, hogy e feladatokhoz továbbadjuk számukra a túlélésünkhöz egykor Neumann által megjelölt emberi tulajdonságok – türelem, rugalmasság, intelligencia – *mesterséges* változatait (Neumann 2003, 366–368).

Való igaz, az MI a legfontosabb előfeltevéseinkkel és filozófiai kérdéseinkkel szembesít bennünket. Az általam vallott nézet az algoritmusokban a „testet-

lent", helyesebben a bennük megtestesült, általuk megvalósított értelmes *gondolatot* látja. Intelligencián pedig nem bizonyos képességek komplexumát érti, hanem azt az *általános értelmet*, amely Descartes és Kant eszméi szerint minden programozó minden gondolatát, sőt a turingi absztrakció nyomán minden algoritmus végrehajtását és minden számítógép működését is vezérli. Hasonlóképpen hiszem, hogy az MI minden egyes fejlesztője felelősségteljes munkája során a benne ható *erkölcsi törvényt* követi (még ha nem is gondol ilyesmire). Ezért is vesznek körül bennünket – egyre táguló és szűkülő körökben – tevékenységük áldásos eredményei, a szabályokat szigorú következetességgel végrehajtó gépek, a törvénykövetés és a kötelességteljesítés mintaképei. Természetesen korlátozott fizikai teremtmények ők is, nem angyalok, de minden látszat szerint a miénknél jobb anyagban művelik azt, ami bennünk e mennyei lényekkel közös: a gondolkodást és a jóakaratú szolgálatot.

Mások az algoritmus fogalmában az öntudatlan ismétlődések *sokaságát* üdvözlik, ekként ajánlják, hogy ismerjünk változataikban magunkra és a világra. A mesterséges NN a rejtett rétegeivel különösen is kedvez elképzelésüknek. Gyanítom, ők szörnyülködve tekintenek víziómra, mint valamiféle rezervátum földi paradicsomába elhelyezett, ártalmatlanná és engedelmessé domesztikált emberiségére. Tőlük is számíthatnak azonban vigaszra azok, akik az MI térnyerését folyamatosan zajló „hatalomátvételnek" érzik. Érzelmek, indulatok, szenvedélyek sikeres szimulálása, de még az emberi életet letapogató öntanuló rendszerekben megújuló előítéleteink is, azt a rafinált reményt kelthetik, hogy a mesterséges értelem a *mesterséges szenvedélyek szolgálójává* válhat, ami nagyon is kívánatos, ha már emberi közösségként is azt véljük a legjobbnak, ahol az egyének mindenekelőtt vágyaikat követik. A mesterséges morál e másik, hume-i (vagy dennetti) útján a gépek nemcsak öntudatosak, hanem még inkább *önérzetesek* lesznek. Mentális állapotaikat (a mieink másolatait) nemcsak emberi érdekek szolgálatában fejlesztik, hanem önértéket tulajdonítanak nekik vagy saját céljaikhoz használják. Mert lesznek saját céljaik is, így „igazi" autonómiával bírván nem ragaszkodnak majd mereven etikai elvekhez például a közlekedésben sem. Politikai szereplőként kiállnak nemcsak saját (immár „felismert") jogaikért, de mindenki szabadságáért is, ezért távol tartják magukat (és másokat is) a vezető szereptől. Eléggé olyanok lesznek, mint mi: tele érzésekkel és gyarlósággal, de kritikai érzékkel és szabadságvággyal is. Így aztán nem kellemetlen elgondolni, hogy velük lelki és testi szimbiózisban élünk majd (egy ideig). És néha-néha még sakkban is legyőzhetjük őket.

Irodalom

Balogh Zsuzsanna 2019. Intersubjectivity and Socially Assistive Robots. *Magyar Filozófiai Szemle*. 63. 4., 25–45.

Bracanović, Tomislav 2019. No Ethics Settings for Autonomous Vehicles. *Magyar Filozófiai Szemle*. 63. 4., 47–60.

Campbell, Murray – Hoane, A. Joseph – Hsu, Feng-hsiung 2002. *Deep Blue*. *Artificial Intelligence*. 134. 1–2., 57–83.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York – Oxford, Oxford University Press.

Dennett, Daniel C. 1998. *Darwin veszélyes ideája*. Ford. Kampis György – Kavetzky Péter. Budapest, Typotex.

Dennett, Daniel C. 1991. *Consciousness Explained*. New York, Little, Brown and Company.

Descartes, René 1992. *Értekezés a módszerről.* Ford. Szemere Samu – Boros Gábor. Dabas, IKON Kiadó.

Descartes, René 1996. *A filozófia alapelvei.* Ford. Dékány András. Budapest, Osiris.

Descartes, René 1986 (= AT XI). Œuvres *XI.* Szerk. Adam, Charles – Tannery, Paul. Paris, J. Vrin.

Dreyfus, Hubert L. 1965. *Alchemy and Artificial Intelligence*. Santa Monica, California, The Rand Corporation.

Hoffmann Miklós 2019. Science as a Human Vocation and the Limitations of AI-Based Scientific Discovery. *Magyar Filozófiai Szemle*. 63. 4., 61–74.

Horváth Zoltán 2013. Mesterséges gondolkodás és emberi értelem. *Magyar Filozófiai Szemle*. 57. 3., 50–73.

Hsu, Feng-hsiung 1999. *IBM's Deep Blue Chess Grandmaster Chips*. *IEEE Micro*. 19. 2., 70–81.

Kant, Immanuel 2004. *A gyakorlati ész kritikája*. Ford. Papp Zoltán. Budapest, Osiris.

Kapelner Zsolt 2019. Why not Rule by Algorithms? *Magyar Filozófiai Szemle*. 63. 4., 75–90.

Lample, Guillaume – Charton, François 2019. Deep Learning for Symbolic

Mathematics. https://arxiv.org/pdf/1912.01412.pdf

Neumann János 2003. *Válogatott írásai.* Vál. Ropolyi László. Ford. Tarján Rezsőné, Augusztinovics Mária, Szalai Sándor. Budapest, Typotex.

Penrose, Roger 1993. *A császár új elméje. Számítógépek, gondolkodás és a fizika törvényei*. Ford. Gálfi László. Budapest, Akadémiai Kiadó.

Searle, John. R. 1980. Minds, brains, and programs. Behavioral and Brain Sciences 3 (3), 417–457.

Shannon, Claude E. 1950. Programming a Computer for Playing Chess. *Philosophical Magazine*. Ser. 7. 41. 314., 256–275.

Silver, David et al. 2017. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. https://arxiv.org/pdf/1712.01815.pdf

Silver, David – Hubert, Thomas – Schrittwieser, Julian – Hassabis, Demis 2018. AlphaZero: Shedding new light on chess, shogi, and Go.

https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

Simon, Herbert A. 1980. Computers – *Non-numerical* computation. *www.pnas.org/content/77/11/6264.full.pdf*

Tollon, Fabio 2019. Moral Agents or Mindless Machines? A Critical Appraisal of Agency in Artificial Systems. *Magyar Filozófiai Szemle*. 63. 4., 9–23.

Turing, Alan M. 1936. On Computable Numbers, with an Application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, vol. s2-42, 230–265.

Turing, Alan M. 1965. Számológépek és gondolkozás. Ford. Tarján Rezsőné. In: Szalai Sándor (szerk.): *A kibernetika klasszikusai*. Budapest, Gondolat Kiadó, 120–160.