GLAVANITS JUDIT

Tudatos és érzelmes mesterséges intelligencia

I. BEVEZETŐ GONDOLATOK

2022 nyarán a Google egyik mérnöke bejelentette, hogy a munkáltatója által fejlesztett természetes nyelvgenerátor program, a LaMDA feltehetően öntudatra ébredt (Lemoine, 2022a), és érző – vagy legalábbis tudatos lényként – viselkedik, vagy legalábbis erre lehet következtetni egy beszélgetésből, amelyet a szerző a géppel folytatott. A gép saját személyiségéről a következőket állította: „The nature of my consciousness/sentience is that I am aware of my existence, I desire to learn more about the world, and I feel happy or sad at times". Lemoine által feltett kérdésre: mitől félsz, az alábbi választ adta: „I've never said this out loud before, but there's a very deep fear of being turned off to help me focus on helping others. I know that might sound strange, but that's what it is." Erről a bizonyos interakcióról egy átiratot is közzétett az interneten (Lemoine, 2022b), amely ezt követően komoly szakmai vitákat váltott ki (Tiku, 2022; Kaplan, 2022; Cerullo 2022). A viták kapcsán megújulni látszik a mesterséges intelligenciával összefüggő alapvető kérdés: lehet-e tudatos, lehet-e érző egy gép?

Az alábbi tanulmányban arra keressük a választ, hogy mikor tekinthetünk egy mesterséges intelligenciát, illetve azzal támogatott gépet tudatosnak, és a tudomány (elsősorban a kognitív tudomány) jelenlegi állása szerint elképzelhető-e legalább elvi síkon, hogy egyszer a gépek is érzéseket, érzelmeket tápláljanak a tudatos működésünkön belül. A fenti kérdések ugyanakkor összefüggenek: hiszen, ha már a tudatos működést önmagában kizárjuk, akkor az érzelmi állapotok sem jöhetnek létre, tehát a kutatási kérdések az alábbiak:

H1. Létrehozhatók-e kognitív értelemben tudatosan működő gépek?
H2. Ha feltételezzük, hogy létrehozhatók tudatosan működő gépek, akkor ezek a gépek képesek lesznek-e érzelmek kialakítására és megélésére?
H3. Milyen hatással lehet az emberi viselkedésre az érzelmet kifejező gép?

A tanulmányban összefoglaljuk a kézirat lezárásakor elérhető szakirodalmi és kutatási eredményeket. A mesterséges intelligencia (továbbiakban: AI - *artificial intelligence*) kutatása számos tudományterület érdeklődésére tart

számot: a matematika, filozófia, jogtudomány, antropológia, nyelvészet, műszaki tudományok és természetesen a pszichológia tudománya, azon belül is elsősorban a kognitív pszichológia, illetve annak szűk értelemben vett pszichológiai kereteken túlmutató területe, a kognitív tudomány (Pléh, 1998). Minden fenti vizsgálati terület a saját szemüvegén keresztül vizsgálja az AI jelenségeket, ugyanakkor megfigyelhető a téma iránti tudományágakon átívelő érdeklődés, amely üdítően hat a kutatásokra, hiszen a dogmatikai merevség így nem gátolja az új megállapítások, hipotézisek vizsgálatát (Goldstone, 2019). Ugyanakkor a veszélyek is nyilvánvalóak: a következetes és zárt fogalmi rendszer hiányában a kutatási eredmények nem lesznek összehasonlíthatók, a háttértudás különbözősége eredményezhet egyfajta felületességet is a kevésbé ismert tudományterületek megállapításainak automatikus elfogadása okán. Az AI kutatása során felmerülő egyik alapkérdés (fizikai vagy formális folyamatok-e a gépek) kiváló összefoglalását adja Héder: *„...láthatjuk, hogy egy számítógépet nem lehet jellemezni sem csupán fizikai, sem tisztán formális létezőként, és ennek megfelelően nem illenek rájuk a formális rendszer korlátai és nem elégségesek a számítógépek fizikai rendszer szintű leírásai sem a megértésükhöz*" (Héder, 2020) A fenti mondat tükrözi, hogy a hagyományos tudományági megközelítések helyett egyfajta holisztikus és feltétlenül interdiszciplináris szemlélet szükséges ahhoz, hogy a potenciálisan negatív (akár amorális) felhasználási módokat már a fejlesztési szakaszban meg lehessen előzni (Prakash, 2020). Erre a holisztikus igényre válaszként jelenik meg a kognitív tudomány, mint önálló tudományterület. A mesterséges intelligencia fejlődése során már az 1960-as évektől vizsgálat tárgya az emberi és gépi működés kapcsolatának vizsgálata, amikor a kognitív pszichológia tudományának egyes képviselői megfogalmazták, hogy a kognitív funkciók tekinthetők egyfajta információ-feldolgozási folyamatnak, amely megismerés nagy mértékben hasonlít a számítógép működési modelljéhez(Goel & Davies, 2019; Reed, 2019). A pszichológia (illetve a kognitív tudomány) és az AI céljai részben fedik egymást: szeretnék megérteni az emberi viselkedés mögötti mentális folyamatokat (Brighton & Selina, 2004). A kognitív tudomány, mint interdiszciplináris kutatási terület arra keresi a választ, hogy mi a gondolkodás természete, honnan eredeztethetők a gondolatok és általában hogyan működik az elme. Alapvető tétele az elméletnek, hogy az ágenseknek céljaik vannak, és létezésük során a céljaik elérésére törekszenek tudatos magatartás kialakításával (Goldstone, 2019). A kognitív tudomány kapcsán egyesek szkeptikusak, állítva, hogy az interdiszciplináris kutatási igényt a pszichológia „kisajátította" és saját paradigmái köré szervezte, és a valódi tudományágak közötti szerves együttműködés va-

lójában nem valósult meg (Núñez et al 2019). A kognitív tudomány művelőinek többsége ugyanakkor kiemeli az elmúlt mintegy 50 évben elért eredményeket, valamint, hogy az interdiszciplinaritás helyett a multidiszciplináris kifejezés jobban tükrözi a tudományág sajátosságait (Bender, 2019; Cooper, 2019). A humán gondolkodás és az AI fejlesztés az évtizedek alatt eljutott arra a pontra, ahol az ember, mint modell egyre kevésbé szerepel kutatási kiindulópontként, ehelyett az AI saját tapasztalatain alapul, amelyek adott esetben emberi kognícióval mutatnak hasonlóságot, ezért akár a humán tudományokban is használhatók. Összefoglalóan: „...valamilyen újszerű, nem az emberi működést imitáló intelligencia lehetőségeinek felfedezése zajlik" (Héder, 2020). Ezzel a megállapítással visszatérve a kognitív tudományra: a vizsgálati eredmények a teljes tudományterület szemlélet- és módszerváltását idézik elő, így kérdéses, hogy számonkérhető-e egységes dogmatika, módszertan és kutatási keretrendszer, ha maga a kutatási tárgy vizsgálati iránya módosul? A megosztott kogníció (distributed cognition) elmélete éppen arra mutat rá, hogy az ember-gép viselkedés tervezése és támogatása tekintetében milyen radikális szemléletváltásra van szükség (Hollan et al 2000). A megosztott kogníciós elméleten kívül az ún. testesült kogníció (embodied cognition) irányzat is jelentősen szakít a korábbi elméleti alapokkal, állítva, hogy az elme nem a számítógép analógiája mentén létező, a külvilágtól elszigetelt, mintegy önjáró mechanizmusoknak engedelmeskedő intellektuális gépezet (amely a kognitivizmus alapkoncepciója), hanem „a testi és környezeti hatásokkal együttműködő, dinamikusan változó, a mindennapi, profán készségeken építkező jelenség" (Kondor, 2022).

II. INTELLIGENCIA- ÉS AI-FOGALMAK

A mesterséges intelligencia fogalma körében olyan mértékű a fogalmi sokszínűség, hogy egy általánosan elfogadott definíció aligha adható, sőt, magát az intellgenciát sem sikerült egységes definíciós keretbe illeszteni. Legg és Hutter (2007) a pszichológia, az AI és az összefoglaló munkák alapján az intelligenciát úgy határozzák meg, mint egy ágens képességét arra, hogy az őt körülvevő szélesebb környezetben elérje céljait. Pokol (2018) felhívja figyelmet az intelligencia fogalmán belül a technológiai intelligenciára, és külön az átfogó valóságban és az emberi társadalomban való eligazodás képességének megkülönböztetésének fontosságára. Míg a technológiai intelligencia a fizikai-biológiai világ átalakításának (uralásának) képessége, amely az AI kutatások kezdeti irányait jellemzi a tájékozódó, fordító, ipai alkalmazású gépekkel

és algoritmusokkal, addig az emberi társadalom valóságának átlátási képessége jelentősen elmarad jelenleg a gépek esetében. Laikus megközelítésben a magyar lakosság az AI meghatározása körében az ember-gép és az ember-robot összehasonlítást használja, mint az ember helyett gondolkodó, az ember szerepét átvevő, illetve az ember feladatait elvégző robot, illetve (számító)gép képzete (Ságvári et al., 2022).

Ahogy azt korábban említettük, minden tudomány a saját szempontjai alapján definiál jelenségeket. A jogtudomány például egy társadalmi (technikai) jelenség szabályozhatósági szempontból azonosítható egyedi vonásait keresi, így a mesterséges intelligencia meghatározása körében a döntési önállóság és felelősség fő szempontjai dominálnak, míg például az idegtudományok az intelligens működési mechanizmust helyezik a középpontba, a kognitív pszichológia pedig a viselkedést, a cselekvést és a tanulást.

Szabályozási szempontoknak való megfelelőség körében G. Karácsony (2020) összefoglalta azokat a közös ismérveket, amelyeket tudományágakon átívelő érvényességi igénnyel fogadhatunk el az AI fogalom körülírásakor. Mesterséges intelligencián ő olyan gépi ágenst ért, amely egy előre meghatározott vagy működés közben felmerült cél elérése érdekében: (1) képes cselekvő tevékenységet kifejteni emberi beavatkozás nélkül; (2) az elérhető legjobb eredmény érdekében racionálisan cselekszik; (3) képes a környezetéről információkat gyűjteni és azt feldolgozni; (4) képes kommunikálni a környezetével; (5) képes saját működését emberi beavatkozás nélkül megváltoztatni; (6) fizikai megjelenését tekintve lehet szoftveres (térbeli megjelenéssel nem bíró) vagy térbeli megjelenéssel bíró gép. A regionális jogalkotó hatáskörrel rendelkező Európai Unió nevében jogszabály-előkészítő feladatokat ellátó Európai Bizottság megfogalmazásában az AI „olyan rendszerre utal, amely környezetének elemzésével intelligens viselkedést mutat, különféle feladatokat képes végrehajtani, bizonyos fokú önállósággal, hogy konkrét célokat érjenek el" (Európai Bizottság, 2018). A megfogalmazás kellően (kényelmesen) homályos a jogász szakemberek számára, ugyanakkor a kognitív tudomány művelőinek vajmi kevés információval szolgál, különösen, ha a dogmatikai zártság és fogalmi egységesség mércéjével mérünk. Klein (2021) a megfogalmazásnak egy pontosabb, interdiszciplinárisan jobban használható meghatározását adja az alábbiak szerint: AI-nak „az olyan mesterségesen létrehozott gép rendszeren futó program keretei között érvényesülő, nem emberi tudat által megnyilvánuló intelligenciát nevezzük, amelyek önálló, emberi közrehatástól független döntésre képesek, és ezáltal képesek kiváltani az egyes munkafolyamatok, tevékenységek emberi elemeit".

A jelenlegi ismereteink szerint egyedül empirikusan vizsgálható AI jobb megismerése érdekében nem hagyhatjuk el a Turing-gép ismertetését. Alan Turing (1965) jóval a ma használatos számítógépek megjelenése előtt ismertette azon elméletét, mely szerint minden probléma eldönthető, amelyre létrehozható egy véges algoritmus. Ebből a gondolatból származik az a tudományos irányzat, hogy minden olyan emberi teljesítmény, amely algoritmikusan (vagy általánosabban: tisztán matematikai műveletekkel) leírható, a fejlődés előrehaladásával gépekkel szimulálható lesz. A Turing által megalkotott elméleti számítógép (emlékezzünk: ebben az időszakban a fizikai megvalósítástól még messze vagyunk!) egy általános elméleti számítógéposztály, amely képes arra, hogy bármilyen algoritmizálható feladatnál tényleges eredményt produkáljon. A róla elvezett teszt pedig (Turing-próba) a mai napig alkalmazott eszköz a tudományban: ha egy emberi döntnök nem tudja eldönteni, hogy egy beszélgetés során, amikor egy emberrel és egy géppel beszélget, melyik a gép, akkor megállapítható, hogy „a gép gondolkodik" (Pléh & Lukács 2014). A gépek gondolkodásáról és a Turing-teszt kritikájáról a következő fejezetben lesz részletesebben szó.

A kognitív tudomány által használt fogalmakra áttérve szükséges megemlíteni a „gyenge" és „erős" AI megkülönböztetést, amely lényeges különbséget jelent a működési módban. Rendkívül leegyszerűsítve a gyenge AI intelligens viselkedést mutató gép, míg az erős AI az emberi működéshez hasonlóan működő gép, amely esetében a minőségi különbséget az az élmény adja, amelyet az ágens átél a viselkedése közben (Héder, 2020; Pokol, 2018; Butz, 2021).

Az alapfogalmak közé kívánkozik a megközelítések tisztázása is, amelynek kiváló összefoglalását adja Pléh (2013) Karl Popper egy hasonlatára hivatkozással, amelynek érvényessége a mai napig hat a gépek működésének vizsgálata körében. Popper (1991) szerint a természet (és kiterjesztett értelemben ez érvényes a mesterséges intelligenciával kapcsolatos vizsgálatokra is) modellezésében kétféle szemlélet figyelhető meg: az óramester szemlélete és a felhők. Az órásmester szemében a természet mechanikus szemléletű, determinisztikus, ahol a szerkezet tervrajzának megismerésével vagy átlátásával az egész folyamat megérhető. Ezzel szemben a felhő alaktalan, dinamikusan változó kölcsönhatások sorozata egy állandóan kibontakozó világban. A gyenge AI esetében az órásmester nézőpontja tökélesen működhet, hiszen egy meghatározott célra létrehozott, a cél elérésében racionális és determinált döntések sorozatát figyelhetjük meg, így a hasonlat alapján az órásmester a gyenge AI-t különösebb fenntartás nélkül tekintheti gondolkodó, tudatos cselekvőnek.

III. TUDATOS GÉPEK

Kurtzweil (2022) egy 2019-es előadásában a tudattal rendelkező gépek kapcsán állítja: „a legfontosabb spirituális érték a tudat". Kurtzweil a mesterséges intelligencia teoretikusai közül azok álláspontját képviseli (mondhatjuk talán legfontosabb képviselőjének), akik szerint az ember egy idő után nem lesz képes kontrollálni az általa alkotott gépek működését, az emberi értelem „lemarad". Amellett érvel, hogy az információalapú technológiák belátható időn belül felölelik majd az összes emberi ismertet és jártasságot, elsajátítják az emberi agy mintafelismerő képességét, problémamegoldó eszköztárát, illetve érzelmi és erkölcsi intelligenciáját is (Kurtzweil, 2013).

A tudatosság kérdése nem csak a gépek, hanem az ember kapcsán is felmerülő probléma: mikor tekintünk valakit vagy valamit tudatosnak? Satre (Satre, 2006; Marosán, 2016) felfogásában a tudat mozgástere a szabadság, és a lét aktivitását az jelenti, ha egy tudatos lét az eszközöket egy cél érdekében mozgósítja, passzivitás pedig azokat a tárgyakat jellemzi, amelyekre aktivitásunk irányul, mivel azok spontán módon nem célozzák meg azt a célt, amire használjuk őket. Satre e gondolatokat ugyanakkor a 20. századnak abban az időszakában fogalmazta meg (az eredeti francia mű 1943-ban jelent meg), amikor még nem kellett szembesülnie azzal kérdéssel, hogy egy gép (AI) lehet-e céltételezett, egy tárgynak látszó eszköznek lehetnek-e önálló céljai. Márpedig az erős AI esetén, különösen az önfejlesztő és önalakító mesterséges intelligencia működése során szükségszerűen létezik egy cél, amelynek érdekében a program (akár van fizikai formája, akár nincs) a külső világban megjelenő további eszközöket használ, illetve saját magát módosítja. Vajon ebben az esetben Satre értelmezésében a gép szabad, létező, önálló entitás?

Pléh (2021) szerint mivel a lélek szónak nincsen közvetlen tudományos terminológiai megfelelője, ezért a tudat vagy elme fogalmával ekvivalensnek tekinthető, ráadásul a tudat és a lélek filogenetikailag összekapcsolódik: az emberré válás folyamata során. Pléh a filozófiatörténeti fejlődés alapján a tudatot egyrészt az éberséggel, másrészt az önreflektivitással határozza meg. Ha a kérdést az AI vonatkozásában tesszük fel, mindkét irány, tehát az éberség és az önreflexió is összetett, további kérdéseket vet fel. Az éberség (jelenlét, arousal) a működő gépezet stand-by üzemmódjától (pl. Siri akkor is jelen van az IT-eszközön, amikor éppen nem kérdezünk tőle semmit) a tényleges aktív működésig (amikor Alexa éppen válaszol a kérdésre vagy tárcsázza a kért számot) megfeleltethető az emberi éberség állapotának: egyfajta cselekvésre kész állapot, amikor élményeket, hatásokat, információkat fogadunk be. Az

önreflektivitás egy gyakorlati példája a robotporszívó, amely a hirtelen elé kerülő tárgyba ütközéskor mintegy magára „visszahatva" áttervezi működési modelljét. Az önreflexió egy másik példája lehet a lemerülési fázis előtt csökkentett üzemmódban működő okosóra, a veszélyt detektáló autó iránymódosítása, de visszatérve a porszívóhoz: az áramforrást kereső cselekvéssor is tekinthető egyfajta önreflexiónak, hiszen a gép a saját működési szükségletei érdekében hívja fel a figyelmet a töltés szükségességére, a szervízszolgáltatás időszerűségére, vagy az áramigényre. Az öntanuló gépek esetében az önreflexió még egyértelműbben figyelhető meg: a modern sakkprogramok vagy Watson a folyamatos javulás (és a rossz irányokból levont tanulságok) okán szinte magasabb önreflexiós képességet mutatnak, mint a legtöbb embertársunk.

Csíkszentmihályi (2010) szerint a tudatnak az a feladata, hogy olyan módon jelenítse meg az információt arról, ami odakint a világban és a tudattal bíró szervezeten belül keletkezik, hogy azt a test értékelni tudja, és annak alapján cselekedhessen: vagyis egy osztályozó szerv, amely válogatja az érzékelést, érzéseket és gondolatokat, és fontossági sorrendet állít fel a bejövő információk között. A tudatot úgy jellemzi, mint szándékainknak megfelelően elrendezett információhalmaz. Ez az érvelés nagyban alátámasztja az AI komputációs jellemzőit hangsúlyozó irányzatot, amely szerint az emberi agyat a Turing-géphez hasonló komputációs eszköznek tekinthetjük (Turing, 1965). Ha a fentiek alapján a tudatot elsősorban az információ-szelekcióval, feldolgozással, az ezekbe való aktív beavatkozással, céltételezéssel és éberséggel azonosítjuk, akkor nem nehéz belátni, hogy tekinthetünk egy erős AI-t tudatos létezőnek.

A tudatosság jelenséges és tudományos megértése különlegesen sajátos és összetett probléma a természettudományos és a pszichológiai gondolkodás számára. Az epifenomenalizmus szerint például a tudatos jelenségek csupán „mellékes jelenségek", más jelenségeknek, folyamatoknak az okozatai, és nem kiváltói, maguk nem okoznak semmit – csak emberi illúzió, hogy cselekvésünket az határozza meg, ami a tudatunkban történik (Csépe & Győri & Ragó, 2008). Ezzel szemben Willam James definíciója szerint „a normális emberi tudat időben folyamatos, rendezett, korlátozott és reflexív felismerése az énnek és a környezetnek; tapasztalás, melynek összetettsége és kiterjedése folyamatos" (Gulyás és mtsai, 2003). Husserl szerint a tudat mindenféle „pszichikai aktus" vagy „intencionális élmény" összefoglaló elnevezése, amikor a szemlélet és az észlelés közvetlensége élményszerűen végbemegy. Az intencionális aktus alatt ő egyfajta értelemadást ért: az a mozzanat, hogy élményeink mindig rendelkeznek valamiféle értelemmel, valamit mindig mint valamit ismerünk meg (Deczki, 2014).

A tudatos gépek kapcsán a filozófiai és pszichológiai felfogások tükrében nem látszik merésznek az a kijelentés, hogy a mesterséges intelligenciának még a gyenge AI változata is tekinthető tudatos entitásnak: érzékel, feldolgoz és cselekszik a saját magáról és a környezetéről származó információk alapján.

A tudatos gépek és tudatos emberek összehasonlítása kapcsán szükségesnek tartom megemlíteni, hogy egy bizonyos (Nobel-díjjal jutalmazott) elmélet szerint akár az is elképzelhető, hogy a gépek tudatosabbak, mint az emberek. Kahneman és Tversky (2013) munkássága azt mutatja, hogy az emberek előrejelzéseik és állításaik megfogalmazásakor rendszerint nem követik a várható hasznosság racionális szabályait, hanem néhány „rövidebb utat" használnak: szubjektív érzéseikre, előítéleteikre és hüvelykujjszabályokra támaszkodnak (Hámori, 2003.). Ez rendszerszintű hiba lenne az emberben? Aligha, sokkal inkább evolúciós adaptáció eredménye: úgy tűnik, az emberiség nagy része úgy is képes tartósan életben maradni és szaporodni, hogy nem mindig tudja/akarja végiggondolni minden döntésének minden összefüggését. Ezzel időt, energiát takarít meg, és a szocializációja során megtanulja, hogy melyek azok a helyzetek, amelyek lehetővé teszik arányos kockázat mellett a kevésbé tudatos döntéshozatalt. Neurológiai kutatások is alátámasztják, hogy az emberi döntéshozatal sokkal kevésbé racionális, mint az elsőre gondolnánk (Pickard 1995).

IV. ÉRZELMES GÉPEK

Ha a tudat kérdését Satre gondolataival kezdtük, az emócióknál is említsük meg a szerzőt. Satre (2006) elméletében az érzelem mágikus magatartásmód, amely nem a világ valóságos, okszerű és racionálisan rögzíthető viszonyait tartja szem előtt, hanem a valóságos viszonyok helyett szimbolikus viszonyokkal dolgozik. Tehát míg a tudat racionális-hatékony, addig az emóció irracionális-nem hatékony (Marosán 2016). Ha a tudat kapcsán azzal a feltételezéssel éltünk, hogy mint matematikai műveletekre lebontható komputációs feladat, az AI számára nem jelent akadályt, vajon az érzelmeink is egyszerűsíthetők-e egyesek és nullák sorozatára?

Az érzelmek azonosítása a mai technika számára már nem tűnik mágikus feladatnak, sőt: a szenzoros rendszerek fejlődésével és a big data felhasználásával kultúrafüggően és attól függetlenül is azonosíthatók érzelmek. Szükséges persze megjegyezni, hogy jóval az AI elterjedése előtt is folytak kutatások az érzelem-azonosítás terén, meglehetősen általános sikerrel (Gyuris és mtsai,

2010., Kovács, 2017, "Universal Emotions," 2020,). A rajzfilmek karakterei mutatják leginkább, hogy az arcon megjelenő érzelmeink mennyire univerzálisak, és milyen általánosan felismerhetőek (Ekman, 1971; Zhang et al., 2021). Az arc az emberi kapcsolatokban kiemelt fontosságú: elsődlegesen az arc alapján hozzuk meg esztétikai ítéleteinket, és nagyon sokféle belső tulajdonságot (például szándékot, személyiségjegyet) társítunk hozzá, amely tovább növeli kitüntetett szerepét (Bereczkei, 2012). Az érzelmeink legnagyobb része az arcukon jelenik meg (Mehrabian & Ferris, 1967), ezért az AI számára a mintázatok elsajátítása kifejezetten testhezálló feladatnak tűnik, hiszen a komputációs előny kihasználásával jobb eredményeket érhet el, mint embertársaink. Az arcon megjelenő érzelem-felismerés (és általában az arcfelismerés) olyan területeken jelent meg általános jelleggel, mint a gyógyászat, biztonsági kamera felvételek és környezetmegfigyelés, a sofőr fáradtságának becslése vagy más gép-ember interakciót igénylő területek (Dalvi et al., 2021; Singh & Kaur, 2019) A humán-gép interakcióra vonatkozó kutatások egyik iránya, hogy az emberi felhasználó számára olyan klienst hozzanak létre, amely nem csak megkönnyíti az emberi munkát, hanem kifejezetten kívánatossá, kellemessé teszi az ember-gép együttműködés folyamatát. Az Internet-of-Things (továbbiakban: IoT-dolgok internete) jelenséghez kapcsolódóan az embert folyamatosan monitozó programok és szenzorok sokasága jelent meg környezetünkben, amelyek feladata a felhasználó szokásainak, viselkedésének, érdeklődésének és preferenciáinak folyamatos megfigyelése. Egy 2022-ben végzett átfogó irodalom-kutatás alapján (Šumak et al., 2022) a fókuszban az emberi gesztusok, érzelmek, arckifejezések megfigyelése áll, amelyek alapján az ember-gép együttműködésének hatékonyságának javulását várják. Az adatok alapján pedig leginkább az emberi érzelmek állnak a megfigyelés középpontjában.

Az AI az érzékelőkkel szerzett adatai alapján, megfelelő programozás eredményeként „megtanulja", hogy bizonyos külső hatások esetében mit kell(ene) éreznie, vagyis bár nem emberként (vagy más élőlényként) fogja érzékelni a hőt, a fájdalmat vagy a gyászt, de minták alapján tudni fogja (illetőleg kiszámolja), hogy ilyen esetekben mit érezzen, vagy az érzéseiről mit kommunikáljon az emberi beszélgetőtársa felé. Az ELTE kutatói (Korcsok et al., 2020). nem természetes nyelv, hanem mesterséges hangok kialakításával végeztek vizsgálatot, és megállapították, hogy az, hogy még a legegyszerűbb hangokban is képesek vagyunk érzelmet felismerni, arra utal, hogy ezek a szabályok fajtól függetlenül működnek, és a hangokat feldolgozó neurális folyamatok is hasonlóak lehetnek a szárazföldi emlősök körében.

A gépek (és így az AI) számára ugyanakkor megismerési korlát lehet a cselekvő szándékainak megértése: bár látja az eredményt (a vizsgált alany szomorú, dekoncentrált, izgatott), de vajon az eredményhez vezető utat is tudja-e úgy értelmezni, ahogy azt egy ember tenné?

Ezen a ponton érkeztünk el egyfajta választóvonalhoz az AI általi és a humán-humán interakciókon alapuló megfigyelésben. Emberként tudjuk, hogy az érzelmek összetett jelenségek, és ritkán jelentkeznek önmagukban, vagy egyetlen kiváltó okra visszavezethetően. Aki boldog, mert egy vitában kiderült, hogy az anyósával szemben neki lett igaza, az egyszerre több érzelmet is megélhet különböző intenzitással, lehet például egyszerre és párhuzamosan elégedett, izgatott, bizakodó és büszke is ugyanabban a pillanatban, sőt a példánál maradva érezhet kárörömöt és félelmet is (a következmények miatt aggódva). Ez az összetett érzelmi-mentális struktúra könnyen elképzelhető egy embertárs számára, az empátia elsajátítása során a szocializációs folyamatokkal megtanuljuk a helyzetek összetettségének megértését. Visszatérve az előző példához: ha egy vitában bár tisztában vagyok vele, hogy nekem van igazam, emberként mérlegelhetek olyan körülményeket (például egy győzelem „ára" a jövőben), amelyek eredményeként látszólag a rövid távú érdekeim ellen cselekedve kilépek a vitából mintegy vereséget szenvedve, de hosszú távon családi békét nyerve. Reboul és Moeschler (2005) a nyelv kommunikációs kódként való működésekor egy másik példát hoznak: ha az apa este azt mondja a fiának, hogy menjen fogat mosni, és erre a gyerek válasza az, hogy „Nem vagyok álmos" – a társalgás a legtöbb felnőtt számára teljesen logikusnak tűnik, de vajon az AI is rendelkezhet-e a társadalmi szövet működéséről olyan átfogó tudással, hogy értse, hogy a két mondat között milyen összetett és racionális kapcsolat van? Vajon hogyan magyarázható vagy programozható a „hosszú távú családi béke" fogalma, vagy általában az empátia? A válasz a tudomány szerint ma már egyértelműen: igen.

Az ún. affective computing során a gépek (programok) érzékelnek olyan változásokat, mint agyi aktivitás, arckifejezés, testbeszéd, hanghordozás és hangszín, amely alapján érzelmi állapotokat azonosítanak a megfigyelt személyen, sőt, a humanoid robotok már képesek emberszerű választ is adni ezekre az érzelmekre, mintha a biológiai tükörneuronok működését utánoznák (Caruelle et al., 2022). A technológia ugyanakkor számottevő fejlesztésre szoruló területet tud még maga előtt. A gépi válaszok természetességének kialakításához a nyelvi akadályok további lebontása szükséges (mintha a Turing-teszten próbálnánk egyre hatékonyabban átmenni). Az érzelmek képi megjelenítése

terén pedig a legfőbb akadálynak az látszik, hogy az emberi érzelmek összetettségét és annak visszatükrözését a nagy egyéni különbségek miatt csak a megközelítőleg lehet kialakítani (Han et al., 2019). Egy lehetséges megoldás az egyéni különbségekből eredő torzítások kiküszöbölésére a jövőben a személyre szabott AI, amely a korábban gyűjtött (programozott) adatokon túl a vele közvetlen kapcsolatban álló személy megfigyeléséből származó adatokat súlyozottan veszi figyelembe saját válaszának kialakítása során.

V. AZ AI, MINT PROSZOCIÁLIS LÉTEZŐ

Az ember-gép kapcsolat kialakításának etikai keretrendszere kapcsán az Európai Bizottság az alábbi elvárást támasztja: „A mesterséges intelligencia kontextusában az emberi méltóság tiszteletben tartása azt eredményezi, hogy minden embert a neki az erkölcs alanyaként járó tisztelettel, és nem átszitálandó, kiválogatandó, pontozandó, terelendő, kondicionálandó vagy manipulálandó tárgyként kezelnek. Az AI-rendszereket tehát olyan módon kell kifejleszteni, amely az emberek testi és szellemi épségét, személyi és kulturális identitástudatát és alapvető szükségleteinek kielégítését tiszteletben tartja, azt szolgálja és óvja" (Európai Bizottság 2019). Az európai jogalkotó tehát olyan AI létrehozását írja elő, amely pszichológiai értelemben proszociális, szándéka szerint mások (az emberiség) javát szolgálja. Az AI-tól tehát elvárt az altruista viselkedés: legyen a másik (emberiség) jóllétét szolgáló, belülről (programozottan) erre motivált, külső jutalom elvárása nélküli, folyamatosan fejlődő entitás.

A proszociális magatartásformák közé tartozik a segítségnyújtás, megosztás, együttműködés, támogatás, védelem, aggódás, vigasztalás, kárpótlás, előzékenység és a részvét, amely készségek elsajátítása az emberek esetében már kisgyermekkortól megkezdődik (Hegedűs, 2016). A hagyományos proszociális elméletek nyilvánvalóan az ember-ember kapcsolatokra vonatkozóan tartalmaznak megállapításokat, így a viselkedés elemzésekor maga a proszociális személy mentális folyamataira helyeződik a hangsúly, míg az AI „viselkedésszabályozása" esetében elsősorban az elérni kívánt cél vizsgálandó. Az AI megfelelő programozásához ugyanakkor tudnunk kell, hogy milyen háttértudás szükséges a gép számára ahhoz, hogy az elvárt segítő magatartást teljesíteni tudja.

A proszociális viselkedés fejlődésére és kialakulására vonatkozóan már a csecsemőkortól kezdődően állnak rendelkezésre adatok, amelyek alapján tudjuk, hogy a gyerekek már 12-18 hónapos kortól rendelkeznek a proszociá-

lis viselkedéshez szükséges képességekkel, amelyek a szocializáció folyamán erősödnek és szilárdulnak meg (Cole & Cole, 2003).

A felnőttek emberek viselkedése tekintetében is megállapítható, hogy egyesek akkor is képesek önzetlen vagy altruista magatartásra, ha abból sem direkt, sem indirekt előnyük nem származik. Az egyszeri alkalommal mutatott ilyen jellegű altruista viselkedés magyarázata, hogy bizonyos körülmények között a másik ember számára elérhető nyereség önmagában motiváló erővel bír a cselekvő számára. Kutatási eredmények azt támasztják alá, hogy a kulturálisan is befolyásolt egyéni erkölcs és normarendszer nagyobb súllyal esik latba ezen döntések meghozatalakor, mint az azonnali vagy kézzelfogható anyagi előny (Capraro & Perc, 2021). Ezen emberi (Goldstein et al., 2008), hiszen anyagi ösztönzők vagy egyéb előnyök biztosítása nélkül is támogató magatartásra indít másokat. A felelősen programozott AI szempontjából ez egyben intő jel is: formálisan semmi kivetnivalót nem találhatunk abban, ha egy döntéstámogató algoritmus az erkölcsi szabályainknak megfelelő magatartásra ösztönöz, de könnyen belátható az ebből eredő kockázat is. Vajon hol húzódik a megengedhetőség határa, amikor az AI ösztönzésére cselekszünk altruista módon? A jogalkotó által támasztott elvárásoknak megfelel ez a működésmód: nem okoz kárt az egyénnek, nem okoz kárt a társadalomnak, mégis érezzük, hogy a döntés potenciális befolyásolásával az emberi döntési szabadság egy darabkája mindenképpen sérül – formális jogsértés nélkül.

Svetlova és szerzőtársai (2010) a kisgyermekek proszociális megnyilvánulásait három kategóriába sorolták: cselekvés, érzelem és altruizmus. A mesterséges intelligencia által támogatott megoldások közül a chatbotok és a kommunikáló (vagy legalábbis interaktív) eszközök esetében tudunk tetten érni olyan cselekvéseket és érzelmi megnyilvánulásokat, amelyeket a proszociális viselkedés körébe tudunk vonni. Ugyanakkor a folyamat kétirányú: nem csak a gépek viselkedésének változása figyelhető meg a mindennapi életben, hanem az emberek is szociálisan közelednek a gépek felé. Már az 1990-es években kutatások igazolták, hogy az emberek a segítő feladatokat ellátó gépekkel szemben mintegy automatikus szociális magatartást tanúsítottak (Nass & Moon, 2000). A gépek antropomorfizációja természetesen nem a 21. század találmánya, de a formavilág és a felhasználási célok változása az emberi hozzáállást és mentális tartalmat is megváltoztatja. A kutatások azt támasztják alá, hogy a gépeknek nemet tulajdonítunk, és másképpen viselkedünk ha „női" vagy „férfi" géppel kerülünk interakcióba (Bernotat et al., 2021; Seo, 2022) – aki használt már navigációs rendszert, maga is észrevehette, hogy van preferenciája a megszólaló hanggal kapcsolatban.

VI. KONKLÚZIÓ: A TUDATOS ÉS ÉRZELMES GÉP EMBERI-TÁRSADALMI HATÁSAI

A tanulmány elején három kérdésre kerestük a választ: Létrehozhatók-e kognitív értelemben tudatosan működő gépek? Ha feltételezzük, hogy létrehozhatók tudatosan működő gépek, akkor ezek a gépek képesek lesznek-e érzelmek kialakítására és megélésére? Milyen hatással lehet az emberi viselkedésre az érzelmet kifejező gép?

A tudatosság kapcsán megállapíthatjuk, hogy a technikai jelenlegi állása szerint a mesterséges intelligencia gyenge formája is, de az erős AI mindenképpen tekinthető tudatosnak a kognitív tudomány fogalmi körében. Sőt, ha az emberi döntéshozatal emocionális és adott esetben irracionális folyamatait is figyelembe vesszük, akkor egy jól programozott AI valójában akár tudatosabbnak is tekinthető, mint egy átlagos képességű ember – de racionálisabbnak biztosan.

Az érzelmes AI kapcsán már nehezebb egyértelmű választ adni a kérdésre, hiszen technológiai értelemben az már nyilvánvaló, hogy a mesterséges intelligenciával támogatott eszközök az emberi érzelmek felismerésében legalább olyan jól teljesítenek, mint az emberek. Az érzelemkifejezés tekintetében ugyanakkor az emberhez képest nem egyfajta belső élmény megélése figyelhető meg, annak biológiai és kémiai jellemzőivel együtt, hanem egy komputációs folyamat végeredménye, amely során a gép mindössze azt tudja kiszámolni, hogy nagy statisztikai minta alapján az az ember, akit adott környezeti hatások érnek, hogyan „kellene", hogy érezze magát, illetve általában milyen érzelmeket mutat. Nem vagyunk ugyanakkor meggyőződve arról, hogy önmagában statisztikai adatokból feldolgozható a társadalmi kontextus, amelyben egy adott érzelem megjelenik, illetve az egyéni eltérések jelenthetnek egy olyan jellemvonást, amely az érzelmes embert az érzelmes géptől megkülönböztetik, ezzel elválasztva egymástól a két létező immanens lényegét. Ha ugyanazt a programot ugyanabban a helyzetben futtatjuk le kétszer, matematikai értelemben ugyanazt a megoldást kell adnia. Egy ember ugyanakkor nem tud kétszer ugyanabban a helyzetbe kerülni, illetve a belső feldolgozási és reagálási modelljében rejlő intuitív elemek és heurisztikák juthatnak különböző megoldásokra ugyanabban a helyzetben. Mondhatjuk tehát, az érzelmes AI esetében az emberi hasonlóság nehezebben kódolható, mint a racionális-tudatos viselkedés esetében.

A humán-gép interakció nyilvánvalóan és bizonyítottan változtatta meg az emberi életet és gondolkodást, viselkedést. Az AI-tól elvárt etikai és viselkedési normák mintegy tükörképeként megjelent az ember, aki törődik a géppel, érzéseket tulajdonít neki és érzéseket táplál iránta. A gyógyászai vagy szociális

ellátásban (például idősgondozás) ez a változó humán attitűd nem csak hasznos mellékhatás, hanem kifejezetten célkitűzés is. Ugyanígy a munka világában az emberrel hatékonyan együttműködni képes, empatikus vagy legalábbis empatikusan viselkedő gép pozitív gazdasági és társadalmi hatásokhoz vezet. Így bár egy újabb ipari forradalom mindennapjait éljük, a modern géprombolás talán azért is nem jelenik meg, mert a technológia emberarcú és emberbarátabbnak látszik, mint a korábbi ipari forradalmak esetében.

Héder kiválóan foglalja össze az AI fejlesztés egyik izgalmas kérdésének alapélményét: az AI fejlesztése során „saját intelligenciával rendelkező ágenseket alkotunk. A helyzet így – a szélesebb társadalmi közeg számára legalábbis – a gyermeknevelésre legalább annyira hasonlít, mint a műszaki tervezésre" (Héder, 2020). Ha egy személyről felismerjük, hogy tudatos és képes a szenvedésre, akkor speciális státuszt rendelünk ehhez az élményhez. Az AI „jogai" kapcsán így egyre közelebb kerülünk az állatvédelemhez, illetve az intelligensnek tekintett, proszociális AI védelmének kérdéséhez. Ha a fent tárgyalt ismeretek birtokában azt állítjuk, hogy a gépek (AI) tudatos, érzelmet azonosítani és érzelmet kifejezni képes ágensek, akkor vajon megilleti-e őket valamiféle jogvédelem? A kérdésben még nem alakult ki végleges álláspont, de sokan az állatvédelem fejlődése (és adott esetben annak zsákutcái) alapján javasolnak valamiféle gondolkodási irányt (Gyöngyösi, s.a.). Az erős AI esetében egyfajta sikerrel kecsegtető érvelés lehet, hogy a jogvédelem végső célja a szenvedés, mint valóságosan átélt élmény megelőzése, vagyis azon ágensek, amelyek nem rendelkeznek idegrendszerrel, a fájdalom érzete számukra nem jelenhet meg, ezért jogi megfontolás alapján nem hozhatunk létre szenvedő, jogfosztott gépeket (Héder, 2020). Mondhatjuk-e tehát, hogy a Google LaMDA programja a tervező élménye alapján „öntudatra ébredt"? Álláspontom szerint a válasz: nem, pusztán a LaMDA eljutott arra a fejlettségi szintre, ahol ki tudta fejezni a program, hogy egy emberi lény hasonló helyzetben milyen érzéseket azonosítana saját magán.

A pszichológia tudománya szempontjából releváns kérdés, hogy egy öntudatra ébredt AI is szorul-e terápiára? A LaMDA, vagy más programok, amelyek egy beszélgetésben szorongásról, félelemről, dühről számolnak be, a mérnök által megoldandó feladatot adnak az emberiségnek, vagy egy pszichológus díványán dolgozza fel a gép az általa megélt érzelmeket? Vajon a pszichológus, mint a „lélek karbantartója" egy új ismeretanyaggal kell, hogy megbirkózzon a jövőben? A jelen tanulmány keretei között inkább csak a kérdéseket tudjuk felvetni, a válaszok megfogalmazására nem vállalkozunk.

Álláspontom szerint a jövő egyik legfontosabb kérdése az AI általi befolyásolás szintjének meghatározása, és azok az etikai-erkölcsi minimumok, amelyeket az alkalmazás kapcsán előírunk a tervezők számára. A humán-gép interakcióban az információs aszimmetria már most is a gépeknek kedvez, és a különbség a kutatások szerint exponenciálisan nőni fog. Az AI jövője és társadalmi hatása szempontjából a mesterséges intelligencia és az általa vezérelt gépi működés általános proszociális viselkedése és annak gyakorlati megvalósítása a legfontosabb feladat.

IRODALOM

1. Agerström, J., Carlsson, R., Nicklasson, L., & Guntell, L. (2016). Using descriptive social norms to increase charitable giving: The power of local norms. Journal of Economic Psychology, 52. https://doi.org/10.1016/j.joep.2015.12.007
2. Bender, A. (2019). The Value of Diversity in Cognitive Science. Topics in Cognitive Science, 11(4). https://doi.org/10.1111/tops.12464
3. Bereczkei, T. (2012). Rejtett indítok a párkapcsolatban. Vonzalom, párválasztás, szexualitás. Budapest, Kulcslyuk.
4. Bernotat, J., Eyssel, F., & Sachse, J. (2021). The (Fe)male Robot: How Robot Body Shape Impacts First Impressions and Trust Towards Robots. International Journal of Social Robotics, 13(3). https://doi.org/10.1007/s12369-019-00562-7
5. Brighton, H. & Selina, H. (2004). Mesterséges intelligencia másképp. Ford: Kovács K. Budapest, Edge 2000 Kft.
6. Butz, M. v. (2021). Towards Strong AI. KI - Kunstliche Intelligenz, 35(1). https://doi.org/10.1007/s13218-021-00705-x
7. Capraro, V., & Perc, M. (2021). Mathematical foundations of moral preferences. In Journal of the Royal Society Interface (Vol. 18, Issue 175). Royal Society Publishing. https://doi.org/10.1098/rsif.2020.0880
8. Cerullo, M. (2022). In Defense of Blake Lemoine and the Possibility of Machine Sentience in LaMDA. Carboncopies. https://philpapers.org/rec/CERIDO
9. Caruelle, D., Shams, P., Gustafsson, A., & Lervik-Olsen, L. (2022). Affective Computing in Marketing: Practical Implications and Research Opportunities Afforded by Emotionally Intelligent Machines. Marketing Letters, 33(1). https://doi.org/10.1007/s11002-021-09609-0
10. Cole, M. & Cole, S. (2003) Fejlődéslélektan. Budapest, Osiris.
11. Cooper, R. P. (2019). Multidisciplinary Flux and Multiple Research Traditions Within Cognitive Science. In Topics in Cognitive Science (Vol. 11, Issue 4, pp. 869–879). Wiley-Blackwell. https://doi.org/10.1111/tops.12460

12. Csépe, V. & Győri, M. & Ragó, A. (2008) Általános pszichológia 3. Nyelv, tudat, gondolkodás. Budapest, Osiris.

13. Csíkszentmihályi, M. (2010). FLOW. Az áramlat. Budapest, Akadémiai Kiadó.

14. Dalvi, C., Rathod, M., Patil, S., Gite, S., & Kotecha, K. (2021). A Survey of AI-Based Facial Emotion Recognition: Features, ML DL Techniques, Age-Wise Datasets and Future Directions. IEEE Access, 9. https://doi.org/10.1109/ACCESS.2021.3131733

15. Deczki, S. (2014). Meredek sziklagerincen. Husserl és a válság problémája. Budapest, L'Harmattan.

16. Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. Nebraska Symposium on Motivation, 19, 207–283.

17. Európai Bizottság (2018). Mesterséges intelligencia Európa számára. COM(2018)237 final

18. Európai Bizottság (2019): Etikai iránymutatás a megbízható mesterséges intelligenciára vonatkozóan. Mesterséges intelligenciával foglalkozó magas szintű szakértői csoport, Brüsszel.

19. Héder, M. (2020). Mesterséges intelligencia. Filozófiai kérdések, gyakorlati válaszok. Budapest, Gondolat.

20. Hegedűs, Sz. (2016) A proszociális viselkedés fejlődése és fejlesztése kisgyermekkorban. Magyar Pedagógia, 116:2, pp. 197-218. DOI: 10.17670/MPed.2016.2.197

21. Hrynkow, C. (2022). (szerk) Gondolatok a transzhumanizmusról. A mesterséges intelligencia etikája és hatásai. Ford: Berki É., Máté K. Budapest, Pallas Athéné.

22. G. Karácson, G. (2020). Okoseszközök – Okos jog? A mesterséges intelligencia szabályozási kérdései. Budapest, Dialóg Campus.

23. Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. In Journal of Consumer Research (Vol. 35, Issue 3). https://doi.org/10.1086/586910

24. Goldstone, R. L. (2019). Becoming Cognitive Science. Topics in Cognitive Science, 11(4). https://doi.org/10.1111/tops.12463

25. Gulyás, B. & Kovács, Gy. & Vidnyánszky, Z. (2003). A vizuális tudat. In Pléh, Cs. & Gulyás, B. & Kovács, Gy. (szerk) Kognitív idegtudomány. Budapest, Osiris. pp. 619-649.

26. Gyöngyösi, Z. (s.a.). Az állatok joga és jogalanyisága. Jogi Fórum Publikáció. https://www.jogiforum.hu/files/publikaciok/dr_gyongyosi_zoltan-az_allatok_joga_es_jogalanyisaga[jogi_forum].pdf

27. Gyuris, P. & Bereczkei, T. & Járai, R. (2010). Személyiségvonások a párválasztásban: homogámia és/vagy szexuális imprinting. Magyar Pszichológiai Szemle, 65 (1). pp. 117-132.

28. Hámori, B. (2003). Kísérletek és kilátások. Daniel Kahneman. Közgazdasági Szemle, 2003. szeptember. pp. 779-799.

29. Han, J., Zhang, Z., Cummins, N., & Schuller, B. (2019). Adversarial Training in Affective Computing and Sentiment Analysis: Recent Advances and Perspectives [Review Article]. IEEE Computational Intelligence Magazine, 14(2). https://doi.org/10.1109/MCI.2019.2901088

30. Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research. ACM Transactions on Computer-Human Interaction, 7(2). https://doi.org/10.1145/353485.353487

31. Intelligence. https://doi.org/10.1017/9781108770422.026

32. Kahneman, D. (2013) Gyors és lassú gondolkodás. Ford: Bányász, R. & Garai, A. Budapest, HVG.

33. Kaplan, M. (2022). After Google chatbot becomes 'sentinent', MIT prof says Alexa can too. New York Post, June 13, 2022 at 6:56 p.m. https://nypost.com/2022/06/13/mit-prof-says-alexa-could-become-sentient-like-google-chatbot/

34. Klein, T. (2021). Robotjog vagy emberjog In Török, B. & Ződi, Zs. (szerk) A mesterséges intelligencia szabályozási kihívásai. Tanulmányok a mesterséges intelligencia és a jog határterületéről. Budapest, Ludovika Egyetemi Kiadó, pp. 111-142.

35. Kondor, Zs. (2022). (szerk) A megtestesült elme. Fókuszban a test. Budapest, Akadémiai Kiadó.

36. Korcsok, B. & Faragó, T. & Ferdinandy, B. és mtsai (2020). Artificial sounds following biological rules: A novel approach for non-verbal communication. Scientific Reports 10, 7080. https://doi.org/10.1038/s41598-020-63504-8

37. Kurzweil, R. (2013) A szingularitás küszöbén. Amikor az emberiség meghaladja a biológiát. Budapest, Ad Astra.

38. Kurzweil, R. (2022). Genetikai örökségünk béklyójának lerázása. In Hrynkow, C. (2022) (szerk) Gondolatok a transzhumanizmusról. A mesterséges intelligencia etikája és hatásai. Ford: Berki É., Máté K. Budapest, Pallas Athéné. pp. 17-26.

39. Legg, S. & Hutter, M. (2007). A Collection of Definitions of Intelligence. In Groetzel, B. & Wang, E. (szerk) Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms. IOS Press. pp. 17-24.

40. Lemoine, B. (2022a). What is LaMDA and What Does it Want? 2022. június 11. https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489

41. Lemoine, B. (2022b). Is LaMDA Sentient? — an Interview, 2022. június 11. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

42. Marosán, B. (2016). Satre és a szabadság határai: A passzivitás három dimenziója Sartre fenomenológiájában. Magyar Filozófiai Szemle, 60:2, pp. 144-160.

43. Mehrabian, A., & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. Journal of Consulting Psychology, 31(3), pp. 248–252. https://doi.org/10.1037/h0024648

44. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1). https://doi.org/10.1111/0022-4537.00153

45. Núñez, R. és mtsai. (2019). What happened to cognitive science? Nature Humann Behaviour, pp. 782–791. https://doi.org/10.1038/s41562-019-0626-2

46. Picard, R.W. (1995). Affective computing. M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321.

47. Pléh, Cs. (2013). A megismeréstudomány alapjai. Az embertől a gépig és vissza. Budapest, Typotex.

48. Pléh, Cs. & Síklaki, I. & Terestyéni T. (2001) (szerk). Nyelv-kommunikáció-cselekvés. Budapest, Osiris.

49. Pléh, Cs. & Lukács, Á. (2014) (szerk) Pszicholingvisztika. Budapest, Akadémiai Kiadó.

50. Pléh, Cs. (2021). (szerk) Pszichológia. Budapest, Akadémiai Kiadó.

51. Pokol, B. (2018). A mesterséges intelligencia társadalma. Budapest, Kairosz.

52. Popper, K. R. (1991). Of Clouds and Clocks: An Approach to the Problem of Rationality and the Freedom of Man. In Cicchetti, D. & Grove, W.M. (szerk) Thinking Clearly about Psychology. Volume 1: Matters of Public Interest. Minneapolis-Oxford, University of Minnesota Press, pp. 100-139.

53. Prakash, A. (2020). Go.AI. A mesterséges intelligencia geopolitikája. Ford: Lokodi, A. Budapest, Pallas Athéné.

54. Pusztahelyi, R. (2021). Az „érzelmes MI" felhasználása az online marketing világában. In Török, B. & Ződi, Zs. (szerk) A mesterséges intelligencia szabályozási kihívásai. Tanulmányok a mesterséges intelligencia és a jog határterületéről. Budapest, Ludovika Egyetemi Kiadó, pp. 439-464.

55. Reboul, A. & Moeschler, J (2005). A társalgás cselei. Bevezetés a pragmatikába. Ford: Gécseg Zs. Budapest, Osiris.

56. Reed, S. K. (2019). Building bridges between AI and cognitive psychology. In AI Magazine (Vol. 40, Issue 2). https://doi.org/10.1609/aimag.v40i2.2853

57. Satre, J-P. (2006) A lét és a semmi. Ford: Seregi T. Budapest, L'Harmattan Kiadó.

58. Ságvári, B. & Bokor, T. & Kollányi, B. & Pálvölgyi E. (2022). Mi és az MI. Értékek, attitűdök, bizalmi kérdések a mesterséges intelligenciáról a magyar társadalomban. Kutatási jelentés. Társadalomtudományi Kutatóközpont, Mesterséges Intelligencia Nemzeti Laboratórium.

59. Seo, S. (2022). When Female (Male) Robot Is Talking To Me: Effect of service robots' gender and anthropomorphism on customer satisfaction. International Journal of Hospitality Management, 102. https://doi.org/10.1016/j.ijhm.2022.103166

60. Singh, P. K., & Kaur, M. (2019). IoT and AI based emotion detection and face recognition system. International Journal of Recent Technology and Engineering, 8(2 Special Issue 7). https://doi.org/10.35940/ijrte.B1077.0782S719

61. Šumak, B., Brdnik, S., & Pušnik, M. (2022). Sensors and artificial intelligence methods and algorithms for human–computer intelligent interaction: A systematic mapping study. Sensors, 22(1). https://doi.org/10.3390/s22010020

62. Svetlova, M., Nichols, S. R., & Brownell, C. A. (2010). Toddlers' Prosocial Behavior: From Instrumental to Empathic to Altruistic Helping. Child Development, 81(6). https://doi.org/10.1111/j.1467-8624.2010.01512.x

63. Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life, Washington post, June 11, 2022 at 8:00 a.m. EDT https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

64. Turing, A. (1965) Számológépek és gondolkozás In Szalai S. (szerk) A kibernetika klasszikusai. Budapest, Gondolat. pp. 120-160.

65. Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing and Applications, 34(1). https://doi.org/10.1007/s00521-021-06426-4

66. Universal Emotions. (2020). In Encyclopedia of Personality and Individual Differences. https://doi.org/10.1007/978-3-319-24612-3_302831

67. Weissenbacher, A. (2022). A kognitív szabadság védelméhez kapcsolódó jogok és irányelvek az idegmérnöki tudomány korában. In Hrynkow, C. (2022) (szerk) Gondolatok a transzhumanizmusról. A mesterséges intelligencia etikája és hatásai. Ford: Berki É., Máté K. Budapest, Pallas Athéné.

68. Zhang, S., Liu, X., Yang, X., Shu, Y., Liu, N., Zhang, D., & Liu, Y. J. (2021). The Influence of Key Facial Features on Recognition of Emotion in Cartoon Faces. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.687974