Magyar Tudomány 182(2021)9, 1203–1209
DOI: 10.1556/2065.182.2021.9.5

ETIKUS ÉS BIZTONSÁGOS MESTERSÉGES INTELLIGENCIA

ETHICAL AND SAFE ARTIFICIAL INTELLIGENCE

Boncz Bettina[1], Szabó Zs. Roland[2]

[1]kutató, Budapesti Corvinus Egyetem, Budapest
bettina.boncz@gmail.com
[2]PhD, habilitált egyetemi docens, Budapesti Corvinus Egyetemen, Budapest
zsoltroland.szabo@uni-corvinus.hu

ÖSSZEFOGLALÁS

A mesterséges intelligencia (MI) megjelenése a 21. század egyik meghatározó eseménye. Munkánk, életünk, életmódunk és vitathatatlanul emberi kapcsolataink is gyökeresen átalakulnak az intelligens gépek korában. Jelen cikk az MI etikai, biztonságossági és jogi dilemmaival foglalkozik. Célunk az, hogy felhívjuk a figyelmet néhány fontos problémára, amelyeket minél hamarabb meg kellene oldanunk. Különben az MI beilleszkedése a komplex és diverz társadalmi rendszereinkbe potenciálisan katasztrófához vezethet.

ABSTRACT

The emergence of artificial intelligence (AI) is one of the defining events of the 21st century. Our work, our lives, our way of life, and indisputably our human relationships are also radically changing in the age of intelligent machines. This article addresses the ethical, security, and legal dilemmas of AI. Our goal is to draw attention to some important issues that we should resolve as soon as possible. Otherwise, the integration of AI into our complex and diverse social systems could potentially lead to disaster.

Kulcsszavak: intelligens gép, kiberbiztonság, robot jogok, algoritmus, öntanuló rendszer

Keywords: intelligent machine, cyber security, robot rights, algorithm, self-learning system

MIT IS ÉRTÜNK MESTERSÉGES INTELLIGENCIA ALATT?

A mesterséges intelligenciát (MI) számos tudományterületen megkísérelték értelmezni; mind a természet- és műszaki tudományok (lásd Poole–Mackworth, 2010), mind a társadalomtudományok (lásd Jarrahi, 2018), mind más területeken. A mesterséges intelligencia egy gyűjtőfogalom. Tudományterületenként más és

más alkalmazásokat, elméleteket takar. Közös jellemzőjük, hogy az MI képes az intelligens viselkedésre, ami azt jelenti, hogy az MI bármilyen eszköz lehet, amely egyszerre (1) modellezi és (2) diagnosztizálja a környezetét, (3) feladatot hajt végre célja minél sikeresebb megvalósítása érdekében, és (4) tanul a korábbi tapasztalatokból (Poole et al., 1998). Vagyis imitálják az emberi intelligenciát, viselkedést, miközben képesek tanulni, megérteni és érzékelni.

Jelenleg úgy vélik, hogy az emberi intelligenciát kiválóan imitáló MI valamikor a 21. század közepén fog megjelenni az életünkben, de a technológiai fejlődés jelenlegi ütemében ez az időpont még előrébb tolódhat. Életünkben viszont már ma is léteznek olyan gépek, szoftverek, alkalmazások, rendszerek, amelyek kezdetleges mesterséges intelligencia algoritmusokat alkalmaznak.

A MI HASZNÁLATÁVAL KAPCSOLATOS BIZTONSÁGOSSÁGI ÉS ETIKAI AGGODALMAK, JOGI DILEMMÁK

Az MI jelenünk és jövőnk meghatározó technológiai újítása lesz, amely eddig nem látott mértékben fogja befolyásolni mindennapi életünket, munkánkat, emberi kapcsolatainkat. Ezen változások hozadéka – reményeink szerint mind – pozitív lesz, azonban megfelelő hátteret kell biztosítani arra, hogy az intelligens rendszer(ek) ne legyen(ek) képes(ek) kárt okozni számunkra. Itt fontos elkülöníteni a technikai biztonságot, az MI etikusságának kérdését és végül a felmerülő jogi problémákat.

A biztonságos MI

Ha megvizsgáljuk a sci-fi irodalmat vagy filmeket, egyértelművé válik, hogy az írók kevésszer képzelnek el kedves, jóindulatú MI-t, ami mindenben a segítségünkre lesz. A leggyakrabban lázadó robotokat, emberi parancsot visszautasító, önállóan eljáró zsarnok gépeket ábrázolnak, amelyek befolyásolni akarják az emberiség életének minden aspektusát, amíg csak lebegő, székekben ülő, túlsúlyos, lebutított lényekké nem válunk, mint a népszerű mesében, a WALL-E-ban (Cave–Dihal, 2019).

Habár a veszély valós, nagyobb eséllyel fordulhat az emberiség ellen egy MI a tervezési fázisban ejtett hiba, szándékos, rosszindulatú beavatkozás vagy környezeti hiba miatt (Yampolskiy, 2016), mint önmagától.

Nemcsak a fantáziairodalom, de a tudományos közösség is tart tőle, és tisztában van ezen veszélyekkel, de a teljes körű megoldás még várat magára (Tegmark, 2017), a legtöbb tanulmány inkább filozófiai, mint számítástechnológiai oldalról próbál megoldást találni (Müller–Bostrom, 2016).

Először is szükséges lenne tisztázni a biztonságos MI technológiai követelményeit. Amikor megalkották az internetet, nem fektettek hangsúlyt a biztonsági

követelményekre, és ma a társadalomra és az egyénre is ártalmas tartalmak milliói jelennek meg rajta. A biztonságos MI megteremtéséhez a tudósok, mérnökök és filozófusok nemzetközi együttműködése szükséges, hogy bárhol is történjen meg az áttörés, az a nemzetközi standardoknak megfelelően történjen (Cave–ÓhÉigeartaigh, 2019). Indokolt lenne egy olyan szigorú szabályozás létrehozása az MI kutatások területén, mint amilyen a génmódosítások körül alakult ki, hogy az alanyokat és tágabb értelemben a teljes emberiséget megvédje a kutatások és kísérletek hátrányos következményeitől (Müller–Bostrom, 2016).

Melyek a legnagyobb biztonsági kockázatok? Egy MI, ami feltörhető, eltéríthető, melynek kódjában biztonsági rések és hibák vannak, az az MI, amelyet nem olyan adatok felhasználásával „tanítottak", amelyek az emberiség többségének viselkedését és szándékait fedik le (például egy szélsőséges terrorista csoport viselkedését veszi alapul), és a lista még nagyon hosszú lehetne (Pistono–Yampolskiy, 2016). A biztonsági kockázatok hatványozódnak, amikor egy önmagát fejlesztő rendszer alapjaiban, első verziójában biztonsági hibák vannak, hiszen a rendszer ezeket tovább hordozza magában, miközben egyre intelligensebbé válik.

Jelenleg sokkal nagyobb a valószínűsége annak, hogy egy rosszul tervezett MI világméretű katasztrófát idéz elő, mint az, hogy egy jól tervezettet meghekkelnek, és emiatt teremtői ellen fordul (Weld, 2016).

Továbbá számos embernek és csoportnak áll érdekében meghekkelni egy MI-t: kormányok, melyek nagyobb hatalomra áhítoznak, vállalatok, melyek monopolhelyzetet akarnak kivívni, gazemberek, akik uralkodni akarnak másokon, világvégehívő szekták, bűnözők, pszichopaták vagy éppen olyan biztonságos MI hívők, akik a technológiával való visszaéléstől való félelmükben maguk okoznak károkat (Glenn–Gordon, 2004; Pistono–Yampolskiy, 2016) a rendszerben.

Jelenleg még a nem intelligens gépek is képesek komoly problémákat okozni, és ezeknek az okait néha sosem ismerjük meg. A General Electric (GE) szoftverében egy kis hiba több mint 50 millió embert hagyott áram nélkül 2003-ban. Sokak által ismert a flash crash 2010-ből, mely alig fél óra alatt az USA tőzsdepiacán okozott több milliárd dolláros veszteséget, és melynek pontos okai máig ismeretlenek, de valószínűsíthetően az automatizált részvénykereskedelemhez köthetőek.

A mai kezdetleges MI-rendszerek működése gyakran kiszámíthatatlan, fekete doboz elven működnek. Ez azt jelenti, hogy a külső szemlélő számára csak az látható, hogy milyen *input*ot adtak a gépnek és milyen *output*ot kaptunk, de a közben lezajló folyamatok szinte feltérképezhetetlenek (Morley et al., 2020). Nem tudja megmagyarázni, miért omlana össze egy épület, egész egyszerűen kijelenti, hogy össze fog omlani. Mindezzel nem is lenne gond, hiszen maguk az emberek sem tudnak sok döntésükre magyarázatot adni. Sokszor érzelmek, intuíciók, megmagyarázhatatlan tudat alatt zajló folyamatok alapján döntünk. Azonban,

ahogy nem szívesen adnánk egyetlen ember kezébe a világ feletti uralmat, úgy egy hasonló módon döntéseket hozó MI kezébe sem kellene.

Mivel a következményeit nem tudjuk megjósolni (Weld, 2016), fontos, hogy ne helyezzük az MI-t kritikus területek, mint például pénzügyi piacok, kritikus infrastruktúra, fegyverek vagy kommunikációs csatornák feletti kizárólagos felügyeletre sem (Pistono–Yampolskiy, 2016).

Külön problémát jelent, hogy egy biztonságosnak vélt MI-rendszert nem tudunk a ma ismert módszerekkel tesztelni. A szoftverfejlesztés egyik módszere, hogy a félkész, de már használható programot a végfelhasználóknak adják, majd a visszajelzéseik alapján végeznek módosításokat. Sajnos, az MI esetében ennek már a gondolata is biztonsági kockázatokat vet fel (Yampolskiy, 2016), hiszen csak akkor lehetne egy ilyen tudású programot tesztelni, ha az egy zárt „térben" létezne, egy tűzfal mögött, internetkapcsolat nélkül, ám erre egyelőre még nem született kielégítő megoldás (Russell, 2016).

Az etikus MI

Az igazi probléma azonban az MI megtanítása a morálra és az etikára. Az emberi társadalmakat rengeteg, néha egymással is ellentmondó morális és etikai szabály irányítja, melyeket néha magunk is figyelmen kívül hagyunk. Az etikai szabályok emellett térben és időben is folyamatosan változnak, és maguk az emberek is sokszor gumiszabályként tekintenek rájuk (Creighton, 2016). Egy MI számára szinte lehetetlen lenne ezen emberi mintázatokból következtetéseket levonni, és saját maga számára cselekvési útmutatót kreálni.

Dönthetnénk úgy is, hogy már az elején beprogramozzuk ezeket az elveket, ám a mesterséges intelligenciát létrehozó tudósok és mérnökök (gyakran) igen csekély ismerettel rendelkeznek a morálról és az etikáról tudományos oldalról nézve (Glenn–Gordon, 2004). Segítségükre lehet, hogy már több mint hetven különböző társadalomtudományi szemszögből megírt MI etikai kódex létezik a világon, ám közös hiányosságuk, hogy arra nem adnak útmutatást, hogyan tudnák a számítástechnikai, mérnöki szakemberek mindezt „megtanítani" az MI-nek (Morley et al., 2020).

Az ellentmondás (egy lehetséges) feloldása az, hogy nincs is szükség etikusan viselkedő gépekre. Az etika, a morál emberi konstrukció, ami erőteljesen köthető érzelmekhez, amivel egy MI nagy valószínűséggel nem fog rendelkezni (Moore, 2010). Elegendő lenne tehát, hogy ha az MI megfelelő célok szerint létezne, morál és etika nélkül. Az emberi kapcsolatokon belül is bizalmat ad, ha tudjuk, hogy bár embertársunk más kultúrában, más etikai és morális iránytű szerint él, de vannak közös pontok az életünkben, mint például, hogy nem fogjuk megölni egymást, nem gyújtunk fel épületeket, és szeretnénk, ha a körülöttünk lévők jóléte növekedne. Ilyen célokat már, mondhatni, „könnyebb" adni, mint morális iránytűt

beépíteni az MI-be. A célok adását is azonban körültekintően kell kezelni, és a célok elérésének módjához szabályokat kell hozni.

Az MI mindig kontroll alatt kell, hogy maradjon, és ha céljai hajhászása közben netalán veszélyessé válna, le kell tudnunk kapcsolni (Arnold–Scheutz, 2018). Egyszerűnek tűnik, de amint az MI rájön, hogy céljának megvalósítása útjában a legnagyobb akadály az, hogy bármikor lekapcsolhatják, ennek minden lehetőségét ki fogja iktatni (Glenn–Gordon, 2004; Russell, 2016). Ezt meg kell akadályozni.

Fontos megemlíteni azt is, hogy jelenleg a legtöbb kutatás a nagy technológiai cégeknél koncentrálódik, és ők fektetik be a legnagyobb összegeket is a kutatásokba. Ezen centralizált kutatások miatt nagyobb az esély egy elfogult MI létrehozására (Montes–Goertzel, 2018), mely csak egy kultúra vagy népességcsoport céljait, elvárásait követi.

Szintén problémát okoz, hogy a gépi tanulás, a kezdetleges MI-algoritmusok tanulási alapját sokszor olyan adatbázisok adják, melyek aránytalanul reprezentálják a világ lakosságát. Az ImagineNet nevezetű adatbázis, melyen a képfelismerés fejlesztése zajlik, 45%-ban az USA-ból és mindössze 3%-ban Indiából vagy Kínából származó képekkel operál, holott lakosságarányosan ennek a számnak legalább 36%-nak kellene lennie (Zou–Schiebinger, 2018).

A jogi dilemmákról röviden

2017 egyik meghatározó híre volt, hogy Sophia, a robot, állampolgárságot kapott Szaúd-Arábiától. Sophia még nem mondható mesterséges intelligenciának, de saját nemében okos, és kialakításának köszönhetően egészen emberi. Képes mosolyogni, elszomorodni és kommunikálni a környezetével, valamint helyváltoztatásra is képes.

A lépés Szaúd-Arábia részéről bár szimbolikus volt, komolyabb következményeket is vonhat maga után. Képzeljük el, hogy az „állampolgár robot" fejében a drótok összekuszálódnak, és egyik napról a másikra embereket kezd bántalmazni. A logikus lépés az lenne, hogy húzzuk ki a konnektorból, és kapcsoljuk ki. Egy robot esetében azonban ez nem tekinthető egynek a halálbüntetéssel? Hogyan nem ítélhetünk halálra egy emberi állampolgárt, de tehetjük ezt meg a géppel, aki jogilag ugyanolyan teljes értékű „ember", mint mi vagyunk?

A lehetséges veszélyek ellenére egy mesterséges intelligenciának, mely potenciálisan (közvetve vagy közvetlenül) emberi életeket befolyásolhat, valamilyen jogi státuszt biztosítani kell, hogy az MI is jogilag felelős személy legyen.

Ám, ahogyan egy gyermek sem rendelkezik teljes jogi felelősséggel, mivel nem feltétlen képes felfogni ésszel vagy érzelmi szinten tetteinek következményeit, úgy ezen analógiára vetítve egy hasonlóan „tudattalan" gép sem kellene, hogy jogilag felelőssé tehető legyen, ám fel kell készülnünk arra a pillanatra, amikor a gép is tudatra ébred.

Kezdetben az ún. robotjogok olyan kérdésekre adnának választ, mint hogy ha egy önvezető autó vezet, és balesetet okoz, ki legyen a jogilag felelősségre vonható: az autó gyártója, az MI, az MI gyártója, vagy az utasok, akik választhatnak, hogy milyen intelligens rendszert kívánnak megvenni az autójukhoz (Bonnefon et al., 2016). Láthatóan a robotjog csak abban az értelmezésben létezik, hogy más természetes vagy jogi személyek jogai hogyan érvényesülnek, azaz nem egy adott robotnak van joga, mint az embernek, hanem közvetve, más jogilag felelős személyek vonatkozásában vannak „jogai". A későbbiekben ez a kör kibővíthető lenne, és elindulhatna a gépek „egyenjogúsítása", például tulajdonjog, gazdasági tevékenység végzéséhez való jog szerzésének biztosítása (Nekit et al., 2020).

ZÁRÓ GONDOLATOK

A fentiekben kifejtett biztonságossági, etikai és jogi dilemmák csupán egy kis részét képezik mindannak a problémacsokornak, amely a mesterséges intelligencia megjelenése körül alakul(t) ki. Ezen kérdéskörök közös pontja, hogy mindegyikükre minél hamarabb, az emberi intelligenciát tökéletesen imitáló (vagy akár felülmúló) MI megjelenése és tömeges elterjedése előtt kell választ találni.

A fő irányvonalakat tekintve a szabályozásnak nemzetközi szinten egységesítettnek kell lennie, majd a helyi igényeknek megfelelően testre szabhatónak, csak így lehet az alkalmazási környezet megbízható, és kiszámítható. Ez nem csupán a felhasználóknak, de a fejlesztőknek, és minden más érintettnek is közös érdeke. Mindemellett valamennyi tudományterület összefogása szükséges, hogy az MI értünk és ne ellenünk legyen!

IRODALOM

Arnold, T. – Scheutz, M. (2018): The "Big Red Button" Is Too Late: An Alternative Model for the Ethical Evaluation of AI Systems. *Ethics and Information Technology,* 20, 4, 59–69. DOI: 10.1007/S10676-018-9447-7, https://hrilab.tufts.edu/publications/arnold2018big.pdf

Bonnefon, J.-F. – Shariff, A. – Rahwan, I. (2016): The Social Dilemma of Autonomous Vehicles. *Science,* 352, 6293, 1573–1576. DOI: 10.1126/Science.aaf2654, https://arxiv.org/ftp/arxiv/papers/1510/1510.03346.pdf

Cave, S. – Dihal, K. (2019): Hopes and Fears for Intelligent Machines in Fiction and Reality. *Nature Machine Intelligence,* 1, 74–78. DOI: 10.1038/S42256-019-0020-9, https://www.repository.cam.ac.uk/handle/1810/288940

Cave, S. – ÓhÉigeartaigh, S. S. (2019): Bridging Near- and Long-Term Concerns about AI. *Nature Machine Intelligence,* 1, 1, 5–6. DOI: 10.1038/S42256-018-0003-2, https://www.repository.cam.ac.uk/handle/1810/293033

Creighton, J. (2016): *The Evolution of AI: Can Morality Be Programmed?* 1 July 2016. https://Futurism.Com/The-Evolution-Of-Ai-Can-Morality-Be-Programmed

Glenn, J. – Gordon, T. J. (2004): Future S&T Management Policy Issues—2025 Global Scenarios. *Technological Forecasting and Social Change,* 71, 913–940. DOI: 10.1016/j.techfore.2003.12.005, https://isiarticles.com/bundles/Article/pre/pdf/15681.pdf

Jarrahi, M. H. (2018): Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making. *Business Horizons,* 61, 577–586. DOI: 10.1016/j.bushor.2018.03.007, https://bit.ly/3xgD43T

Montes, G. A. – Goertzel, B. (2018): Distributed, Decentralized, and Democratized Artificial Intelligence. *Technological Forecasting and Social Change,* 141, 354–358. DOI: 10.1016/j.techfore.2018.11.010, https://www.researchgate.net/publication/329379212_Distributed_decentralized_and_democratized_artificial_intelligence

Moore, D. (2010): *Critical Thinking and Intelligence Analysis*. Books Express Publishing

Morley, J. – Floridi, L. – Kinsey, L. et al. (2020): From What to How: An Overview of AI Ethics Tools, Methods and Research to Translate Principles Into Practices. *Science and Engineering Ethics,* 26, 2141–2168. https://link.springer.com/article/10.1007/s11948-019-00165-5

Müller, V. C. – Bostrom, N. (2016): Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In: Müller, V. C. (ed.): *Fundamental Issues of Artificial Intelligence*. (*Synthese Library*) Berlin: Springer International Publishing, 571. DOI: 10.1007/978-3-319-26485-1_33, https://www.nickbostrom.com/papers/survey.pdf

Nekit, K. – Tokareva, V. – Zubar, V. (2020): Artificial Intelligence as a Potential Subject of Property and Intellectual Property Relations. *Revista De Derecho,* 9, 1, 23–28. DOI: 10.31207/Ih.V9i1.227, http://iushumani.org/index.php/iushumani/article/view/227

Pistono, F. – Yampolskiy, R. V. (2016)*: Unethical Research: How to Create a Malevolent Artificial Intelligence*. In proceedings of Ethics for Artificial Intelligence Workshop (AI-Ethics-2016). New York, NY. July 9–15, 2016. 1–7. https://arxiv.org/ftp/arxiv/papers/1605/1605.02817.pdf

Poole, D. – Mackworth, A. K. – Goebel, R. (1998): *Computational Intelligence: A Logical Approach*. Oxford University Press, https://www.researchgate.net/publication/220689680_Computational_Intelligence_A_Logical_Approach

Poole, D. – Mackworth, A. (2010): *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press, 2nd ed. 2017. https://artint.info/2e/html/ArtInt2e.html

Russell, S. (2016): Should We Fear Supersmart Robots? *Scientific American,* 314, 6, 58–59. DOI: 10.1038/Scientificamerican0616-58, https://people.eecs.berkeley.edu/~russell/papers/sciam16-supersmart.pdf

Tegmark, M. (2017): *Life 3.0: Being Human in the Age of Artificial Intelligence.* Knopf

Weld, S. D. (2016): Guest Commentary: *The Real Threat of Artificial Intelligence*. 23 May 2016. https://www.geekwire.com/2016/Guest-Commentary-Real-Threat-Artificial-Intelligence/

Yampolskiy, R. V. (2016): *Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures*. http://arxiv.org/abs/1610.07997

Zou, J. – Schiebinger, L. (2018): AI Can Be Sexist and Racist — It's Time To Make It Fair. *Nature,* 559, 7714, 324–326. DOI: 10.1038/D41586-018-05707-8, https://www.nature.com/articles/d41586-018-05707-8