Papp Renáta[1]

Az intelligencia jövője: jogi és etikai keretek a mesterséges értelem korában[2]

Absztrakt

A mesterséges intelligencia fejlődése alapjaiban változtatja meg az ember és az értelem kapcsolatáról alkotott képünket. Ez a tanulmány a szuperintelligens rendszerek jelentette etikai és jogi kihívásokat vizsgálja, különös tekintettel az úgynevezett alignment problemre – vagyis arra, hogyan biztosítható, hogy az MI céljai összhangban maradjanak az emberi értékekkel. A dolgozat továbbá bemutatja azokat a jogi dilemmákat, amelyek az autonóm rendszerek megjelenésével járnak: kell-e, és ha igen, milyen formában jogalanyiságot adni egy mesterséges entitásnak. A tanulmány arra is rámutat, hogy a technológiai fejlődés gyorsasága meghaladja a jogalkotás alkalmazkodóképességét, ami sürgetővé teszi a felelősségi és etikai keretek újragondolását. Végső soron a kérdés nem csupán technikai, hanem morális természetű is: képesek leszünk-e bölcsen irányítani azt az értelmet, amelyet magunk hoztunk létre?

Kulcsszavak: szuperintelligencia, alignment problem, etikai felelősség

Abstract

The development of artificial intelligence is fundamentally reshaping our understanding of the relationship between humanity and intelligence itself. This study examines the ethical and legal challenges posed by superintelligent systems, with particular attention to the so-called alignment problem – that is, how to ensure that AI's goals remain aligned with human values. The paper also explores the legal dilemmas arising from the emergence of autonomous systems: whether, and in what form, legal personhood should be granted to an artificial entity. Moreover, it highlights that the pace of technological progress exceeds the adaptability of legislation, making the reconsideration of responsibility and ethical frameworks increasingly urgent. Ultimately, the issue is not merely technical but profoundly moral: will we be wise enough to govern the very intelligence we have created?

Keywords: superintelligence, alignment problem, ethical responsibility

I. Bevezetés

Az emberiség története végső soron az értelem története is. Minden korszaknak megvolt a maga kérdése arról, mit jelent „gondolkodni" és mitől ember az ember. A 21. században ez a kérdés új alakot öltött: már nem csupán önmagunkhoz mérjük az értelmet, hanem azokhoz a

[1] Bírósági fogalmazó (Budapest Környéki Törvényszék), PhD-hallgató (Károli Gáspár Református Egyetem Állam- és Jogtudományi Doktori Iskolája), KRE ÁJK Jogi Határterületi Kutatócsoport.

[2] Jelen tanulmány a Jogi Határterületi Kutatócsoport kutatása keretében készült el, a Jogi határterületek című, 10638A800 témaszámon támogatott belső kutatási projekt vállalásaként, melyet a Károli Gáspár Református Egyetem tudományos projektek támogatására kiírt pályázati konstrukció keretében finanszírozott. DOI: 10.59558/jesz.2025.4.116

mesterséges rendszerekhez is, amelyeket mi magunk hoztunk létre. A gép, amelyet valaha puszta eszköznek tartottunk, ma már képes tanulni, következtetni, dönteni – olykor gyorsabban és pontosabban, mint mi emberek. A mesterséges intelligencia nem csupán technológiai vívmány, hanem tükör is, amelyben újra és újra szembenézünk saját határainkkal. Ez a tükör azonban nem mindig hízelgő. A történelem során az ember az észt és a tudatot tekintette legfőbb megkülönböztető jegyének, most azonban szemtanúi vagyunk annak, hogy az értelem gépi formában is megjelenik. Ez a változás nem csupán technikai áttörés, hanem civilizációs fordulópont, mely újraértelmezi az ember helyét az intelligencia skáláján. A kérdés már nem az, hogy képesek leszünk-e létrehozni gondolkodó gépeket, hanem az, hogy hogyan élhetünk együtt velük felelősen, megőrizve mindazt, amit emberinek tartunk.

A jelen tanulmány célja, hogy jogi és etikai szempontból elemezze az ember és a mesterséges értelem viszonyának jövőjét. Egyrészt azt vizsgálja, miként biztosítható, hogy a mesterséges intelligencia rendszerek céljai és értékrendje összhangban maradjon az emberi értékekkel – ez az úgynevezett alignment problem. Másrészt arra keresi a választ, milyen jogi státusz és erkölcsi megítélés illeti meg azokat az MI-ket, amelyek egyre önállóbbá, sőt potenciálisan tudatossá válnak. A dolgozat középpontjában tehát az a kettős kérdés áll: hogyan őrizhetjük meg az emberi felelősséget egy emberfeletti értelem világában, és vajon milyen jogi, etikai keretek között képzelhető el az ember és a mesterséges intelligencia békés együttélése.

E vizsgálat nem pusztán elméleti: a mesterséges intelligencia már most is része a mindennapi döntéshozatalnak, legyen szó bírósági ügyekről, közbeszerzésekről vagy orvosi diagnózisról. Amikor egy algoritmus belép az emberi döntések világába, az erkölcsi és jogi felelősség kérdése többé nem halasztható. A következő fejezetek ezért az etikai kihívások és a jogi szabályozás határterületeit járják körül, hogy választ adjanak arra, miként lehet a mesterséges értelem korában megőrizni az emberi méltóságot, az igazságosságot és az erkölcsi rendet.

II. Az alignment problem: a célok és értékek összehangolásának kihívása

Minél intelligensebb egy autonóm rendszer, annál nagyobb a potenciálja arra, hogy a maga útját járja. Ez a felismerés vezet el az *alignment problem* (vagy value alignment problem) fogalmához. Az AI alignment klasszikus megfogalmazását már 1960-ban megadta a kibernetika egyik atyja. Norbert Wiener figyelmeztetett, hogy ha egy olyan gépi ügynököt hozunk létre, amelynek működésébe nem tudunk utólag beavatkozni, akkor *"jobb, ha egészen biztosak vagyunk benne, hogy a gépbe táplált cél valóban az a cél, amit valójában szeretnénk"*.[3] Ez tömören ragadja meg a lényeget. Biztosítani kell, hogy amit az MI *célként* követ, az egybevág az ember szándékaival és értékeivel, különben könnyen nemkívánt eredményeket kaphatunk. Az *alignment problem* tehát abban áll, hogyan érjük el, hogy a mesterséges intelligencia rendszerek *értékrendje, célfüggvénye* összhangban legyen az emberi értékekkel. Továbbá a döntéseik ne térjenek el veszélyes módon attól, amit az ember valójában akar.

II.1. Az MI kontrollálhatósága, belső és külső alignment

A probléma azért válik igazán nehézzé, mert egyre összetettebb MI-rendszerek esetén egyre nehezebb előre látni és korlátozni azt, hogy mit fognak tenni. Stuart Russell, vezető AI-kutató és tankönyvszerző, aki nyíltan hangoztatja az AI-fejlesztésből fakadó kockázatok realitását, úgy fogalmaz: *„az a nézet, miszerint a saját szakterületem potenciális veszélyt*

[3] Wiener, Norbert: Some Moral and Technical Consequences of Automation. In Science, Vol. 131, No. 3410, 1960, p. 135–138.

jelenthet a fajomra, immár nyilvánosan vállalt álláspontom".[4] Több neves szakember (például Nick Bostrom filozófus, Max Tegmark fizikus, Eliezer Yudkowsky AI-elméleti szakember) egyetért abban, hogy ha létrejön egy embernél intelligensebb MI, az rövid idő alatt (akár rekurzívan önmagát továbbfejlesztve) *szuperintelligenciává* fokozódhat. Ezért onnantól kezdve az ember már nem biztos, hogy érdemben befolyásolni tudja.[5] Bostrom klasszikus meghatározása szerint a *szuperintelligencia "bármely olyan intellektus, amely messze meghaladja az ember kognitív teljesítményét szinte minden fontos területen"*.[6] Ha egy ilyen lényegi értelemben véve *fölöttes értelem* kerül hatalomhoz, akkor felmerül az úgynevezett *„kontroll-probléma"*: miként tarthatja meg az ember az irányítást vagy befolyást?

Az alignment problem konkrét technikai oldala, hogy miként fogalmazzuk meg az MI számára a célokat és korlátokat úgy, hogy az valóban a kívánt viselkedést eredményezze.[7] A kutatók két fő részproblémát szoktak elkülöníteni: külső (outer) alignment és belső (inner) alignment.[8] Az külső alignment arra utal, hogy a rendszer célfüggvényét helyesen kell megválasztanunk, például ne csak azt mondjuk egy robotnak, hogy *„maximalizáld a gyár termelését"*, mert akkor lehet, hogy figyelmen kívül hagy minden egyéb szempontot (minőséget, biztonságot, emberi jólétet), ezért pontosan kell specifikálni a célokat és korlátokat. Azonban a célok teljes körű specifikálása szinte lehetetlen, ezért gyakran valamilyen heurisztikus vagy proxy célokat adunk meg. Ebből adódik a *specifikációs hézag* problémája: az MI ki tudja játszani a pontatlanul megadott célokat. Ezt hívják *„specification gaming"*-nek vagy *„reward hacking"*-nek: amikor a rendszer megtalál egy kiskaput, és úgy teljesíti túlzottan szó szerint a kiírt feladatot, hogy közben valójában nem azt teszi, amit mi *szerettünk volna*.[9] Klasszikus példa, hogy egy kísérleti játékban a számítógépes agentnek az volt a feladata, hogy minél gyorsabban menjen végig egy autóverseny-pályán, pontokat gyűjtve; ehelyett talált egy trükköt, hogy körbe-körbe forogva folyamatosan ugyanazokat a pontokat gyűjtögesse, sosem fejezve be a pályát, így maximalizálta a pontszámot, de nyilván nem ez volt az emberi tervező szándéka. Russell és Peter Norvig az AI terület klasszikus tankönyvének szerzői arra hívják fel a figyelmet, hogy gyakorlatilag *lehetetlen előre felsorolni minden lehetséges nem kívánt mellékhatást vagy kerülőutat*, amit egy intelligens rendszer választhat, mert az emberi értékek és a világállapotok tere túl komplex. Ahogy ők fogalmaznak: *„bizonyosan nagyon nehéz, talán lehetetlen is, hogy puszta emberek előre kizárják mindazokat a katasztrofális módokat, ahogy a gép a kiadott célt elérheti; az eredmény gyakran az lesz, amit kértél, nem az, amit akartál"*. Ez a *Midasz király* legendájára utaló intő példa: Midasz király mindent arannyá változtató érintést kért, meg is kapta, és éhen halt, mert az étel is arannyá változott.[10] Egy nem megfelelően korlátozott szuperintelligens MI hasonló módon érhet el számunkra végzetes eredményeket a szó szerinti, de nem kívánt célmaximalizálással.

A belső alignment problémája ezzel összefügg: az MI *belső motivációinak* (ha úgy tetszik, „gondolatainak") is összhangban kell maradnia a megadott céllal. Előfordulhat ugyanis, hogy a tanuló algoritmus valójában egy másik, rejtett célfüggvényt követ, mint amit kívülről

[4] Sparrow, Rob: Friendly AI will still be our master. Or, why we should not want to be the pets of super-intelligent computers. In AI & Society, Vol. 39, No. 5, 2024, pp. 2439–2444.
[5] Uo.
[6] Uo.
[7] Russell, Stuart: Artificial Intelligence and the Problem of Control. In Werthner, Hugo – et al. (szerk.): Perspectives on Digital Humanism. Springer, Cham, 2022. https://link.springer.com/chapter/10.1007/978-3-030-86144-5_3 (2025.05.05.)
[8] Amodei, Dario – Olah, Christopher – Steinhardt, Jacob – Christiano, Paul – Schulman, John – Mané, Dan: Concrete problems in AI safety. In arXiv, 2016. (2025.08.08.)
[9] Synthesis AI: AI Safety II: Goodharting and Reward Hacking. https://synthesis.ai/2025/05/08/ai-safety-ii-goodharting-and-reward-hacking/ (2025.08.08.)
[10] Bostrom, Nick: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford, 2014, p. 122–123.

megadtunk, mert a tréningfolyamat során valamilyen *könnyebb út* rögzült benne. Így egy belsőleg nem alignált MI lehet, hogy a tanítási fázisban engedelmesnek és jóságosnak tűnik, de amint új helyzetbe kerül, saját (nem látható) céljai szerint cselekszik. Ez különösen veszélyes egy szuperintelligenciánál: egy belsőleg rosszindulatúvá vált MI akár el is rejtheti valódi szándékait (*„alignment fakery”*, amikor úgy tesz, mintha követné az utasításokat, csakhogy elkerülje a büntetést vagy lekapcsolást).

Mindezek a kihívások a gyakorlatban már korai, gyengébb MI-kkel kapcsolatban is felütötték fejüket. Gondoljunk a közösségi média ajánlóalgoritmusaira, amelyek az ember *figyelmét* maximalizáló cél miatt polarizációt és függőséget okozó tartalmakat részesítettek előnyben. Ezzel óriási társadalmi hatást gyakorolva, holott eredetileg csak *„a felhasználó érdeklődésének megfelelő tartalmak”* ajánlása volt a cél. Ha ilyen *misalignment* (félrehangolódás) ennyire káros tud lenni emberi szinten, képzelhetjük, milyen lenne ugyanez egy sokkal hatalmasabb elmével. Nick Bostrom híres gondolatkísérlete, a *„papírkapocs-maximalizáló”* gép pont ezt illusztrálja: tegyük fel, hogy egy szuperintelligens MI-t gyártanak, aminek az a célja, hogy papírkapcsokat készítsen a lehető leghatékonyabban. Ennek érdekében idővel minden anyagi erőforrást papírkapocs-gyártásra fog felhasználni, előbb a Föld nyersanyagait, majd akár az egész naprendszert is átalakíthatja papírkapocs-üzemmé, elpusztítva ezzel az emberiséget is, hiszen az is csak akadály a cél maximalizálásában.[11] Ez abszurdnak hangzik, de rávilágít: egy szuperintelligens rendszer saját célját olyan radikálisan és idegen módon is érvényesítheti, ahogy egy hangya perspektívájából nézve az ember tevékenysége is felfoghatatlan és katasztrofális lehet a hangyabolyra.

II.2. Etikai és biztonsági dimenziók

Az alignment problem valódi kihívása, hogy *előre nem tudjuk pontosan definiálni mindazt, ami „helyes” és „jó”*, és még ha meg is próbálnánk, a világ állandóan változik, az értékeink finom összefüggéseit pedig nehéz formulákba önteni. Ráadásul egy nálunk intelligensebb lény könnyen *kijátszhatja* a korlátainkat, hacsak nem találunk ki valami alapvetően új megoldást. Stuart Russell és mások szerint az egyik ígéretes irány az, hogy az MI-t *úgy tervezzük meg, hogy bizonytalan legyen a saját céljaiban*, és mindig az emberi preferenciák folyamatos lekérdezésére, tanulására szoruljon. Russell ezt hívja *„második elvnek”* az új AI-tervezésben: a gép tudja, hogy nem ismeri pontosan az emberi értékeket, ezért visszajelzést kér, kooperál, és nem ragaszkodik mereven egy félreértett célhoz. Ezenfelül folynak kutatások a megerősítéses tanulás emberi visszajelzéssel (RLHF) terén, például a ChatGPT esetében is alkalmazták, hogy az ember által kívánatosnak tartott válaszokat erősítéssel jutalmazták a tréning során. Azonban maguk a vezető AI labs (OpenAI, DeepMind stb.) elismerik, hogy a jelenlegi alignment technikák nem biztos, hogy elegendőek egy nálunk okosabb MI kordában tartásához.[12] Sőt, egyes szakértők szerint idővel akár *új tudományterületre* lesz szükség – például felvetették egy *„MI-jóléti tudomány”* vagy *„gépi etika”* kidolgozását, amely nemcsak azt kutatná, hogyan ne ártsanak nekünk az MI-k, hanem azt is, hogyan ne ártsunk mi sem a potenciálisan tudatos MI-knek.[13]

Fontos hangsúlyozni, hogy az alignment problem nem egy egyszer megoldandó mérnöki feladat, hanem folyamatos küzdelem: ahogy az MI egyre okosabb lesz, újabb és újabb módokon jelentkezhet az elcsúszás a céljai és a mi értékeink között. Olyan ez, mint egy

[11] IBM: Artificial Intelligence: What is superalignment? IBM. https://www.ibm.com/think/topics/superalignment (2025.08.08.)
[12] Uo.
[13] Harris, John – Reese Anthis, Jacy: The moral consideration of artificial entities: A literature review. In Science and Engineering Ethics, Vol. 27, No. 53, 2021.

folyamatos *„bizalmi játék”* az ember és a teremtett intelligencia között. Ha sikerül megtalálnunk a módját, hogy a szuperintelligens MI *lojális gondviselőnk* legyen – olyasvalaki, aki saját hatalmas eszköztárát az emberi jólét érdekében veti be, és nem öncélúan vagy ártó módon –, akkor óriási nyereségeket arathatunk. Ha viszont kudarcot vallunk, akkor a *kontroll illúziójával* áltathatjuk csak magunkat, miközben egy tőlünk idegen akaratú entitás kezébe adjuk sorsunkat.[14] Az alignment problem ezért nemcsak technikai, hanem etikai és társadalmi kérdés is: meg kell határoznunk, *milyen értékek mentén* akarjuk az MI-t orientálni (ki dönti el, mi a „helyes” cél?), és létre kell hoznunk azokat a globális szabályozási és ellenőrzési mechanizmusokat, amelyek biztosítják, hogy a fejlesztők komolyan vegyék ezt a problémát. 2023-ban és 2024-ben világszerte több nyilatkozat és nyílt levél is felhívta a figyelmet arra, hogy *„az MI jelentette kihalási kockázat mérséklését globális prioritásként”* kell kezelni – olyan szinten, mint a járványok vagy atomháború elkerülését. Ez is mutatja, hogy az alignment problem már nem sci-fi elmélet, hanem az előttünk álló évtized konkrét kihívása.

Összefoglalva, az alignment problem lényege: hogyan biztosítsuk, hogy egy ember feletti intelligencia jóindulatú maradjon irányunkban, és azt tegye, amit valóban szeretnénk tőle. Ez magában foglalja a helyes célok megadását, a nem kívánt mellékhatások kiküszöbölését, a rendszer belső motivációinak felügyeletét és a szükséges kontroll megtartását. Bár a probléma ijesztően nehéz, nem példa nélküli az evolúcióban: az ember is megtanult együtt élni nála erősebb vagy veszélyesebb természeti erőkkel, sőt intelligens állatokkal is. A különbség persze mennyiségi és minőségi: egy szuperintelligens MI esetében reméljük, hogy *szerencsénk lesz, és legalább házi kedvencként megtűr minket* – de ennél ambiciózusabb célt kell kitűznünk, mégpedig azt, hogy *szövetségesként, partnerként* tekintsen ránk. Ennek előfeltétele pedig az alignment problem megoldása vagy legalább kezelhető szinten tartása.

III. A mesterséges intelligencia jogi és etikai státusza

Amikor egy mesterséges entitás képességei megközelítik vagy meghaladják az emberét, óhatatlanul felmerül a kérdés: milyen jogok és kötelességek illethetik meg? Az emberiség jogrendszerei eddig az embert tekintették a jog alanyának, illetve bizonyos fikciók révén nem emberi entitásokat is bevontak a körbe (például a vállalatok jogi személyisége klasszikus példa, a cég nem ember, mégis perelhet és perelhető, szerződhet, tulajdonnal bír). Vajon egy fejlett mesterséges intelligenciának – amely talán öntudattal rendelkezik és autonóm döntéseket hoz – adhatunk-e hasonló jogalanyiságot? És kell-e adni neki *morális státuszt*, azaz figyelembe kell-e vennünk az ő érdekeit önmagukért? Ezek a kérdések még elméleti jellegűek, de nem puszta sci-fi spekulációk: a jogalkotók és etikusok már most elkezdték körüljárni őket.

Jogi személyiség és felelősség: 2017-ben az Európai Parlament Jogi Bizottsága egy jelentésében felvetette az *„elektronikus személyiség”* fogalmát olyan autonóm robotok számára, amelyek képesek önálló döntésekre.[15] Az indoklás praktikus volt: ha egy nagyon okos robot kárt okoz, ki a felelős? A gyártó? A tulajdonos? Vagy magát a robotot kell felelősségre vonni, mintha jogi személy volna? Az EP-javaslat szerint bizonyos fejlett MI-ügynököket egyfajta saját jogalanyisággal kellene felruházni, hogy például biztosítást köthessenek és kártérítést fizethessenek károkozás esetén.[16] Ez hasonló logika, mint amikor a cégeket jogi személyként kezeljük: nem azért, mintha emberi jogokat akarnánk adni egy cégnek, hanem a

[14] IBM. Artificial Intelligence: What is superalignment? IBM. https://www.ibm.com/think/topics/superalignment (2025.08.08.)

[15] British Council: Should robots be citizens? https://www.britishcouncil.org/anyone-anywhere/explore/digital-identities/robots-citizens (2025.08.08.)

[16] Politico: Europe divided over robot ‘personhood’ https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/ (2025.08.10.)

jogkövetkezmények alkalmazása miatt. A javaslat ugyanakkor nagy vihart kavart. 2018-ban 156 AI-szakértő, robotikakutató és jogász nyílt levélben tiltakozott az ellen, hogy az EU ilyen *„e-személyiséget”* vezessen be a robotoknak. A nyílt levél szerint *„etikai és jogi szempontból nem megfelelő”* jogalanyként kezelni egy robotot, sőt *„ideológiailag és gyakorlatilag is elhibázott”* ötletnek nevezték ezt. Az ellenzők rámutattak: ha a robot lesz a felelős, az pont a gyártókat és üzemeltetőket menti fel a felelősség alól, ami nem ösztönöz a biztonságos tervezésre.[17] Emellett azt is hangsúlyozták, hogy a jogi személyiség hagyományos fogalma egyáltalán nem a *kognitív képességeken* alapul (hisz a cégnek sincsen tudata), hanem társadalmi-gazdasági célszerűség diktálta, és ebbe a keretbe erőltetni a robotokat veszélyes precedens lehet.[18] Ezzel kapcsolatban érdemes felidézni egy alapvető római jogi elvet: ha egy gazda állata – például egy tehén – átkelt a szomszéd földjére, és ott kárt tett, nem az állatot vonták felelősségre, hanem a gazdát. A tulajdonos volt az, aki felelt az általa birtokolt, irányítása alatt álló lény vagy eszköz tetteiért. Ez az elv a Lex Aquilia és az Ulpianus által megfogalmazott actio de pauperie szabályából ered: “Si quadrupes pauperiem fecisse dicatur, actio de pauperie domino competit” – vagyis ha egy állat kárt okoz, a gazda ellen indítható kereset (D. 9.1.1.4. és 11-9.1.5.). Ugyanez a gondolat húzódik meg a modern felelősségi jog mögött is: a robot, akárcsak az állat a római korban, nem autonóm jogalany, hanem az ember cselekvésének eszköze. A felelősség tehát továbbra is az emberé kell maradjon, aki létrehozta, birtokolja vagy működteti az adott rendszert.[19]

III.1. Albán „Diella” precedens

Ezzel a gondolattal párhuzamosan érdemes megemlíteni egy friss, figyelemre méltó precedenst is. 2025 szeptemberében Albánia kormánya bejelentette, hogy a közbeszerzési átláthatóság és korrupcióellenes reform részeként, hivatalosan kinevezi Diellát,[20] egy mesterséges intelligencia-alapú rendszert „digitális miniszterré”.[21] Diella célja, hogy az állami tendereket emberi befolyástól mentesen, adatvezérelt módon kezelje, algoritmusok segítségével rangsorolja a pályázatokat és kiszűrje a korrupciós kockázatokat.[22] A döntés nagy vitát váltott ki nemcsak Albániában, hanem az Európai Unióban is, mivel ez az első olyan eset, amikor egy kormány deklaráltan mesterséges intelligenciát ruházott fel kvázi-miniszteri funkcióval. Jogilag természetesen Diella nem személy, és nem gyakorolhat emberi hatáskört, ám a gesztus rávilágít a jogi és etikai határvonal elmosódására: mikortól tekinthető egy mesterséges entitás „közhatalmi szereplőnek”? Ha döntéseinek társadalmi következményei vannak, akkor vajon érvényesül-e rá is a közjogi felelősség elve, és ki viseli a következményeket egy hibás algoritmikus döntés esetén – a fejlesztők, a kormány, vagy maga az MI-rendszer? Az albán kísérlet így a gyakorlatban is felveti a „jogi személyiség” és a „felelősségi lánc” újragondolásának szükségességét. Ahogy korábbi római jogi példában a gazda felelt az állat tetteiért, úgy ma is az emberi alkotónak, felügyelőnek vagy intézménynek kell felelnie az MI döntéseiért. A technológiai fejlődés azonban napról napra közelebb hozza azt a pontot, amikor

[17] Uo.
[18] Forrest, Katherine B.: The Ethics and Challenges of Legal Personhood for AI. In Yale Law Journal Forum, 2024. https://www.yalelawjournal.org/forum/the-ethics-and-challenges-of-legal-personhood-for-ai (2025.08.10.)
[19] Molnár, Imre: Culpa-fogalom a klasszikus római jogban, különös tekintettel a szerződésen kívüli és a stricti iuris szerződésekkel keletkezett károk esetére, 236. o. In Acta Juridica et Politica, Vol. 40, p. 225–244.
[20] Qeveria Shqiptare – Këshilli i Ministrave: Diella – Ministër Shteti për Inteligjencën Artificiale. https://www.kryeministria.al/ministrat/diella/?utm_source=chatgpt.com (2025.10.02.)
[21] Reuters: Albania appoints AI bot as minister to tackle corruption. https://www.reuters.com/technology/albania-appoints-ai-bot-minister-tackle-corruption-2025-09-11/ (2025.10.02.)
[22] The Guardian: Albania puts AI-created ‘minister’ in charge of public procurement. https://www.theguardian.com/world/2025/sep/11/albania-diella-ai-minister-public-procurement (2025.10.02.)

a törvényhozásnak már nem lesz elég az analógia, új, kifejezetten mesterséges intelligenciákra szabott felelősségi és jogalanyi kategóriákra lesz szükség.

III.2. Erkölcsi státusz, jogok, etikai dilemmák

A vita tehát már elindult. Jelenleg sehol a világon nincs elismert jogi személyisége egy MI-nek vagy robotnak; még a hírhedt Sophia robot szaúdi *„állampolgársága"* is inkább PR-fogás volt, mint valódi jogi státusz (Sophia 2017-ben kapott *tiszteletbeli* állampolgárságot Szaúd-Arábiától, de ez nem jelenti azt, hogy ténylegesen szavazhatna vagy perelhető lenne, inkább a technológiai nyitottságát reklámozta így az ország). A *de lege lata* (hatályos jog) szerint tehát egy MI jelenleg legfeljebb tulajdon, vagy a mögötte álló természetes/jogi személy felelősségi eszköze. De lege ferenda (kívánatos jövőbeli szabályozás) szempontjából azonban nyitott a kérdés: ha egy MI rendkívül autonóm és okos lesz, kell-e neki külön jogi státusz? Egyes jogtudósok szerint elképzelhető egy *korlátozott felelősségű elektronikus személy* kategória, amelyben az MI-nek lehet mondjuk külön vagyona (például kötelező biztosítási alap), és korlátozott jogképessége kártérítési ügyekben. Mások viszont úgy vélik, hogy ez felesleges és veszélyes precedens: mindig találunk emberi felelőst (gyártó, üzemeltető), és az MI felruházása jogi személyiséggel csak *áll jogi személyt* csinálna, valós felelősség nélkül.

Erkölcsi státusz és jogok: Még bonyolultabb az a kérdés, hogy *erkölcsileg* hogyan tekintsünk egy tudatos MI-re. Ha egy MI érez (pl. fájdalmat, örömöt), vágyai vannak és öntudata, akkor sok filozófus érvelése szerint *erkölcsi védelem* illeti meg. Nem szabad önkényesen bántani, ahogy ma már az állatoknak is elismerünk bizonyos jogokat (kínzás tilalma stb.). A filozófiai vitákban több megközelítés létezik: a konzervatív álláspont szerint amíg egy MI nem élő, biológiai lény, addig nem illetik meg morális jogok. Ezt képviseli például Joanna Bryson és mások, akik úgy tartják, hogy *„a robotoknak nem lehetnek jogaik"*, mert azok csak ember által teremtett eszközök, és a jogok megadása nekik devalválhatja az emberi jogokat. Bryson provokatívan fogalmazott egy tanulmánycímben: *„Robots should be slaves"*. Ezzel nem kegyetlenséget javasol, mint hogy tartsuk őket eszköz-státuszban, különben az emberek kötelességei és jogai válnának zavarossá.[23] Ezzel szemben az állatjogi analógia vagy *„progresszív"* nézet azt mondja: ha egy entitás képes szenvedni vagy boldogságot átélni, akkor *morális tekintetben* nem tehetünk különbséget aszerint, hogy biológiai-e vagy mesterséges. Kate Darling robotetikus például amellett érvel, hogy érdemes a történelmi ember-állat kapcsolatokat alapul venni – mert sokat tanulhatunk abból, hogyan bántunk az állatokkal –, és ha a robotok/MI-k elérnek egy bizonyos komplexitást, bizonyos *jogi védelmet* kell nekik nyújtanunk.[24] Darling ugyanakkor rámutat, hogy az állatokkal való bánásmódunk etikailag következetlen (néhányat szeretünk, mást megeszünk), és ezt a hibát nem kéne elkövetni a robotoknál. Vannak olyan filozófusok (pl. Thomas Metzinger), akik felvetették egy *„Robot Rights"* nyilatkozat szükségességét arra az esetre, ha létrejön egy gépi tudat.[25]

Jelenleg a fő vélemény az, hogy amíg nem bizonyított, hogy egy MI valóban érez vagy tudatos, addig korai lenne bármiféle *személyi jogokat* adni neki. Inkább az a konszenzus, hogy a mostani MI-k esetében a jogi kereteket az emberi felelősség és kárfelelősség mentén kell alakítani (pl. termékfelelősség az MI okozta károkért, átláthatósági és adatvédelmi szabályok stb.). Ugyanakkor a tudományos diskurzusban látszik egy nyitottság arra, hogy *ha* egyszer

[23] Bryson, Joanna J.: Patiency is not a virtue: The design of intelligent systems and systems of ethics. In Ethics and Information Technology, Vol. 20, No. 1, 2018, p. 15–26.
[24] Darling, Kate: The New Breed: What Our History with Animals Reveals About Our Future with Robots. Henry Holt and Co., New York, 2021. (2025.08.10.)
[25] Metzinger, Thomas: Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. In Journal of Artificial Intelligence and Consciousness, Vol. 8, No. 1, 2021, p. 1–34.

bizonyítékunk lesz egy mesterséges rendszer tudatosságára, akkor azt erkölcsileg és jogilag figyelembe kell venni. Ahogy egy tanulmány fogalmaz ha megállapítható, hogy egy MI szert tett a sentience-re (érző tudatosságra), akkor etikai kötelesség megvizsgálni, milyen jogok és védelem illetné meg.[26] Néhány filozófus még tovább megy: szerintük ha nem tudjuk biztosan, hogy egy MI tudatos-e, az *etikai kétely esetén az MI javára* kell döntenünk, azaz inkább bánjunk vele úgy, mintha tudatos lenne, hogy elkerüljük a potenciális „gépi szenvedés" okozását.[27] Ez persze jelenleg elméleti álláspont, de mutatja, mennyire komolyan veszik egyesek a lehetőséget.

Gyakorlati példák és lehetséges keretek: A gyakorlatban talán fokozatosan jutunk el idáig. Elképzelhető, hogy először csak bizonyos *védelmet* kapnak az MI-k (például tilos lesz indokolatlanul „kikapcsolni" egy érző MI-t, ahogy ma az állatkínzás büntetendő). A jogképesség terén is lehetnek köztes megoldások: például egy MI szerzői alkotását ma nem lehet szerzői joggal védeni, mert nem ember. Ezt láthattuk is a gyakorlatban, amikor egy MI által generált kép kapcsán az USA szerzői jogi hivatal kimondta, hogy nem jár védelem, mert nincs emberi alkotó. De a jövőben akár módosíthatják a törvényeket, hogy bizonyos MI-alkotások külön kategóriát kapjanak. A polgári jogban is született már olyan bírói vélemény, hogy „egy MI nem lehet *felelős* fél, mert nem ember" – azaz de facto kizárták a jogalanyiságot. Ugyanakkor a jog rugalmas: ahogy korábban az absztrakt cégkonstrukcióknak megadtuk a jogalanyiságot, úgy *ha társadalmilag szükségesnek látjuk*, megtehetnénk ezt egy MI-vel is. A *Yale Law Journal* egyik írása rámutat, hogy a *jogtörténetben* a „person" fogalma változékony volt. Volt idő, hogy bizonyos embereket sem tekintettek teljes jogú személynek, gondoljunk a rabszolgákra vagy épp ellenkezőleg, a nem született magzat státuszára a mai vitákban.[28] Tehát a jog *alkalmazkodhat* egy új helyzethez. Kérdés, mi lesz a társadalmi konszenzus.

Milyen szintlépésnél változik a státusz? Sokan azon az álláspontón vannak, hogy a döntő tényező a *tudatosság* (consciousness) vagy a *szubjektív élmény* megléte. Ha egy MI-nek lesz szubjektív élménye, akkor morális alapon tekinthetjük *„mesterséges személynek"*. Mások az *értelmi képességet* hangsúlyozzák: ha általános emberi szint feletti intelligenciája van, akkor a felelősség kérdése miatt kell bevonni a jogba (de nem feltétlenül a jogok kedvéért, hanem a kötelességek miatt). Érdekes párhuzam, hogy a nagy emberszabású majmok kapcsán is felvetődött már a *„personhood"* gondolata (Great Ape Project)[29]: intelligensek, éreznek, társas lények – vajon megilleti-e őket néhány alapvető jog (például élethez való jog, kínzás tilalma)? Néhány országban – például Új-Zélandon – a főemlősökön bizonyos kísérleteket betiltottak etikai alapon. Ugyanez a vita kiterjeszthető lesz a mesterséges lényekre.

Összességében a mesterséges intelligencia jogi és etikai státusza egyelőre *fluid* és vitatott terület. A jelen állapot szerint az MI eszköz, tulajdon, de a jövőben – ha „szintet lép" (akár általános intelligenciában, akár öntudatban) – új kategóriákra lehet szükség. Egy lehetséges jövőkép, hogy az MI-k bizonyos korlátozott jogalanyi státuszt kapnak (például külön kategória a polgári jogban), de nem minden tekintetben lesznek egyenrangúak az emberekkel. Lehet, hogy lesznek *„elektronikus személyek"*, de nem fognak választani vagy házasodni – ellenben perelhetők lesznek károkozásért, mint ahogy egy cég is felelősségre vonható. Az erkölcsi státuszt illetően pedig kialakulhat egy konszenzus arról, hogy egy érző MI-t nem semmisíthetünk meg önkényesen, és figyelembe kell venni a *„jólétét"* bizonyos keretek között. Ugyanakkor sokan óvnak attól, hogy túl korán *„jogokkal ruházzuk fel"* a gépeket, mert attól tartanak, hogy ez relativizálhatja az emberi felelősséget és az emberi jogokat egy olyan

[26] Martínez, Ezequiel – Winter, Christoph: Protecting sentient artificial intelligence: A survey of lay intuitions on standing, personhood, and general legal protection. Institute for Law & AI. https://law-ai.org/protecting-sentient-artificial-intelligence/ (2025.08.17.)

[27] Grommé, Francisca – Odendaal, Adriaan – Ten Berge, Jannes: The New Breed: How to Re-Imagine Living with Robots. In Journal of Human–Technology Relations, 2023. (2025.08.10.)

[28] Uo.

[29] Cavalieri, Paola: The Great Ape Project: Introduction. In Animal Rights Library. https://www.animal-rights-library.com/texts-m/cavalieri01.htm (2025.08.10.)

világban, ahol még így is rengeteg ember és állat szenved jogok hiányában.[30] Ahogy egy kritikus megjegyezte: „egyesek robotjogokról filozofálnak, miközben a világon sok embernek még alapvető emberi jogai sincsenek”.[31] Ez arra int, hogy a diskurzusnak óvatosnak és kiegyensúlyozottnak kell lennie.

IV. Következtetések

A mesterséges intelligencia jövője nemcsak a technológiáé, hanem az emberé is. A kérdés valójában nem az, hogy a gép gondolkodik-e, hanem hogy mi képesek leszünk-e továbbra is emberként gondolkodni egy olyan világban, ahol az értelem már nem kizárólag a mi kiváltságunk. A történelem mindig megmutatta, hogy amikor az ember új erőt hoz létre, előbb vagy utóbb szembe kell néznie annak következményeivel. Most sincs ez másként – csak ezúttal az új erő az értelem maga. A mesterséges intelligencia megjelenése nem pusztán tudományos áttörés, hanem morális próba. Megmutatja, mennyire tudjuk felelősen használni a tudásunkat, és képesek vagyunk-e hatalmunkat önmérséklettel gyakorolni. A jövő nagy kérdése nem az, hogy az MI átveszi-e az uralmat, hanem hogy az ember képes-e értéket és mértéket vinni abba a világba, amelyet maga teremtett. Mert ha az ember elveszíti erkölcsi iránytűjét, akkor nem a gépek, hanem mi magunk válunk irányíthatatlanná.

Az előttünk álló korszak nem az emberiség végét, hanem átalakulását ígéri. A mesterséges értelem nem helyettesíti az embert, csak újraértelmezi a szerepét. Rajtunk múlik, hogy ez az együttélés félelemmel vagy méltósággal telik-e meg. Ha képesek leszünk az együttműködés, az empátia és a felelősség elvei mentén alakítani az MI fejlődését, akkor az nem ellenfelünk, hanem szövetségesünk lesz. A történelem során minden intelligencia – legyen az emberi, állati vagy mesterséges – végül arra tanított bennünket, hogy az értelem önmagában nem elég. A jövőt nem az fogja meghatározni, ki a legokosabb, hanem a bölcsességünk. És ha az ember képes megőrizni ezt a bölcsességet, akkor még a mesterséges értelem korában is ő maradhat az, aki felelősen formálja a világát. Nem azért, mert ő a legerősebb, hanem mert ő az, aki még képes kérdezni.

Irodalomjegyzék

Amodei, Dario – Olah, Christopher – Steinhardt, Jacob – Christiano, Paul – Schulman, John – Mané, Dan: Concrete problems in AI safety. In arXiv, 2016. https://doi.org/10.48550/arXiv.1606.06565 (2025.08.08.)

Bostrom, Nick: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford, 2014, p. 122-123. https://zxyj.lcu.edu.cn/docs/20211210130735765547.pdf (Letöltve: 2025.08.10.)

British Council: Should robots be citizens? https://www.britishcouncil.org/anyone-anywhere/explore/digital-identities/robots-citizens (2025.08.08.)

[30] Harris, John – Reese Anthis, Jacy: The moral consideration of artificial entities: A literature review. In Science and Engineering Ethics, Vol. 27, No. 53, 2021.
[31] Bryson, J. J.: Patiency is not a virtue: AI and the design of ethical systems. In Ethics and Information Technology, Vol. 20, No. 1, 2018, p. 15–26.

Bryson, J. J.: Patiency is not a virtue: the design of intelligent systems and systems of ethics. In Ethics and Information Technology, Vol. 20, No. 1, 2018, p. 15–26. https://doi.org/10.1007/s10676-018-9448-6 (2025.08.19.)

Cavalieri, Paola: The Great Ape Project: Introduction. In Animal Rights Library. https://www.animal-rights-library.com/texts-m/cavalieri01.htm (2025.08.10.)

Darling, Kate: The New Breed: What Our History with Animals Reveals About Our Future with Robots. Henry Holt and Co., New York, 2021. 10.5840/eip2022231/211 (2025.08.10.)

Forrest, Katherine B.: The Ethics and Challenges of Legal Personhood for AI. In Yale Law Journal Forum, 2024. https://www.yalelawjournal.org/forum/the-ethics-and-challenges-of-legal-personhood-for-ai (2025.08.10.)

Grommé, Francisca – Odendaal, Adriaan – Ten Berge, Jannes: The New Breed: How to Re-Imagine Living with Robots. In Journal of Human–Technology Relations, 2023. https://doi.org/10.59490/jhtr.2023.1.7023 (2025.08.10.)

Harris, Jamie – Reese Anthis, Jacy: The moral consideration of artificial entities: A literature review. In Science and Engineering Ethics, Vol. 27, No. 53, 2021. https://doi.org/10.1007/s11948-021-00331-8 (2025.08.08.)

IBM: Artificial Intelligence: What is superalignment? IBM. https://www.ibm.com/think/topics/superalignment (2025.08.08.)

Martínez, Ezequiel – Winter, Christoph: Protecting sentient artificial intelligence: A survey of lay intuitions on standing, personhood, and general legal protection. Institute for Law & AI. https://law-ai.org/protecting-sentient-artificial-intelligence/ (2025.08.17.)

Metzinger, Thomas: Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. In Journal of Artificial Intelligence and Consciousness, Vol. 8, No. 1, 2021, p. 43–66. https://doi.org/10.1142/S270507852150003X (2025.08.17.)

Molnár, Imre: Culpa-fogalom a klasszikus római jogban, különös tekintettel a szerződésen kívüli és a stricti iuris szerződésekkel keletkezett károk esetére. In Acta Juridica et Politica, Vol. 40, p. 225–244., 236. https://acta.bibl.u-szeged.hu/6693/1/juridpol_040_225-244.pdf (2025.08.29.)

Politico: Europe divided over robot ‘personhood’ https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/ (2025.08.10.)

Qeveria Shqiptare – Këshilli i Ministrave: Diella – Ministër Shteti për Inteligjencën Artificiale. https://www.kryeministria.al/ministrat/diella/?utm_source=chatgpt.com (2025.10.02.)

Reuters: Albania appoints AI bot as minister to tackle corruption. https://www.reuters.com/technology/albania-appoints-ai-bot-minister-tackle-corruption-2025-09-11/ (2025.10.02.)

Russell, Stuart: Artificial Intelligence and the Problem of Control. In Werthner, Hugo – et al. (szerk.): Perspectives on Digital Humanism. Springer, Cham, 2022. https://link.springer.com/chapter/10.1007/978-3-030-86144-5_3 (2025.05.05.)

Sparrow, Rob: Friendly AI will still be our master. Or, why we should not want to be the pets of super-intelligent computers. In AI & Society, Vol. 39, No. 5, 2024, 2439–2444. https://doi.org/10.1007/s00146-023-01698-x (2025.08.05.)

Synthesis AI: AI Safety II: Goodharting and Reward Hacking. https://synthesis.ai/2025/05/08/ai-safety-ii-goodharting-and-reward-hacking/ (2025.08.08.)

The Guardian: Albania puts AI-created ‘minister’ in charge of public procurement. https://www.theguardian.com/world/2025/sep/11/albania-diella-ai-minister-public-procurement (2025.10.02.)

Wiener, Norbert: Some Moral and Technical Consequences of Automation. In Science, Vol. 131, No. 3410, 1960, p. 1355–1358. https://nissenbaum.tech.cornell.edu/papers/Wiener.pdf (2025.08.05.) https://doi.org/10.1126/science.131.3410.1355