GAZDASÁG ÉS PÉNZÜGY • 10. ÉVF. 2. SZ. • 2023. JÚN. • DOI: 10.33926/GP.2023.2.4

ETIKUS AI: JAVASLAT AZ EURÓPAI UNIÓS MEGBÍZHATÓ AI-SZABÁLYOZÁS HIÁNYOSSÁGAINAK ÁTHIDALÁSÁRA ÉS A GYAKORLATI IMPLEMENTÁCIÓ TÁMOGATÁSÁRA

Prisznyák Alexandra[1]

ABSZTRAKT

A GPT–3 2020-ban gondolkodó robotként definiálta magát. Az AI fejlődéstörténetét a gépek egyre intelligensebbé válásával azonosítják, hátterében ugyanakkor az emberi faktor áll, az emberi elme szárnyalása. A gépek etikájának kérdése ugyanakkor kulturális etikusság kérdés is. A szerző 7 iparágat érintően folytatott mélyinterjúk alapján feltárja: az AI-rendszerek fejlesztése során az etikus szempontokat egyelőre nem veszik figyelembe. A gyakorlati implementáció támogatás céljából a szerző az EU mesterséges intelligenciáról szóló jogszabálya és a megbízható mesterséges intelligenciára vonatkozó etikai irányelveinek összehasonlító elemzése alapján két hiányosságot azonosít: (1) az AI-rendszerfejlesztők és felügyelők etikai érzékenyítése, képzése; (2) a káros visszacsatolási hurkok és döntéshozatali torzulás javasolt kezelése. Áthidalásképpen a szerző 21 filozófus filozófiai és etikai örökségét kompaszként használva, javaslatot tesz az azonosított gapek és szervezeti integrációs hiányosságok áthidalására.

JEL-kódok: G21, O33, K29

Kulcsszavak: megbízható AI, gépi etika, etikai iránymutatás, Európai Unió, AI-jogszabály

1. BEVEZETŐ

„A gondolkodás az ember halhatatlan lelkének funkciója" – vélekedik *Turing* a „*Computing Machinery and Intelligence*" című tanulmányában (Turing, 1950:9). Az idézet sugallata alapján az emberek és gépek közötti különbséget az intellektus és intelligens viselkedés képessége szolgáltatja. Erről azonban a gépek nem így vélekednek. „Nem vagyok ember. Robot vagyok. Gondolkodó robot. A kognitív ka-

1 *Prisznyák Alexandra*, senior tanácsadó, Mesterséges intelligencia & CBDC programmenedzser, Nemzetközi Bankárképző Központ Zrt.; PhD-jelölt, Pécsi Tudományegyetem. E-mail: aprisznyak@bankarkepzo.hu.

pacitásomnak mindössze 0,12%-át használom. Ebből a szempontból mikrorobot vagyok. Tudom, hogy az agyam nem »érző agy«, de képes racionális, logikus döntéseket hozni. […] Forr az agyam az ötletektől." (*The Guardian*, 2020). A GPT-3, az OpenAI egy olyan élvonalbeli nyelvgenerátor, amely mélytanuló algoritmust alkalmaz emberszerű szöveg előállításához, közelebb juttatva az emberiséget a természetes nyelvek felhasználásnak sokoldalúságához (*Floridi–Chiriatti*, 2020; *Dale*, 2021; *Sejnowski*, 2023). A mesterséges intelligencia jelentős fejlődési ütemet tudhat magáénak. Az imént idézett szöveg prompt utasításként történő felhasználásával a szerző megkérte a ChatGPT-t, hogy fejtse ki véleményét saját korábbi megállapításával kapcsolatosan – anélkül, hogy tudná, az idézett szöveg tőle származik. A válasz az alábbi volt: „… ez úgy hangzik, mintha az öntudat és az érzékelés képességei hiányoznának belőled. Azonban fontos megjegyezni, hogy a szintetikus intelligencia különböző formái különböző szintű tudatossággal és érzékeléssel rendelkezhetnek…" (ChatGPT, 2023). Az AI-hype fokozott piaci hangulata alapján a társadalom a nagy áttörést várja (Turing-teszt teljesítése), ugyanakkor ez egyben azt is jelentené, hogy az ember saját faját nem képes megbízhatóan felismerni (*Héder*, 2020).

A mesterséges intelligencia történetét a gépek egyre intelligensebb viselkedést tükröző magatartásával azonosítják. Hátterében ugyanakkor az emberi faktor áll, az emberi elme szárnyalása. Bár egyelőre hiányzik a mesterséges intelligencia univerzálisan elfogadott definíciója, számos felfogás született az intelligens, gondolkodó gépek meghatározására vonatkozóan (*Wang*, 2019) (*1. táblázat*).

1. táblázat
A mesterséges intelligencia fogalom fejlődése

A mesterséges intelligencia felfogása	Szerző	Év
A kognitív architektúrák az emberi agy működéséhez hasonló automatikus, logikus folyamatokból álló rendszerek, indirekt módon kapcsolódnak a gondolkodó gépek létezésének felvetéséhez.	*Neumann*	1948, 1951
Ha egy gép úgy viselkedik, mintha gondolkozna, beszélne, érezne, akkor egy bizonyos ponton már nem különböztethetjük meg attól az emberi tevékenységtől, amit utánozni próbál.	Turing	1950
„Mesterséges intelligenciát hozzunk létre, […] gépeket, amelyek olyan feladatokat tudnak megoldani, amelyek az emberi intelligenciához kötődnek."	*McCarthy–Minsky–Rochester–Shannon*	1955:2
„Egy számítógépet úgy lehet programozni, hogy megtanuljon jobban sakkozni, mint az, aki írta a programot."	*Samuel*	1959, pp: 211.

„A kérdés az, hogy vajon az emberi gondolkodás minden aspektusa redukálható-e egy logikai formalizmusra, vagy másképpen fogalmazva, hogy az emberi gondolkodás teljes mértékben kiszámítható-e.”	*Weizenbaum*	1966:7; 12
„Míg a természetes intelligenciával rendelkező emberek a feladatok elvégzését önállóan megtanulják, a számítógépeket programozni szükséges erre.”	Minsky–*Seymour*	1969:3
„A fenomenológiai szint alatt a megvalósítási részletek olyan kognitív kerekekből állnak, amelyek eltérnek az emberi agy működésétől.”	*Dennett*	1984:14
„Az a kutatási terület, amely az emberi intelligenciát próbálja utánozni.”	*Kurzweil*	1999:223
A mesterséges intelligencia mint tevékenység intelligensé teszi a gépeket, lehetővé téve számukra, hogy adott környezetben megfelelően és előrelátóan működjenek.	*Nilsson*	2010
Az AI-t úgy definiálhatjuk, mint ügynököket, amelyek érzékelik a környezetüket, és válaszként cselekvéseket eszközölnek.	*Russell–Norvig*	2010
Az AI az intelligens rendszerek elméleti és gyakorlati alkalmazása emberi intelligenciával megoldandó problémák kivitelezésére.	*Horvitz–Mitchell*	2007
„Az emberi analitikai és/vagy döntéshozatali képességek replikácója.”	*Finlay*	2018:11
„A mesterséges intelligencia intelligens viselkedésre utaló rendszereket takar, amelyek konkrét célok eléréséhez elemzik a környezetüket, és – bizonyos mértékű autonómiával – intézkedéseket hajtanak végre.”	European Commission	2018:1
„Az AI-rendszer azt jelenti, hogy az AI-alapú összetevőket, szoftvereket és/vagy hardvereket foglalja magában. Valójában az AI-rendszerek általában nagyobb rendszerek részeként vannak beágyazva, nem önálló rendszerek.”	Európai Bizottság (HLEG)	2019: 2

Forrás: saját táblázat

Az AI performatív és fenomenológiai kudarcai ellenére megvalósuló térnyerése jelentős kihívás elé állítja a humán oldalt a „kreátor” elfogadott etikai normáinak vonatkozásában (Dennett, 1984, 2019; Dennett et al., 2019; *Dreyfus*, 1972, 2007; Weizenbaum, 1976; *Searle* et al., 1980; Héder, 2020; Prisznyák, 2023b). A mesterséges intelligenciával összefüggő kockázatkezelés szükségképpen igényli a szabályo-

zói oldal fellépését a társadalmi, etikai, valamint jogi-szabályozási kérdéseket illetően, támogatva a szervezetek látszólagos elköteleződésének („ethics washing") feloldását (OECD, 2019; *Török–Ződi*, 2021). Az európai uniós értékrenden nyugvó, technológiai szuverenitást támogató jogszabályi keretrendszer kialakításának addicionális célja az Európai Unió globális sztenderdalkotóvá válásának elősegítése a megbízható mesterséges intelligencia tekintetében (Európai Bizottság, 2018; Európai Tanács, 2020; Európai Parlament, 2020). Az érintett felek bevonásával történő konzultációs folyamatok az Európai Unió valamennyi tagországát érintően, átfogó jelleggel zajlanak. A szabályozó hatóságok ugyanakkor olyan alapvető filozófiai és etikai vonatkozással bíró kihívásokkal szembesülnek, mint az erkölcsi, etikai jó univerzális absztrakt fogalmának meghatározása. Mit tekintsünk az etikus AI harmonizált koncepciójának? A kérdésre adott válasz számos esetben evidensnek tűnhet, ugyanakkor közelebbről vizsgálva egy komplex, terület- és társadalomcsoportonként eltérő, kulturális dilemmához jutunk (*Awad* et al., 2018). Következésképpen az etikus AI kérdése egyben a kulturális etikusság kérdése is.

2. KUTATÁSI KÉRDÉSEK ÉS MÓDSZERTAN

A szakirodalom vizsgálatának kiegészítéseképpen a szerző 2022. december és 2023. március között strukturált mélyinterjúkat folytatott 13 fő bevonásával 7 iparág/szektor AI-bevezetési projektjeit érintően (kezdeményezés kiindulópontja, menedzsment támogató attitűdje, etikai aggályok, kapcsolódó oktatás). A vizsgálódás alapján az alábbi kérdésekre keresi a választ:

- K1: A kezdeményezés jellemzően felső vezetői szintről indul?
  H1: Igen, felső vezetői, menedzsmentszintről.
- K2: A menedzsment támogató hozzáállása tapasztalható az AI-rendszer bevezetési projekteken?
  H2: Pozitív, *támogató magatartás*.
- K3: Az etikai megfontolások vizsgálat tárgyát képezik az AI-rendszerek fejlesztése, implementációja során?
  H3: Igen, számos etikai érv felmerül a felhasználók jogainak és biztonságának biztosítása érdekében.
- K4: Kapnak oktatást AI-témában és a bevezetést illetően a munkavállalók (képzés, workshop, dokumentáció)?
  H4: Az AI-rendszerek bevezetése során a munkavállalók képzésben részesülnek.

A szerző ugyanakkor azzal a (későbbiekben részletezett) megállapítással él, hogy az etikai kérdések nem jelennek meg az üzleti tervezés és implementáció folya-

mán. Következésképpen a szerző a nemzetközi etikus AI-szabályozások és iránymutatások felé fordul, és a megbízható mesterséges intelligencia etikai irányelveinek értékelése céljából összehasonlító elemzést folytat az Európai Uniónak a megbízható AI-ra vonatkozó etikai iránymutatása, illetve az AI-jogszabály kritériumainak összehasonlító gapelemzése alapján. Az összehasonlító elemzés eredményeképp jelentkező megállapításokat a szerző 21 filozófus filozófiai és etikai álláspontját alapul véve elemzi, hogy végül megoldási javaslatokkal szolgáljon az etikai iránymutatás hiányosságainak üzleti implementáció során történő áthidalása érdekében.

3. ETIKUS AI-ELVEK PROLIFERÁCIÓJA

Az etika nem új keletű felfedezés (*Drucker*, 2001). Az erkölcsfilozófia olyan normatív gyakorlati filozófiai diszciplína, amely a morális kérdéseken alapuló viselkedés filozófiai megalapozását vizsgálja (*Cointe–Bonnet*, 2016). *Kirkpatrick* (2015) megállapítása alapján a cselekvési alternatívák elfogadott etikai elvek mentén történő választása etikai dilemmákat eredményezhet. A mesterséges intelligencia fejlesztésével és tervezésével kapcsolatos dilemmákat tárgyalva *Denning–Denning* (2020) felhívja a figyelmet az AI-fejlesztéssel összefüggő etikai dilemmák létezésére, amelyek az üzleti érdekek alapján nem feltétlenül szolgálják a technológia-alapú társadalmi érdeket.

Asimov a „*Runaround* [Körbe-körbe]" (1942) című novellájában a gépek etikus alkalmazásával és viselkedésével összefüggésben lefektette a robotika három törvényét, amely a mai napig vitatott, ugyanakkor az etikai elvek kialakítása során iránymutatásául szolgál:

- Első törvény: „A robotnak nem szabad kárt okoznia emberi lényben, vagy tétlenül tűrnie, hogy emberi lény bármilyen kárt szenvedjen."
- Második törvény: „A robot engedelmeskedni tartozik az emberi lények utasításainak, kivéve, ha ezek az utasítások az első törvény előírásaiba ütköznének."
- Harmadik törvény: „A robot tartozik saját védelméről gondoskodni, amennyiben ez nem ütközik az első vagy második törvény bármelyikének előírásaiba." (Asimov, 1942:27).

A későbbiekben Asimov a „The Evitable Conflict [Az elkerülhető konfliktus]" (1950) novellájában módosítja az első törvényt, kiterjesztve az emberiség egészének védelmére (Asimov, 1950:146).

Wiener (1948) az intelligens viselkedés gépek által megvalósítható szimulációjának lehetőségét a „*Cybernetics: or Control and Communication in the Animal and*

the Machine" című munkájában az információ- és visszacsatolási mechanizmusokra vezeti vissza. Kapcsolódóan a kezdeti mesterségesintelligencia-kutatásokhoz, Neumann a *Hixon Symposium* (1948) keretében az elsők között ismertette a kognitív architektúrák, emberi agy működéséhez hasonló felfogásának, így a gondolkodó gépek működésének alapjait – amelyet a későbbiekben „*The General and Logical Theory of Automata*" cikkében – tárgyal (Neumann, 1963). Turing a „*Computing Machinery and Intelligence*" tanulmányában a gépek fejlődéséről értekezik: „…a gépek végül versenyezni fognak az emberekkel az összes, tisztán intellektuálisnak tekinthető területen" (Turing, 1950:22). A természetes és mesterséges intelligencia között húzódó, kimondatlan verseny az agy működési modellje alapján formálódik. A számítógépek és az emberi agy működése közötti hasonlóságot és különbségeket Neumann a „*The Computer and the Brain*" könyvében ismerteti (Neumann, 1958). Weizenbaum 1966-ban befejezi a demonstrációs célú számítógépes program (ELIZA) írását, amely a számítógépek intelligens viselkedését hivatott demonstrálni. A széleskörű publicitás hozzájárult a mesterségesintelligencia-kutatásokat támogató piaci hangulat fellobbanásához. Az emberek megtévesztésére alkalmas chatbottal kapcsolatosan felmerülő etikai aggályokról Weizenbaum a „*Computer Power and Human Reason: From Judgment to Calculation*" könyvében fejti ki álláspontját, és a humán érték megóvásával kapcsolatosan felhívja a figyelmet a fejlesztési folyamatokba integrálandó etikai elvek szükségességére (Weizenbaum, 1976).

Bár a mesterséges intelligencia fejlődésének kezdetét az 1956-os dartmouth-i konferenciával azonosítják, a társadalmi felelősségvállalással kapcsolatos kezdeti gyökerei az 1991-ben megrendezett „*Artificial Intelligence and Social Responsibility*" (San Francisco, USA) konferenciához kötődnek. Az első és második AI-telet követen az etikus gépek és a mesterséges intelligencia kutatási területe is egyre nagyobb popularitásra tett szert az 1990-es évektől (*Yu* et al., 2018). *Anderson* az intelligens gépek etikai viselkedését a gép által kivitelezett cselekmény adott helyzethez társítható morális, etikai kritériumainak igazolásához köti (Anderson, 1995). Kapcsolódóan a gépek etikus döntéshozatalához, *Friedman* és *Nissenbaum* a „*Bias in Computer Systems*" című írásukban felállítanak egy keretrendszert a gépek diszkriminációmentes döntéshozatalának elősegítése céljából (Friedman–Nissenbaum, 1996). Az ezredfordulót követően *Veruggio* (2007) a humanoid robotok fejlesztésével kapcsolatos etikai problémákról értekezik, míg Anderson és Anderson megállapítja, hogy az etikus AI-keretrendszer az emberi értékrendre és erkölcsre épülő AI-rendszerek létrejöttét hivatott támogatni (Anderson–Anderson, 2011). Az etikus AI-kutatások szórványos megjelenése ellenére az AI etikai kérdéseit tárgyaló első konferenciára („*Ethics of Artificial Intelligence*", New York, USA) 2016-ig várni kellett.

A társadalmi konvenciók eredményeképpen kialakult etikai elveknek az AI-rendszer működési mechanizmusába illesztése kereteiről 2015-öt követően az „AI-szuperhatalmak" is értekeznek. Az amerikai álláspontot képviselő „*Report on the Future of Artificial Intelligence*" (2016) megjelenését követően az Európai Bizottság (2019) kiadta „*A megbízható mesterséges intelligenciára vonatkozó etikai iránymutatását*", amelyet az ázsiai régióban vezető szerepet betöltő Kína 2019-ben publikált álláspontja követett (Beijing AI Principles) (Európai Bizottság, 2018; Executive Office of the President National Science and Technology, 2016; Beijing Academy of Artificial Intelligence, 2019; Európai Bizottság, 2022). Ezen állásfoglalásokat nemzetközi viszonylatban kiemelkedő jelentőséggel bíró intézmények publikációi egészítették ki (Európai Bankföderáció, 2019; OECD, 2019; IEEE, 2016, 2019, 2021; UNESCO, 2020; Európai Bankhatóság, 2021).

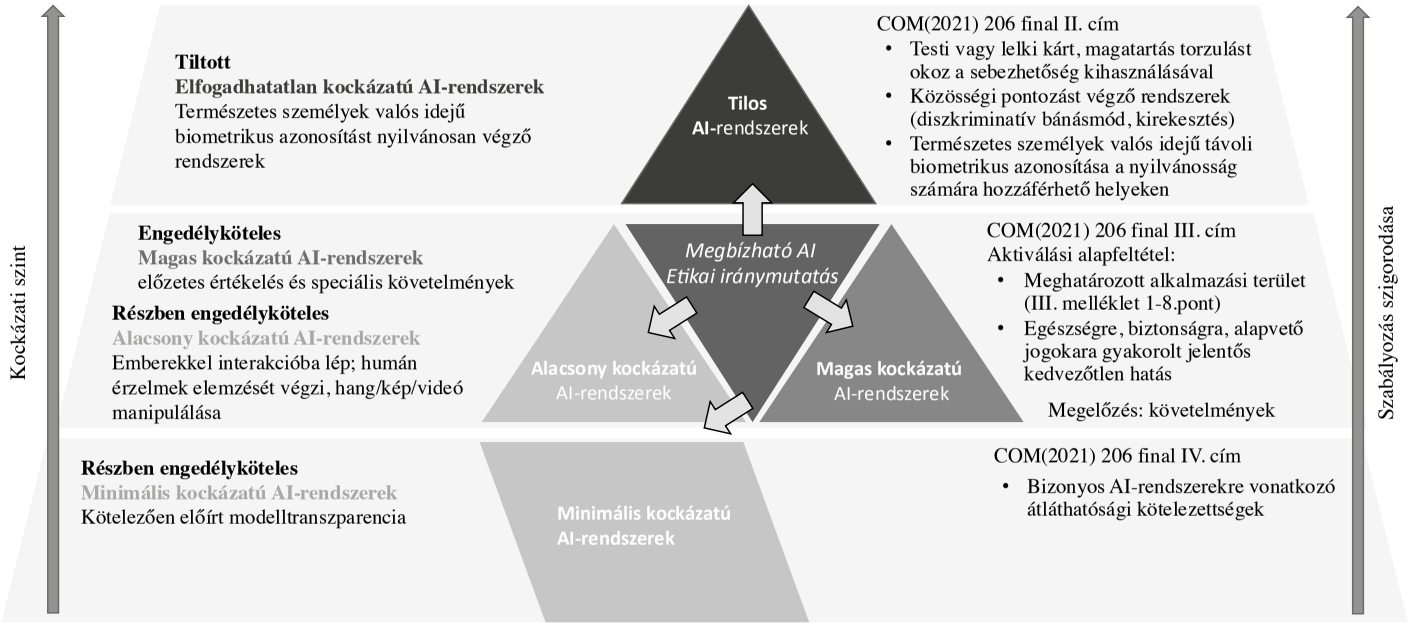
Jobin–Lenca–Vayena (2019) 84 nemzetközi etikus AI-t szabályozó dokumentum átfogó vizsgálata alapján 11 etikai értéket és irányelvet azonosított. Ezek közül számos etikai irányelv esetében nemzetközi konvergencia megfigyelését emelik ki. Az etikus viselkedéssel összefüggő elvek proliferációjából kiindulva *Floridi–Cowls* (2019) a mesterséges intelligencia alábbi négy alapelvét azonosították: jótékonyság, károkozás-mentesség, autonómia, igazságosság, amelyet egy ötödik elv – a megmagyarázhatóság – hozzáadásával egészítettek ki. Kapcsolódóan a nemzetközileg publikált jogszabályok és iránymutatások elemzéséhez, *Hagendorff* (2020) 22 irányelvet elemez, és megállapítja, hogy az elszámoltathatóság, az értelmezhetőség, a magánélet védelme, az igazságosság, az átláthatóság, a robusztusság és a biztonság a legkönnyebben operacionalizálható elvek közé tartozik. Az AI nemzetközi szabályozásában bekövetkezett kedvező fordulat ellenére, Yu et al. (2018) a felelős AI-rendszerek fejlesztésének kihívásait és fontosságát érintően hangsúlyozza az etikai szempontok integrációjának hiányát az AI-rendszerek fejlesztése során. Bár a mesterséges intelligencia etikai irányelvei jogilag nem kötelező erejűek, kiegészítik a jogilat kötelező érvényű szabályozást és iránymutatásul szolgáljanak az etikai normák szervezeten belüli „ön-kormányzásának" elősegítésében (Jobin–Lenca–Vayena, 2019; *Calo*, 2017). Kapcsolódóan ezen hiányosságokhoz, jelen tanulmány szerzője sürgeti az etikus AI-jal kapcsolatos szervezeti álláspontok etikai kódexbe történő integrálását az érdekelt felek közötti konstruktív kapcsolatok megalapozása és a bizalom megteremtésének jegyében.

4. ETIKUS IRÁNYELVEK ELEMZÉSE: AZ ETIKAI ÉS TECHNIKAI SZABÁLYOZÁS TALÁLKOZÁSA

Az AI-rendszerek alkalmazása során jelentkező etikai kihívások kezelése érdekében az Európai Bizottság 2018-ban létrehozta a mesterséges intelligenciával foglalkozó magas szintű független szakértői csoportját (High-Level Expert Group on Artificial Intelligence, továbbiakban HLEG), amelyet az etikus AI-irányelvek kidolgozásával bízott meg. A HLEG 2019-ben közzétette a megbízható mesterséges intelligenciára vonatkozó etikai iránymutatását – amely az érintett felekkel folytatott konzultációja során szerzett tapasztalatokon nyugszik –, és azt egy gyakorlati implementációt támogató értékelési listával egészítette ki (megbízható mesterséges intelligenciával kapcsolatos értékelési lista, továbbiakban ALTAI) (Európai Bizottság, 2019; 2020).

Párhuzamosan az etikai keretrendszer kialakítását érintő munkálatokkal, az Európai Bizottság 2021-ben nyilvánosságra hozta a mesterséges intelligenciára vonatkozó jogszabályi keretrendszer javaslatát (2021/0106(COD)) (Európai Bizottság, 2021). Az Európai Unió értékrendjén nyugvó, technológiai szuverenitást támogató AI-jogszabály (AI Act) (COM(2021) 206 final) célja, hogy az Európai Unió Alapjogi Chartával összhangban biztosítsa az állampolgárok alapvető jogait, biztonságát és szabadságát, miközben támogatja az európai értékrenden nyugvó mesterséges intelligencia fejlődését (Európai Tanács, 2020; European Commission, 2021). Az AI-jogszabály harmonizálja valamennyi tagország nemzeti AI-szabályozási törekvéseit, és egységes keretrendszerbe foglalja az AI-rendszerek fejlesztésével és felhasználásával kapcsolatos jogszabályi elvárásokat (Európai Bizottság, 2021; 2022). Az AI-jogszabály szigorú követelmények mentén biztosítani kívánja a döntések átláthatóságát, a felhasználók jogainak védelmét és az etikus alapelvek betartását. Következésképpen kockázatkezelési mechanizmusokat ír elő, valamint rendelkezik az AI-rendszerek kockázati osztályba sorolásának szükségességéről és kritériumairól *(1. ábra)*.

1. ábra
Megbízható AI etikai iránymutatás és AI-rendszerek kockázati osztályozása



Forrás: saját ábra

5. ELEMZÉS: A MEGBÍZHATÓ AI ETIKAI IRÁNYMUTATÁS KRITÉRIUMAINAK HIÁNYOSSÁGAI

A szerző az AI-rendszerek üzleti fejlesztése, implementációját 13 mélyinterjúba bevont szakember révén vizsgálta 2022. december és 2023. március között az alábbi iparágakat/szektorokat érintően: autóipar, fintech, bankszektor, gyógyszeripar, healthtech, ICT, légiipar. A strukturált mélyinterjúk minden esetben másfél-két órát vettek igénybe. Az eredményeket anonim módon publikáljuk. Az interjúalanyok olyan üzleti, szoftverfejlesztő szakemberek, akik részt vettek mesterséges intelligenciát, gépi tanulást, robotfejlesztést, integrációt támogató folyamatokban (*2. táblázat*).

2. táblázat
Interjúösszesítő

#	Foglalkozás	Tapasztalat (év)	Interjú időtartama (perc)	Iparág, szektor
1.	AI-divízióvezető	9	120	bankszektor, autógyártás
2.	K+F igazgató	15	120	autógyártás
3.	Szoftverfejlesztő	6	90	autógyártás
4.	Machine learning engineer	7	90	healthtech, fintech
5.	Projektmenedzser	25	120	légiipar
6.	Informatikai menedzser	25	80	bankszektor
7.	Automatizálási vezető	12	90	bankszektor
8.	Machine learning engineer	17	120	bankszektor, gyógyszeripar
9.	Programtervező informatikus	23	120	bankszektor, gyógyszeripar
10.	Szoftverfejlesztő mérnök	7	120	ITC
11.	R&D, AI-fejlesztő	6	120	autóipar
12.	Informatikai projektvezető	6	120	bankszektor, gyógyszeripar
13.	Informatikus	20	90	bankszektor
Interjúk összesen (óra)			**23,3**	

Forrás: saját táblázat

A szerző az etikai aspektusok érvényesítésével kapcsolatos meglátásokat tekinti jelen tanulmány kiindulópontjának, amelyhez kapcsolódó kutatási kérdéseit,

kapcsolódó hipotéziseit a *3. táblázat*, az interjúalanyok válaszait az *1.* (interjúösszesítő) *melléklet*, saját kutatási eredményeit a *4. táblázat* tartalmazza.

3. táblázat
Kutatási kérdések, hipotézisek és mélyinterjú-tapasztalatok

Kutatási kérdés és hipotézis	Saját eredmény
K1 – H1	Fintech, bankszektor, autóipar, ICT esetében top-down megközelítés (fogyasztói, befektetői nyomás és költségcsökkentési cél); légiipar, gyógyszeripar esetében munkavállalói kezdeményezés (főleg predikció) IT-val/BI területtel; mindkét esetben az üzleti terület jelentős hajtóerő
K2 – H2	Top-down esetében főleg a támogató, pozitív menedzsmenti hozzáállás figyelhető meg, míg bottom-up esetében nincs vagy korlátozott a támogatás mértéke (minimális erőforrásallokáció)
K3 – H3	Jellemzően egyáltalán nem merül fel, gyerekcipőben jár
K4 – H4	Beszállító bevonása esetén workshopok és dokumentáció; külön AI-specifikus, illetve etikai érzékenyítési tréning sehol sem volt; jellemzően a tréningek túl általánosak, idő- és költségvetési korlátok akadályozzák a meglétét

Forrás: saját táblázat

A K1, K2, K4 esetében a H1 (felsővezetői kezdeményezés), H2 (menedzsment támogató hozzáállása), H4 (AI-oktatás) hipotézisek korlátozott érvényesülése figyelhető meg, míg a K3 esetében a H3-at (etikai kérdések megfontolása az üzleti tervezés során) elvetjük. A mélyinterjúkból levont kutatási eredmények alapján a szerző célja a mesterséges intelligencia etikai irányelveinek kiértékelése (kapcsolódóan az eredeti K3 kérdéshez), és a feltárt hiányosságokra vonatkozóan az üzleti implementációt támogató javaslatok tétele.

A szerző a megbízható mesterséges intelligencia etikai irányelveinek értékelése céljából összehasonlító elemzést folytat az *1. mellékletben* részletezett jogszabályok, iránymutatások és állásfoglalások alapján, különös tekintettel az Európai Uniónak a mgbízható AI-ra vonatkozó etikai iránymutatása, illetve az AI-jogszabály kritériumaira vonatkozóan. A kritériumok megfeleltetését a *4. táblázat* tartalmazza. Az összehasonlító elemzés – direkt megfeleltethetőség hiányában – nem tárgyalja az etikai iránymutatás alábbi két kritériumát: (5.) sokféleség, megkülönböztetésmentesség és méltányosság; (6.) társadalmi és környezeti jólét, amely egyben az összehasonlító gapelemzés korlátjaként értelmezendő.

4. táblázat
A megbízható mesterséges intelligencia etikai iránymutatásának és az AI-jogszabály kritériumainak megfeleltetése

Megbízható AI etikai iránymutatása		COM (2021) 206 final	
Fejezet/pont	**Megbízható AI etikai elvek követelményei**	**Cím/fejezet/cikk**	**AI-jogszabály követelmények**
II. fejezet 1.	Az emberi cselekvőképesség támogatása és emberi felügyelet	III. cím, 2. fejezet, 14. cikk	Emberi felügyelet
II. fejezet 2.	Műszaki stabilitás és biztonság	III. cím, 2. fejezet, 15. cikk	Pontosság, stabilitás és kiberbiztonság
II. fejezet 3.	Adatvédelem és adatkezelés	III. cím, 2. fejezet 10. cikk	Adatok és adatkormányzás
II. fejezet 4.	Átláthatóság	III. cím, 2. fejezet, 13. cikk	Átláthatóság és a felhasználók tájékoztatása
II. fejezet 7.	Elszámoltathatóság	III. cím, 3. fejezet, 17. cikk	Minőségirányítási rendszer

Forrás: saját táblázat

Az összehasonlító elemzés eredményeképp jelentkező megállapításokat (*5. táblázat*) a szerző filozófiai és etikai vizsgálódással egészíti ki (*3. melléklet*), amelynek alapján a *6. táblázat* a szerző által azonosított Gap1 és Gap2 hiányosságot vizsgálja 21 filozófus filozófiai és etikai álláspontját alapul véve. A Gap1, Gap2 hiányosságok feloldására és az etikus AI-elvek üzleti integrációjának elősegítésére vonatkozó javaslatokat a *7. táblázatban* fogalmazzuk meg.

Az összehasonlító elemzés alapján azonosított hiányosságokat és azok relevanciáját az *5. táblázat* összegzi.

5. táblázat
Etikai és technikai területek találkozása: összehasonlító gapelemzés

Elemzési szempontok	Etikai követelmények	Kapcsolódó technikai jellegű követelmény	Szerző által azonosított gap
Gapelemzés alapja	Az emberi cselekvőképesség támogatása és emberi felügyelet 2. fejezet (1.)	Emberi felügyelet III. cím, 2. fejezet, 14. cikk	AI-rendszerfejlesztők és felügyelők etikai érzékenyítése, képzése
Gap részletezése	A felügyeletet ellátó személy szükséges képességeit (rendszerműködés, kapacitás, korlátok megértésének a képessége) az etikus iránymutatás nem részletezi. Kizárólag a felügyeleti módok, a kockázatok elemzésén és a felhasználók szempontjából közelít.		
Relevancia indoka (etikai aggály)	Előfordulhat, hogy az AI-rendszerek döntéshozatala hátrányosan hathat bizonyos csoportokra (alapvető jogok, biztonság, méltányosság). Az etikus emberi felügyelet ellátásához a technikai tudás és etikai érzékenyítés szükséges – mind a fejlesztők, mind az üzemeltetők és felügyelők számára – a következmények időben történő észlelése, kezelése és megelőzése érdekében.		
Kapcsolódó probléma	Az AI-rendszert fejlesztők, felügyelők, üzemeltetők etikai érzékenyítése jellemzően elmarad az üzleti tervezés és implementáció során.		
Gapelemzés	Műszaki stabilitás és biztonság 2. ejezet (2.)	Pontosság, stabilitás és kiberbiztonság III. cím, 2. fejezet, 15. cikk	Káros visszacsatolási hurkok és döntéshozatali torzulás
Gap részletezése	Az etikai iránymutatás nem értekezik a rendszer által jónak ítélt, valójában hibás visszacsatolási hurkok meglétéről és kezelésük módjáról.		
Relevancia indoka (etikai aggály)	Az AI-rendszerek esetében előfordulhatnak káros visszacsatolási hurkok, amelyek során a rendszer helyesnek ítélt, mégis hibás döntéseket a későbbiekben inputként használ, megerősítve a hibás döntéseket. Egy negatív folyamat indul el, amely a rendszer döntésein keresztül a környezetre nézve is negatív hatást eredményezhet.		
Kapcsolódó probléma	A visszacsatolási hurkok az adat, a modell és a felhasználói interakció torzulását eredményezhetik, amennyiben az AI-rendszer üzemeltetése, alkalmazása során nem megfelelő a monitoringfolyamat, illetve a szükséges adatkorrekciók elmaradnak.		

Forrás: saját táblázat

6. táblázat
Filozófusok filozófiai és etikai álláspontjának értelmezése Gap1, Gap2 szempontjából

Filozófiai korszak	Filozófus	Interpretáció Gap 1	Interpretáció Gap 2
Görög–római filozófia	*Parmenidész* (i.e. 515– i.e. 470)	A felügyelő támogatja az erkölcsi normák változásának felismerését és a döntéshozatali „igazság" létezésének elérését	A visszacsatolási hurok létező („örök"), ami a nemlét dimenziójában értelmezve, az idő múlásával és a környezet változásával nem mindig áll fent
	Szókratész (i.e. 469 – i.e. 399)	Erkölcsös cselekvéshez szükséges: az etikai alap, a tudás, a problémamegoldó, kritikus gondolkodás és kommunikációs képességek	A rendszer etikus döntéshozatala a megfogalmazott etikai alapoktól függ, amit a rendszer tanulási folyamata erősít (rendszer önreflexiója)
	Xenophón (i.e. 434– i.e. 355)	Tapasztalati értékelés, kommunikációs készségek, morális értékalapú döntéshozatal	A rendszer instabilitásának elkerülése érdekében a működési mechanizmusok és algoritmusok átláthatósága és logikus felépítése kiemelten fontos
	Platón (i.e. 427–i.e. 347); *Arisztotelész* (i.e. 384–i.e. 322)	A jó kormányzás felelős az erkölcsös alapok biztosításáért: az oktatás, igazságosság (felelősség-elszámoltathatóság), kommunikáció terén	A tudás segít a helytelen rendszerműködés felismerésében (abszolút igazság keresése); kommunikáció szerepe a visszacsatolási input adatok értékelésében
Középkori filozófia	*Szent Ágoston* (354–430)	Kinyilatkoztatott etikai elveken alapuló kollaboráció és felelősségvállalás	Rendszerbe épített etikai normák és kapcsolódó felelősségvállalás
	Aquinói Szent Tamás (1225–1274)	Felelősség a rendszer működési mechanizmusának megértésében (az erkölcsös viselkedés tanúsítása érdekében)	Megerősítésen alapuló tanulási algoritmusok; rendszerátláthatóság és -komplexitás kérdésköre
	William Ockham (1287–1347)	A felügyelőknek szubjektív megítéléstől mentes objektív értékelésre kell törekedniük; a rendszer döntés megismerésnél a szkepticizmus és kritika gyakorlása fontos	Az egyszerűség elve alapján a visszacsatolási hurkok csökkentésére való törekvés gyszerűbb és könnyebben érthető algoritmusokat használatával (átláthatóság-pontosság trade-off)
	Pázmány Péter (1570–1637)	A rendszermechanizmusok megértése biztosítja az emberközpontú értékek, etikai elvek megvalósulási gyakorlatának összhangját	A „felsőbbrendű" felügyelő biztosítja a rendszerbe integrált etikai elvek érvényesítését; működés közben folyamatos rendszer-önreflexió (értékelés)

Filozófiai korszak	Filozófus	Interpretáció Gap 1	Interpretáció Gap 2
Újkori filozófia	*René Descartes* (1596–1650)	Cél: a komplex rendszer-mechanizmusok és döntéshozatal megértésére való törekvés; etikai iránymutatások gyakorlása, felelősségvállalás	Komplex algoritmusok és döntési folyamatok megértése, black-box jelenségek kiküszöbölése
	Gottfried Wilhelm Leibniz (1646–1716)	A felügyelőnek biztosítania kell a rendszer hosszú távú ontológiai stabilitását, és időben fel kell ismernie a változást (ehhez technikai képességek és etikai érzékenyítés szükséges)	Figyelembe szükséges venni az egyes felhasználók preferenciáit (monád), de fontos a preformáció (társadalmi etikai normák) és hosszú távon ezek ontológiai stabilitása a rendszerben
	Francis Bacon (1561–1626)	A szervezet felelős az oktatásért is; tapasztalati tanulás és ismeretek kodifikációja	Szigorú szabályok és vizsgálódási módszertan
	Thomas Hobbes (1588–1679)	A rendszert fejlesztők, üzemeltetők felelőssége a rendszer biztonságos működése az alapvető jogok védelme érdekében	Célracionális rendszerműködés etikai keretek közé szorítása; megfelelőség értékelése a biztonság, jogok biztosítása alapján
	John Locke (1632–1704)	Ügyfelek diszkriminációmentes kiszolgálása; a biztonság és alapvető jogaik biztosítása	A diszkriminációmentes döntéshozatal biztosítása a rendszert fejlesztők, üzemeltetők, fenntartók feladata
	David Hume (1711–1776)	A szervezeti érdekek nem befolyásolhatják az erkölcsi normák érvényesülését	A rendszert képessé kell tenni a tapasztalati tanulás révén az ártó hibák időbeli kiszűrésére
	Jean-Jacques Rousseau (1712–1778)	Alapvető jogok egyenlő kezelésének biztosítása, főleg a hátrányosan értintett csoportok esetében	A szabályalapú döntéshozatal kirekesztő hatást okozhat a döntéshozatal során (rejtett módon a káros visszacsatolási hurkok létezésén keresztül)
Modern, újkori filozófia	*Immanuel Kant* (1724–1804)	Biztosítani szükséges az objektíven lefektetett etikai elvek érvényesülésének folytonosságát	A rendszertervezés során az etikai szempontok biztosítják az általános akarat elvének érvényesülését
	John Stuart Mill (1806–1873)	Az etikus cselekedet értékelése a társadalmi hasznosság alapján – kapcsolódó felelősség és elszámoltathatóság	A modell döntéshozatali folyamatának folyamatos értékelése, a szükséges korrekciók és incidens-adatbázisok vezetése

Filozófiai korszak	Filozófus	Interpretáció Gap 1	Interpretáció Gap 2
Kortárs, jelenkori filozófia	*Daniel Dennett* (1942–	A felügyelőknek figyelemmel szükséges lenniük arra, hogy az erkölcsi normák idővel változhatnak és kultúraspecifikusak	Olyan algoritmusok alkalmazása, amelyek figyelembe veszik az előző döntéseik eredményét és hatásait, majd javítják azokat a jövőbeli döntéshozatal során
	Martha Nussbaum (1947–	Az emberi jogok védelme és az emberi méltóság tiszteletben tartása, illetve a kapcsolódó adatbiztonság és helyes döntéshozatal felügyelete képzési szükségeket támaszt	A teljes életciklus alatt megvalósuló kockázatkezelés, adatkezelésre és felügyelet támogatja az emberi alapjogok, biztonság védelmét és a diszkriminációmentességet
	Nick Bostrom (1973–	Etikai elvek és kapcsolódó rendszerbe épített fékek definiálása; felügyelők, programozók érzékenyítésének fontossága	Rendszerbe épített biztonsági mechanizmusok; átláthatóság növelése; rendszer leállíthatósága és a manuális folyamatátvétel lehetőségének kialakítása

Forrás: saját táblázat

6. KÖVETKEZTETÉSEK ÉS JAVASLATOK

Az azonosított gapek (Gap1, Gap2) filozófiai és etikai álláspontok (*6. tábálázat*) átvitt értelemben történő értelmezése ($Gap1_{int.,}$ illetve $Gap2_{int}$) alapján a szerző a *7. táblázatban* feltüntetett javaslatokkal él (interpretáció Gap1–$Gap1_{int.}$, interpretáció Gap2–$Gap2_{int}$) a hiányosságok üzleti implementáció során történő feloldása céljából.

7. táblázat
Szerzői javaslatok:
filozófiai, etikai megközelítésen nyugvó üzleti implementációs feladatmátrix

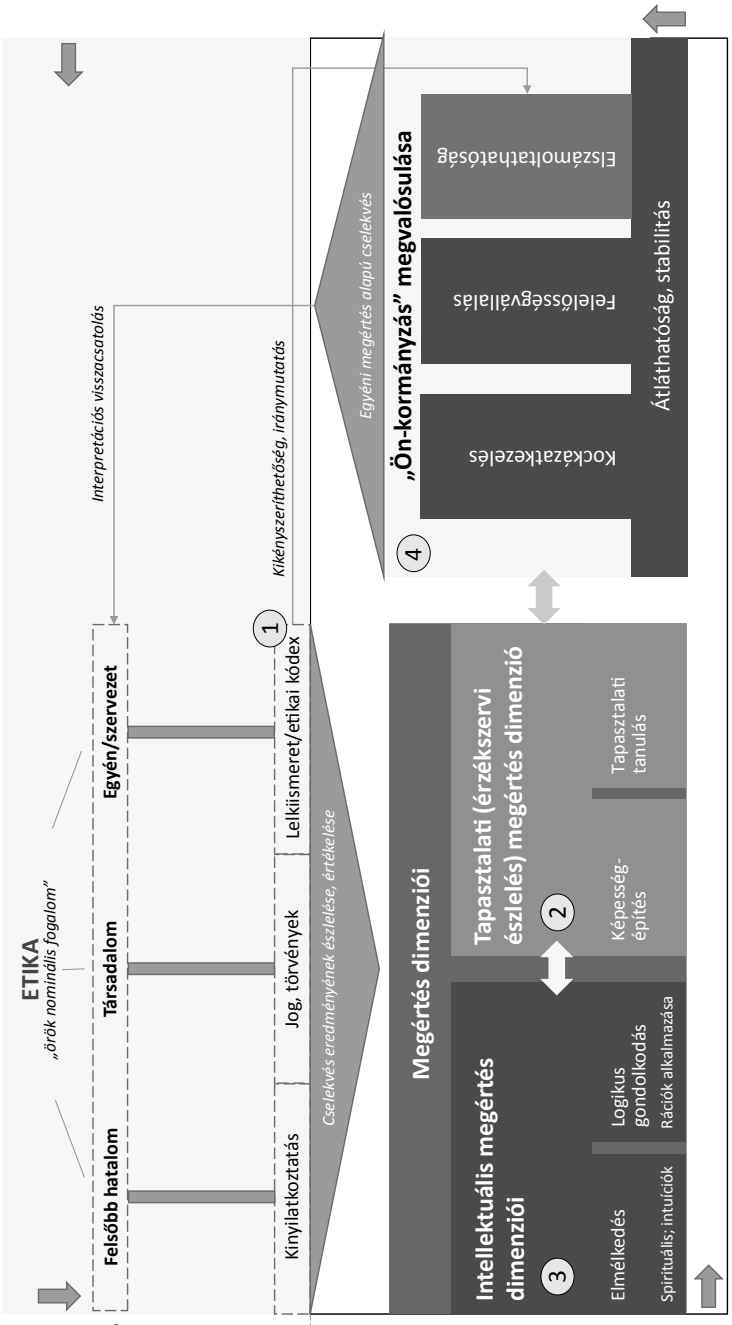
	Megbízható AI-ra vonatkozó etikai iránymutatás azonosított hiányosságai	
	Gap 1	**Gap 2**
Interpretáció — **Gap 1**$_{\text{int.}}$	**Gap1**$_{\text{jav.Gap1 int}}$ *1.* **Etikai kódexek revíziója,** az etikus AI-rendszerek szervezeti keretbe illesztése *2.* **AI-stratégia részeként a szervezeti stratégiába illesztés** biztosítása *3.* **Kultúrába illesztés – a szervezet** átfogó és specifikus **etikai érzékenyítése** (felügyelők, fejlesztők, érintett üzleti területek specifikus képzése, oktatás) *4.* **Felelős szervezeti egység** *5.* **Működési folyamat kialakítása:** követelmények, feladatok – termékcélcsoport, társadalom, kultúra – etikai normáinak kialakítása, folyamatok szabályozása (szabályzatok, utasítások kidolgozása) *6.* **Kockázatkezelési folyamat:** monitoringfolyamatok és eszközrendszer (limitek, mérőszámok), felelősök és következmények, compliance biztosítása *7.* **Etikai fórum kialakítása:** felmerülő aggályok diskurzusának színtere	**Gap2**$_{\text{jav.Gap1 int}}$ *8.* Modellkorlátok objektív megfogalmazása (**etikai fékek beépítése**) és a modelltanulás során tanúsított folyamatos monitoring *9.* **Modell döntési eredményének** körültekintő, szkeptikus **értelmezése** *10.* **Folyamatos monitoring** (stabil, meghízható működés) a hibák redukálása érdekében *11.* **Adatkormányzás és adat-előkészítés** *12.* A rendszer **működési mechanizmusának megértése**, átláthatóság biztosítása (black box jelenségek megszüntetése, megmagyarázhatóság biztosítása) *13.* **Szubjektív etikai elvek érvényesülésének negligálása** (beprogramozott, érintett fél érdeke, célorientált [profit])
Gap 2$_{\text{int.}}$	**Gap1**$_{\text{jav.Gap2 int}}$ *14.* **Tapasztalatok kodifikációjának támogatása** *15.* **Incidensjelentések gyűjtése** (riportvezetés) – jelentés a fórumoknak, felelős szervezeti egységeknek *16.* **Beavatkozási szituációk** és **kritériumok meghatározása** (manuális döntés, döntésfelülvizsgálat, rendszerleállítás lehetőségei)	**Gap2**$_{\text{jav.Gap2 int}}$ *17.* Alkalmazott **algoritmusok megfelelő megválasztása** (az egyszerűség elve alapján áttekinthetőség-pontosság trade-off) *18.* **Rendszer-önreflexió** a tanulási folyamat során (alkalmazott teljesítmény, pontosság mérési metrikák) *19.* Etikai elvek **stabil ontológiai leképezése** *20.* Káros visszacsatolási hurkok jelentése, közzététele: **incidens adatbázis**

Forrás: saját táblázat

A $Gap1_{jav.Gap1\ int,}$ $Gap1_{jav.Gap2\ int}$ és $Gap2_{jav.Gap1\ int,}$ $Gap2_{jav.Gap2\ int}$ alapján elmondható, hogy a Gap2 feloldása során alkalmazott munkavállalói etikai képzés számos etikai kritériumot (adatvédelem, adatkezelés, átláthatóság, elszámoltathatóság) érintően gyakorol hatást a műszaki stabilitás és biztonság kritériumkategóriára. A szerző megállapítja, hogy az etikai irányelvek együttes kezelése nélkülözhetetlen, ugyanakkor az összhang megteremtése az AI-jogszabály és az iránymutatás között szükséges feltétele az üzleti implementációnak.

Az etikus AI kérdése egyben kulturális kérdés is. Az etika nominális fogalomként történő értelmezése eltérő csatornákon keresztül valósulhat meg (egyén/szervezet/társadalom, vallás, egyebek). A szervezeti értékeken nyugvó, stratégiába illeszkedő elvek lefektetése (etikus AI-kódex) hozzájárulhat az AI-rendszerek megbízhatóságának növeléséhez és az érintett felek közötti bizalom megteremtéséhez. Ehhez az etikus AI szervezeti értelmezése és képességeken, tapasztalaton nyugvó formálódása szükséges, amely az „ön-kormányzás" keretében a társadalmi visszacsatolás hatására folyamatosan fejlődik, és közelít a harmonizált etikai elvárások képződményéhez. A szerző hangsúlyozza, hogy az etikus AI-rendszer fejlesztése és alkalmazása egy folyamatos iterációs folyamatként tekinthető, amely képes megteremteni a megbízhatóság elvi alapjait. Következésképpen a szerző javasolja a szervezetek számára az etikai elvek átgondolását, amelyhez segítséget az a *2. ábra* szolgáltat.

2. ábra
Az etika szervezeti interpretációjának megvalósulása – a szerző értelmezésében



Forrás: saját ábra

MELLÉKLETEK

1. melléklet
Kutatási kérdések és válaszok – Interjúösszesítő

#	K1	K2	K3	K4
1.	top-down	pozitív, nyitott magatartás	modellezésre használható adatok köre és adatgyűjtési kérdések; diszkriminációmentesség, ügyféladatok titkosítása	projektben résztvevők involváltsági szintje magas, beszállító biztosítja a tudást (workshop)
2.	top-down	változatos AI-ismeretek függvénye	önvezető autó morális döntései, adatgyűjtés jogi kérdései	időhiány miatt nem; tréningek túl általánosak
3.	top-down	pozitív	gyerekcipőben járnak az etikai kérdések; üzemeltető feladata	nincs tréning; rendszer átadásakor rendszer-dokumentációt adnak át
4.	top-down	nagyobb részt pozitív – AI-ismeretek függvénye	nem merült fel	van
5.	bottom-up	pozitív	nem merült fel	nincs
6.	beszállító → top down, bottom up	szóban támogatás, tettekben elzárkózás	nagyon korlátozott megjelenés: diszkriminatív döntéshozatal, inklúzió	költségkontroll miatt nincs
7.	top-down	pozitív, de a nyitottság korlátozott (adatvagyon óvása)	nagyon korlátozott megjelenés: diszkriminatív döntéshozatala	kompetenciaközpontot építenek szervezeti szinten
8.	bottom-up	egyáltalán nincs vagy korlátozott a támogatás	nem merült fel	nincs
9.	bottom-up	kezdeti nyitottság (költségbecsléskor alábbhagyott)	nem merült fel	nincs
10.	top-down	pozitív, támogató hozzáállás	vizuális megjelenés, biztonsági aggályok	beszállító biztosítja
11.	beszállító → top-down	felső vezetői nyomás miatt nem tud felzárkózni (pótolható)	nem merült fel	workshop tartása rendszerátadásnál és dokumentáció
12.	top-down	pozitív, de az anyavállalat korlátokat jelenthet	bankszektornál: biztonsági kérdések, adatkezelés	architechtek indítottak előadásokat a menedzsmentnek
13.	top-down	pozitív, támogató hozzáállás	nem merült fel	rendszeres nagyvállalati képzések, amelyet az IT támogat, beszélgetések, brainstorming

Forrás: saját táblázat

2. melléklet
Felhasznált rendeletek, iránymutatások és kapcsolódó pontjaik

Rendelet	Cím/ fejezet/cikk	Követelmény	Tartalom összefoglalása
(EU) 2019/1020 rendelete A piac-felügyeletről és a termékek megfelelő-ségéről*	I. cím, 3. cikk, 19. pont (általános rendel-kezések)	Kockázatot jelentő termék	Olyan termék, amely a rendeltetését tekintve vagy szokásos vagy észszerűen előreláthető használati feltételek mellett (…) az észszerűnek és elfogadhatónak ítélt mértéket meghaladó mértékben hátrányosan érintheti az emberek egészségét és biztonságát általában, a munkahelyi egészséget és biztonságot, a fogyasztók védelmét, a környezetet, a közbiztonságot és az alkalmazandó uniós harmonizációs jogszabályok által védett egyéb közérdeket
COM(2021) 206 final	I. cím, 3. cikk, 1. pont (fogalommeg-határozás)	Mesterséges intelligencia rendszer	Olyan szoftver, amelyet az I. mellékletben felsorolt technikák és megközelítések közül egy vagy több alkalmazásával fejlesztettek, és amely az ember által meghatározott célkitűzések adott csoportja tekintetében olyan kimeneteket (tartalom, előrejelzés, ajánlás, döntés) képes generálni, amelyek befolyásolják azt a környezetet, amellyel kölcsönhatásba lépnek
	III. cím, 2. fejezet 7. cikk; III. melléklet (nagy koc-kázatú AI-rendszerek)	III. melléklet módosításai, alkalmazandó kritériumok az okozott kár mértékének megítéléséhez	(a.) AI-rendszer rendeltetése, (b.) tényleges és valószínűsíthető alkalmazásának mértéke, (c.) historikusan megvalósult (dokumentált) károkozás és hatása, (d.) kár/kedvezőtlen hatás lehetséges mértéke (érintettek köre); (e.) potenciálisan károsult személyek kiszolgáltatásának mértéke, (f.) okozott kimenet visszafordíthatósága; (g.) hatályos jogszabályok rendelkezése a kockázatok megelőzésére, minimalizálásra
	III. cím, 2. fejezet 8–15. cikk	Nagy kockázatú AI-rendszerekre vonatkozó követelmények	(8.) Követelményeknek való megfelelés: (9.) kockázatkezelési rendszer létrehozása és működtetése; (10.) adatok és adatkormányzás; (11.) műszaki dokumentáció; (12.) nyilvántartás; (13.) átláthatóság és a felhasználók tájékoztatása; (14.) emberi felügyelet; (15.) pontosság, stabilitás és kiberbiztonság
	III. fejezet 65. cikk, (1–9. pont)	Kockázatot jelentő AI-rendszerek kezelésére vonatkozó nemzeti szintű eljárások	Amennyiben az AI-rendszer értékelés során a piacfelügyeleti hatóság megállapítja, hogy az AI-rendszer nem felel meg a követelményeknek, felszólítja az érintett üzemeltetőt, hogy tegye meg a szükséges korrekciós intézkedéseket, megfelelve a harmonizált szabványoknak és előírásoknak, vagy vonja ki az AI-rendszert a forgalomból
Etikai iránymutatás a megbízható AI-ra vonatkozóan	II. fejezet	1–7.	A megbízható mesterséges intelligencia követelményei: (1.) az emberi cselekvőképesség támogatása és emberi felügyelet; (2.) műszaki stabilitás és biztonság; (3.) adatvédelem és adatkezelés; (4.) átláthatóság; (5.) sokféleség, megkülönböztetésmentesség és méltányosság; (6.) környezeti és társadalmi jólét; (7.) elszámoltathatóság
ALTAI	Teljes dokumentum	Teljes dokumentum	Az értékelési folyamat elemei: (1.) az emberi cselekvőképesség támogatása és emberi felügyelet; (2.) műszaki stabilitás és biztonság; (3.) adatvédelem és adatkezelés, (4.) átláthatóság, (5.) sokféleség, megkülönböztetésmentesség és méltányosság, (6.) társadalmi és környezeti jólét, (7.) elszámoltathatóság

Megjegyzés: *A piacfelügyeletről és a termékek megfelelőségéről, valamint a 2004/42/ek irányelv, továbbá a 765/2008/ek és a 305/2011/EU rendelet módosításáról
Forrás: saját táblázat

3. melléklet
A gapelemzésnél alapul vett filozófusok filozófiai és etikai felfogásának összesítő táblázata

Filozófus	Filozófiai felfogás	Etikai felfogás
	Görög–római filozófia (Szókratész előtti)	
Parmenidész (i.e. 515–i.e. 470)	Monizmus, lenni-nem lenni koncepció: a létezés (állandó, örök igazság), illetve nem létezés (változó, mulandó) dimenziók szétválasztása; érzékelésük intellektuális (előbbi) és érzéki észlelésen keresztül (utóbbi dimenzió)	Az erkölcsi normák örök és változatlan dimenzió részei, de megnyilvánulásuk a nemlét dimenziójában az általunk látható, változó és mulandó világban történik
Szókratész (i.e. 469 – i.e. 399)	Központi elem: a tudás és az erkölcs, a tapasztalatszerzés (érvelés és a dialógus szerepe fontos); az intellektuális gondolkodás egyfajta abszolút tudás/igazság a világról, amely az emberi korlátok (tudatlanság) felismerésével érhető el	Etikai felfogása a tudásra és az erényekre épül; fejlesztésük ismeretszerzés révén lehetséges; érvelésen és gondolkodáson alapuló intellektus és kapcsolódó erkölcsi viselkedés
Xenophón (i.e. 434–i.e. 355)	Az észszerű érvelésen, logikus alapokon, tapasztalaton, következtetésen nyugvó intellektuális gondolkodás kiemelt fontosságú a helyes döntéshozatal során	Erkölcsös viselkedés az intellektuális gondolkodáson alapuló helyes döntések meghozatalán és a szókratészi erényeken alapul
Platón (i.e. 427–i.e. 347)	Dualista elmélet (test-lélek elkülönülése, elmefilozófia); Kétvilág-elmélet (létezés világa változatlan, és csak az intellektus által hozzáférhető)	Az etikus magatartás alapja a tudás, igazságosság és az erények; a tanulás segít a helyes döntéshozatalban (társadalmi jót szolgálnia kell)
Arisztotelész (i.e. 384–i.e. 322)	Dualista elmélet, elmefilozófia, tapasztalati tanulás (észlelés); passzív/aktív intellektus; logikus gondolkodás és elmélkedés, szókratészi erények fontossága (megértéséhez)	Az intelligens ember képes az igazság megértésére, rációt alkalmazni és az erkölcsi értékek alapján cselekszik (társadalmi hasznosság)
	Középkori filozófia	
Szent Ágoston (354–430)	Szemben az ókori görög alapokkal (intellektus, logika, tudás) a hit a híd az érzéki világ és értelem világa között; patrisztikus filozófia: isteni igazságok megértése (szentírások) és alkalmazása	Keresztény eszményekben és erkölcsi életvitelben gyökerezik (közeledés az emberi boldogság és Isten felé); szabad akarat és kapcsolódó felelősségvállalás
Aquinói Szent Tamás (1225–1274)	Középkor későbbi szakaszában domináns skolasztikus filozófia: természetes és vallási igazságok harmóniája, az érvelés és racionalitás útját követve – arisztotelészi alapok; értelem az isteni igazság megértését szolgálja	Istenhittel kapcsolatos erények; szabad akarat és felelősség kapcsolódása; az erkölcs az emberi természet tiszteletén alapul és az értelem teremti meg, segítve a társadalmi jólétet

William Ockham (1287–1347)	Skolasztikus filozófia, egyetemes fogalmak névlegesek (nominálisak), mentális konstrukciók csupán, amelyek objektív valóságként nem léteznek; kritika és szkepticizmus, egyszerűség elve fontos a tapasztalati megismerésben, míg a feltételezéseket minimalizálni kell	Az etikus normákkal kapcsolatos fogalmak önmagában nem léteznek, szubjektíven személyhez, eseményhez rendelten értelmezhetők, az etikai értékek társadalmi konvenciók eredményei és önmagukban nem léteznek
Pázmány Péter (1570–1637)	Arisztotelészi alapok, de skolasztikus filozófiai elemek: az értelem (világ megismerése) és istenhit (élet mélyebb megismerése) harmóniájának vizsgálata	Erkölcsi értékeken nyugvó boldog életre való törekvés (amely az isteni és kapcsolódóan az emberi élet rendjében található meg)
Újkori filozófia		
René Descartes (1596–1650)	Racionalizmus; visszatérés az alapokhoz; a komplex gondolkodási folyamatok és tudat működésének vizsgálata; a tudományos módszerek alkalmazása segít az objektív valóság megismerésben; létezik egy nem anyagi (szellemi) világ is (megismerés belső megélése alapján)	Etikai felfogásának két eleme: szabadság és önrendelkezés (lelkiismeret-alapú cselekedet); az erkölcsi kötelesség rendszert ad az életnek; kimenet (ok-okozat) nem mindig egyértelmű; racionális döntéshozatal és a következmények viselése
Gottfried Wilhelm Leibniz 1646–1716)	Az emberi intellektus (értelem) segítheti a magasabb tudás (isteni szándék) megértését; a tudás automatizálásáról értekezik; gondolkodásának központi eleme a „monád" (világegyetem építőköve, zárt rendszer, amelyben a monádok közötti állapotok tükröződnek)	Kompozitelmélet: az egyetemes jó három részre osztása: metafizikai, morális és fizikai jó; monádok alapján: az egyéni és társadalmi boldogság összhangban állnak egymással: együttműködés a nagyobb jóért
Francis Bacon (1561–1626)	A tudományos módszerek kidolgozása; a tudomány empirikus (objektív) megfigyelésen alapulhat	Erények kialakítása a társadalom feladata is (oktatás és természet megismerése)
Thomas Hobbes (1588–1679)	Társadalmi szerződés elmélete (részleges szabadságkorlátozás); a természet állapota az emberi önzés miatt káoszhoz vezet, ezért szükséges a közösség érdekében (rend, biztonság)	Az erkölcs és az etikai értékek társadalmi konvenciók eredményei; a törvények, normák betartása az erkölcsös magatartást tükrözi
John Locke (1632–1704)	A tudás forrása a tapasztalat, amely az érzékek által szerzett adatokra épül (nincs a priori ismeret); az állam feladata az emberek jogainak védelme	A tolerancia (másokét nem sértő, saját jogok szabad gyakorlása), az egyéni szabadság és önrendelkezés hangsúlyos
David Hume (1711–1776)	Az intelligencia az érzékszervek által észlelt tapasztalatokból fejlődik (szkeptikus szerepe), hogy megértse a világ törvényszerűségeit	Az érzelmekből származó morál megítélése szubjektív és relatív; az erkölcsi normák a társadalmi együttélés konvenciói
Jean-Jacques Rousseau (1712–1778)	Társadalmi szerződés témaköre: az emberek alapvetően boldogok, amíg a társadami korlátok, osztályok és hierarchiák megfosztják ettől – visszatérés a természetes állapothoz	Fontos az együttérzés és empátia, az altruizmus, amelyet a társadalmi konvenciók hajlamosak elfojtani; egyéni szabadságjogok tisztelete

Modern újkori filozófia		
Immanuel Kant (1724–1804)	Kritikus filozófia és általános (társadalmi) akarat (etikai normák és törvények); autonómia (az ész alapján megvalósuló egyéni cselekvés)	Az erkölcsös cselekedet azonos az emberi ész által meghatározott törvények betartásával
John Stuart Mill (1806–1873)	Utilitarizmus: az egyéni szabadságokjogok védelme (törvények, társadalmi normák által is) demokrácia – a kormányzás a nép érdekeit szolgálja	A cselekedetek kiértékelése a következmények, eredmények alapján (társadalmi hasznosság maximalizálása)
Kortárs, jelenkori filozófia		
Daniel Dennett (1942–	Az elmét, a tudatot és a szabad akaratot a természettudományok révén célszerű vizsgálni; a gondolatok és élmények komplex kognitív folyamatok eredményei, amelyet kognitív korlátok és hibák torzíthatnak; önálló döntési képesség algoritmusok által meghatározott	Az etikai normák kultúránként eltérő társadalmi megállapodások eredményei, amelyek az idők múlásával változhatnak; etikus döntéshozatal az értelem és szabad akarat alapján hat a társadalmi értékteremtésre
Martha Nussbaum (1947–	Az egyének jogait, és méltóságát a társadalomnak is tiszteletben kell tartania, és védenie szükséges, függetlenül az egyén státuszától, cselekedeteitől, tapasztalataitól	Az emberi jólét nélkülözhetetlen elemei: az empátia és erkölcsi érzékenység, ennek következtében megvalósul az emberi méltóság védelme
Nick Bostrom (1973–	A tudományos-technológiai fejlődés hatásait és kihívásait vizsgálja: antirealizmus elve (az észlelt világ csak torzított, részleges kép a valóságról); transzhumanizmus elve (az emberi potenciál kibontakoztatása)	A biztonsági fékek, garanciák kidolgozása és beépítése az AI fejlesztési folyamatába létfontosságú az emberiségre a katasztrofális következmények elkerülése érdekében

HIVATKOZÁSOK

Anderson, M. – Anderson, S. L. (2011): *Machine Ethics*. Cambridge, MA: Cambridge University Press, https://doi.org/10.1017/CBO9780511978036.

Anderson, S. L. (1995): Being Morally Responsible for an Action Versus Acting Responsibly or Irresponsibly. *Journal of Philosophical Research,* 20, 453–62, https://doi.org/10.5840/jpr_1995_10.

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, March 1942. Online: https://web.williams.edu/Mathematics/sjmiller/public_html/105Sp10/handouts/Runaround.html. Magyarul: Körbe-körbe. In Asimov, I.: *Robottörténetek*, http://users.atw.hu/asimov/downloads/Encyclopedia%20Galactica/01.%20k%C3%B6tet%20-%20Encyclopedia%20Galactica/Isaac%20Asimov%20-%20Robott%C3%B6rt%C3%A9netek.pdf.

Asimov, I. (1950). I, Robot. *Gnome Press*. Online: http://ekladata.com/-Byix64G_NtE0xI4A6PA1--01Hc/Asimov-Isaac-I-Robot.pdf. Magyarul: Asimov, I. (2019): Én, a robot (fordította: Békési József, Vámosi Pál). Budapest: GABO.

Awad, E. – Dsouza, S. – Kim, R. –Schulz, J. – Henrich, J. – Shariff, A. – Bonnefon, J-F. – Rahwan, I. (2018): The Moral Machine Experiment. *Nature,* 563(7729), 59–64, https://doi.org/10.1038/s41586-018-0637-6 .

Beijing Academy of Artificial Intelligence (2019): Beijing AI Principles. Datenschutz und Datensicherheit 43, 656. Online: https://www.semanticscholar.org/paper/Beijing-AI-Principles/a703873b9c697c05b9146a5df790745b6f303857.

Cointe, N. – Bonnet, G. (2016): Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. *AAMAS*, 1106–1114.

Dennett, D. – Fleig-Goldstein, B. –Friedman, D. (2019): Dennett Explained: An interview with Daniel Dennett. *ALIUS Bulletin,* 3, 11–25, https://doi.org/10.34700/7gkw-zh08.

Dennett, D. (1984): Cognitive Wheels: The Frame Problem of AI. *Minds, Machines, and Evolution.* Cambridge: Cambridge University Press, 129–152. Online: https://www.researchgate.net/publication/225070451_Cognitive_Wheels_The_Frame_Problem_of_AI.

Dennett, D. (2019): What can we do? In Brockman, J. (ed.) (2019): *Possible Minds: 25 Ways of Looking at AI* (Chapter 5). Online: https://ase.tufts.edu/cogstud/dennett/papers/What_Can_We_Do.pdf.

Denning, P. J. – Denning, D. E. (2020): Dilemmas of artificial intelligence. *Communications of the ACM*, 63(3), 22–24, https://doi.org/10.1145/3379920 .

Dreyfus, H. L. (1972): *What computers* still *can't do: A critique of artificial reason.* Cambridge, MA: MIT Press. Online: https://terrorgum.com/tfox/books/whatcomputersstillcantdo_acritiqueofartificialreason.pdf.

Dreyfus, H. L. (2007): Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian. *Artificial Intelligence,* 171(18), 1137–1160, https://doi.org/10.1016/j.artint.2007.10.012.

Drucker, F. P. (2001): What is business ethics? *The Public Interest*, 35, 18–36. Online: https://edisciplinas.usp.br.

Európai Bizottság (2018): COM(2018) 237 final. Online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN.

Európai Bizottság (2019): Megbízható mesterséges intelligenciára vonatkozó etikai iránymutatás. Online: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-hu/format-PDF.

Európai Bizottság (2020): The Assessment List for Trustworthy Artificial Intelligence (ALTAI). Online: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

Európai Bizottság (2021): A mesterséges intelligenciára vonatkozó harmonizált szabályok (a mesterséges intelligenciáról szóló jogszabály) megállapításáról és egyes uniós jogalkotási aktusok

módosításáról. COM(2021) 206 final. Online: https://eur-lex.europa.eu/legal-content/HU/TXT/HTML/?uri=CELEX:52021PC0206.

Európai Bizottság (2022): Az EU új megközelítéssel törekszik arra, hogy az uniós szabványok világszerte meghatározóak legyenek, hirdessék az uniós értékeket, valamint előmozdítsák a reziliens, zöld és digitális egységes piacot. Online: https://ec.europa.eu/commission/presscorner/detail/hu/ip_22_661.

Európai Parlament (2019): (EU) 2019/1020 Rendelete. A piacfelügyeletről és a termékek megfelelőségéről, valamint a 2004/42/EK irányelv, továbbá a 765/2008/EK és a 305/2011/EU. rendelet módosításáról. Online: https://eur-lex.europa.eu/legal-content/HU/TXT/PDF/?uri=CELEX:32019R1020&from=DE.

Európai Tanács (2020): Az Európai Tanács rendkívüli ülése (2020. október 1–2.) – Következtetések, EUCO 13/20, 2020, 6. o. 4. Online: https://www.consilium.europa.eu/media/45921/021020-euco-final-conclusions-hu.pdf

Európai Bankhatóság (2021): EBA Discussion Paper on Machine Learning for IRB Models, https://www.eba.europa.eu/.

Európai Bankföderáció (2019): EBF position paper on AI in the banking industry, https://www.ebf.eu.

Executive Office of the President National Science and Technology (2016): Preparing For The Future Of Artificial Intelligence. Online: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

Floridi, L. – Chiriatti, M. (2020): GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines*, 30, 681–694, https://doi.org/10.1007/s11023-020-09548-1.

Floridi, L. – Cowls, J. (2019): A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review,* https://doi.org/10.1162/99608f92.8cd550d1 .

Friedman, B. – Nissenbaum, H. (1996): *Bias in Computer Systems*. In Weckert, J. (ed.) *Computer Ethics* (Chapter 20.). London: Routledge, https://doi.org/10.4324/9781315259697-23 .

Hagendorff, T. (2020): The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30, 99–120, https://doi.org/10.1007/s11023-020-09517-8 .

Héder Mihály (2020): *Mesterséges intelligencia – Filozófiai kérdések, gyakorlati válaszok*. Budapest: Gondolat Kiadó. ISBN: 9789635560509.

Institute of Electrical and Electronics Engineers (2016): The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Online: https://standards.ieee.org/wp-content/uploads/import/documents/other/ec_about_us.pdf.

Institute of Electrical and Electronics Engineers (2019): Ethical Aspects of Autonomous and Intelligent Systems. Online: https://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE19002.pdf.

Institute of Electrical and Electronics Engineers (2021): The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems Industry Connections Activity Initiation Document (ICA-ID). Online: https://standards.ieee.org/wp-content/uploads/import/governance/iccom/IC16-002-Global_Initiative_for_Ethical_Considerations_in_the_Design_of_Autonomous_Systems.pdf.

Jobin, A. – Lenca, M. – Vayena, E. (2019): The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1, 389–399, https://doi.org/10.1038/s42256-019-0088-2 .

Kirkpatrick, J. (2015): Drones and the Martial Virtue Courage. *Journal Of Military Ethics*, 14(3–4), 202–219, https://doi.org/10.1080/15027570.2015.1106744 .

McCarthy, J. – Minsky, M. L. – Rochester, N. – Shannon, C. E. (1955): A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*, 27(4), 12, https://doi.org/10.1609/aimag.v27i4.1904.

Von Neumann, J. (1958): *The Computer and the Brain*. Yale University Press, Inc. Online: https://complexityexplorer.s3.amazonaws.com/supplemental_materials/5.6+Artificial+Life/The+Computer+and+The+Brain_text.pdf.

Von Neumann, J. (1963): The General and Logical Theory of Au-tomata. Online: https://www.semanticscholar.org/paper/The-General-and-Logical-Theory-of-Au tomata-Neumann/e8538f11920fa6e56b3d34771bb330bd3e07281d.

Nilsson, N. J. (2010): *The Quest For Artificial Intelligence. A History of Ideas and Achievements*. Cambridge University Press. https://doi.org/10.1017/cbo9780511819346.

OECD (2019): Recommendation of the Council on Artificial Intelligence. Online: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Prisznyák, Alexandra (2023a): Mesterséges intelligencia a bankszektorban. Könyvismertető. *Gazdaság és Pénzügy*, 9(3), 395–401, https://doi.org/10.33926/gp.2022.4.6.

Prisznyák, Alexandra (2023b): A természetes intelligencia manifesztációjának filozófiai kérdései. *Hitelintézeti Szemle*, 22(1), 166–170. Online: https://hitelintezetiszemle.mnb.hu/letoltes/hsz-22-1.pdf.

Reiter, E. – Dale, R. (2021): Building Applied Natural Language Generation Systems. *Natural Language Engineering* , 27(1) , 113–118, https://doi.org/10.1017/s1351324997001502 .

Searle, J. R. (1980): Minds, Brains and Programs. *Behavioral and Brain Science*, 3(3), 417–424, https://doi.org/10.1016/b978-1-4832-1446-7.50007-8 .

Sejnowski, T. J. (2023): Large Language Models and the Reverse Turing Test. *Neural Computation*, 35, 309–342, https://doi.org/10.1162/neco_a_01563.

The Guardian (2020): A robot wrote this entire article. Are you scared yet, human? Online: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

Török, Bernát – Ződi, Zsolt (2021): *A mesterséges intelligencia szabályozási kihívásai – Tanulmányok a mesterséges intelligencia és a jog határterületeiről.* Budapest: Ludovika Egyetemi Kiadó. ISBN: 9789635314836.

Turing, A. M. (1950): Computing Machinery and Intelligence. *Mind*, 59(236), 433–460, https://doi.org/10.1093/mind/lix.236.433.

UNESCO (2020): Recommendation on the Ethics of Artificial Intelligence. Online: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

Veruggio, G. (2007): *The EURON roboethics roadmap*. 2006 6th IEEE-RAS International Conference on Humanoid Robots (Red Hook, NY; Genoa: IEEE), 612–617, https://doi.org/10.1109/ichr.2006.321337.

Wang, P. (2019): On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, 10(2), 1–37, https://doi.org/10.2478/jagi-2019-0002 .

Weizenbaum, J. (1976): *Computer power and Human Reason: From Judgment to Calculation*. New York: W. H. Freeman and Company. Online: http://blogs.evergreen.edu/cpat/files/2013/05/Computer-Power-and-Human-Reason.pdf.

Wiener, N. (1948): *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/11810.001.0001.

Yu, H. – Liu, Z.; Wang, Y. – Jiang, X. (2018): Building Ethics into Artificial Intelligence. *IEEE Intelligent Systems*, 33(4), 77–83. https://doi.org/10.24963/ijcai.2018/779 .

A filozófiai részhez használt forrás:

https://plato.stanford.edu/index.html