A biztonság értelmezése emberek nélkül – Az emberi tényező szerepe a mesterséges intelligencia korában

Molnár Balázs

PhD-hallgató, Óbudai Egyetem, Biztonságtudományi Doktori Iskola
molnar.balazs@bgk.uni-obuda.hu

Absztrakt: A tanulmány központi kérdése, hogy értelmezhető-e a biztonság fogalma emberek nélkül. A biztonság hagyományosan az emberi élet, egészség, jogok és bizalom védelméhez kapcsolódik, azonban a technológiai fejlődés – különösen a mesterséges intelligencia és az autonóm rendszerek térhódítása – új megközelítést igényel. A kutatás hipotézise szerint a biztonság két szinten értelmezhető: rendszerszintű stabilitásként és emberi értékalapú védelemként. A tanulmány bemutatja, hogy míg az előbbi a technikai működés hibamentességére és integritására épül, az utóbbi a bizalomra, etikai felelősségre és jogi megfelelésre. A tanulmány újszerűsége abban áll, hogy különválasztja ezt a két szintet, miközben rámutat: a valódi biztonság fogalma az emberi érdekekhez kötött, de a technológiai fejlődés hatására megjelenhet egy ember nélküli, önállóan működő biztonságfelfogás is.

Kulcsszavak: biztonság, mesterséges intelligencia, szervezeti integritás, bizalom

1 A biztonság emberi és technikai értelmezésének dilemmája

Ahogy Giddens rámutat, a modern társadalmakban a biztonság és a bizalom egyre inkább rendszerekbe vetett hitként jelenik meg, nem pedig közvetlen emberi tapasztalatként. Ez a „reflexív biztonság” a technológiai működés stabilitásába vetett bizalomra épül, amely az emberi felelősséget részben kiszervezi a rendszerekbe (Giddens, 1990).

A biztonság emberek nélküli értelmezése különválasztható az emberekhez kötődő biztonságtól, amely a bizalom és a védelem dimenziójában értelmezhető. Míg az emberhez kapcsolódó biztonság középpontjában a védelem erkölcsi és társadalmi célja – az emberi élet, méltóság és jogok megőrzése – áll, addig az emberek nélküli biztonság technikai értelemben a rendszer önfenntartó stabilitását jelenti. Ez utóbbi esetben a biztonság nem értékhez, hanem működéshez kötődik: a

rendszer célja önmaga fenntartása, nem pedig az ember védelme. E két megközelítés közötti határvonal az autonóm technológiák fejlődésével egyre inkább elmosódik. A biztonság így kettős értelmű fogalommá válik: egyszerre jelenti a technikai működés folytonosságát és az emberi érdekek védelmét. A kihívás abban rejlik, hogy a technikai stabilitás ne váljon öncélúvá, hanem továbbra is az emberi bizalom és felelősség szolgálatában maradjon (Beck, 1992).

A biztonság valódi társadalmi és szervezeti jelentése mindig az emberekhez kapcsolódik. Itt a biztonság az egyének életét, egészségét, jogait és adatait védi, és szorosan kapcsolódik a bizalomhoz. Ez a dimenzió adja meg a biztonság normatív tartalmát: mit és kit kell védeni, mi számít veszélynek, és hogyan kapcsolódik mindez az etikai, jogi és társadalmi normákhoz.

A technikai, algoritmikus vagy gépi biztonság felfogható úgy, mint a hibamentes működés, az önfenntartás és az integritás fenntartása. Jó példa erre egy autonóm rendszer, amely képes önmagát védeni a hibás bemenetekkel szemben, vagy egy kiberbiztonsági architektúra, amely akkor is működik, ha nincs közvetlen emberi felügyelet. Ez a biztonság-értelmezés „önmagáért való", és a rendszer fennmaradását célozza. Az autonóm rendszerek képesek saját integritásuk fenntartására.

A biztonság ez esetben fogalmi kettősség: értelmezhető funkcionális és normatív értelemben. Ugyanakkor emberi perspektíva nélkül nincs érték: a biztonság emberi értelmezés nélkül puszta technikai jelenség. Durkheim szerint a társadalmi stabilitás az emberi kapcsolatok biztonságán alapul (Durkheim, 1893). A szervezeti biztonság és a bizalom összefüggése: a technikai védelem az emberek bizalmát szolgálja. A biztonság értelmezése emberek nélkül erősen korlátozott, mert maga a „biztonság" fogalom az emberi élethez, működéshez és bizalomhoz kapcsolódik.

A két értelmezés (rendszerstabilitás és emberi biztonság) egymásra épül. A technikai rendszerek önálló biztonsága önmagában még nem indokolható, csak akkor válik jelentőssé, ha az emberek biztonságát és bizalmát szolgálja. Ugyanakkor egyre inkább látszik, hogy egyes mesterséges intelligencia rendszerek képesek lehetnek saját „működési biztonságukat" megőrizni, függetlenül az emberi jelenléttől. Ez a tendencia új kihívást jelent a jog és az etika számára.

A szervezeti biztonság – legyen szó fizikai védelemről, kiberbiztonságról, objektumvédelemről vagy munkavédelemről – mindig az emberek érdekeit szolgálja: az alkalmazottakét, ügyfelekét vagy társadalmi közösségekét. Az emberi vonatkozás és beavatkozás nélkül a biztonság absztrakció maradna. A szervezeti biztonság lényege az arányosság: mennyit áldozunk a védelemre, milyen szabadságot korlátozunk, milyen adatot gyűjtünk és hogyan használjuk fel. E mérlegelés értékelvű – nélkülözhetetlen hozzá az emberi kontextus. A tisztán technikai optimalizáció könnyen vált ki bizalomvesztést és ellenállást, ami végső soron épp a biztonságot gyengíti.

2 A technikai stabilitás mint biztonsági paradigma

A rendszerszintű biztonság a technikai rendszerek önazonosságának megőrzésére, hibamentes működésére és működési céljaik elérésére vonatkozó képesség. E dimenzióban a biztonság a megbízhatóság (reliability), rendelkezésre állás (availability), integritás (integrity) és ellenállóképesség (resilience) mérhető mutatóival írható le. Egy modern vállalati környezetben ezt a szemléletet tipikusan az információbiztonság (CIA-triád), a kiberbiztonsági védelem, a fizikai hozzáférés-vezérlés és az üzemeltetési folyamatok automatizált monitorozása testesíti meg (Stallings, 2019; ISO/IEC, 2024; Von Solms & Van Niekerk, 2013).

Az autonóm védekezésre képes rendszerek – például önbeavatkozó behatolás-megelőző rendszerek (IPS), anomália-alapú észlelők idővel a saját működési stabilitásukat tekintik elsődleges célnak. Ez létrehozza az „önvédelmi rendszerek paradoxonát”: minél jobban védi magát a rendszer, annál inkább elszakadhat az emberi céloktól. Egy túl agresszív automatizált blokkolás például üzletileg kritikus forgalmat is elvághat (false positive), miközben a mérőszámok szerint a rendszer „biztonságosabbá” vált. A technikai biztonság tehát könnyen elszakadhat az emberi értéktől.

A gépi tanulással működő védelmi rendszerek sokszor nem transzparensek. Ha egy modell döntését nem lehet visszafejteni, akkor a szervezet nem képes jogilag és etikailag felelősen elszámolni a következményekről. Ez különösen kockázatos olyan szabályozott területeken, ahol a döntések auditálhatósága jogszabályi kötelezettség. A technikai verifikálhatóság hiánya közvetlenül rombolhatja az alkalmazottak és ügyfelek bizalmát.

A szervezetek gyakran előbb vezetnek be automatizált védelmeket, mintsem megfogalmaznák, pontosan milyen értékekhez kell ezeknek igazodniuk (pl. szolgáltatás-folytonosság kontra adatintegritás). Ez „értékillesztési adósságot” hoz létre: a rendszer formálisan biztonságos, de a rosszul priorizált célfüggvény miatt a szervezeti értékeket – hosszú távú reputáció, jogkövetés, felhasználói élmény – sértheti.

A technikai biztonság minimális feltétele az önkorrekció és a reziliencia: a rendszer képes legyen hibatűrésre, gyors helyreállításra és kontrollált degradációra. A reziliencia-centrikus tervezés (graceful degradation, chaos engineering) abban segít, hogy a rendszer emberi beavatkozás nélkül is visszanyerje működőképességét (Basiri et al, 2016; Hollnagel et al, 2006) – de ez önmagában továbbra is csak stabilitás, nem pedig morális értelemben vett biztonság. Ez a fajta önkorrekció nem helyettesíti az etikai felelősséget, csupán technikai stabilitást biztosít.

3 Az autonóm döntéshozatal felelősségi kérdései

A technikai rendszerek stabilitása csak addig nevezhető biztonságnak, amíg az emberi felelősség a döntési folyamat része marad. A támadó–védő dinamika napjainkban sokszor gépek közötti. Automatizált exploit-kampányokra automatizált észlelők és válaszok reagálnak. A túlautomatizálás azonban az egész rendszerre kiterjedő kockázatot hoz: egy rosszul tanított detektor láncreakciót indíthat, amely a teljes szolgáltatást megbénítja. A szervezeti felelősség ilyenkor nem hárítható a gépre: az ellenőrizhetőség és a visszavonhatóság („human override") minden esetben követelmény kell, hogy legyen (Floridi & Cowls, 2021; OECD, 2021; European Union, 2024.)

A „humán-in-the-loop" (HITL) fogalom a mesterséges intelligencia és az automatizált döntéshozatal egyik központi elve, amely szerint az ember a döntési folyamatban nem kizárható, hanem aktív felügyeleti és korrekciós szerepet kell betöltenie (Floridi & Cowls, 2021; OECD, 2021; European Union, 2024). A biztonság szempontjából ez a modell azt jelenti, hogy a technológiai rendszerek autonómiája csak addig terjedhet, amíg az ember megőrzi a végső döntési jogot és felelősségét (OECD, 2021; DARPA, 2019).

A „human-in-the-loop-by-design" megközelítés szerint a rendszert már tervezéskor úgy kell kialakítani, hogy beépített legyen az emberi döntés, korrekció és etikai mérlegelés lehetősége. Ez a szemlélet nemcsak technológiai, hanem társadalmi elvárás is: a gép és ember közötti biztonsági felelősség közös, nem alternatív.

A humán kontroll három szintje különíthető el: a „human in the loop" (az ember minden döntést megelőzően beavatkozik), a „human on the loop" (felügyeletet gyakorol és szükség esetén módosít), valamint a „human out of the loop" (a döntés teljesen autonóm). A biztonsági gyakorlatban az első két szint a kívánatos: az EU mesterséges intelligenciáról szóló rendelete (AI Act, 2024) is előírja, hogy a kritikus döntésekben az embernek meg kell őriznie a beavatkozási lehetőséget.

A szervezeti biztonságban a HITL három fő funkciót tölt be: (1) kontrollmechanizmus, amely az emberi jóváhagyást és auditot biztosítja, (2) tanulási visszacsatolás, mivel a gép az emberi korrekciókból fejlődik, és (3) etikai garancia, amely a döntések morális hátterét adja. A „human feedback loop" ezért nem csupán technikai, hanem etikai és jogi kategória is (Floridi & Cowls, 2021).

A biztonság hatékonyságát nem a gép–ember arány maximalizálása, hanem a szerepek helyes megosztása növeli. A gép gyors, ismételhető és fáradhatatlan; az ember képes kontextust, értéket és arányosságot mérlegelni. A „human-in-the-loop" szerep azt jelenti, hogy az ember dönt a határhelyzetekben, és visszavonhatja a gépi döntéseket, amelyek aránytalan sérelmet okoznának.

A gépi döntések emberi felelősség alól nem mentesítenek. A szervezeti integritást támogató mechanizmusok – például a visszaélés-bejelentési csatornák – épp azt

biztosítják, hogy a biztonsági incidensek és etikai vétségek következménnyel járjanak. Az elszámoltathatóság a bizalom előfeltétele: ha nincs felelős, nincs biztonság sem.

4 Emberi dimenziójú biztonság

Az emberi dimenzió a biztonság morális tengelye: itt dől el, hogy a rendszer stabilitása valóban jót jelent-e valakinek. A szervezeti biztonság nemcsak külső fenyegetések kivédéséről szól, hanem arról is, hogy a szervezet képes-e kiszámítható, értékalapú módon működni – ez hozza létre az alkalmazotti és ügyfélbizalmat. A bizalom hiánya közvetlen termelékenységi és reputációs kockázat.

A bizalom több szinten is értelmezhető: személyes (munkatárs–munkatárs), vertikális (vezető–munkavállaló) és intézményi (szervezet–stakeholderek). Mayer, Davis és Schoorman integrált bizalommodellje alapján a kompetencia, jóindulat és integritás észlelése teremti meg a bizalmat (Mayer et al, 1995). Ha a biztonsági gyakorlatok nem transzparensek, a munkavállalók a kontrollt „felügyeletként" élik meg, ami csökkenti az elköteleződést és a bizalmat. Az ideális biztonság nem láthatatlan, hanem érthető és arányos.

Schein szerint a kultúra mélyben lévő feltételezésekből, deklarált értékekből és látható mesterségesen létrehozott tárgyakból áll (Schein, 2010). Ha a mélyben az a feltevés él, hogy „az emberek jelentik a legnagyobb kockázatot", a biztonsági rendszer ellenállást és rejtett megkerülési praktikákat szül (Schein, 2010). Ezzel szemben az etikus, részvételi kultúra a biztonság felelősségét megosztja: a munkavállaló nem tárgya, hanem alanya a biztonságnak.

5 A biztonság értékalapú újraértelmezése

A biztonság nem pusztán állapot, hanem érték: valami jó, mert valakit véd. A működési stabilitás – legyen az technikai, fizikai vagy kiberbiztonsági rendszer – önmagában semleges jelenség. Csak akkor válik biztonsággá, ha az emberi érdek, a méltóság vagy a közösség védelme társul hozzá. A biztonság tehát relációs fogalom: mindig valakihez, valamihez viszonyítva nyer értelmet.

Ez a megközelítés elválasztja a technikai biztonságot a morális biztonságtól. Egy rendszer lehet hibamentes, mégsem tekinthető biztonságosnak, ha közben emberi jogokat, bizalmat vagy társadalmi értékeket sért. A biztonság lényegéhez ezért hozzátartozik a „kinek az érdekében" és a „mitől" kérdése. Emberi jelenlét nélkül nincs, akinek az érdekei sérülnének vagy védelmet igényelnének; a „veszély" és

„kár" fogalmaknak nincs értelmezési referenciája. A gép számára a hiba csak eltérés a specifikációtól, nem morális rossz.

Ahogy Hans Jonas megfogalmazza: „A felelősség mindig valakiért van, soha nem önmagáért." (Jonas, 1997) E gondolat a biztonságra is érvényes: nincs felelősség, ha nincs védendő létező. A biztonság így nem csupán cél, hanem értékalapú ítélet – az emberi közösség döntése arról, mit akar megőrizni a jövő számára.

Az értékek – élet, méltóság, szabadság, tulajdon – emberi konstrukciók és tapasztalatok. A biztonság ezek intézményesített védelme. Ha kivesszük az embert a képletből, a biztonság fogalma a rendszerelmélet „stabilitására" zsugorodik. A két fogalom szinonimaként kezelése kategóriahibához vezet.

A biztonság definíciója magában foglalja a felelősséget: ki dönt, ki viseli a következményeket, ki kap jóvátételt? Ezek a kérdések csak az emberek világában értelmezhetőek. A gép nem felelős, így a „gépi biztonság" legfeljebb műszaki minősített állapot. Ebből következően a technikai biztonság önmagában nem értékelhető, csak az emberi célrendszeren keresztül nyer jelentést. A humán-in-the-loop hiánya jogi és bizalmi deficitet okoz. A GDPR, az Ibtv. és az ISO 27001 szabvány is előírja a döntési folyamatok elszámoltathatóságát, amit csak emberi részvétellel lehet biztosítani. Ha az ember kikerül a döntésből, a rendszer működőképes maradhat, de elveszíti a biztonság morális és társadalmi értelmét – „biztonságos lesz, de nem emberi módon". A humán-in-the-loop ezért nem egyszerű technikai kontroll, hanem a biztonság emberi és erkölcsi dimenziójának garanciája.

Következtetések

A filozófiai és ontológiai érvelés alapján a biztonság, mint fogalom ember nélkül elveszíti értékdimenzióját, mert hiányzik belőle az, akinek az érdekeit védené. A szervezeti biztonság példáján keresztül a dolgozat rávilágít arra, hogy a humán tényező a biztonság morális középpontja, és hogy az emberi döntésképesség és felelősség nem váltható ki algoritmusokkal. A tanulmány részletesen elemzi a humán-in-the-loop koncepció szerepét, amely az emberi kontroll és beavatkozás különböző szintjein keresztül biztosítja a technológiai rendszerek etikus működését.

Az ember mint morális kontrollmechanizmus biztosítja, hogy a technikai biztonság ne váljon értéksemlegessé. A gép statisztikai, az ember erkölcsi megfontolás alapján dönt. Az ember képes felismerni az aránytalanságot, a méltányosság hiányát, és érzékelni a társadalmi következményeket. „Az emberi felelősség nem zárható ki a döntésekből anélkül, hogy el ne veszne az emberi méltóság maga" (Arendt, 1958).

A biztonság emberek nélkül legfeljebb technikai stabilitásként értelmezhető, de valódi társadalmi értékként csak az ember által nyer jelentést, vagyis emberek nélkül nem létezik biztonság. Ha a képletből kikerülne az ember, akkor gépek

biztonságáról sem beszélhetnénk, így a biztonság kizárólag emberi szinten lehet hitelesen megfogható és értelmezhető.

Összefoglalás

Az eredmények szerint a technikai biztonság önmagában értéksemleges, és csak akkor válik valódi biztonsággá, ha emberi érdekhez, felelősséghez és döntéshozatalhoz kötődik. A tanulmány különválasztja a technikai és morális biztonság fogalmát, és rámutat, hogy a humán-in-the-loop szerepe a jövő biztonságrendszereinek alapvető etikai és jogi garanciája. Az ember a biztonság értelmének kulcsa, amíg a döntés a kezében marad

A mesterséges intelligencia és autonóm rendszerek kora arra kényszerít bennünket, hogy újraértelmezzük a biztonság fogalmát: a valódi biztonság az emberi értékek megőrzésében mérhető. A technológia képes stabilitást teremteni, de nem képes erkölcsi döntéseket hozni. A biztonság ezért mindig az emberi döntésképesség, felelősség és bizalom központi szerepe marad.

A jövő biztonságrendszereinek nem az a céljuk, hogy kizárják az embert, hanem hogy az emberi értékeket integrálják a technológiai folyamatokba. Ez az etikai integráció nem pusztán morális igény, hanem a rendszerek fenntartható működésének feltétele.

Végső soron a biztonság nem a gépek, hanem az emberek közötti kapcsolat stabilitása. A technológiai rendszerek csak addig nevezhetők biztonságosnak, amíg a mögöttük álló társadalom azokat igazságosnak, elszámoltathatónak és emberközpontúnak tartja.

A jövő biztonságfilozófiája nem az ember kizárására, hanem az emberi értékek technológiai integrálására kell, hogy épüljön.

Hivatkozások

[1] Arendt, H. (1958). The Human Condition. University of Chicago Press.

[2] Basiri, A., Behnam, N., de Rooij, R., Hochstein, L., Kosewski, L., Reynolds, J., & Rosenthal, C. (2016). Chaos engineering. IEEE Software, 33(3), 35–41. https://doi.org/10.1109/MS.2016.60

[3] Beck, U. (1992). Risk Society: Towards a New Modernity. Sage Publications.

[4] DARPA (2019). Explainable Artificial Intelligence (XAI) Program. U.S. Defense Advanced Research Projects Agency. https://www.darpa.mil/research/programs/explainable-artificial-intelligence?utm_source=chatgpt.com (utolsó megtekintés dátuma: 2025. október 21.)

[5] Directive (EU) 2022/2555 of the European Parliament and of the Council (NIS2). Official Journal of the European Union. https://eurlex.europa.eu/eli/dir/2022/2555/oj/eng?utm_source=chatgpt.com (utolsó megtekintés dátuma: 2025. október 24.)

[6] Durkheim, É. (1893/1984). The Division of Labour in Society (W. D. Halls, Trans.). Macmillan. (Eredeti mű megjelent 1893-ban.)

[7] European Union. (2024). Artificial Intelligence Act (Regulation (EU) 2024/1689 of the European Parliament and of the Council). Official Journal of the European Union, L 202, 1–157. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689 (utolsó megtekintés dátuma: 2025. október 24.)

[8] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

[9] Giddens, A. (1990). The Consequences of Modernity. Stanford University Press.

[10] Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.). (2006). Resilience engineering: Concepts and precepts. Ashgate.

[11] Jonas, H. (1997). A felelősség elve: Kísérlet az etika új megalapozására. Osiris Kiadó. (Eredeti mű 1979-ben: The Imperative of Responsibility: In Search of an Ethics for the Technological Age.)

[12] Magyarország. (2013). 2013. évi L. törvény az állami és önkormányzati szervek elektronikus információbiztonságáról. Magyar Közlöny. https://net.jogtar.hu/jogszabaly?docid=A1300050.TV (utolsó megtekintés dátuma: 2025. október 21.)

[13] Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

[14] MSZ ISO/IEC 27001:2023. Információbiztonság, kiberbiztonság és a magánélet védelme. Információbiztonság-irányítási rendszerek. Követelmények. Magyar Szabványügyi Testület, Budapest.

[15] OECD. (2021). State of implementation of the OECD AI principles. OECD Digital Economy Papers, No. 312. OECD Publishing. https://doi.org/10.1787/1cd40c44-en (utolsó megtekintés dátuma: 2025. október 22.)

[16] Schein, E. H. (2010). Organizational Culture and Leadership (4th ed.). Jossey-Bass.

[17] Stallings, W. (2019). Information security: Principles and practice (3rd ed.). Pearson Education.

[18] Von Solms, B., & Van Niekerk, J. (2013). From information security to cyber security. Computers & Security, 38, 97–102. https://doi.org/10.1016/j.cose.2013.04.004