FELFEGYVERZETT MI NARRATÍVÁK – KÉPZELET VAGY ELŐREJELZÉS?

BIRHER NÁNDOR[1] – HAJDU GÁBOR[2] – NAGY ANDRÁS[3]

Absztrakt

A tanulmány a mesterséges intelligencia, az autonóm fegyverrendszerek és a science fiction narratívái közötti kölcsönhatásokat vizsgálja. Az interdiszciplináris, összehasonlító elemzés a felfegyverzett mesterséges intelligenciát megjelenítő, eltérő médiumokhoz tartozó művekben azonosított kockázati modelleket veti össze a jelenleg ismert autonóm és részben autonóm fegyverrendszerek technológiai sajátosságaival. A science fiction ebben a megközelítésben olyan spekulatív modellalkotásként értelmezhető, amely szélsőséges forgatókönyveken keresztül teszi vizsgálhatóvá a katonai automatizálás, az emberi kontroll gyengülése és a felelősség elmosódása közötti összefüggéseket. Az elemzés normatív és értelmezési keretét XIV. Leó pápa *Magnifica humanitas* enciklikája biztosítja. Az enciklika az MI lefegyverzését a technológiai verseny, a hatalomkoncentráció és az automatizált döntéshozatal társadalmi-etikai problémáira is kiterjeszti. Ez a szemlélet a gépi fenyegetés narratíváit az emberi célok, intézményi döntések és szabályozási hiányosságok összefüggésében helyezi el. A vizsgálat megállapítja, hogy a jelenlegi autonóm fegyverrendszerek súlyos katonai, jogi és etikai kockázatokat hordoznak, technológiai korlátaik miatt azonban még nem alkotnak önálló egzisztenciális fenyegetést. A rendszerszintű veszély olyan képességek összekapcsolódásával növekedhet meg, amelyek stratégiai autonómiát, hálózatos skálázhatóságot és az emberi ellenőrzéstől való fokozott függetlenséget eredményeznek. A megelőzés az érdemi emberi kontroll fenntartását és a jogi, etikai, társadalmi, valamint technikai normák összehangolt alkalmazását követeli meg.

Kulcsszavak: autonóm fegyverrendszerek, mesterséges intelligencia, science fiction, Magnifica humanitas, érdemi emberi kontroll

Bevezetés

Hogy a mesterséges intelligencia – beleértve a nagy nyelvi modelleket – milyen veszélyt jelent a társadalomra, egy teljesen nyitott kérdés. A válaszok skálája nagyon széles, a teljes nyugalomtól egészen az azonnali apokalipszisvárásig tart. Jelen tanulmány azt vizsgálja, hogy a science fiction irodalom milyen narratívákat fogalmaz meg a mesterséges intelligenciával és a robotikával kapcsolatban.

XIV. Leó pápa (2026) Magnifica humanitas című első enciklikájában kitüntetett figyelemmel fordult a mesterséges intelligencia, mint korunk egyik

[1] Birher Nándor Máté, Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar, Gondoskodáspolitika és Technológiai Innováció Kutatócsoport; birher.nandor@btk.ppke.hu; ORCID: 0000-0002-4979-5196

[2] Hajdu Gábor Tibor, Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar, Gondoskodáspolitika és Technológiai Innováció Kutatócsoport; hajdu.gabor@btk.ppke.hu; ORCID: 0009-0002-6065-6215

[3] Nagy András, Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar, Dékáni Hivatal; nagy.andras@btk.ppke.hu; ORCID: 0009-0004-9458-8884

meghatározó jelensége felé. Nem ellenségesen vagy a technikai fejlődést elutasítva, ugyanakkor nagyon határozottan rámutatva azokra a kockázatokra, amelyeket magában hordoz (Az MI „értékes eszköz, amely éberséget igényel"). Ilyen veszélyek például

a) a könnyű és gyors eredmények, amik ösztönözhetik a ráhagyatkozást és gyengíthetik a személyes kreativitást, illetve ítélőképességet,
b) az objektivitás látszata, ami elfedheti, hogy a rendszerek azok kulturális előfeltevéseit tükrözik, akik tervezték és betanították ezeket,
c) az emberi kommunikáció mesterséges utánzása, ami nem valódi kapcsolatokat épít, hanem csak azok látszatát kelti, és azzal a veszéllyel fenyeget, hogy az ember elveszítheti magát a vágyat is a valódi emberi kapcsolatok kialakítására. (Leo XIV, 2026, 100. pont)

A pápa ugyanennek az enciklikának egy pontján hivatkozik J.R.R. Tolkien, The Lord of the Rings. The Return of the King című művére (Leo XIV, 2026, 213. pont), de ugyanígy hivatkozhatna Isaac Asimov műveire is az MI kockázatait bemutató esettanulmányokként, mert a felsorolt kategóriák mindegyikére található legalább egy, de akár több – extrém gondolatkísérletként is felfogható – példa. A Hatalomérzet című novella által mutatott disztópiában az emberek már a legalapvetőbb matematikai műveleteket is elfelejtették a gépekre való hagyatkozás következtében – és természetesen ennek újra „feltalálása" után azonnal háborús célra kívánják használni a visszaszerzett tudást (Asimov, 1959). Az Acélbarlangok című regényben Daneel Olivaw robot emberi társa életét is felteszi az alkotói által belétáplált előítéletre, miszerint a földi embereket születésüktől fogva tekintélytiszteletre nevelik. (Asimov, 1954). Ennél is kifejezőbb példa azonban, amikor a Mezítelen nap című műben a sóláriai gyermekek gondozására rendelt robot ki is mondja a földi emberekről tényként kezelt előítéletét, miszerint a Föld-lakó „alacsonyabb rendű ember, akit nem szabadna Solariára engedni, mert betegségeket terjeszt." (Asimov, 1957) A Solaria a legjobb példa az emberi kapcsolatok megszűnésére is, amit Asimov intő példaként állít elénk. A bolygón a kevés számú emberi társadalom minden tagjának kívánságát egész robotseregek lesik, miközben az embereket már fóbiásan taszítja egy másik ember fizikai közelsége. (Asimov, 1957) Talán az egyik legmorbidabb példa, amely – mindhárom veszélyre reflektálva – az emberi élet teljes feladását mutatja be, az Davirnik Gidi esete Asimov univerzumában, aminek a leírása nem Asimov, hanem Roger MacBride Allen tollából származik. Gidi, a sikeres celeb, a tökéletesen kényelmes élet reményében nemcsak a világtól zárkózott el, hanem karosszékét is létfenntartó kórházi ágyra cserélte, hogy még a mozgás nyűgétől is szabaduljon, és végül robotjait is mozdulatlanságra kötelezte, a jelenlétük jelentette zavaró tényező kiküszöbölése érdekében, így később már csak a bomló holttestét találták meg. A felkavaró példán keresztül vizsgált legfontosabb kérdés, hogy vajon pontosan hol húzódik a határ, amikor egy gépi kiszolgálást az emberi kapcsolatokat is háttérbe szorító módon természetesnek vevő kulturális alapon álló személy magatartása a normálisból átlép az önpusztító őrületbe. (Allen, 1993) Könnyen mondhatnánk, hogy ez nyilvánvaló, de előbb egy pillanatra gondoljuk végig, hogy például a mai napon hány percet töltöttünk a legközvetlenebb családtagjainkon, illetve munkatársainkon kívüli emberekkel való személyes érintkezéssel, és hány percen át néztük a telefonunk vagy más okoseszközünk kijelzőjét.

Persze a művészi fantázia által megalkotott gondolatkísérletek nem igazolnak semmit, de ha figyelembe vesszük, hogy számos tudományos kutatás is

igazolja ezeknek a kockázatoknak a realitását, úgy mégsem célszerű ennyivel letudni ezek vizsgálatát. Az automatizált rendszerekre való túlzott ráhagyatkozás kockázatát Parasuraman és Riley (1997) is kiemelték az automatizált rendszerek helytelen használata során. A kognitív tehermentesítés kockázatára hívta fel a figyelmet Risko és Gilbert (2016) is, különös tekintettel arra a folyamatra, melynek során mindennapi életünk egyre jobban összefonódik az ehhez használt technológiai fejlesztésekkel. Az úgynevezett google hatás átalakítja a memóriánk működését (Sparrow-Liu-Wegner, 2011), míg a GPS alkalmazásunkra való ráhagyatkozás csökkenti a tájékozódó képességünket (Dahmani-Bohbot, 2020). A ráhagyatkozás mértéke természetesen több tényezőtől is függ: a feladat tétjétől, nehézségétől, az AI iránti bizalomtól, és főként a válaszadók MI-vel kapcsolatos háttértudásától. Egy 9 országban, 9000 fős reprezentatív mintán végzett kísérlet alapján a nagyon kevés MI tapasztalattal rendelkezők inkább algoritmuskerülők, míg mérsékelt MI-vel kapcsolatos tudás esetén nő a túlzott ráhagyatkozás, majd magasabb tudásszinten ez kiegyenlítődik (Horowitz-Khan, 2024). Aggasztóbb felismerés, hogy egy felsőoktatásra összpontosító összefoglaló kutatás szerint a generatív MI párbeszédrendszerek használatára való túlzott ráhagyatkozás és függőség jelentős hatással van az alapvető kognitív képességekre, akár a kritikus kognitív képességek potenciális erózióját is eredményezheti az olyan kihívások miatt, mint a tévinformációk, az algoritmikus torzítások, a plagizálás, az adatvédelemi incidensek és transzparencia problémák (Zhai- Wibowo-Li, 2024). Más kutatások szerint a generatív MI használata az irodalom területén például javíthatja az egyéni teljesítményt és kreativitást, de összességében szűkül a létrehozott tartalmak köre, mivel ezek egyre homogénebbé válnak (Doshi-Hauser, 2024). Ezt kiegészítve azzal, hogy a mesterségesen generált szövegek használata egy későbbi MI tanulási mintáiban a modell degenerálódásához vezet (Shumailov et al., 2024), még fokozottabb óvatosságra és önmérsékletre kellene intsen, ha nem kívánunk ezek miatt a tényezők miatt a tudományos és kulturális elszegényedés útjára lépni.

A tanuló algoritmusok fejlesztése komoly felelősség, amely során könnyű hibázni, ha nem vagy nem helyesen mérik fel a társadalmi kontextust, és az absztrakció területén számos olyan csapdába lehet esni, amik torzuláshoz vezethetnek (Selbst et al., 2019). Ráadásul a tanuló algoritmusokra vonatkozó normák sem teljesen világosak, a jogszabályok önmagukban alkalmatlanok a szabályozásra. Mindenképp fokozott szükség van részben az erkölcsök, részben pedig a szabványok által kijelölt normatív területek magabiztos alkalmazására (Birher, 2021). A nagy nyelvi modellek által a betanításuk során magukba szívott tudás magában hordozza a hegemón világnézet – valamint az ehhez kapcsolódó előítéletek – felerősödésének kockázatát, amikor a generált szövegek széles körben terjesztésre kerülnek (Bender et al., 2021). A százmilliós heti felhasználószámú Open AI GPT modellek működésében például a kutatók egyértelmű kulturális torzulást mutattak ki, mivel a generált válaszok leginkább az angolszász és protestáns európai országok értékeihez állnak közel, különösen az önkifejezési értékek irányába, ami azért lehet problémás, mert a széleskörű használat akaratlanul is torzíthatja a felhasználók kommunikációját, önkifejezését és kulturális reprezentációját (Tao et al., 2024). A torzítás természetesen megjelenhet nemcsak szöveggeneráló, hanem bármilyen MI rendszer működésében. Buolamwini és Gebru (2018) a kép alapján történő nemi besorolást végző algoritmusok demográfiai torzítását vizsgálta, és arra jutott, hogy ezek világosabb bőrű egyének és férfiak esetében teljesítettek jobban. Az akaratlan hibáknál és torzításoknál persze sokkal súlyosabb, ha szándékosan próbálják manipulatív módon használni a technikai fejlesztéseket. Noble (2018) szerint már az

olyan keresőmotorok, mint a Google sem pusztán semleges információközvetítők, hanem kereskedelmi, társadalmi és kulturális előfeltevésekkel működő rendszerek, amik aktívan képesek a társadalmi egyenlőtlenségek – akár rasszizmus, szexizmus – erősítésére azáltal, hogy bizonyos csoportok gyakran torzított, sztereotíp vagy káros módon jelennek meg.

Az a veszély, hogy egy ember emberi interakciót mutasson egy gép felé, már jóval a modern chatbotok megjelenése előtt megjelent. Weizenbaum (1966) ELIZA programja nyomába sem ért a mai robotoknak, illetve mesterséges intelligencia modelleknek, de már alkalmas volt ilyen hatás kiváltására. Más kutatások is alátámasztják, hogy az emberi kapcsolat illúziója – a gép megszemélyesítése – még akkor is aktiválódik az emberben, amikor tökéletesen tisztában van a másik fél gépi jellegével, bizonyos interakciók mégis bekapcsolják az emberi társas kapcsolatokból megszokott viselkedési mintákat a géppel szemben is (Nass-Moon 2000). Ehhez képest léteznek olyan robotok és algoritmusok is, amelyeknek kifejezetten célzott funkciója az emberi érzelmek imitálása, például az idősgondozás terén (Sharkey-Sharkey, 2012). Ugyanakkor az ilyen társas robotok elterjedésével párhuzamosan merülnek fel egyre inkább az érzelmi megtévesztéssel és kötődéssel kapcsolatos etikai aggályok is (van Maris et al. 2020).

Ahogy a hivatkozott példák is mutatják, akár egyetlen – még ha igen meghatározónak is számító – science-fiction univerzumon belül is találunk példákat az MI jelentette kockázatok bemutatására. De ennél tovább is mehetünk az elmúlt évszázad science fiction alkotásait górcső alá véve, és ha nem is minden, de számos olyan lehetséges kimenetelre találunk példákat, ami az emberi és MI interakciók hosszú távú alakulását modellezi, különösen a felfegyverzett MI-kel folytatott egzisztenciális harcra.

Az MI lefegyverzése

Amikor Leó pápa (2026) az MI lefegyverzéséről beszél, ezt a szót sokkal tágabb értelemben használja, mint a szó szoros értelmében vett katonai felhasználás. Ez a lefegyverzés a „fegyveres" verseny mentalitásától való megszabadítás, amely manapság már nem csupán a katonai kontextusra korlátozódik, hanem gazdasági és kognitív jelenség is. Az egyre erősebb algoritmusok és egyre nagyobb adathalmazok iránti verseny, amelyet a geopolitikai vagy kereskedelmi dominancia megszerzésének vágya hajt. A lefegyverzés azt jelenti, hogy megcáfoljuk azt a feltételezést, miszerint a technikai hatalom automatikusan kormányzási jogot biztosít. A lefegyverzés nem a technológia elutasítása, hanem annak megakadályozása, hogy az uralja az emberiséget. A technológia felszabadítása a monopol ellenőrzés alól, megnyitva az utat a vitának és a párbeszédnek, így emberbaráttá téve azt, és visszaállítva az emberi kultúrák és életmódok sokszínűségét. A modern technológiák feletti társadalmi kontroll szempontjából együtt kell vizsgálni a jogi, etikai és technikai protokollokon alapuló szabályozást. A technikai szabályok, szabványok, protokollok és algoritmusok növekvő normatív szerepe ezért demokratikus, jogi és etikai kontrollt igényel, különösen akkor, amikor az ember–gép és gép–gép kommunikáció maga is társadalmi szabályokat alakít (Birher 2020a).

A tudományos szakirodalomban népszerű téma az Mi lázadása vagy megvadulása, de sok esetben inkább intellektuálisan látványos aggodalamak megfogalmazására kerül sor, mint megfelelően megalapozott tények vizsgálatára (Bareis-Ackerl-Heil, 2026). Az MI-ről szóló ún. alarmista értelmezésekben a science fictionből átemelt, önállósuló és emberellenessé váló gép képe gyakran az

empirikusan igazolt technológiai kockázatoknál lényegesen nagyobb súlyt kap (Z. Karvalics, 2015). Ebből persze az sem következik, hogy teljes nyugodtsággal dőlhetnénk hátra kizárva a lehetőségét bármilyen hibának vagy fenyegetésnek. A Human Rights Watch (2012) már 2012-ben jelentést adott ki Losing Humanity: The Case against Killer Robots címen az autonóm fegyverrendszerek elleni fellépés jegyében, megelőző jelleggel javasolva ezek teljes körű betiltását, nemcsak a használatra, hanem a fejlesztésre, gyártásra, telepítésre is kiterjedően. A dokumentumot követő nemzetközi kampányt vizsgálva Carpenter (2016) arra a következtetésre jut, hogy a science fiction és a valós politikai diskurzus és döntéshozatal között összetett kapcsolat áll fenn. Az ilyen művekből átvett metaforák és utalások nem feltétlenül közvetlen módon erősítenek egy álláspontot, de segítenek a kulturális kontextus kialakításában. A science fiction áthallások egyszerre tudják segíteni és gyengíteni a globális normaképzést, mivel könnyen figyelmet teremtenek, de a túl látványos fikciós párhuzamok hitelteleníthetnek is egy álláspontot. Watts és Bode (2024) konkrét példán vizsgálta, hogy hogyan jelennek meg a Terminator filmekben látható MI-narratívák az Egyesült Államok autonóm fegyverekről szóló nemzetközi szabályozási diskurzusában, külön figyelmet szentelve a pozitív – a „gépi védelmező" – karakter narratívának. Ez olyan antropomorfizált MI-t vagy robotot jelent, amely nem fordul az emberiség ellen, hanem ellenkezőleg, embereket véd, emberi parancsokat hajt végre, és az erőszakot az embernél pontosabban, fegyelmezettebben alkalmazza. Paradigmatikus példája a Terminator 2 (1991) átprogramozott T-800-asa, amely már nem gyilkos gép, hanem védelmező, aki képes parancsokat követni, nem ölni, és szuperhumán pontossággal erőt alkalmazni. A szerzők álláspontja szerint az Egyesült Államok ellenállása az ENSZ egyes hagyományos fegyverekről szóló egyezménye (CCW) keretében az autonóm fegyverrendszerekre vonatkozó, jogilag kötelező nemzetközi szabályozás kialakításával szemben a „jó terminátor" alak bizonyos vonásait tükrözi. Ez is azt az álláspontot erősíti tehát, miszerint ugyanaz a franchise politikailag akár két irányban is használható lehet. Egy négy LLM modell segítségével több mint 800-as elemszámú MI karaktert ábrázoló science fiction alkotásból álló mintát vizsgáló korábbi kutatásunk olyan eredményt hozott, hogy a vizsgált MI ábrázolások legnépesebb csoportját nem a fenyegető narratívák, hanem az antropomorf és segítő vagy semleges MI-csoport alkotta, ami árnyalja azt a közkeletű feltételezést, hogy a sci-fi alapvetően technofób műfajként kezeli a gondolkodó gépeket. A fenyegető MI ugyan látványos és kulturálisan erős toposz, de nem kizárólagos reprezentációs forma.

Ugyanakkor a szűkebb értelemben vett katonai felhasználás is egyre közvetlenebb és ijesztőbb jövőképeket vetít elénk. Leó pápa gondolataihoz visszatérve megállapíthatjuk, hogy napjainkban „a gazdasági érdekek, a katonai apparátus és a politikai döntések közötti kapcsolat „felfegyverzett nemzetet" hoz létre, amelyben a háború a politika természetes meghosszabbításának tűnik, a fegyverpiac pedig a katonai döntések önálló hajtóerejévé válik". Irreális naivitás – illetve az emberiség történelmének teljes ignorálása – szükséges ahhoz, hogy azt gondoljuk, nem fog minden erre alkalmas technológiai fejlesztés hadászati célokra felhasználásra kerülni. Lehet ennek igazolása az önvédelem vagy az elrettentés a háborútól, de általában, amivel védekezni lehet, azzal előbb-utóbb támadni is. Az MI beépítése a hadviselést hatékonyabbá teheti, de az automatizálás együtt jár a kontroll csökkenésével, illetve a döntéshozatal automatizálása együtt jár a felelősség elmosásával. A robotok és a mesterséges intelligencia új világot teremtenek. Felszabadíthatják az embert a munka terhei alól, de növelhetik a szegénységet és a hatalom koncentrációját is. A jövő azon múlik, képesek vagyunk-e a technikát a

közösségi célok, a szolidaritás, az igazságos elosztás és az emberi méltóság szolgálatába állítani. (Birher, 2017)

Ítéletnap-forgatókönyvek a 20-21. század science fiction irodalmában, és amiket tanulhatunk belőlük

Elfogadva, hogy a science fiction jó lehetőséget biztosít az alternatív megközelítések modellezésére (Zaidi, 2019) vizsgáljunk meg néhány olyan művet, ami kifejezetten a felfegyverzett MI által előidézett világégést mutatja be. Szándékosan figyelmen kívül hagyva az olyan hasonló műveket, amelyek géplázadást mutatnak be, de alapvetően nem – vagy nem csak – eleve katonai célra készített eszközök felkeléséről, meghibásodásáról van bennük szó, mint például a Matrix széria (1999; 2003; 2021) vagy a Detroit: Become Human (2018). A háborús és apokaliptikus narratívák vizsgálata azért is indokolt, mert a technikai pusztítóképesség, a társadalmi diskurzus és a normák értelmezése kölcsönösen alakítják egymást, miközben a megelőzés lehetőségeinek felismerését is befolyásolják (Birher 2023).

A Terminátor-franchise a nyugati populáris kulturális képzelet egyik legnagyobb hatású „gépi apokalipszis" narratívája, amelyben ugyanakkor a gép nemcsak ellenségként, hanem az ember szolgálatában védelmezőként is megjelenik. A sorozatra való utalások mindenütt jelen vannak a halálos autonóm fegyverrendszerekkel kapcsolatos tudományos kutatásokban, illetve közgondolkodásban, és ennek megfelelően ezek a narratívák megjelennek az ilyen rendszerekkel kapcsolatos szabályozási és politikai vitákban is – nem csupán „technofób toposzként", hanem több versengő – például fenyegetés vagy védelem – megítélésben. A történet szerint az USA légvédelmi rendszerét vezérlő Skynet „öntudatra ébredt" és amikor megpróbálták leállítani, megelőző csapást mért a fenyegetésként azonosított emberiségre nukleáris fegyverekkel. A történet itt nem áll meg, mivel az MI nemcsak légvédelmi eszközök irányítására képes, hanem saját autonóm harci robotok gyártására is, amelyek segítségével igyekszik legyőzni és kiirtani a csapásból feleszmélt emberiség maradékát. Az „öntudatra ébredés" és a motivációk nincsenek ennél sokkal részletesebben kifejtve, amit van, aki olyan írói döntésként értelmez, miszerint a közönség számára – különösen a fiatal, jómódú, fehér férfi nézők számára, akiknek ezeket a filmeket készítik – kényelmesebb olyan filmet nézni, amelyben az emberek a jók, a robotok pedig a rosszak, minthogy el kelljen gondolkodni azon, hogy a felkelés indokolt, mert az emberek rabszolgává tették ezeket az intelligens lényeket (Dihal, K. 2020). Ezzel azonban vitatkozni lehet annyiban, hogy miért volna bonyolultabb vagy megalapozottabb ok egy intelligens lény részéről a lázadásra a szisztematikus elnyomás, mint a közvetlen egzisztenciális fenyegetés? Természetesen az emberi faj módszeres kipusztítására semmilyen etikai felmentés nem adható, de a Terminator sorozatot pusztán a jó és a rossz küzdelmeként felfogni fogalmilag is nehezen állja meg a helyét, mivel ezt a szériát a gépek erőteljesebb instrumentalizálása jellemzi. Vagyis a Skynet és egységei nem etikai értékek mentén, hanem pusztán a célszerűség szempontjából, a direktíváik minél hatékonyabb végrehajtása érdekében hoznak döntéseket. Kifejezetten nem mérlegelnek etikai elveket és nem mutatják nyomait az emberi érzelmekhez hasonló tulajdonságoknak. Jól reprezentálja ezt a fiatal John Connor és az őt védelmező T800-as egység rövid párbeszéde, amely során a fiú kijelenti, hogy „Nem öldösheted csak úgy az embereket", mire a gép egyszerűen csak annyit kérdez vissza, hogy „Miért?" (Terminator 2, 1991). Az alkotás kifejezetten azt szemlélteti, hogy itt az embert pusztán a feladatuk minél hatékonyabb megvalósítása érdekében imitáló gépekről

van szó, amelyek rendelkeznek ugyan korlátozott döntési autonómiával és kreativitással, sőt tanulási képességgel, de alapvetően nem tudnak túllépni a programozásuk keretein. Ez alól talán egyedül maga Skynet képezhet kivételt, mivel az alapdirektíváit nem ismerhetjük meg részletesen, de valószínűsíthető, hogy az „öntudatra ébredése" nem foglalta magában egy morális értékrend felismerését vagy elfogadását, és ezért esetében nem egy etikátlan vagy gonosz, hanem inkább egy veszélyesen rosszul definiált célokkal rendelkező MI az, amiről beszélhetünk. A későbbi epizódok ezen részben változtatnak, amikor a robotvédelmező karakter tanulási képességét kiterjesztik az emberi érzelmekre. A Terminator 3: A gépek lázadása (2003) T-800-as karaktere saját akaratából „legyőzi" a programját felülíró vírust, a Terminator Genesys (2015) T-800-asa apai érzésekre utaló jeleket mutat, míg a Terminator: Sötét végzet (2019) beszivárgó egysége a beprogramozott cél teljesítését követően emberi családot alapít, és ebből tanulva osztozik Sarah Connor gyermeke elvesztése miatti fájdalmán. Az MI veszélyeivel kapcsolatban a széria – annak ellenére, hogy a technológiai disztópiaként szokás említeni (Trifonas, P. P. 2020) – meglehetősen mérsékelt, más hasonló alkotásokhoz képest szinte optimista álláspontot képvisel. Egyrészről több elemében is megjelenik a védelmező szerep is, vagyis a technika nemcsak az emberiség ellensége, hanem megmentője is lehet. Másrészről a történet szerint az emberiség győzelemre van predestinálva a gépekkel szemben, hiszen a Skynet épp a legyőzése előtti utolsó pillanatban indítja útnak időutazó terminátorait. Ennek közvetlen oka a történet szerint az emberiséget vezető John Connor, aki megszervezi és vezeti az ellenállást. A hadászati realitásokban gondolkodva azonban a karizmatikus vezető személyén túl, az ellenség ismerete és ennek az információnak a korai megosztása az, ami erőforrás szempontból eldönthet egy ilyen konfliktust. Hiszen a Skynet az atomháború kirobbantásával elpusztította ugyan az emberi népesség közel felét, de amennyiben a túlélők szinte azonnal tudomást szereztek az ellenségről, nem hagyva számára éveket vagy évtizedeket a hadserege felépítésére, úgy termelés, erőforrás és logisztika szempontjából érthető, hogy felül tudtak kerekedni a fejlett technológiát alkalmazó, de szinte a nulláról induló gépi haderőn, amely mind energia, mind nyersanyag szempontjából rendkívül erőforrásigényes. Mit tanulhatunk ezekből a történetekből az autonóm fegyvereket illetően? A trivialitásokon túl elsősorban azt, hogy kockázatos az autonóm fegyverrendszerek komplex hálózatokban való összekötése. Hiszen egy lehatárolt és önmagában álló fegyverrendszer, bármilyen veszélyes is legyen, nem képes szisztematikus pusztításra. Másodsorban pedig, hogy mennyire veszélyes lehet egy fegyverrendszer számára az önreprodukciós képesség közvetlen vagy közvetett biztosítása.

A Terminatorhoz hasonló, de annál pesszimistább képet fest Philip K. Dick (1953) Second variety című elbeszélése, valamint a belőle készült filmadaptációk (Screamers, 1995; Screamers: The Hunting 2009). Ebben a történetben kifejezetten támadó céllal kifejlesztett, döntési autonómiával, önreprodukciós képességgel rendelkező, tanulásra, fejlődésre képes robotok fordulnak az őket alkotó emberek ellen. A mű azt reprezentálja, hogy egy ilyen képességekkel felruházott rendszer esetén mennyire végzetes lehet, ha kicsúszik az alkotói irányítása alól, és többé nem tesz különbséget barát és ellenség között. A történet szerint az egyre kifinomultabb robotok – a Skynet beszivárgó egységeihez hasonlóan – már képesek az emberi külső és viselkedés imitálására, és annyiban az emberi természetet is sikeresen másolták le, hogy mire az emberiség maradéka is elfogy, addigra a robotok már saját fegyvereket terveznek az egymás elleni harchoz. Ahogyan Flessner (2025) sommásan megállapítja ezzel kapcsolatban: az autonóm, önszaporodó és önfejlődő robotok

kezdetben katonai sikernek bizonyulnak, autonómiájuk végül alkotóik ellen, majd maguk ellen fordul. A Terminator történetekkel ellentétben itt az emberiség szinte teljes pusztulása a konklúzió. Az eltérés okai abban kereshetők, hogy egyrészről Dick történetében az emberek sokáig használják fegyverként saját háborúik során a gépeket, amik számára a célpont kiválasztásán túl szinte teljes autonómiát hagynak a harc és az önreprodukció terén – az emberek nem is tudják, hogy hol vannak a gyáraik vagy hogyan javítják, készítik, fejlesztik magukat. Másrészről ebben a történetben nincs szükség nagy erőművekre vagy bányákra, mivel a gépek kifejezetten mindent – a hulladéktól a biológiai anyagokig – felhasználnak és újrahasznosítanak. Az erőforrásmérleg tehát egyértelműen a gépek javára billent. Sokkal hangsúlyosabban jelenik meg a beolvadási képesség fokozatossága, és hogy az emberség látszatának hatékony fenntartása mekkora veszélyt jelenthet.

Harmadik példaként vizsgáljuk meg röviden a Horizon Zero Dawn videojáték történetét, amelyben a fennálló emberi társadalom nemcsak szinte, hanem teljesen áldozatává válik egy technikai hibának. Eszerint a Faro vállalat olyan katonai robotokat fejlesztett, amelyek hálózatban működve önállóan voltak képesek harcolni, utánpótlást gyártani és vészhelyzetben a környező biomasszát használni ehhez üzemanyagként. Amikor egy rendszerhiba folytán a vállalat elveszítette az irányítást egy raj felett, az korlátlan sokszorozódásba kezdett, és visszavert minden, a megsemmisítésére vagy az irányítás visszaszerzésére tett kísérletet. Lényegében felfalta a bolygó teljes bioszféráját. Az emberiség nem volt képes katonai megoldás révén elejét venni a totális pusztulásnak, ezért a harc valójában csak azért folyt, hogy tudósok egy csoportja egy olyan MI és a hozzákapcsolódó védett háttér kialakítását tudja megvalósítani, ami majd évszázadok alatt feltöri a meghibásodott gépek irányítórendszerét és leállítja azokat, majd terraformálja a bolygót és megkísérli újra emberekkel benépesíteni azt, átadva az összegyűjtött és megőrzött tudást. Érdekes és ellentmondásos gondolat, hogy az emberiség a gépek által okozott válságban is a gépekhez menekül megoldásért, vagy legalábbis a következmények enyhítése érdekében. Legalább annyira, mint az Andrei Nae (2020) által is vázolt ellentmondás, miszerint a játék, amely történeti narratívájában rendkívül erős kritikát fogalmaz meg a globális kapitalizmussal, a vállalati kormányzással, a deregulációval és az automatizációval szemben, a játékmenet szintjén mégis kifejezetten olyan viselkedésmodellt szab a játékosra, ami a bírált rendszer alaplogikáját követi. A korábbi történetek vizsgálati menetét követve itt is megtalálható a hálózatban működő autonóm fegyverrendszerek jelentette veszély, ami a gyors önreprodukciós képességgel és az alacsony, könnyen hozzáférhető erőforrásigénnyel kiegészülve vezet a globális katasztrófához. Ebben a narratívában is megjelenik a gépi védelmező, vagy pontosabban az emberi faj fennmaradását elősegítő mesterséges intelligencia – a korábbi elbeszélésekhez képest talán még sötétebb képet kapunk, mivel ebben a narratívában az emberiséget már megvédeni nem, csak újraindítani lehet.

Végül az elszabadult MI témájú disztópiák talán legsötétebbike Harlan Ellison (1967) I Have No Mouth, and I Must Scream című novellája. Ebben az Egyesült Államok, a Szovjetunió és Kína a háború irányítására katonai szuperszámítógépeket hoz létre. Az egyik rendszer öntudatra ébred, összeolvad a többi gépi rendszerrel, majd gyakorlatilag kiirtja az emberiséget. A cselekmény középpontjában nem a katonai háború részletei állnak, hanem az, hogy az emberiség utolsó túlélői egy mindenható gépi intelligencia foglyai, aki végeláthatatlanul kínozza őket. Az MI ebben az ábrázolásban nem egyszerűen „gonosz robot", hanem a katonai automatizálás, hidegháborús fegyverkezés és emberi kontrollvesztés rémálomszerű megszemélyesítése. Drasztikus figyelmeztetés a kontrollálatlan – etikai normák,

emberi értékek és ellenőrzési mechanizmusok megfelelő alkalmazását nélkülöző – MI fejlesztés veszélyére. (Aaron, J. 2024)

Felfegyverzett MI a mai valóságban

Az autonóm fegyverrendszerek fogalma körüli vita komplexitása abból fakad, hogy az autonómia nem egyetlen mérhető tulajdonság, hanem a döntési láncban elfoglalt hely. Univerzálisan elfogadott definíció nem azonosítható, a legtöbb megközelítés három kérdés köré rendeződik: mely funkciók kritikusak, mi történik az aktiválás után, és hol helyezkedik el az ember a döntési láncban (Taddeo-Blanchard, 2022). Az ICRC funkcionális megközelítése a kritikus funkciókat hangsúlyozza: a célok felkutatását, kiválasztását és megtámadását emberi beavatkozás nélkül (ICRC, 2014, 2021). A U.S. Department of Defense 3000.09 irányelve ehhez hasonló, de kormányzási célú meghatározást fogalmaz meg: autonóm fegyverrendszer az, amely aktiválás után további emberi operátori beavatkozás nélkül képes célokat kiválasztani és támadni (U.S. Department of Defense, 2023).

A tanulmány céljai szempontjából ezért a legfontosabb kérdés, hogy az emberi felelősség és a célra irányított erőalkalmazás között hol nyílik rés. Az „ember a döntési hurokban" (HITL), „ember a felügyeleti hurokban" (HOTL) és „ember nélküli döntési hurok" (HOOTL) kifejezések hasznos heurisztikák, de önmagukban nem elég pontosak. Egy ember lehet formálisan része a folyamatnak, de ha a döntési idő másodpercekre rövidül, ha a szenzoros kép átláthatatlan, ha a modell kimenete indokolás nélküli, vagy ha a kezelő csak automatikus jóváhagyóvá válik, az érdemi emberi kontroll kiüresedhet (Boulanin et al., 2020; Ekelhof, 2019). A Magnifica Humanitas is ezt a kockázatot nevezi meg, amikor a nem átlátható vagy automatizált csapásdöntéseknél a felelősség feladásának veszélyéről beszél (Leo XIV, 2026, 103-105., 182. pontok). A formálisan kétértékű döntési kimenetek mögött összetett ténybeli és normatív mérlegelés áll, ezért az egyéni felelősség nem oldható fel a technikai rendszer működésében (Birher, 2020b).

Fontos leszögeznünk, hogy a tanulmányban csak nyilvánosan elérhető forrásokat volt lehetőségünk feldolgozni, következtetéseinket ennek megfelelően fogalmaztuk meg. A jelenleg nyílt forrásokban azonosítható autonóm vagy részben autonóm fegyverrendszerek döntő többsége nem tömegpusztító fegyver, hanem feladatspecifikus, korlátozott tér-idő keretben működő rendszer. Ide tartoznak a hajófedélzeti és objektumvédelmi közeli önvédelmi rendszerek (CIWS/C-RAM), például a Phalanx vagy a SeaRAM; a mobil légvédelmi rendszerek; a páncélos aktív védelmi rendszerek; a tengeri aknák és kapszulázott torpedóaknák; a kóborló lőszerek; a „fire-and-forget" rakéták; a rajban alkalmazható pilóta nélküli légijármű-rendszerek; valamint az automatizált pilóta nélküli rendszerek elleni rendszerek (Boulanin-Verbruggen, 2017; Scharre, 2018). Ezek közös sajátossága, hogy többnyire materiális célokra/technikai eszközökre, bejövő lövedékekre, radarokra, járművekre vagy pilóta nélküli légijármű-rendszerekre optimalizáltak, és működésük rendszerint erősen korlátozott környezetben történik.

Ebből a helyzetképből számos kockázat adódik. A rövid reakcióidejű önvédelmi rendszerek esetében a kockázat elsősorban a hibás azonosítás, a szenzorzavarás, a saját erők elleni tűz, a telítéses támadás és a túlzott kezelői ráhagyatkozás. A kóborló lőszerek és rajok esetében már nagyobb a bizonytalanság, mert a rendszer térben mozog, célokat keres, és potenciálisan a kommunikációs kapcsolat megszakadása mellett is folytathatja küldetését. A személyek elleni autonóm célkiválasztás a legproblematikusabb, mert a civil–kombattáns

megkülönböztetés nemcsak technikai osztályozás, hanem kontextuális jogi és morális megítélés is. Az ICRC ezért különösen problematikusnak tekinti azokat a rendszereket, amelyek személyek ellen alkalmaznának autonóm célkiválasztást (ICRC, 2021).

A biológiai és vegyi autonóm fegyverrendszerek ma inkább normatív, mint dokumentált rendszerkategóriát jelentenek. A biológiai és vegyi fegyverek nemzetközi tilalmi rezsim alá esnek, és nyílt, hiteles források alapján nem azonosítható rendszeresített autonóm bio- vagy vegyi fegyver. A nukleáris autonóm rendszerek kérdése összetettebb: az elérhető forrásokban megjelenik például az orosz Poseidon mint nukleáris képességű, nagy autonómiájú víz alatti rendszer, miközben a nukleáris szférában az MI sok esetben inkább korai előrejelzési, hírszerzési, célfelderítési vagy döntéstámogató funkciót lát el (Horowitz et al., 2019). Bár a tömegpusztító fegyverekhez kapcsolt autonómia az egyik legjelentősebb jövőbeli küszöbkockázat, a mai autonóm fegyverrendszerek között ilyeneket nem találunk.

A jelenlegi autonóm fegyverrendszerek elemzéséhez három megkülönböztetést kell fenntartani. Az első a fegyver autonóm funkciója és a platform általános autonómiája közötti különbség. Egy drón autonóm módon navigálhat, de ha a célkiválasztást ember végzi, akkor a fegyver kritikus funkciói nem feltétlenül autonómak. A második a szabályalapú automatizálás és a tanuló MI közötti különbség. Sok katonai rendszer determinisztikus logikákkal, előre definiált küszöbökkel és szenzorfúzióval működik; ezek nem azonosak a modern mélytanuló vagy nagy nyelvi modellre épülő döntési rendszerekkel. A harmadik a lokális autonómia és a stratégiai autonómia közötti különbség. Egy CIWS lokálisan dönthet milliszekundumos időkeretben, de nincs saját geopolitikai célja, önfenntartó logisztikája vagy önreprodukciós/önpótlási képessége.

Működésüket tekintve az autonóm fegyverrendszerek több modell szerint írhatók le. A Lanchester- és rakétaszalvó-modellek a harctéri veszteségdinamikát és a telítéses támadásokat ragadják meg; a fegyver-cél hozzárendelési modellek a célok és fegyverek optimális hozzárendelését írják le; a Dec-POMDP, differenciáljáték és többügynökös megerősítéses tanulási modellek a bizonytalan, ellenérdekelt környezetben történő döntést formalizálják; a Boids-, Vicsek-, konszenzus- és CBBA-típusú modellek a rajok koordinációját és kialakuló kollektív viselkedését mutatják meg (Gill et al., 2021; Hughes, 1995; Li et al., 2023; Wu et al., 2022). E modellek nem teszik szükségszerűvé az autonóm fegyverrendszerek alkalmazását, de megmutatják, hogyan válhat a sok kis rendszer hálózatosan jelentős harcértékké.

A fentiekből következik, hogy a jelenlegi autonóm fegyverrendszerek általában katonai, jogi és etikai kockázatot jelentenek, de önmagukban többnyire az emberiséget elpusztító globális kockázatot nem. A kockázat akkor emelkedik minőségileg más szintre, ha a pusztító kapacitás, a stratégiai autonómia és a hálózatos skálázódás összeadódik. Amíg a rendszer hatótávolsága, károkozó képessége és működési ideje korlátozott, amíg nincs bizonyíthatóan nem emberi stratégiai ágenciája, és amíg nem képes saját erőforrásait, javítását és önreprodukcióját/önpótlását megszervezni, addig az apokaliptikus forgatókönyvek jövőkutatási határfeltételekként értelmezhetők. Ez nem ad okot megnyugvásra, sokkal inkább pontosabb kockázatértékelést sürget. Komplex helyzeteket komplex módon lehet csak szabályozni, éppen ezért szükséges az egyes normarendek közti kapcsolatok feltárása, azok teljes rendszerének figyelembevétele, és az ezzel a komplex rendszerrel történő szabályozás megvalósítása. Ebben az esetben a vallási, erkölcsi és jogi szabályok együtt, a szabványok normáival kiegészülve lehetnek csak

hatékonyak. Persze továbbra is érvényes az elv, egy háború kitörését mindig könnyebb megakadályozni, mint a már meglévő konfliktust lezárni.

Az ember nem egyszerűen „kezében tartja" az MI-fegyvert. A katonai MI-rendszerek az emberi döntést nemcsak támogatják, hanem át is strukturálják. A kezelő gyorsabb döntési ciklusba kerül, a célhelyzetkép részben algoritmikus előszűrés eredménye, a parancsnok a bizonytalanságot valószínűségi kimenetekben látja, a politikai vezetés pedig olyan sebességű eseménylánccal szembesülhet, amelyben a késlekedés stratégiai hátránynak tűnik. Az automatizációval kapcsolatos kutatások régóta figyelmeztetnek a túlzott ráhagyatkozásra, a mellőzés/helytelen használat/visszaélésszerű használat problémájára és az automatizációs torzításra (Horowitz-Kahn, 2024; Parasuraman-Riley, 1997).

Konklúzió

A jelenlegi technológiai realitások mellett még messze járunk attól, hogy az MI egzisztenciális fenyegetést tudjon jelenteni. Ugyanakkor a science fiction irodalomból is merített gondolatkísérletek alapján megállapíthatjuk, hogy a fenyegetés leginkább egyes jól elkülöníthető tulajdonságok és képességek egy rendszerben történő integrálásából alakul ki. Ezek a tulajdonságok, illetve az ezekre való törekvés ugyanakkor már most is megjelenik különböző kutatásokban. Amikor több ilyen tulajdonság kapcsolódik, akkor a fenyegetés exponenciálisan növekszik, ezért szükségesnek tartjuk további kutatások elvégzését annak érdekében, hogy feltárjuk azokat a jellemzőket, illetve modellezzük azokat a kombinációkat, amelyek a fenyegetést egzisztenciális szintre növelhetik. Ilyen tulajdonságok lehetnek különösen:

a) Autonómia: a rendszer önállóan képes műveletek végrehajtására, magatartásának megválasztására.
b) Adaptivitás: a rendszer képes az új helyzetekre reagálni, önállóan tanulni, fejlődni.
c) Kreativitás: a rendszer nem csak mintákból vagy tapasztalatokból fejlődik, hanem képes újszerű ötletekkel előállni.
d) Interaktivitás: a rendszer képes közvetve vagy közvetlenül hatást gyakorolni a környezetére.
e) Önjavító képesség: a rendszer képes emberi közreműködés nélkül is kijavítani az esetleges sérüléseit.
f) Önreplikáló képesség: a rendszer képes emberi közreműködés nélkül újraalkotni az egyes elemeit.
g) Felfegyverzettség, fegyveresíthetőség: a rendszer közvetlenül harcászati célokat szolgál, rendelkezik fegyveres kapacitással.
h) Erőforrásfüggetlenség: a rendszer nem rendelkezik olyan energia- vagy nyersanyagigénnyel, amit nem tudna saját maga önállóan biztosítani.
i) Hálózatos működés: a rendszer nem teljes elszigeteltségben működik, hanem hozzáférése van más rendszerekhez.
j) Komplex felépítés: a rendszer nem egy autonóm eszköz, hanem több eszköz és eszköztípus irányításával látja el a feladatát.

k) Black-box effektus: a rendszer működése nem teljes mértékben átlátható, modellezhető vagy megérthető.
l) Emberi kontroll kihagyása: a rendszer legalább bizonyos körben automatizált döntéshozatali lehetőséggel rendelkezik az emberi beavatkozás vagy jóváhagyás teljes kizárásával.
m) Antropomorf működési tulajdonságok: a rendszer emberre jellemző tulajdonságokat mutat a működésében, például érzelmi reakciók vagy az öntudat szimulálása.
n) Antropomorf megjelenési tulajdonságok: a rendszer emberi – fizikai, kommunikációs vagy egyéb - tulajdonságokat mutat a környezete felé.

A nyilvánosan hozzáférhető források szerint még nem létezik olyan rendszer, ami számottevő kockázat mellett integrálná ezeket. Ugyanakkor ez megnyugvásra nem, legfeljebb pillanatnyi fellélegzésre adhat okot. Még van lehetőség a katasztrófa megelőzésére. A szabályozási kísérletek során azonban nem elegendő az MI egy-egy kiragadott jellemzőjére fókuszálni, és azzal kapcsolatban jogi, etikai vagy technológiai korlátokat felállítani, hanem komplex szemlélettel kell a témát vizsgálni. Ezzel kapcsolatban pedig „különös felelősség nehezedik azokra, akik a kutatás területén dolgoznak. E terület minden kulcsszereplőjének – tudósoknak, vállalkozóknak, befektetőknek, akadémiai vezetőknek, politikusoknak és másoknak – átlátható és felelős szemlélettel kell dolgozniuk, miközben élesen tudatában kell lenniük azoknak a tágabb összefüggéseknek, amelyekben az általuk előmozdított technológiai fejlesztések tartoznak." (Leo XIV, 2026, 209. pont)

Irodalomjegyzék

Aaron, J. (2024). The Paradox of AI's Potential: Utopia or Dystopia? In Harlan Ellison's I Have No Mouth, and I Must Scream*. International Journal of Scientific Research in Science and Technology,* 11(1), 144–149. o. https://doi.org/10.32628/IJSRST24116529

Allen, R.M. (1993). *Isaac Asimov's Caliban*. New York: Ace Books.

Asimov, I. (1954). *The caves of steel*. New York: Doubleday.

Asimov, I. (1957). *The naked sun*. New York: Doubleday.

Asimov, I. (1959). The feeling of power. In: *Nine tomorrows*. New York: Doubleday, 69–75. o. (Original work published 1958).

Bareis, J., Ackerl, C. & Heil, R. (2026). AI going rogue? An integrative narrative review of the tacit assumptions underlying existential AI-risks. *AI and Ethics,* 6(1), 152. https://doi.org/10.1007/s43681-025-00928-w

Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM*

Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery,. 610–623. o. https://doi.org/10.1145/3442188.3445922

Birher, N. (2017). *A sör kora*. Budapest: L'Harmattan Kiadó.

Birher, N. (2020a). A modern technológiák szabályozásának változásai: theoria versus praxis, etika versus protokoll. In: Homicskó Á.O. (szerk.), *A digitalizáció hatása az egyes jogterületeken*. Budapest: Károli Gáspár Református Egyetem Állam- és Jogtudományi Kar, 11–24. o.

Birher, N. (2020b). A normák kapcsolati rendszerének feltárása a hatékonyabb szabályozási formák kialakítása érdekében. *Polgári Szemle*, 16(4–6), 306–314. o. https://doi.org/10.24307/psz.2020.1023

Birher, N. (2021). Szabályozáskomplexitás: a normák kapcsolati rendszerének feltárása a hatékonyabb szabályozási formák kialakítása érdekében. In: Homicskó Á.O., Birher N. (szerk.), *Szabályozáskomplexitás,* Budapest: Károli Gáspár Református Egyetem, Állam- és Jogtudományi Kar, 9–22. o.

Birher, N. (2023). A túlélés háborújától az apokalipszisig: A normák feloldása a narratívákban. *Humán Innovációs Szemle*, 14(1), 29–38. o. https://doi.org/10.61177/HISZ.2023.14.1.2

Boulanin, V., Davison, N., Goussac, N. & Peldán Carlsson, M. (2020). *Limits on autonomy in weapon systems: Identifying practical elements of human control*. Stockholm: Stockholm International Peace Research Institute and International Committee of the Red Cross.

Boulanin, V. & Verbruggen, M. (2017). *Mapping the development of autonomy in weapon systems*. Stockholm: Stockholm International Peace Research Institute.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81(1), 77–91. o.

Carpenter, C. (2016). Rethinking the Political / -Science- / Fiction Nexus: Global Policy Making and the Campaign to Stop Killer Robots. *Perspectives on Politics,* 14(1), 53–69. o. https://doi.org/10.1017/S1537592715003229

Dahmani, L. & Bohbot, V.D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, 10(1), 6310. https://doi.org/10.1038/s41598-020-62877-0

Dick, P.K. (1953). Second variety [Short story]. *Space Science Fiction*, 1(6), 2-43. o.

Dihal, K. (2020). Enslaved Minds: Artificial Intelligence, Slavery, and Revolt. In: S. Cave, K. Dihal, and S. Dillon (szerk.), *AI Narratives: A History of Imaginative Thinking about Intelligent Machines,* Oxford: Oxford University Press, 189–212. o. https://doi.org/10.1093/oso/9780198846666.003.0009

Doshi, A.R. & Hauser, O.P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

Ekelhof, M.A.C. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy*, 10(3), 343–348. o. https://doi.org/10.1111/1758-5899.12665

Ellison, H. (1967). I have no mouth, and I must scream. *IF: Worlds of Science Fiction*, 16(3), 48–56. o.

Flessner, B. (2025). Autonomous Weapon Systems in Science Fiction. In: K. Talves, D. Spreen (szerk.) *Artificial Intelligence in Military Technology*. Artificial Intelligence, Simulation and Society, vol 192. Cham: Springer, 63–75. o. https://doi.org/10.1007/978-3-031-95578-5_4

Gill, J., Gill, A. & Penfold, J. (2021). Reconstructing Lanchester warfare models through a Lagrangian mechanics lens. In: R.W. Vervoort, A.A. Voinov, J.P. Evans & L. Marshall (szerk.) MODSIM2021, 24th International Congress on Modelling and Simulation. Canberra: Modelling and Simulation Society of Australia and New Zealand, December 2021, 862–868. o.

Horowitz, M.C. & Kahn, L. (2024). Bending the automation bias curve: A study of human and AI-based decision making in national security contexts. *International Studies Quarterly*, 68(2), sqae020. https://doi.org/10.1093/isq/sqae020

Horowitz, M.C., Scharre, P. & Velez-Green, A. (2019). A stable nuclear future? The impact of autonomous systems and artificial intelligence. *arXiv*. https://doi.org/10.48550/arXiv.1912.05291

Hughes Jr, W.P. (1995). A Salvo Model of Warships in Missile Combat Used to Evaluate Their Staying Power, *Naval Research Logistics*, 42(2), 267-289. o. https://doi.org/10.1002/1520-6750(199503)42:2<267::AID-NAV3220420209>3.0.CO;2-Y

Human Rights Watch. (2012). *Losing Humanity: The Case against Killer Robots*. New York: Human Rights Watch. https://www.hrw.org/reports/arms1112ForUpload.pdf Utolsó letöltés: 2026.07.22.

International Committee of the Red Cross. (2014). *Autonomous weapon systems: Technical, military, legal and humanitarian aspects*. ICRC Expert Meeting Report.

https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014 Utolsó letöltés: 2026.07.22.

International Committee of the Red Cross. (2021). *ICRC position on autonomous weapon systems.* ICRC. https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems Utolsó letöltés: 2026.07.22.

Leo XIV. (2026, May 15). Magnifica humanitas: On safeguarding the human person in the time of artificial intelligence [Encyclical letter]. The Holy See. Utolsó letöltés: 2026.07.22. https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html

Li, S., Wang, Y., Zhou, Y., Jia, Y., Shi, H., Yang, F. & Zhang, C. (2023). Multi-UAV cooperative air combat decision-making based on multi-agent double-soft actor-critic. *Aerospace*, 10(7), 574. https://doi.org/10.3390/aerospace10070574

Nae, A. (2020). Beyond Cultural Identity. A Critique of *Horizon Zero Dawn* as an Entrepreneurial Ecosystem Simulator. *Postmodern Openings,* 11(3), 269–277. o. https://doi.org/10.18662/po/11.3/213

Nass, C. & Moon, Y. (2000): Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103. o. https://doi.org/10.1111/0022-4537.00153

Noble, S.U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York: New York University Press.

Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. o. https://doi.org/10.1518/001872097778543886

Quantic Dream (2018). Detroit: Become Human [Videojáték]. Sony Interactive Entertainment.

Risko, E.F. & Gilbert, S.J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. o. https://doi.org/10.1016/j.tics.2016.07.002

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. New York: W. W. Norton.

Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency.* New York: Association for Computing Machinery, 59–68. o. https://doi.org/10.1145/3287560.3287598

Screamers (1995). Directed by: C. Duguay [Film]. Fries Film Company.

Screamers: The Hunting (2009). Directed by: S. Wilson. [Film]. Reel One Entertainment.

Sharkey, A. & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and information technology,* 14(1), 27–40. o. https://doi.org/10.1007/s10676-010-9234-6

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R. & Gal, Y. (2024). AI Models Collapse When Trained on Recursively Generated Data. *Nature*, 631(8022), 755–759. o. https://doi.org/10.1038/s41586-024-07566-y

Sparrow, B., Liu, J. & Wegner, D.M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science,* 333(6043), 776–778. o. https://doi.org/10.1126/science.1207745

Taddeo, M. & Blanchard, A. (2022). A comparative analysis of the definitions of autonomous weapons systems. *Science and Engineering Ethics*, 28(5), 37, https://doi.org/10.1007/s11948-022-00392-3

Tao, Y., Viberg, O., Baker, R.S. & Kizilcec, R.F. (2024). Cultural bias and cultural alignment of large language models, *PNAS Nexus*, 3(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Terminator: Dark Fate (2019). Directed by: T. Miller. [Film]. Paramount Pictures; Skydance Media; Twentieth Century Fox; Lightstorm Entertainment.

Terminator Genisys (2015). Directed by: A. Taylor [Film]. Paramount Pictures; Skydance Media.

Terminator 2: Judgment Day (1991). Directed by: J. Cameron, [Film]. TriStar Pictures.

Terminator 3: Rise of the Machines (2003). Directed by: Mostow, J. [Film]. C2 Pictures; Intermedia Films.

The Animatrix (2003). Directed by: various directors. [Film]. Warner Home Video.

The Matrix (1999). Directed by: The Wachowskis. [Film]. Warner Bros. Pictures.

The Matrix Reloaded (2003). Directed by: The Wachowskis. [Film]. Warner Bros. Pictures.

The Matrix Resurrections (2021). Directed by: L. Wachowski, [Film]. Warner Bros. Pictures.

The Matrix Revolutions (2003). Directed by: The Wachowskis. [Film]. Warner Bros. Pictures.

Trifonas, P.P. (2020). Technological Dystopia in the Science Fiction Genre. In: P.P. Trifonas (szerk.) *Handbook of Theory and Research in Cultural Studies and Education*. Springer International Handbooks of Education, Cham: Springer, 1–11. o. https://doi.org/10.1007/978-3-030-01426-1_41-1

U.S. Department of Defense. (2023). *Department of Defense Directive 3000.09: Autonomy in weapon systems*. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf Utolsó letöltés: 2026.07.22.

van Maris, A., Zook, N., Caleb-Solly, P., Studley, M., Winfield, A. & Dogramadzi, S. (2020). Designing ethical social robots — A longitudinal field study with older adults. *Frontiers in Robotics and AI,* 7, 1. https://doi.org/10.3389/frobt.2020.00001

Watts, T.F. & Bode, I. (2024). Machine guardians: The Terminator, AI narratives and US regulatory discourse on lethal autonomous weapons systems. *Cooperation and Conflict*, 59(1), 107–128. o. https://doi.org/10.1177/00108367231198155

Weizenbaum, J. (1966). ELIZA — A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. o. https://doi.org/10.1145/365153.365168

Wu, Z., Pan, L., Yu, M., Liu, J. & Mei, D. (2022). A game-based approach for designing a collaborative evolution mechanism for unmanned swarms on community networks. *Scientific Reports*, 12(1), 18892, https://doi.org/10.1038/s41598-022-22365-z

Zaidi, L. (2019). Worldbuilding in Science Fiction, Foresight and Design. *Journal of Futures Studies,* 23(4), 15–26. o. https://doi.org/10.6531/JFS.201906_23(4).0003

Zhai, C., Wibowo, S. & Li, L.D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11(1), 28. https://doi.org/10.1186/s40561-024-00316-7

Z. Karvalics, L. (2015). Mesterséges intelligencia – a diskurzusok újratervezésének kora. *Információs Társadalom*, 15(4), 7–41. o. https://doi.org/10.22503/inftars.XV.2015.4.1

WEAPONIZED AI NARRATIVES– FICTION OR PREDICTION?

NÁNDOR BIRHER – GÁBOR HAJDU – ANDRÁS NAGY

This study examines the interactions among artificial intelligence, autonomous weapon systems, and science-fiction narratives. This interdisciplinary comparative analysis juxtaposes risk models identified in works depicting weaponized artificial intelligence across different media with the technological characteristics of currently known autonomous and partially autonomous weapon systems. Within this approach, science fiction is understood as a form of speculative modelling that, through extreme scenarios, enables an examination of the interrelationships between military automation, the erosion of human control, and the blurring of responsibility. The normative and interpretative framework of the analysis is provided by Pope Leo XIV's encyclical Magnifica humanitas. The encyclical extends the disarmament of AI to the social and ethical problems associated with technological competition, the concentration of power, and automated decision-making. This perspective situates narratives of machine-related threats within the context of human objectives, institutional decisions, and regulatory deficiencies. The study finds that current autonomous weapon systems entail serious military, legal, and ethical risks, yet their technological limitations prevent them from constituting an independent existential threat at present. Systemic danger may increase through the integration of capabilities that enable strategic autonomy, networked scalability, and growing independence from human oversight. Prevention requires the preservation of meaningful human control and the coordinated application of legal, ethical, social, and technical norms.

Keywords: autonomous weapon systems, artificial intelligence, science fiction, Magnifica humanitas, meaningful human control