információ

MDPI

Cikk

Középtávú Mesterséges intelligencia és társadalom

Seth D. Baum

Global Catastrophic Risk Institute, P.O. Box 40364, Washington, DC 20016, USA;

seth@gcrinstitute.org Megérkezett: Elfogadva: 2020. május 26.; Közzétéve: 2020. május 29.

Összefoglaló: A rövid és hosszú távú mesterséges intelligencia-technológia és a vele járó társadalmi kérdések nagy figyelmet kaptak, de a középtávú kérdéseket nagyrészt figyelmen kívül hagyták. Ez a tanulmány kidolgozza a középtávú mesterséges intelligencia fogalmát, értékeli annak fontosságát, és elemez néhány középtávú társadalmi kérdést. A középtávú mesterséges intelligencia önmagában is fontos lehet, és olyan témaként is, amely áthidalhatja az olykor éles szakadékot azok között, akik a rövid távú és a hosszú távú mesterséges intelligenciát részesítik előnyben. A tanulmány a középtávú mesterséges intelligencia hipotézisét javasolja: a középtáv fontos azok szempontjából, akik a rövid távú mesterséges intelligenciára fordítanak figyelmet, és azok szempontjából is, akik a hosszú távú mesterséges intelligenciára. A tanulmány elemzi a középtávú mesterséges intelligenciát a kormányzati intézmények, a kollektív cselekvés, a vállalati mesterséges intelligenciafejlesztés és a katonai/nemzetbiztonsági közösségek szempontjából. E négy terület egyes részein némi támogatást találunk a középtávú mesterséges intelligencia hipotézisének, bár néhány esetben a kérdés nem egyértelmű.

Kulcsszavak: rövid távú mesterséges intelligencia; hosszú távú mesterséges intelligencia; középtávú mesterséges intelligencia; középtávú mesterséges intelligencia; a mesterséges intelligencia társadalmi hatásai

---

1. Bevezetés

A mesterséges intelligenciával kapcsolatos technológiákra és az azokat kísérő társadalmi kérdésekre irányuló figyelem általában a rövid vagy a hosszú távú mesterséges intelligenciára összpontosító csoportokba tömörül, és némi éles vita folyik közöttük arról, hogy melyik a fontosabb. Baum [1] nyomán a közeljövőben gondolkodók táborát "prezentistáknak", a hosszú távon gondolkodókat pedig "futuristáknak" nevezhetjük.

A dolgok jelenlegi állása két okot vet fel a közeli és a hosszú távú időszak közötti köztes időszak figyelembevételére. Először is, a középtáv (vagy felváltva: középtáv vagy középtáv) a benne rejlő jelentőségéhez képest elhanyagolt. Ha vannak fontos témák, amelyek a közeli és hosszú távú mesterséges intelligenciát érintik, akkor talán a középtáv is fontos témákkal rendelkezik. Másodszor, a középtáv közös pontot jelenthet a jelen- és a jövőkutatók között. Amennyiben mindkét fél fontosnak tartja a középtávot, ez egy konstruktív témát kínálhat, amelybe olyan energiát lehet irányítani, amelyet egyébként a nézeteltérések tisztázására fordítanának.

Ritka példa a középtávú mesterséges intelligenciával foglalkozó korábbi tanulmányokra a Parson és munkatársai [2,3]. (Rengeteg olyan munka létezik, amely érinti a középtávú mesterséges intelligencia témáit, amelyek egy részét ebben a tanulmányban idézzük. A Parson et al. [2,3]-on kívül azonban nem tudok olyan publikációról, amely kifejezetten a középtávú mesterséges intelligenciát dedikált figyelmet érdemlő témaként jelölte volna meg). Mindkét tanulmány [2,3] fontosnak és elhanyagoltnak ismeri el a középtávú mesterséges intelligenciát. Parson et al. [2] elismeri, hogy a mesterséges intelligenciával kapcsolatos korábbi munkák egy része olyan témákkal foglalkozik, amelyek minden időszakban fontosak, és így a középtávú időszak szempontjából is relevánsak. Megadja a középtávú mesterséges intelligencia fogalmát, amelyet alább tárgyalunk, és elemzi a középtávú mesterséges intelligencia témáit. Parson et al. [3] azt állítja, hogy a középtáv elhanyagolása részben a mesterséges intelligenciával foglalkozó kutatók tudományágaiból és módszertanából eredhet, amelyek a kutatókat a közeli vagy a hosszú távú, de nem a középtávra irányítják. A jelen tanulmány kibővíti

Parson et al. [2] definíciókkal kapcsolatos munkáját, és eredeti elemzést mutat be a középtávúak másfajta keverékéről

AI témák. A jelen tanulmány a középtávot is vizsgálja, mint a jelen- és a jövőkutatók közötti lehetséges közös pontot.

Korábban több kísérletet is tettek a jelen- és a jövőkutatók közötti szakadék áthidalására [1,4,5]. A szakirodalom egyik átfogó témája az, hogy az előrelépéshez szükséges gyakorlati lépések gyakran (bár nem mindig) ugyanazok a közeli és a hosszú távú mesterséges intelligencia esetében. Ahelyett, hogy energiát fordítanánk a közeli és a hosszú távú mesterséges intelligencia relatív fontosságának megvitatására, gyakran eredményesebb lehet a figyelmet azokra a gyakorlati lépésekre összpontosítani, amelyekben a vita mindkét oldala egyetért. Ez a gyakorlati szinergia két különböző okból is létrejöhet, mindkettőnek van hatása a középtávú mesterséges intelligenciára.

Először is, bizonyos intézkedések javíthatják a rövid távú mesterséges intelligenciát és a hosszú távú mesterséges intelligenciáról szóló rövid távú beszélgetést. Az ilyen intézkedések gyakran javítják a középtávú mesterséges intelligenciáról szóló rövid távú megbeszéléseket is. Például az informatikusok és a politikai döntéshozók közötti párbeszéd elősegítésére tett erőfeszítések javíthatják a közeli, közép- és hosszú távú mesterséges intelligenciáról szóló politikai viták minőségét. Emellett a mesterséges intelligencia fejlesztőit arra ösztönző erőfeszítések, hogy nagyobb felelősséget vállaljanak munkájuk társadalmi és etikai következményeiért, befolyásolhatják a közeli, közép- és hosszú távú mesterséges intelligenciával kapcsolatos munkát. Például az etikai elvek, amelyeket számos mesterséges intelligencia csoport a közelmúltban hozott létre [6], gyakran elég általánosak, és alkalmazhatók a közeli és hosszú távú mesterséges intelligenciával kapcsolatos munkára, ahogyan az ezen elvek korlátairól szóló elemzések is [7]. Itt meg kell magyarázni, hogy vannak olyan közeli munkák, amelyek olyan rendszerek kifejlesztésére irányulnak, amelyek csak közép- vagy hosszú távon válhatnak működőképessé, különösen az AI képességek nagy áttörései felé irányuló alapkutatásból álló munkák.

Másodszor, bizonyos intézkedések javíthatják a rövid távú mesterséges intelligenciát, és végül a hosszú távú mesterséges intelligenciát. Ezek az intézkedések gyakran a középtávú mesterséges intelligenciát is javíthatják. Például a rövid távú mesterséges intelligencia-rendszerek biztonságosabb kialakításának kutatásai megalapozhatják a közép- és hosszú távú mesterséges intelligencia-rendszerek biztonságosabbá tételét is. Ez látható az Amodei et al. [8] AI biztonsági tanulmányában, amely a közeljövőbeli mesterséges intelligencia szempontjából készült; a vezető szerző, Amodei szerint a munka a hosszú távú mesterséges intelligencia szempontjából is releváns [9]. Ezen túlmenően a rövid távon létrehozott mesterséges intelligencia irányítási intézmények közép- és hosszú távon is fennmaradhatnak, tekintettel számos politikai intézmény tartósságára. Természetesen a rövidtávon és hosszú távon is fennmaradó mesterséges intelligencia rendszertervek és irányítási intézmények a középtávon is jelen lesznek. Ezen túlmenően a hosszú távú fennmaradásuk értékeléséhez szükség lehet annak megértésére, hogy mi történik középtávon.

A középtávra fordított figyelem egy másik közös pontot kínálhat a jelen- és a jövőkutatók között: mindkét fél fontosnak tarthatja a középtávot. A prezentisták a középtávot elég korainak találhatják az ízlésüknek, míg a futuristák a sajátjuknak elég későinek. Amint azt az alábbiakban részletezzük, a prezentistáknak a rövid távú mesterséges intelligenciát előnyben részesítő okai más típusúak, mint a futuristákéi. A prezentisták inkább az azonnali megvalósíthatóságot, a bizonyosságot és a sürgősséget hangsúlyozzák, míg a futuristák inkább a szélsőséges mesterséges intelligencia képességeket és következményeket. Lehetséges, hogy a középtávú időszak a megvalósíthatóság, a bizonyosság, a sürgősség, a képességek és a következmények széles körben vonzó keveréke. Vagy nem: az is lehetséges, hogy a középtáv egy "holt zónában" helyezkedik el, túl átláthatatlan ahhoz, hogy a jelen idejűek érdeklődését kivívja, és túl jelentéktelen ahhoz, hogy a futuristák érdeklődését kivívja. Ez a kérdés végigvonul majd a dokumentumon, és érdemes formálisan is kifejteni:

> *A középtávú mesterséges intelligencia hipotézis*: Van egy köztes időszak, amelyben a mesterséges intelligencia technológia és az azt kísérő társadalmi kérdések mind a jelen-, mind a jövőbelátás szempontjából fontosak.

A középtávú mesterséges intelligencia hipotézis empirikus vagy normatív szempontból is vizsgálható. Empirikus hipotézisként azt javasolja, hogy a jelen- és jövőkutatók valóban fontosnak tartják a középtávot, vagy hogy hajlamosak lennének egyetérteni azzal, hogy a középtáv fontos, ha

lehetőségük lenne elgondolkodni rajta. Normatív hipotézisként azt javasolja, hogy a prezentistáknak egyet kellene érteniük azzal, hogy a középtáv fontos, tekintettel a prezentista és a futurista nézőpontok értékorientált elkötelezettségére. Tekintettel arra a gyakorlati célra, hogy áthidaljuk a prezentisták és a futuristák közötti szakadékot, az empirikus forma végső soron fontosabb: az számít, hogy a szakadék ellentétes oldalán álló konkrét személyek megfontolás után közös nevezőre jutnának-e a középtávban. (Nem valószínű, hogy jelenleg középtávon közös nevezőre jutnak, mivel nem fordítanak rá figyelmet.) Empirikus vizsgálat a prezentista és

a középtávú futurista reakciókat jelen írás keretein kívül esik. Ehelyett a cél az, hogy tisztázzuk a jelen- és a jövőkutatói nézőpontok természetét a középtáv azon tulajdonságai szempontjából, amelyeket fontosnak kell tartaniuk, majd megvizsgáljuk, hogy a középtáv valószínűleg rendelkezik-e ezekkel a tulajdonságokkal. A dolgozat ezért elsősorban normatív módon fogalmaz, bár a tényleges jelen- és jövőkutatók által megfogalmazott perspektívák empirikus megfigyelésén alapul.

Pontosabban, a középtávú mesterséges intelligencia hipotézis azt javasolja, hogy a két csoport alapjául szolgáló perspektíváknak a középtávot fontosnak kell minősíteniük. Ez azt feltételezi, hogy a "perspektívák" akkor is fontosnak minősíthetnek dolgokat, ha elszakadnak az őket birtokló emberektől. Ez a függetlenedés itt egyszerűen azért megengedett, hogy az elemzés anélkül folytatódhasson, hogy a jelen- és jövőképpel rendelkező emberekkel való konzultáció bonyolultabb (de végső soron fontos) folyamatán keresztülmenne.

A középtávú mesterséges intelligencia hipotézis értékelése e dokumentum egyik célja. Először is azonban többet kell mondani arról, hogy miként határozzuk meg a középtávot.

2. A középtáv meghatározása

A középtáv természetesen a közeli és a hosszú távú időszak közötti időszak. A közeli és hosszú távú mesterséges intelligenciáról szóló viták azonban gyakran nem határozzák meg pontosan, hogy mi számít közeli és hosszú távúnak. A mesterséges intelligencia jövőbeli fejlődésével kapcsolatos bizonytalanság miatt elkerülhetetlen némi kétértelműség. Emellett különböző kontextusokban és célokra különböző definíciók lehetnek megfelelőek - például az, hogy mi minősül rövid távúnak, más lehet egy programozó számára, mint egy döntéshozó számára. Mindazonáltal érdemes röviden megvizsgálni, hogyan lehet a közeli, a közép- és a hosszú távú fogalmakat meghatározni a mesterséges intelligencia esetében. A közeli, közép- és hosszú távú fogalmak mindegyike a jelen írás időpontjában (2019-2020) érvényes nézőponthoz képest került meghatározásra. Az idő előrehaladtával a közeli, a közép- és a hosszú távú besorolás változhat.

Az első dolog, amit meg kell jegyeznünk, hogy a közeli vs. közép- vs. hosszú távú meghatározás több dimenzió mentén történhet. Az első az időrendi: a közeljövő A évtől B évig tart, a középtávú B évtől C évig, a hosszú távú pedig C évtől D évig. A második a mesterséges intelligencia megvalósíthatósága vagy ambiciózussága szempontjából: a közeljövő az, ami már megvalósítható, a hosszú távú az, amit a legnehezebb lenne elérni, a középtávú pedig valahol a kettő között van. Harmadszor, és a másodikhoz kapcsolódóan, a mesterséges intelligenciával kapcsolatos bizonyosság foka: a közeli táv az, ami egyértelműen megvalósítható, a hosszú távú a legbizonytalanabb és legspekulatívabb, a középtávú pedig valahol a kettő között van. Negyedszer a mesterséges intelligencia kifinomultságának vagy képességének mértéke: a közeljövő a legkevésbé alkalmas, a hosszú távú a leginkább alkalmas, a középtávú pedig valahol a kettő között van. Ötödször, és a negyedikhez kapcsolódóan, a hatások tekintetében: a közeljövő (vitathatóan; lásd alább) a legenyhébb hatással van az emberi társadalomra és a világ egészére, a hosszú távú a legszélsőségesebb hatással jár, a középtávú pedig valahol a kettő között van. A hatodik a sürgősség: a közeljövő (vitathatóan) a legsürgősebb, a hosszú távú a legkevésbé sürgős, a középtávú pedig valahol a kettő között van.

A hatások dimenziója némileg összetett, és érdemes röviden kibontani. A közeljövőben a mesterséges intelligenciának lehetnek a legenyhébb hatásai, abban az értelemben, hogy ha a mesterséges intelligencia egyre nagyobb képességekkel rendelkezik, és egyre szélesebb körben és egyre következetesebb környezetben alkalmazzák, akkor általában nagyobb hatással lesz az akkor létező emberi társadalomra. Másképp fogalmazva, ha A = a közeli mesterséges intelligencia hatása a közeli társadalomra, B = a középtávú mesterséges intelligencia hatása a középtávú társadalomra, és C = a hosszú távú mesterséges intelligencia hatása a hosszú távú társadalomra, akkor (feltételezhetően) A < B < C. Vannak azonban alternatív módjai a hatások fogalmának. Lehet egyfajta prezentista nézetet képviselni, és azzal érvelni, hogy az erkölcsi értékelés szempontjából csak a jelenben élő emberek számítanak, ahogyan azt Arrhenius [10] is tárgyalja, vagy hogy a jövőbeli hatásokat diszkontálni kell, ahogyan azt számos gazdasági költség-haszon értékelésnél teszik. Ezekben az esetekben a rövid távú mesterséges intelligenciát lehet úgy értékelni, hogy az a legnagyobb hatással jár, mivel a közép- és hosszú távú mesterséges intelligencia hatásai kevésbé vagy egyáltalán nem számítanak. Vagy figyelembe vehetjük a mesterséges intelligencia időszakának hatásait az összes

időszakra: a rövid távú mesterséges intelligencia hatásait a közeli, közép- és hosszú távú időszakra, a középtávú mesterséges intelligencia hatásait a közép- és hosszú távú időszakra, és a hosszú távú mesterséges intelligencia hatásait a hosszú távú időszakra. Ez a nézőpont elismeri a mesterséges intelligencia technológia tartós hatásainak lehetőségét, és hajlamos lenne növelni a rövid és középtávú hatások értékelt méretét.

AI. Bár elismerjük a hatások ezen alternatív felfogásainak előnyeit, ez a dokumentum az első felfogást használja, amely az A vs. B vs. C felfogást foglalja magában.

A közeli/közepes/hosszú távú meghatározáshoz nem biztos, hogy létezik egyetlen helyes dimenzióválasztás. A különböző körülmények eltérő meghatározásokat vonhatnak maguk után. Például Parson et al. [2] különösen a társadalmi hatások és a kormányzásra gyakorolt hatások iránt érdeklődik, ezért elsősorban a hatásokban gyökerező meghatározásokat használ. Azt javasolják, hogy a rövid távú mesterséges intelligenciához képest a középtávú mesterséges intelligencia "nagyobb alkalmazási skálán mozog, amihez a hatókör, a komplexitás és az integráció terén bekövetkező változások társulnak" [2] (8-9. o.), és a hosszú távú mesterséges intelligenciához képest a középtávú mesterséges intelligencia "nem önvezérelt vagy önállóan akaratlagos, hanem még mindig jelentős mértékben emberi ellenőrzés alatt fejlesztik és alkalmazzák" [2] (9. o.). (Lehet vitatkozni ezekkel a definíciókkal. Vitatható, hogy a rövid távú mesterséges intelligencia már nagy léptékű alkalmazásban van, és lehet, hogy nincs egyértelmű határvonal a közeli és a középtávú mesterséges intelligencia között. Továbbá, bár a javaslat szerint a hosszú távú mesterséges intelligencia kikerülhetne az emberi ellenőrzés alól, ez nem feltétlenül lenne így. Sőt, a hosszú távú mesterséges intelligenciáról szóló viták néha kifejezetten arra a kérdésre összpontosítanak, hogy hogyan lehet egy ilyen mesterséges intelligenciát irányítani [11]). A középtáv egy olyan időszak, amikor a mesterséges intelligenciát lényegesen nagyobb mértékben használják a döntéshozatalban, és ez olyan mértékű lehet, hogy "a kormányzás értelme" megkérdőjeleződik [2] (9. o.), de végső soron az ember marad az irányító. Ez a középtávú mesterséges intelligencia ésszerű meghatározása, különösen a hatások és a kormányzás szempontjából.

A jelen írás inkább a jelenlévők/futuristák vitájára összpontosít, ezért érdemes megvizsgálni a vitában használt definíciókat. A hat dimenzió mindegyikének elemei megtalálhatók, de nem egységesen. A prezentisták gyakran a megvalósíthatóságot és a bizonyosság mértékét hangsúlyozzák. Andrew Ng informatikus emlékezetesen úgy hasonlította a hosszú távú mesterséges intelligenciára fordított figyelmet, mintha a "Mars túlnépesedése" [12] miatt aggódnánk, ami alatt Ng azt értette, hogy ez végül is fontos lehet, de túl átláthatatlan és a jelenlegi mesterséges intelligenciától elszakadt ahhoz, hogy megérje a jelenlegi figyelmet. Egy másik jelen idejű téma a sürgősség, különösen a közeljövőbeli mesterséges intelligencia társadalmi következményeit illetően. Ryan Calo jogász [13] (27. o.) szerint "a mesterséges intelligencia rövid távon számos sürgető kihívást jelent az egyének és a társadalom számára", ezért a hosszú távú mesterséges intelligenciához képest figyelmet érdemel. A futuristák a maguk részéről gyakran a képességekre és a hatásokra helyezik a hangsúlyt. Gyakran idézik I.J. Good [14] (33. o.) korai megjegyzését, miszerint az "ultraintelligens" mesterséges intelligencia (az emberek intelligenciáját jelentősen meghaladó intelligenciával rendelkező mesterséges intelligencia) lehet "az utolsó találmány, amelyet az embernek valaha is meg kell tennie, feltéve, hogy a gép elég engedelmes ahhoz, hogy megmondja nekünk, hogyan tartsuk ellenőrzés alatt". A kronológiai meghatározások kevésbé gyakoriak. Kivételt képez Etzioni [15], aki a hosszú távú mesterséges intelligenciát lebecsüli azzal az indokkal, hogy 25 éven belül valószínűleg nem fog bekövetkezni. (Válaszul a futuristák, Dafoe és Russell [16] azzal érvelnek, hogy a lehetséges jövőbeli eseményekkel akkor is érdemes törődni, ha azok nem a következő 25 évben következnek be).

A fentieket figyelembe véve ez a dokumentum a rövid távú mesterséges intelligenciára a megvalósíthatóság, a hosszú távú mesterséges intelligenciára pedig a képességek meghatározását használja. A dokumentum a közeljövőben megvalósítható mesterséges intelligenciát *olyan mesterséges intelligenciaként* határozza meg, *amely már létezik vagy aktívan fejlesztés alatt áll, és egyértelmű útja van a megvalósítás és a bevezetés felé*. E meghatározás szerint a közeli jövőbeni mesterséges intelligencia nem igényel jelentős kutatási áttörést, hanem a meglévő technikák egyszerű alkalmazásaiból áll. Az "egyértelmű", "jelentős" és "egyszerű" fogalmak homályosak, és ésszerű lehet, hogy különböző kontextusokban eltérő módon határozzák meg őket. (Ez a homályosság a középtávú mesterséges intelligencia-hipotézis szempontjából is fontos; erről bővebben alább). Mindazonáltal ez a meghatározás a jelenlegi mesterséges intelligencia rendszerekre és a lehetséges jövőbeli mesterséges intelligencia rendszerekre utal, amelyek valószínűleg hamarosan megépülnek, és nem függnek a kutatási áttörésektől, amelyek vagy megnyilvánulnak, vagy nem.

A dokumentum a hosszú távú mesterséges intelligenciát *olyan mesterséges intelligenciaként*

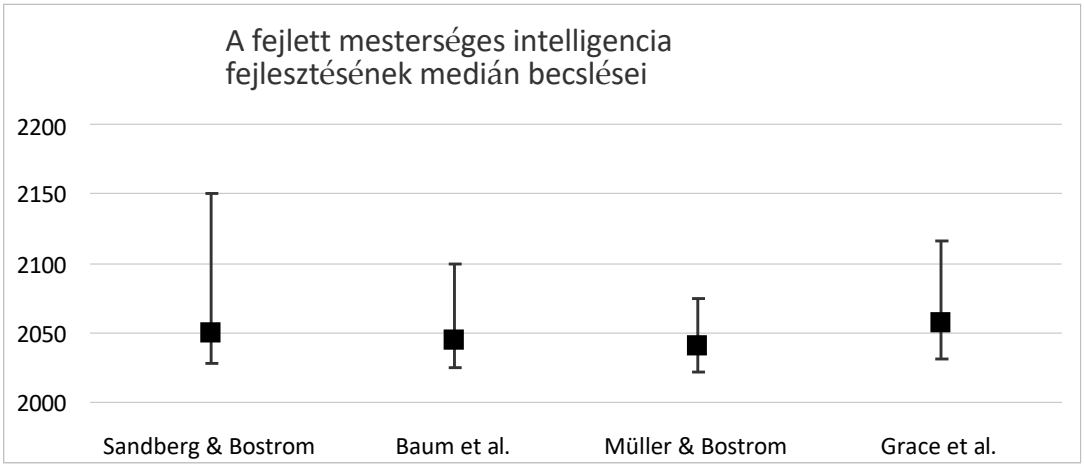
határozza meg, *amely legalább emberi szintű általános intelligenciával rendelkezik*. A hosszú távú mesterséges intelligencia iránti érdeklődés gyakran az emberi szintű mesterséges intelligencia (HLAI), a mesterséges általános intelligencia (AGI), az erős mesterséges intelligencia és a mesterséges szuperintelligencia (ASI) köré összpontosul. Létezhetnek azonban olyan szűk értelemben vett mesterséges intelligencia rendszerek is, amelyeket célszerű hosszú távúnak minősíteni. Cave és ÓhÉigeartaigh [4] (5. o.) például a szuperintelligencia kilátásától elkülönítve a "munkahelyek széles körű elvesztését" sorolja a hosszú távú AI-problémák közé. (Megjegyzendő, hogy a munkahelyek legszélesebb körű elvesztéséhez AGI-ra lehet szükség. Ford [17] (3. o.) például azt írja: "Ha egy napon a gépek képesek lesznek az emberi gondolkodás és új ötletek kigondolásának képességével megegyezni, sőt meghaladni azt - miközben a számítógép minden előnyét élvezhetik olyan területeken, mint a számítási képességek, akkor az intelligenciát nem lehet megkérdőjelezni.

sebesség és adathozzáférés - akkor nehéz elképzelni, hogy milyen munkák maradhatnak még a legképzettebb emberi munkaerő számára is."). A hosszú távú mesterséges intelligencia egy elfogadható alternatív definíciója az *olyan mesterséges intelligencia, amely jelentős szellemi mérföldköveket ér el és/vagy nagy és átalakító hatással bír*. Ez inkább egy gyűjtőfogalom, amely magában foglalhatja a kellően fontos, szűk értelemben vett mesterséges intelligencia-rendszereket, például a munkahelyek elvesztésében érintetteket. Ebben a meghatározásban a "jelentős", "nagy" és "átalakító" kifejezések homályosak. A jelenlegi mesterséges intelligencia rendszerek vitathatatlanul megfelelnek ennek a meghatározásnak. Ezért a tanulmány a hosszú távú mesterséges intelligenciát a HLAI szempontjából határozza meg, miközben az alternatív meghatározások mellett szóló érveket is megemlíti.

Az, hogy a dokumentum a megvalósíthatóságot rövid távon, a képességet pedig hosszú távon határozza meg, összhangban állhat a mesterséges intelligenciáról szóló viták során általánosan használt fogalmakkal. Azonban a rövid távú (megvalósíthatóság) és a hosszú távú (képesség) dimenzió eltérő használata két fontos szempontból is okozhat némi kronológiai elmosódást.

Először is, az azonnal megvalósítható mesterséges intelligencia-projektek hosszú időhorizontúak lehetnek. Ez különösen gyakori lehet az olyan projektek esetében, amelyekben a mesterséges intelligencia csak egy összetettebb és tartósabb rendszer egyik összetevője. A katonai rendszerek egyike a hosszú élettartamú területeknek. Egy 2016-os jelentés szerint egyes amerikai nukleáris fegyverrendszerek még mindig az 1970-es évekbeli 8 hüvelykes floppylemezeket használnak [18]. A mesterséges intelligenciát jelenleg a legkülönfélébb katonai rendszerekhez használják és fejlesztik [19]. Ezek közül néhány elképzelhető, hogy még sok évtizedig fennmaradhat a jövőben is - talán a B-52H bombázóban, amelyet az 1960-as években építettek, és a tervek szerint a 2050-es évekig szolgálatban marad [20]. (A mesterséges intelligenciát a bombázókban például a célzás javítására használják [21]. A mesterséges intelligenciát szélesebb körben használják a vadászgépekben, amelyek gyors sebességgel hajtanak végre összetett légi manővereket, és jelentős taktikai előnyre tehetnek szert a megnövekedett számítási teljesítmény és az emberi pilótákkal szembeni autonómia révén [22]). Elképzelhető, hogy a B-52H-t a jelenlegi mesterséges intelligencia algoritmusokkal szerelik fel, és ezeket az algoritmusokat a 2050-es években is megtartják, ahogyan a 8 hüvelykes floppylemezeket is megtartották más amerikai katonai rendszerekben. A jelen tanulmány definíciói szerint ez a B-52H mesterséges intelligencia a közeljövőbeli mesterséges intelligenciának minősülne, amely történetesen hosszú időn keresztül használatban marad, jóval azon a 25 éven túl, amelyet Etzioni [15] "előrelátható horizontként" kezel, és amely figyelmet érdemel.

Másodszor, a nagy és átalakító hatású mesterséges intelligencia rendszerek, beleértve az AGI-t is, potenciálisan viszonylag rövid idő alatt megépíthetők. Az, hogy az AGI és a mesterséges intelligencia rokon formái mikor épülnek meg, jelentős bizonytalanságot és nézeteltérést okoz. Több tanulmány is megkérdezte a mesterséges intelligencia kutatóit - elsősorban informatikusokat -, hogy mikorra várják az emberi vagy emberfeletti képességekkel rendelkező mesterséges intelligencia megépülését [23-26]. (Megjegyzendő, hogy ezeket a tanulmányokat általában szakértők megkérdezésének tekintik, de nem egyértelmű, hogy a felmérésben résztvevők szakértők-e abban a kérdésben, hogy mikor fog megépülni az AGI. Az AI-val kapcsolatos korábbi előrejelzések gyakran megbízhatatlanok voltak [27]. Lehet, hogy ez egy olyan téma, amelynek nincsenek szakértői; erről a kérdésről lásd Morgan [28]). A kutatók több évtizedre kiterjedő becsléseket mutatnak be, némelyik becslés meglehetősen korai. Az 1. ábra e tanulmányok medián becsléseit mutatja be. A mediánbecslések elrejtik a felmérésben résztvevők becsléseinek tartományát, de a teljes tartományt nem lehetett könnyen bemutatni az 1. ábrán, mert sajnos csak Baum et al. [23] tartalmazta a teljes felmérés adatait. Ha az 1. ábrán bemutatott korai becslések helyesek, akkor a jelen tanulmány definíciói szerint a hosszú távú mesterséges intelligencia viszonylag hamar, esetleg a következő 25 éven belül megjelenhet.



1. ábra. Becslések arra vonatkozóan, hogy a mesterséges intelligencia mikor éri el az emberfeletti képességeket (Baum et al.) [23] és az emberi szintű képességeket (Sandberg és Bostrom, Müller és Bostrom, Grace et al.) [24-26]. Az ábrán a mérföldkő elérésének 10%-os (alsó jelölés), 50%-os (négyzet) és 90%-os (felső jelölés) valószínűségére vonatkozó becslések láthatók. Minden tanulmány esetében a felmérésben résztvevőkre vonatkozó medián becslések vannak ábrázolva.

3. A középtávú mesterséges intelligencia hipotézis

A fenti meghatározásokat szem előtt tartva érdemes felülvizsgálni a középtávú mesterséges intelligencia hipotézist. Ha a prezentistákat definíciójuk szerint csak a jelen érdekli, akkor egyáltalán nem törődnének a középtávon. A közeli és a középtáv közötti határvonal azonban elmosódik. A fenti meghatározás szerint a közeljövőbeli mesterséges intelligenciának egyértelmű utat kell mutatnia a megépítéshez és alkalmazáshoz, de a "egyértelműség" fok kérdése. Ahogy a megépítéshez és alkalmazáshoz vezető út egyre kevésbé lesz egyértelmű, a mesterséges intelligencia a rövid távúból a középtávúvá válik, és a jelenlévők egyre kevésbé érdeklődnek iránta. Ebből a szempontból a jelenlévőket valamelyest érdekelheti a középtáv, különösen annak korábbi részei, de nem olyan mértékben, mint amennyire a közeli táv érdekli őket.

Alternatív megoldásként a jelenlévők azért törődhetnek a középtávval, mert a mögöttes dolgok, amelyekkel törődnek, szintén középtávon merülnek fel. Néhány jelenlévőt érdekelnek a mesterséges intelligenciának a társadalmi igazságosságra, a fegyveres konfliktusokra, a közlekedésre stb. gyakorolt hatásai. Míg a hosszú távú mesterséges intelligencia következményeiről nehéz lehet koherens módon gondolkodni, addig a középtávú mesterséges intelligencia esetében ez nem olyan nehéz. Például az autonóm fegyverekről (a mesterséges intelligenciát a célpontok kiválasztására és a célpontokra való tüzelésre használó gépekről) szóló viták egyik fő tényezője az, hogy ezek a fegyverek képesek-e megfelelően megkülönböztetni az elfogadható és az elfogadhatatlan célpontokat (pl. ellenséges harcosok és civilek) [29,30]. A rövid távú mesterséges intelligencia nem képes megfelelően megkülönböztetni; a középtávú mesterséges intelligencia talán képes lesz rá. Ezért az autonóm fegyverek miatt aggódó jelenlévők okkal érdeklődnek a középtávú mesterséges intelligencia iránt. Azt, hogy ez az érdeklődés kiterjed-e más presentista aggodalmakra (társadalmi igazságosság, közlekedés stb.), eseti alapon kell mérlegelni.

A futuristák számára a középtáv azért lehet fontos, mert megelőzi és befolyásolja a hosszútávot. Ha a hosszútáv az emberi szintű AGI megjelenésével kezdődik, akkor ezt a mesterséges intelligenciát középtávon tervezik és építik meg. Az AGI-val kapcsolatos munka már folyamatban van [31], de még viszonylag korai stádiumban lehet. Az 1. ábra szemlélteti a bizonytalanságot: az AGI (és a mesterséges intelligencia hasonló formáinak) megjelenésére vonatkozó legkorábbi becslések a közeljövőre eshetnek, míg a legújabb becslések sokkal, de sokkal későbbre. A jövőkutatókat leginkább a közvetlenül a hosszú távú időszakot megelőző időszak érdekli, mivel ez van a legnagyobb hatással az AGI-ra. A korábbi időszakok iránti érdeklődésük attól függhet, hogy mennyire jelentős az AGI-ra gyakorolt ok-okozati hatása.

Ebből következik, hogy a középtávú mesterséges intelligencia hipotézis értékelésének két alapja van. Először is, a hipotézis akkor állhat fenn, ha a rövid távú mesterséges intelligenciához hasonló mesterséges intelligencia a hosszú távú mesterséges intelligenciát is befolyásolja. Ebben az esetben

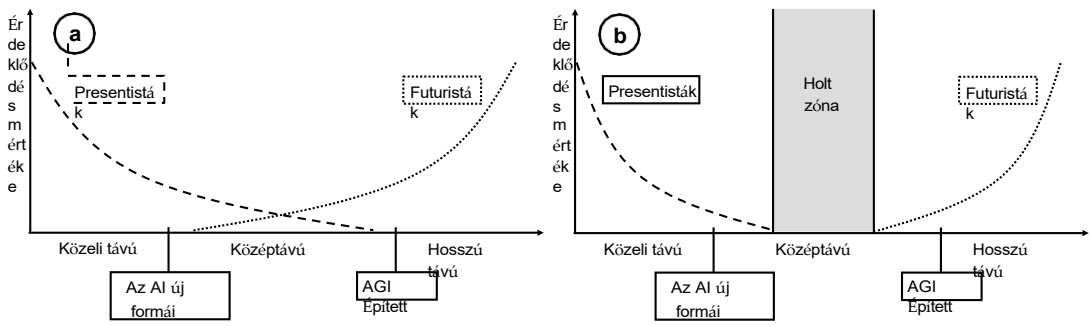
maga a technológia mind a jelen-, mind a jövőkutatók számára érdekes lehet. Másik lehetőség, hogy a hipotézis akkor áll fenn, ha a középtávú mesterséges intelligencia társadalmi következményei hasonló kérdéseket vetnek fel, mint a közeli mesterséges intelligencia, és ha a középtávú

a társadalmi kontextus is befolyásolja a hosszú távú mesterséges intelligenciát. Például a középtávú autonóm fegyvertechnológia hasonló problémákat vethet fel a célpontok megkülönböztetésével kapcsolatban, mint a rövid távú technológia esetében, és a hosszú távú mesterséges intelligencia fegyverkezési versenyét is táplálhatja. (A félreértések elkerülése végett meg kell jegyezni, hogy a hosszú távú mesterséges intelligenciáról szóló viták során a "fegyverkezési verseny" kifejezés néha arra az általános versenyre utal, hogy ki építi meg elsőként a hosszú távú mesterséges intelligenciát, anélkül, hogy ez feltétlenül a katonai fegyverkezéshez kapcsolódna [32]. Ennek ellenére a hosszú távú mesterséges intelligenciáért néha katonai fegyverkezési versenyt is feltételeznek [33].

Mindkettő a közeli, a közép- és a hosszú távú időszakok közötti bizonyos mértékű folytonosságból ered. A folyamatosságot a mesterséges intelligencia rendszerekben és a kapcsolódó társadalmi kérdésekben bekövetkező változások mértékével lehet meghatározni. Ha a közeli mesterséges intelligencia technikák és társadalmi dimenziók jelentős mértékben fennmaradnak a középtáv végén (amikor a hosszú távú mesterséges intelligencia létrejön), akkor a középtávú mesterséges intelligencia hipotézis valószínűleg érvényesül.

Fontos tényező lehet a középtáv kronológiai időtartama. Az 1. ábra a hosszú távú kezdetre vonatkozó becslések széles skáláját tartalmazza. Ha a későbbi becslések helytállónak bizonyulnak, akkor a középtáv meglehetősen hosszú lehet. A hosszú időtartam valószínűleg kisebb folytonosságot jelentene a közeli, a középső és a hosszú távú időszakok között, és ezért kevésbé támogatná a középtávú mesterséges intelligencia hipotézisét. Ez nem feltétlenül van így. Elképzelhető például, hogy a mesterséges intelligenciának csak egy további technikai áttörésre van szüksége ahhoz, hogy a jelenlegi képességekből AGI-vé váljon, és hogy ennek az áttörésnek az elérése sok évtizedet vesz igénybe. Azt is elképzelhető, hogy a mesterséges intelligenciával kapcsolatos kérdések meglehetősen állandóak maradnak, amíg ez az áttörés meg nem történik. Ebben az esetben a közeljövő technikái és problémái még középtávon is fennmaradnának. Valószínűbb azonban, hogy egy hosszú középtávon kevésbé lenne folyamatos, és nagyobb lenne a holt zóna időszaka, amely sem a jelen-, sem a jövőkutatók érdeklődését nem keltené fel. Ha az AGI mondjuk csak 500 év múlva épül meg, a jelenlévők valószínűleg nem fognak érdeklődni iránta.

A 2. ábra két vázlatot mutat be arról, hogy a jelen- és a jövőkutatók középtávon milyen mértékben lehetnek érdekeltek. A 2a. ábra egy olyan átfedési időszakot mutat, amelyben mind a jelen-, mind a jövőkutatóknak van némi érdeklődésük; itt a középtávú AI-hipotézis érvényesül. A 2b. ábra egy holt zónát mutat, ahol nincs átfedés az érdeklődésben; itt a középtávú AI-hipotézis nem érvényesül. A 2. ábra szigorúan szemléltetés céljából került bemutatásra, és nem jelzi a jelen- vagy jövőkutatók tényleges érdeklődésének szigorúan levezetett becslését. Azt szemlélteti, hogy a prezentisták érdeklődésének mértéke idővel csökkenhet, a futuristáké pedig növekedhet, ami hatással lehet a középtávú mesterséges intelligencia hipotézisre. A 2. ábra azt mutatja, hogy a prezentisták/futuristák érdeklődése az idő múlásával megközelítőleg exponenciálisan csökken/növekszik. Ennek nincs különösebb alapja, és a görbéket akár másképp is megrajzolhatták volna.



2. ábra. A közeli, közép- és hosszú távú jelen- és futurista érdeklődés szemléltető vázlatai. (**a**) átfedő érdeklődést mutat: a középtávú AI-hipotézis érvényesül; (**b**) holt zónát mutat, ahol nincs átfedő érdeklődés: a középtávú AI-hipotézis nem érvényesül. A vázlatok szigorúan csak szemléltető célokat szolgálnak. A "mesterséges intelligencia új formái" kifejezést a főszövegben szereplő, a közeljövőre vonatkozó mesterséges intelligencia definíciójára való hivatkozással határoztuk meg.

Összefoglalva, a mesterséges intelligencia középtávú hipotézisének értékeléséhez meg kell vizsgálni, hogyan nézhetnek ki a középtávú mesterséges intelligencia technikák és társadalmi

dimenziók, valamint a rövid, közép- és hosszú távú időszakok közötti folytonosság mértékét.

4. A középtávú mesterséges intelligencia alapvető jelentősége

Eddig az írás a középtávú mesterséges intelligencia potenciális értékét hangsúlyozta, mint a jelen- és jövőkutatók közös érdekét. Ez a "konszenzusos érték" az alábbi szakaszok egyik fő témája marad. Érdemes azonban megállni, hogy megismételjük, hogy a középtávú mesterséges intelligencia a maga nemében is fontos lehet, függetlenül a jelen- és jövőkutatókra gyakorolt hatásától. Annak megítéléséhez, hogy mennyire fontos önmagában véve, szükség van valamilyen mérőszámra az önmagában vett fontossághoz. A részletes mérőszám meghatározása meghaladja e dokumentum kereteit. A jelen célokra elegendő, ha úgy véljük, hogy a mesterséges intelligencia és a vele járó társadalmi kérdések középtávon fontosak lehetnek a világ számára, ahogyan az középtávon létezik. Érdemes továbbá azt is felvetni, hogy a mai embereknek lehetőségei lehetnek arra, hogy jelentősen befolyásolják a középtávot, így a középtávra ma figyelmet kell fordítani, mivel az önmagában véve is fontos. Ezt szem előtt tartva a tanulmány most rátér a középtávú mesterséges intelligencia és a társadalom részleteire.

5. Középtávú mesterséges intelligencia technikák

A saját szakterületem nem a mesterséges intelligencia informatikai tudománya, így viszonylag keveset tudok mondani arról, hogy a számítástechnikai mesterséges intelligencia technikák középtávon hogyan nézhetnek ki. Ezért ez a szakasz helyfoglalóként szolgál, hogy megjegyezzem, hogy a lehetséges középtávú mesterséges intelligencia technikák területe olyan téma, amely megérdemli a figyelmet azok számára, akiknek megvan a szakértelme ahhoz, hogy elemezzék és kommentálják azt.

6. Középtávú mesterséges intelligencia társadalmi dimenziók

Bár a mesterséges intelligencia középtávú társadalmi dimenziói legalábbis bizonyos mértékig a középtávú mesterséges intelligencia technikák képességeitől függnek, a technikák tisztázása nélkül is lehetséges legalább részleges képet alkotni a társadalmi dimenziókról. Az alábbiakban valóban egy részleges képet adok, amelyet jelentős mértékben saját szakterületem alakított ki. Célom, hogy számos területen bemutassam a lehetséges középtávú forgatókönyveket, és megvitassam a közeli és hosszú távú mesterséges intelligenciára gyakorolt hatásaikat, valamint a jelen- és jövőkutatók közötti szakadék áthidalásának kilátásait.

6.1. Irányítási intézmények

A kormányzati intézmények meglehetősen tartósak lehetnek. Az Egyesült Nemzetek Szervezetét például 1945-ben alapították, és a reformra irányuló számos felhívás ellenére az ENSZ Biztonsági Tanácsa állandó tagja Kína, Franciaország, Oroszország, az Egyesült Királyság és az Egyesült Államok. A "P5-ök" a második világháborúból származó ereklye, amely vitathatatlanul nem felel meg a jelenlegi nemzetközi ügyeknek, de a tagság megváltoztatásához olyan konszenzusra lenne szükség, amely meglehetősen nehezen elérhető. Például Brazília és India felvétele mellett lehetne érvelni, de akkor Argentína és Pakisztán tiltakozhatna, így nem történne változás. Nem minden kormányzati intézmény ennyire megcsontosodott, de sok közülük elég tartós. Ez a folyamatosság teszi a kormányzati intézményeket a középtávú mesterséges intelligencia hipotézis meggyőző jelöltjévé.

A közeljövő izgalmas időszak a mesterséges intelligencia irányítása szempontjából. Az intézmények tervezése és elindítása folyamatban van. A most meghozott döntéseknek hosszú távú hatásai lehetnek, amelyek a középtáv végén és a hosszú távú időszak elején is érvényesülhetnek. (Nehezebb előre megjósolni, hogy az AGI/ASI/HLAI megépül-e, és ha igen, mikor, milyen formában fognak működni az irányítási intézmények. Egy kísérlet ilyen előrejelzések készítésére Hanson [34]).

Az egyik figyelemre méltó példa erre a mesterséges intelligenciával foglalkozó nemzetközi testület (IPAI) és a globális partnerség a mesterséges intelligenciáról (GPAI). Az IPAI/GPAI-t a közelmúltban a kanadai és a francia kormány javasolta, először IPAI, majd GPAI néven [35,36]. Az IPAI/GPAI dokumentumai olyan kérdésekre helyezik a hangsúlyt, amelyek rövid távon relevánsak, és középtávon is relevánsak maradhatnak. A szemléltetés céljából felsorolt kérdések egyike a következő:

"adatgyűjtés és hozzáférés; adatellenőrzés és adatvédelem; bizalom a mesterséges intelligencia iránt; a mesterséges intelligencia elfogadása és elfogadása; a munka jövője; kormányzás, jogszabályok és igazságszolgáltatás; felelős mesterséges intelligencia és emberi jogok; méltányosság, felelősség és közjó" [35].

Az IPAI/GPAI-ról közzétett dokumentumok nem utalnak arra, hogy az AGI-vel kapcsolatos hosszú távú kérdésekre összpontosítanának. (A munka jövője vitathatóan hosszú távú kérdésnek minősülhet.) Azonban,

az IPAI/GPAI mindazonáltal hosszú távon is releváns lehet. Ha az IPAI/GPAI érvényesül, akkor hosszú ideig fennmaradhat. Összehasonlításképpen: az éghajlatváltozással foglalkozó kormányközi munkacsoport (IPCC) 1988-ban alakult, és továbbra is aktív és fontos intézmény. Az IPAI/GPAI hasonló modellt követ, mint az IPCC, és hasonlóan tartósnak bizonyulhat. Továbbá, bár a hosszú távú kérdések nem szerepelnek az IPAI/GPAI eddig közzétett korai szakaszbeli dokumentumaiban, ez nem zárja ki, hogy az IPAI/GPAI a hosszú távú kérdéseket is bevonja a hatáskörébe, ha egyszer már működőképes lesz. A hosszú távú kérdések bevonása attól függ, hogy a hosszú távú kérdések iránt érdeklődők kezdeményezik-e a részvételt az IPAI/GPAI folyamataiban. Az IPAI/GPAI-ról szóló egyik legátgondoltabb vitát a mai napig Nicolas Miailhe [37] folytatja a The Future Society [38] nevű szervezetről, amely kifejezetten azon dolgozik, hogy "holisztikusan kezelje a rövid, közép- és hosszú távú kormányzási kihívásokat" a mesterséges intelligencia területén. Ez a tevékenység azt sugallja, hogy az IPAI/GPAI olyan intézmény lehet, amely az időskálák széles skáláján működik, és jelentős mértékben a jövőben is fennmarad.

6.2. Kollektív fellépés

A mesterséges intelligencia társadalmi hatásai szempontjából fontos dinamika, hogy a mesterséges intelligencia fejlesztési projektek sikeresen tudnak-e együttműködni a kollektív cselekvési problémákban: olyan helyzetekben, amikor az összes projekt kollektív érdeke eltér a projektek egyéni érdekeitől. A kollektív cselekvés jelentős téma volt a hosszú távú mesterséges intelligenciáról szóló vitákban, és arra a kilátásra összpontosított, hogy a projektek a biztonságot illetően megszorításokat tesznek, hogy elsőként érjenek el fontos technológiai mérföldköveket [32,39]. A kollektív cselekvés problémái a rövid távú mesterséges intelligencia esetében is felmerülhetnek. A közeljövő egyik aggodalma a katonai mesterséges intelligencia fegyverkezési versenyével kapcsolatos [40] (bár ez az aggodalom nem általános [41]).

A kollektív cselekvési problémákkal foglalkozó társadalomtudományi kutatások három széles körű megoldási csoportot azonosítanak arra vonatkozóan, hogyan lehet a szereplőket együttműködésre bírni: kormányzati szabályozás, magántulajdon és közösségi önszerveződés [42]. Mindegyiket érdemes röviden megvizsgálni középtávon.

A kormányzati szabályozás talán a leggyakrabban javasolt megoldás a mesterséges intelligencia kollektív cselekvési problémáira. Míg egyes javaslatok a hazai intézkedésekre összpontosítanak [43], a globális rendszerek kedvezőek lehetnek, mivel a mesterséges intelligencia világszerte fejlődik. Ezt tükrözik a nemzetközi szerződésekre vonatkozó javaslatok [44], vagy - ami még ambiciózusabb - olyan globális kormányzási rendszerek, amelyek széles körű felügyeleti hatáskörrel és a potenciálisan veszélyes mesterséges intelligenciaprojektek erőszakkal történő megelőző leállításának képességével rendelkeznek [45]. Ez az ambiciózusabb megközelítés elméletileg vonzó lehet a mesterséges intelligencia kollektív fellépésének biztosítása szempontjából, ugyanakkor nem vonzó a visszaélések lehetősége miatt, egészen a katasztrofális totalitarizmusig [46]. Ettől függetlenül a gyakorlatban egy beavatkozó globális kormányzat jelenleg és a belátható jövőben, valószínűleg még középtávon is, valószínűleg nem fog működni. A nemzetek túlságosan valószínűtlenek, hogy hajlandóak lennének nemzeti szuverenitásukat egy globális rendszernek átengedni, különösen egy olyan kérdésben, amely jelentős gazdasági és katonai jelentőséggel bír. (Lehet, hogy bizonyos jövőbeli körülmények ezt megváltoztathatják, de a szuverenitás megőrzésének vágya, különösen a riválisés ellenséges államokkal szemben, tartósan jellemzi a nemzetközi rendszert). Még egy szerényebb nemzetközi szerződés is túl nagy kérés lehet. Szerződéseket nehéz létrehozni, különösen akkor, ha egyetemes nemzetközi konszenzusra van szükség (például azért, mert a mesterséges intelligenciát bárhol ki lehet fejleszteni), és ha a technológiához való hozzáférés és az azzal kapcsolatos képességek egyenlőtlenül oszlanak meg a nemzetközi közösségen belül (mint ahogy ez a mesterséges intelligencia esetében nagyon is jellemző; az új technológiák szerződéses kihívásainak általános megvitatását lásd [47]). Ehelyett a kormányzati szabályozás valószínűleg szerényebb lesz, és legfeljebb részleges szerepet játszik a kollektív cselekvés elősegítésében. Bármit is tesznek végül a kormányok, a 6.1. szakaszban tárgyaltak szerint nagy lehetőség van a középtávon tartós intézmények létrehozására. A természeti erőforrásokkal való gazdálkodásban általában magántulajdont alkalmaznak. A természeti erőforrást birtokló szervezetet ösztönzi a természeti erőforrás fenntartása, és rendelkezik az ehhez szükséges eszközökkel is: a

felhasználóknak kellően magas díjat kell fizetniük a hozzáférésért. A magántulajdonosi rendszerek nehezen alkalmazhatók a mesterséges intelligencia szoftverekre a hozzáférés korlátozásának nehézségei miatt. A hardverek életképesebb megoldást kínálhatnak, mivel a hardvergyártó létesítmények földrajzilag rögzített és jól látható helyszínek, amelyek jelentős ipari infrastruktúrát jelentenek, szemben a szoftverek efemer mivoltával (kapcsolódó vitát lásd [48]). A hardvergyártás jellemzően magántulajdonban van [49]. A mesterséges intelligencia kollektív fellépését elképzelhető, hogy a

a gyártók, különösen a legképzettebb mesterséges intelligencia-projektekben használt fejlett hardverek kiválasztott gyártói. A mesterséges intelligenciával kapcsolatos kollektív cselekvés előnyeit azonban számos szervezet tapasztalja, ezért a hardvergyártók szempontjából túlnyomórészt externáliáknak minősülnének, abban az értelemben, hogy az előnyöket más emberek, nem pedig a gyártók élveznék. Ez csökkenti a gyártók ösztönzését a kollektív cselekvés előmozdítására, és hasonlóképpen csökkenti a mesterséges intelligenciával kapcsolatos kollektív cselekvés magántulajdonosi rendszereinek életképességét. Mindazonáltal, amennyiben a hardvergyártás szerepet játszhat, az tartós szerepet játszhat. A hardvergyártást olyan viszonylag tartós vállalatok vezetik, mint az Intel (1968-ban alapították), a Samsung Electronics (1969-ben alapították), az SK Hynix (korábban Hyundai Electronics, 1983-ban alapították) és a Taiwan Semiconductor Manufacturing Company (1987-ben alapították). Ezek a vállalatok valószínűleg középtávon és potenciálisan hosszú távon is fontosak maradnak.

A mesterséges intelligenciával kapcsolatos kollektív cselekvésre irányuló közösségi önszerveződés több fontos területen is megfigyelhető. Az egyik ilyen terület a mesterséges intelligencia fejlesztőinek összefogására irányuló kezdeményezések, amelyek célja az etikai elvek előmozdítása. A Partnerség a mesterséges intelligenciáról (Partnership on AI) figyelemre méltó példa erre. Fontos, hogy a partnerség nemrégiben fogadta első kínai tagját, a Baidut [50]. Ez arra utal, hogy az emberi jogokra helyezett hangsúly (a partnerek között van az Amnesty International és a Human Rights Watch is) nem fogja a nyugati szervezetekre korlátozni a hatókörét. Egy másik terület a mesterséges intelligencia projektek közötti együttműködés. Baum [31] például számos kapcsolatot dokumentál az AGI-projektek között a közös személyzet és az együttműködések révén, ami egy együttműködő közösségre utal. A közösségi önszerveződés talán nem rendelkezik a kormányzati szabályozás vagy a magántulajdon elméleti eleganciájával, de a gyakorlatban gyakran sikeres. Hogy a mesterséges intelligencia esetében is sikeres lesz-e, még nem tudjuk. A mesterséges intelligenciával kapcsolatos közösségi kezdeményezések viszonylag fiatalok, így még bizonytalanabb, hogy közép- és hosszú távon hogyan fognak működni.

6.3. Vállalati AI fejlesztés

A profitorientált vállalatok pénzügyi ösztönzői minden időszakban komoly kihívást jelenthetnek a mesterséges intelligencia biztonságos és etikus fejlesztése szempontjából. Hogyan lehet meggyőzni a vállalatokat, hogy a közérdeknek megfelelően cselekedjenek, ha pénzügyi önérdekük más irányba mutat? Ez természetesen számos ágazatot érintő fontos kérdés, nem csak a mesterséges intelligenciát érintő kérdés. Jelenleg a mesterséges intelligencia szempontjából is kérdés, a közösségi médiabotokban, a megfigyelőrendszerekben és a fegyverekben használt mesterséges intelligenciával kapcsolatos aggodalmak "techlash" közepette. Közép- és hosszú távon is kérdés lehet a mesterséges intelligencia számára.

A hosszú távú mesterséges intelligenciával kapcsolatban Baum [31] (19. o.) bevezeti az "AGI profit-R&D szinergia" kifejezést, amelyet úgy határoz meg, mint "minden olyan körülmény, amelyben a hosszú távú AGI K+F rövid távú nyereséget eredményez". Ha jelentős AGI-profit-R&D szinergia áll fenn, akkor ez jelentősen megnehezítheti az AGI irányítását, mivel olyan pénzügyi ösztönzőket hoz létre, amelyek nem feltétlenül igazodnak a közérdekhez. Az AGI nyereség és a K+F szinergia a hosszú távú mesterséges intelligenciára vonatkozik, de ez természeténél fogva középtávú jelenség, mivel az AGI fejlesztése során jelentkezne. Az AGI nyereség és a kutatás-fejlesztés közötti szinergia kilátásainak értékelése megköveteli a mesterséges intelligencia műszaki informatikai részleteinek megértését, amint az a középtávon átmegy hosszú távra, ami meghaladja e dokumentum kereteit. Ha a középtávú részletek valamilyen szoros kapcsolatban állnak a közeljövőben megvalósuló mesterséges intelligenciával, az jelentősen megerősítheti a középtávú mesterséges intelligencia hipotézisét.

Ha a mesterséges intelligenciával foglalkozó vállalatok pénzügyi önérdeke eltér a közérdektől, hogyan fognak viselkedni? Ideális esetben a közérdeknek megfelelően cselekednének. Bizonyos esetekben talán így is tesznek, különösen, ha a vállalatokon belüli és kívüli emberek erre ösztönzik őket. Sajnos más ágazatok tapasztalatai azt mutatják, hogy a vállalatok gyakran a közérdek ellenében cselekszenek, ahogyan például a dohányipar a rákkockázat csökkentését célzó szabályozás ellen, a fosszilis tüzelőanyag-ipar a globális felmelegedés kockázatának csökkentését

célzó szabályozás ellen [51], és az ipari vegyipar az idegrendszeri betegségek kockázatának csökkentését célzó szabályozás ellen [52]. Érdemes megfontolni, hogy a mesterséges intelligenciával foglalkozó vállalatok is hasonlóan (félre)viselkedhetnek.

Felvetődött, hogy a mesterséges intelligenciával foglalkozó vállalatok politizálhatnak a mesterséges intelligenciával és annak kockázataival kapcsolatos szkepticizmussal, hogy elkerüljék a nyereséges tevékenységüket korlátozó szabályozást [53]. Az ilyenfajta politizált szkepticizmusnak hosszú története van, kezdve a dohányipar szkepticizmusával a cigaretta és a rák közötti összefüggéssel kapcsolatban.

és a mai napig folytatódik, például a fosszilis tüzelőanyag-ipar szkepticizmusa a globális felmelegedéssel kapcsolatban. Ennek a munkának az egyik mechanizmusa a névleg független agytrösztök finanszírozása, hogy olyan kiadványokat készítsenek, amelyek a vállalatok pénzügyi önérdekeinek megfelelő politikákat és álláspontokat támogatnak.

Ennek a mintának néhány jellemzője látható a Center for Data Innovation agytröszt nemrégiben megjelent írásában, amely egy "ellenőrizetlen technopánikra" figyelmeztet, amely visszafogja a közvélemény lelkesedését a mesterséges intelligencia iránt, és kormányzati szabályozást motivál [54]. Nem világos, hogy ez mennyiben tekinthető a politizált szkepticizmus esetének. Konkrétan a Center for Data Innovation ipari kapcsolatainak mértékét nem lehetett megállapítani e tanulmányhoz. Hasonlóképpen nem célja ennek a tanulmánynak, hogy ezt a szervezetet összeférhetetlenséggel vádolja. Az sem áll szándékunkban, hogy az ellenkezőjét állítsuk - hogy ebben az esetben nincs összeférhetetlenség. (Valójában az összeférhetetlenség jelenléte gyakran rejtve marad - ezért finanszírozzák az ipari cégek a névlegesen független agytrösztök munkáját, ahelyett, hogy házon belül végeznék azt). Ehelyett a szándék csupán az, hogy egy olyan példát mutassunk be, amely a politizált szkepticizmus mintájának néhány aspektusát illusztrálja. Fontos, hogy míg a politizált AI-szkepticizmusra vonatkozó javaslat a hosszú távú AI-val kapcsolatos szkepticizmusra összpontosít [53], addig a Center for Data Innovation szkepticizmusa a közeljövőre összpontosít [54]. Hasonlóképpen, a politizált AI-szkepticizmus mintázata potenciálisan több időszakon átívelően is érvényesülhet, különösen akkor, ha jelentős profit és a kutatás-fejlesztés szinergiája, valamint a kormányzati szabályozás egyidejű kilátása áll fenn.

6.4. Katonaságok és nemzetbiztonsági közösségek

A fejlett hadseregek már régóta részt vesznek a mesterséges intelligencia élvonalában, mint a kutatás finanszírozói és egyre inkább mint a technológia felhasználói. A fejlett hadseregek gyakran jelentős műszaki szakértelemmel is rendelkeznek, csakúgy, mint a tágabb értelemben vett nemzetbiztonsági politikai közösségek, amelyekkel kapcsolatban állnak. Továbbá a hadseregek néha különböző időszakokra kiterjedő műveletekkel és tervezéssel vannak megbízva, és a nemzetbiztonsági közösségek is néha ilyen időszakokban való gondolkodásra orientálódnak. Ezt mutatja a fent említett példa, a B-52H bombázógép 2050-ig történő üzemben tartásának terve. Értelemszerűen a fejlett hadseregek és a nemzetbiztonsági közösségek érdeklődhetnek a középtávú mesterséges intelligencia és annak a közeli és hosszú távú időszakok közötti kapcsolatai iránt.

A hadsereg már most is figyelmet fordít az AGI-ra. Egyértelmű példa erre a JASON jelentése: *Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD* [55], amely az Egyesült Államok Védelmi Minisztériumának az AGI-vel kapcsolatos megkeresésére készült. A másik a kiváló könyv [19], amely egy teljes fejezetet szentel az AGI-nek és a mesterséges intelligenciának. Mindkét kiadvány árnyaltan számol be a hosszú távú mesterséges intelligenciáról. A kiadványokat olyan elemzők készítették, akik különösen technikailag jártasak, és nem reprezentálják a teljes katonai és nemzetvédelmi közösséget. Mindazonáltal ezek a kiadványok azon kiadványok közé tartoznak, amelyeket az e közösségekben dolgozó emberek tanulmányozhatnak, és jelzik a hosszú távú mesterséges intelligenciával kapcsolatos tudatosság bizonyos fokát.

Amint azt Baum [31] dokumentálta, van néhány jelenlegi AGI K+F projekt katonai kapcsolatokkal. Ezek többsége olyan amerikai akadémiai csoport, amelyek katonai kutatási ügynökségektől, például a DARPA-tól és a Haditengerészeti Kutatási Hivataltól kapnak támogatást. Az egyik egy kis csoport Szingapúr elsődleges nemzeti védelmi kutatási ügynökségénél. Egyiküknél sincs semmi jele annak a fajta nagyszabású stratégiai kezdeményezésnek, amelyet a hosszú távú mesterséges intelligenciával foglalkozó szakirodalomban néha feltételeznek [33].

A dolgok jelenlegi állása alapján nagyon valószínű, hogy a fejlett hadseregek és a nemzetbiztonsági közösségek középtávon is foglalkozni fognak a mesterséges intelligenciával. Ez felveti a kérdést, hogy milyen szerepet játszhatnak. A mesterséges intelligenciával foglalkozó közösségeken belüli általános aggodalmak ellenére - mint például a Google alkalmazottainak a Maven projekt miatti tiltakozása - a hadseregek valóban konstruktív hangot adhatnak az etika és a biztonság kérdésében. Az [55] jelentés egyik fő témája például az, hogy az általa "ilities"-nek nevezett "megbízhatóság, karbantarthatóság, elszámoltathatóság, ellenőrizhetőség, fejleszthetőség,

támadhatóság és így tovább" [55] (2. o.) komoly aggodalomra ad okot a katonai alkalmazások tekintetében, és "potenciális akadálya annak, hogy a DoD használja ezeket a modern AI rendszereket, különösen, ha figyelembe vesszük a felelősséget és az elszámoltathatóságot a mesterséges intelligencia halálos rendszerekben való alkalmazásával kapcsolatban" [55] (27. o.). A hadseregek igyekeznek elkerülni a nem szándékolt következményeket, különösen a nagy tétet jelentő harctéri technológiák esetében.

Fontos figyelembe venni azt a geopolitikai kontextust is, amelyben a hadseregek működnek. A hadseregek megengedhetik maguknak, hogy visszafogottabban fejlesszenek és használjanak kockázatos technológiákat, ha a nemzetük

békében vannak. Egy interjúban Larry Schuette, a Haditengerészeti Kutatási Hivatal munkatársa az autonóm fegyvereket a tengeralattjárókhoz hasonlítja [19] (100-101. oldal). Schuette elmeséli, hogy az 1920-as és 1930-as években az USA ellenezte a korlátlan tengeralattjáró-háborút, de ez azonnal megváltozott az 1941. december 7-i Pearl Harbor elleni támadást követően. Hasonlóképpen, az USA jelenleg is ellenzi az autonóm fegyvereket, és arra a kérdésre, hogy továbbra is ellenzi-e, Schuette így válaszol: "December nyolcadika vagy december hatodika van"?

Ebből következik, hogy a hadseregek szerepe a középtávú mesterséges intelligenciában nagyban függhet a nemzetközi kapcsolatok állapotától ebben az időszakban. Értelemszerűen az óvatos és etikus mesterséges intelligencia fejlesztésének kilátásai sokkal nagyobbak békeidőben, mint háborús időkben. Ahogyan Danzig [56] is megfogalmazta, a stratégiai előnyök érdekében történő technológiai előretörés és a nem szándékolt következményekkel szembeni óvatosság között eredendő feszültség áll fenn. A békés nemzetközi kapcsolatok az óvatosság felé billentik a számítást, és felhatalmazhatják a hadseregeket és a nemzetbiztonsági közösségeket, hogy fontos hangot adjanak a biztonság és az etika kérdésében.

7. Következtetések

Parson et al. [2] azzal érvelt, hogy a mesterséges intelligencia és a hozzá kapcsolódó társadalmi kérdések középtávon önmagukban is fontosak. A jelen tanulmány elemzése ugyanerre a következtetésre jut. Az itt vizsgált problématerületek mindegyikénél - kormányzati intézmények, kollektív cselekvés, vállalati fejlődés és katonai/nemzetbiztonság - a középtávú időszak fontos folyamatokat fog tartalmazni. Bizonyos értelemben ez nem is olyan nagy következtetés. Az már most is világos, hogy a mesterséges intelligencia rövid távon fontos, és bőven van okunk azt hinni, hogy a technológia és alkalmazásai további fejlődésével a mesterséges intelligencia még fontosabbá válik.

Mi a helyzet tehát a jelen- és jövőkutató vitával? Ez az írás a középtávú mesterséges intelligencia hipotézisét javasolja, amely szerint létezik egy köztes időszak, amely mind a prezentisták, mind a futuristák szempontjából fontos. Mivel a közeljövő a megvalósíthatóság, a hosszú távú pedig a képességek szempontjából meghatározott, ebből következik, hogy a középtávú mesterséges intelligencia hipotézis nagyobb valószínűséggel állja meg a helyét, ha a közeljövő mesterséges intelligencia technikái és társadalmi dimenziói jelentős mértékben fennmaradnak a középtáv végén, amikor a hosszú távú mesterséges intelligencia kiépül. Amennyiben a hipotézis igaz, a középtávra fordított figyelem fontos szerepet játszhat a jelen- és jövőkutató közösségek közötti szakadék áthidalásában.

A tanulmány vegyes alátámasztást talál a középtávú mesterséges intelligencia hipotézisre. A támogatás erős a mesterséges intelligencia irányítási intézményei esetében, amelyek jelenleg fejlesztés alatt állnak, és középtávon is fennmaradhatnak, ami a hosszú távú mesterséges intelligenciára is hatással lehet. A támogatottság nem egyértelmű a mesterséges intelligenciával kapcsolatos kollektív cselekvés esetében: a kollektív cselekvés előmozdítására irányuló kormányzati kezdeményezések viszonylag kevés szerepet játszhatnak mindenkor, a magántulajdonosi rendszereket nehéz megszervezni a mesterséges intelligencia számára, és a közösségi önszerveződésben van potenciál, amely vagy megvalósul, vagy nem. A kollektív cselekvés megvalósításának e három rendszere közül mindhárom rövid és középtávon potenciálisan megvalósulhat, ami hatással lehet a hosszú távú mesterséges intelligenciára, de hogy ez valószínűsíthető-e, az nem világos. Ami a vállalati AI-fejlesztést illeti, az egyik kulcskérdés, hogy a közeli és középtávú AI-technológia az AGI nyereséges előfutára lehet-e, ami AGI-profit és K+F szinergiát hozhat létre. Hogy ez a szinergia létrejön-e, az a jövőbeni kutatások fontos kérdése. Végezetül, a fejlett hadseregek és a nemzetbiztonsági közösségek már most is figyelmet fordítanak az AGI-re, és középtávon valószínűleg továbbra is aktívak maradnak a különböző AI-technológiák terén. Bár nem világos, hogy a katonai/nemzetbiztonsági közösségek fontos szereplői lesznek-e az AGI fejlesztésének, jelentős potenciál van benne, ami alátámasztja a középtávú AI-hipotézist.

Végezetül, ez az írás megmutatta, hogy középtávon valószínűleg legalább néhány fontos mesterséges intelligencia-folyamat játszódik le, és hogy ezek a maguk nemében és mind a jelen-, mind a jövőbelátás szempontjából fontosak lesznek. A középtávú mesterséges intelligencia pontos természete és jelentősége a jövőbeni kutatások méltó tárgya. Amennyiben a középtávú mesterséges

intelligenciát meg lehet érteni, ez rámutathat azokra a lehetőségekre, amelyek pozitívan befolyásolhatják azt, ami jobb általános eredményeket eredményezhet a társadalom számára.

Finanszírozás: Irlam Jótékonysági Alapítvány finanszírozta.

Köszönetnyilvánítás: Maas, Jun Hong Yap, Steven Umbrello, Richard Re, Ted Parson, két névtelen véleményező, valamint a Berkeley Egyetem emberekkel kompatibilis mesterséges intelligenciával foglalkozó központja és a Global Catastrophic Risk Institute által szervezett szemináriumok hallgatósága. Robert de Neufville is nyújtott kutatási segítséget. Dakota Norris segítséget nyújtott a kézirat elkészítésében. A fennmaradó hibák kizárólag a szerzőt terhelik.

Összeférhetetlenség: A szerző nem jelent összeférhetetlenséget.

Hivatkozások

1. Baum, S.D. A közeli és hosszú távú mesterséges intelligenciára összpontosító frakciók közötti megbékélés. *AI Soc.* **2018**, *33*, 565-572. [CrossRef]
2. Parson, E.; Re, R.; Solow-Niederman, A.; Zeide, E. Mesterséges intelligencia stratégiai kontextusban: An Introduction. *AI Pulse*. 2019. február 8. Online elérhető: https://aipulse.org/artificial-intelligence-in-strategic-context-an- bevezetés (elérés: 2020. február 2.).
3. Parson, E.; Fyshe, A.; Lizotte, D. A mesterséges intelligencia társadalmi hatásai, irányítás és etika: Bevezetés a 2019. évi nyári intézetbe az AI és a társadalom és annak gyors kimeneteleiről. *AI Pulse*. 2019. szeptember 26. Online elérhető: https://aipulse.org/artificial-intelligences-societal-impacts-governance-and-ethics-introduction-to- the-2019-summer-institute-on-ai-and-society-and-its-rapid-outputs (hozzáférés: 2020. február 2.).
4. Cave, S.; Ó hÉigeartaigh, S.S. A mesterséges intelligenciával kapcsolatos közeli és hosszú távú aggodalmak áthidalása. *Nat. Mach. Learn.* **2019**, *1*, 5-6. [CrossRef]
5. Prunkl, C.; Whittlestone, J. Túl a közeli és a hosszú távú: A mesterséges intelligencia etikája és társadalma kutatási prioritásainak világosabb bemutatása felé. In Proceedings of the Third AAAI/ACM Annual Conference on AI, Ethics, and Society, New York, NY, USA, 2020. február 7. New York, NY, USA.
6. Zeng, Y.; Lu, E.; Huangfu, C. A mesterséges intelligencia elveinek összekapcsolása. In Proceedings of the AAAI Workshop on Artificial Intelligence Safety, Honolulu, HI, USA, 2019. december 12.
7. Whittlestone, J.; Nyrup, R.; Alexandrova, A.; Cave, S. Az elvek szerepe és korlátai a mesterséges intelligencia etikájában: A feszültségek középpontba állítása felé. In Proceedings of the Second AAAI / ACM Annual Conference on AI, Ethics, and Society, Honolulu, HI, USA, 2019. január 27. Honolulu, HI, USA, 2019. január 27.
8. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Konkrét problémák az AI biztonságában. 2016. Online elérhető: https://arxiv.org/abs/1606.06565 (elérés: 2020. február 2.).
9. Conn, A. Átirat: Dario Amodeival és Seth Baummal: Konkrét problémák a mesterséges intelligencia biztonságában. *Az Élet Jövője Intézet*. 2016. Online elérhető: https://futureoflife.org/2016/08/31/transcript-concrete-problems-ai-safety- dario-amodei-seth-baum (elérés: 2020. február 2.).
10. Arrhenius, G. A személyre ható korlátozás, a komparativizmus és a potenciális emberek erkölcsi státusza. *Etikai nézőpont.* **2005**, *10*, 185-195. [CrossRef] [PubMed]
11. Bostrom, N. *Szuperintelligencia:* Oxford University Press: *Paths, Dangers, Strategies:* Oxford, UK, 2014.
12. Garling, C. Andrew Ng: Ng: Miért a "mélytanulás" nem csak a gépek, hanem az emberek számára is kötelező. *Wired*. 2015. Online elérhető: https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans- not-just-machines (elérés: 2020. február 2.).
13. Calo, R. Mesterséges intelligencia politika: A Primer and Roadmap. Elérhető online: https://www.ssrn.com/ abstract=3015350 (hozzáférés: 2020. február 2.).
14. Good, I.J. Spekulációk az első ultraintelligens géppel kapcsolatban. In *Advances in Computers*; Alt, F.L., Rubinoff, M., Eds.; Academic Press: New York, NY, USA, 1965; pp. 31-88.
15. Etzioni, O. Nem, a szakértők szerint a szuperintelligens mesterséges intelligencia nem jelent fenyegetést az emberiségre. *MIT Technology Review*. 2016. szeptember 20. Online elérhető: https://www.technologyreview.com/s/602410/no-the-experts-dont-. think-superintelligent-ai-is-a-threat-to-humanity (hozzáférés: 2020. február 20.).
16. Dafoe, A.; Russell, S. Igen, aggódunk a mesterséges intelligencia egzisztenciális kockázata miatt. *MIT Technology Review*. 2016. november 2. Online elérhető: https://www.technologyreview.com/s/602776/yes-we-are- worried-about-the-existential-risk-of-artificial-intelligence (elérés: 2020. február 2.).
17. Ford, M. A mesterséges intelligencia munkanélküliségi válságot okozhat? *Commun. ACM* **2013**, *56*, 1-3. [CrossRef]
18. A szövetségi ügynökségeknek foglalkozniuk kell az elöregedő örökölt rendszerekkel. United States Government Accountability Office, GAO-16-468. 2016. Elérhető online: https://www.gao.gov/assets/680/677436.pdf (elérés: 2020. február 2.).
19. Scharre, P. *Army of None:* W. W. Norton: *Autonóm fegyverek és a háború jövője:* New York, NY, USA, 2018.

20. Mizokami, K. Hogyan fognak repülni a B-52-es bombázók a 2050-es évekig. *Popular Mechanics*. 2018. szeptember 10. Elérhető online: https://www.popularmechanics.com/military/aviation/a23066191/b-52-bombers-fly-until-the-2050s (elérés: 2020. február 2.).
21. Roblin, S. Bombák távolodnak: Oroszország "új" Tu-22M3M bombázója ismerős lehet (és még mindig halálos). *The National Interest*. 2018. október 13. Online elérhető: https://nationalinterest.org/blog/buzz/bombs-away- russias-new-tu-22m3m-bomber-might-look-familiar-and-still-deadly-33381 (elérés: 2020. február 2.).
22. Byrnes, M.W. Nightfall: Gépi autonómia a levegő-levegő harcban. *Air Space Power J.* **2014**, *május-június*, 48-75.
23. Baum, S.D.; Goertzel, B.; Goertzel, T.G. Mennyi idő van az emberi szintű mesterséges intelligenciáig? Egy szakértői értékelés eredményei.
Technol. Előrejelzés. Soc. Chang. **2011**, *78*, 185-195. [CrossRef]
24. Sandberg, A.; Bostrom, N. Machine Intelligence Survey. Technikai jelentés #2011-1, Future of Humanity Institute, Oxfordi Egyetem. 2011. Online elérhető: https://www.fhi.ox.ac.uk/wp-content/uploads/2011-1.pdf (elérés: 2020. május 27.).
25. Müller, V.C.; Bostrom, N. A mesterséges intelligencia jövőbeli fejlődése: A szakértők körében végzett közvélemény-kutatás. In *Fundamental Issues of Artificial Intelligence*; Müller, V.C., Ed.; Springer: Berlin, Germany, 2016; pp. 555-572.
26. Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. Mikor fogja az AI meghaladni az emberi teljesítményt? Bizonyítékok a mesterséges intelligencia szakértőitől. *J. Artif. Intell. Res.* **2018**, *62*, 729-754. [CrossRef]
27. Armstrong, S.; Sotala, K.; Ó hÉigeartaigh, S.S. A híres mesterséges intelligencia előrejelzések hibái, felismerései és tanulságai - És mit jelentenek a jövőre nézve. *J. Exp. Theor. Artif. Intell.* **2014**, *26*, 317-342. [CrossRef]
28. Morgan, M.G. A szakértői véleménynyilvánítás használata (és visszaélése) a közpolitikai döntéshozatal támogatására. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 7176-7184. [CrossRef] [PubMed]
29. Arkin, R. Halálos autonóm rendszerek és a nem harcosok helyzete. In *The Political Economy of Robots*; Kiggins, R., szerk.; Palgrave Macmillan: Cham, Svájc, 2018; pp. 317-326.
30. Rosert, E.; Sauer, F. Az autonóm fegyverek tiltása: Az emberi méltóság előtérbe helyezése. *Glob. Policy* **2019**, *10*, 370-375. [CrossRef]
31. Baum, S.D. A mesterséges általános intelligencia etikai, kockázati és politikai projektjeinek áttekintése. *GCRI Work. Pap.* **2017**,
2017. [CrossRef]
32. Armstrong, S.; Bostrom, N.; Shulman, C. Verseny a szakadék felé: A mesterséges intelligencia fejlődésének modellje.
AI Soc. **2016**, *31*, 201-206. [CrossRef]
33. Shulman, C. Fegyverzetellenőrzés és hírszerzési robbanások. In Proceedings of the 7th European Conference on Computing and Philosophy, Bellaterra, Spanyolország, 2009. július 2-4.
34. Hanson, R. *A korszak Em: Work, Love, and Life When Robots Rule the Earth*; Oxford University Press: Oxford, UK, 2016.
35. Kanada miniszterelnöke. Megbízás a mesterséges intelligenciával foglalkozó nemzetközi testület számára. *Kanada*. 2018. december 6. Elérhető online: https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-
artificial-intelligence (hozzáférés: 2020. február 2.).
36. Kohler, K.; Oberholzer, P.; Zahn, N. *Making Sense of Artificial Intelligence:* Swiss Forum on Foreign Policy: *Why Switzerland Should Support a Scientific UN Panel to Assess the Rise of AI:* Genf, Svájc, 2019; Elérhető online: https://www.foraus.ch/wp-content/uploads/2019/10/20191022_Making-Sense-of-AI_WEB-.
1.pdf (hozzáférés: 2020. február 2.).
37. Miailhe, N. AI & Global Governance: Miért van szükség a mesterséges intelligenciával foglalkozó kormányközi testületre. Centre for Policy Research, United Nations University. 2018. december 20. Online elérhető: https://cpr.unu.edu/ai-global-governance-why-we-need-an-intergovernmental-panel-for-artificial- intelligence.html (elérés: 2020. február 2.).
38. A mesterséges intelligencia kezdeményezés. A jövő társadalma. 2018. Elérhető online: http://thefuturesociety.org/the-ai-initiative (elérés: 2020. február 2.).
39. Cave, S.; Ó hÉigeartaigh, S.S. An AI race for strategic advantage: Retorika és kockázatok. In Proceedings of the AAAI/ACM Annual Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2-3 February 2018.
40. Geist, E.M. Már túl késő megállítani a mesterséges intelligencia fegyverkezési versenyét - ehelyett

irányítanunk kell. *Bull. At. Sci.* **2016**, *72*, 318-321. [CrossRef]
41. Roff, H.M. A keretprobléma: Az AI "fegyverkezési verseny" nem az. *Bull. At. Sci.* **2019**, *75*, 95-98. [CrossRef]
42. Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action*; Cambridge University Press: Cambridge, Egyesült Királyság, 1990.

43. Scherer, M.U. A mesterséges intelligencia rendszerek szabályozása: Kockázatok, kihívások, kompetenciák és stratégiák.
Harv. J. Law Technol. **2016**, *29*, 353-400. [CrossRef]
44. Wilson, G. Az új technológiákból eredő globális katasztrofális és egzisztenciális kockázatok minimalizálása a nemzetközi jog segítségével. *Va. Environ. Law J.* **2013**, *31*, 307-364.
45. Bostrom, N. A sebezhető világ hipotézise. *Glob. Policy* **2019**, *10*, 455-476. [CrossRef]
46. Caplan, B. A totalitárius fenyegetés. In *Global Catastrophic Risks*; Bostrom, N., C´ irkovic´, M.M., Oxford University Press: Oxford, UK, 2008; pp. 504-519. szerkesztők;
47. Picker, C.B. Kilátás 40 000 láb magasból: A nemzetközi jog és a technológia láthatatlan keze. *Cardozo Law Rev.* **2001**, *23*, 149-219.
48. Hwang, T. Számítási teljesítmény és a mesterséges intelligencia társadalmi hatása. 2019. Online elérhető: https://www.ssrn.com/abstract=3147971 (elérés: 2020. február 2.).
49. Félvezetőgyártó üzemek listája. *Wikipedia*. Elérhető online: https://en.wikipedia.org/wiki/List_of_semiconductor_fabrication_plants (hozzáférés: 2020. február 2.).
50. Bemutatjuk az első kínai tagunkat a mesterséges intelligenciával foglalkozó partnerségben. Partnerség a mesterséges intelligenciáról. 2018. október 16. Elérhető online: https://www.partnershiponai.org/introducing-our-first-chinese-member-to-the-partnership-on-ai (hozzáférés: 2020. február 2.).
51. Oreskes, N.; Conway, E.M. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*; Bloomsbury: New York, NY, USA, 2010.
52. Grandjean, P. *Csak egy esély: the Brains of the Next Generation*; Oxford University Press: Oxford, UK, 2013.
53. Baum, S.D. A szuperintelligencia-szkepticizmus mint politikai eszköz. *Information* **2018**, *9*, 209. [CrossRef]
54. Castro, D. Az USA elveszítheti a mesterséges intelligencia versenyét a féktelen technopánik miatt. Center for Data Innovation. 2019. március 5. Online elérhető: https://www.datainnovation.org/2019/03/the-u-s-may-lose-the-ai-race-.
because-of-an-un-checked-techno-panic (hozzáférés: 2020. február 2.).
55. Potember, R. *A mesterséges intelligencia és a mesterséges általános intelligencia kutatásának a védelmi minisztérium szempontjából releváns perspektívái*; The MITRE Corporation: McLean, VA, USA, 2017.
56. Danzig, R. Technológiai rulett: Az irányítás elvesztésének kezelése, miközben számos hadsereg technológiai fölényre törekszik. Center for a New American Security. 2018. május 30. Online elérhető: https://www.cnas.org/ publications/reports/technology-roulette (elérés: 2020. február 2.).