Kutatási prioritások a megbízható és hasznos mesterséges intelligencia létrehozásáért

A mesterséges intelligenciával kapcsolatos kutatások sikeressége magában hordozza a lehetőségét annak, hogy az emberiség példa nélküli előnyökhöz jusson. Érdemes ezért feltárni azokat a kutatási területeket, amelyek az esetleges buktatók elkerülésével egy időben segíthetnek maximalizálni az elérhető eredményeket. Jelen tanulmány számos ilyen témakört és példát mutat be (a teljesség igényének hajszolása nélkül), amelyek biztosíthatják, hogy az mesterséges intelligencia a jövőben is robusztus és az ember számára előnyös maradjon.

Kulcsszavak: *mesterséges intelligencia, rövid és hosszú távú hatások, jog és etika, kutatási irányok*

Így hivatkozzon erre a cikkre:
„Kutatási prioritások a megbízható és hasznos mesterséges intelligencia létrehozásáért”.
Információs Társadalom XV, 4. szám (2015): 60–76.
https://dx.doi.org/10.22503/inftars.XV.2015.4.7

Kutatási prioritások a megbízható és hasznos mesterséges intelligencia létrehozásáért[1]

A mesterséges intelligenciával kapcsolatos kutatások sikeressége magában hordozza a lehetőségét annak, hogy az emberiség példa nélküli előnyökhöz jusson. Érdemes ezért feltárni azokat a kutatási területeket, amelyek az esetleges buktatók elkerülésével egy időben segíthetnek maximalizálni az elérhető eredményeket. Jelen tanulmány számos ilyen témakört és példát mutat be (a teljesség igényének hajszolása nélkül), amelyek biztosíthatják, hogy az mesterséges intelligencia a jövőben is robusztus és az ember számára előnyös maradjon.

1. A MESTERSÉGES INTELLIGENCIA NAPJAINKBAN

A mesterséges intelligencia (MI) kutatása már a kezdetektől fogva számos különböző problémát és megközelítést vetett fel, de az elmúlt 20 év során főként az *intelligens ágensek* – olyan rendszerek, amelyek bizonyos környezetben képesek az érzékelésre és a cselekvésre – megalkotása körüli problémákra koncentrált. Ebben a kontextusban az intelligencia jellegzetesen a racionalitás statisztikai és gazdasági fogalmaihoz kapcsolódik – egyszerűbben fogalmazva a helyes döntések meghozatalára, a tervezésre vagy következtetések levonására való képességet jelenti. A valószínűségi megközelítés és a statisztikai tanulási metódusok alkalmazása nagyfokú integrációhoz vezetett, ahol termékenyítőleg hatott egymásra az MI, a gépi tanulás, a statisztika, az irányítás-elmélet, a neurológia és más, kapcsolódó területek. A közös elméleti keretek létrehozása a jelenleg rendelkezésre álló adatmennyiséggel és számítási kapacitással kombinálva figyelemre méltó sikereket eredményezett az olyan területeken, mint például a beszédfelismerés, a képek osztályozása, az autonóm járművek, a gépi fordítás, a robotok lábbal történő helyváltoztatása és a kérdés-válasz rendszerek.

Ahogy a lehetőségek ezeken és más területeken lehetővé tették, hogy a megoldások átlépjenek a laboratóriumi kísérleteken és elinduljanak a gazdaságilag is hasznosítható technológiák fejlesztése felé, úgy a teljesítmény kismértékű javulása is jelentős pénzt és nagyobb kutatási beruházásokat mozgat meg. Széleskörű egyetértés mutatkozik abban, hogy az MI kutatás folyamatos fejlődése egyre nagyobb hatással lehet a társadalomra. A potenciális előnyök óriásiak, hiszen minden, amit a civilizáció nyújtani tud, az az emberi intelligencia terméke; nem tudjuk megjósolni, hogy mit érhetünk el, ha ezt az intelligen-

[1] A tanulmány első verzióját Stuart Russell, Daniel Dewey és Max Tegmark írta Janos Kramar és Richard Mallah anyagainak felhasználásával. Észrevételeikkel, visszajelzéseikkel a tanulmányt segítették: Anthony Aguirre, Erik Brynjolfsson, Ryan Calo, Tom Dietterich, Dileep George, Bill Hibbard, Demis Hassabis, Eric Horvitz, Leslie Pack Kaelbling, James Manyika, Luke Muehlhauser, Michael Osborne, David Parkes, Heather Roff, Francesca Rossi, Bart Selman, Murray Shanahan. A fordítást a legutolsó elérhető verzió (http://futureoflife.org/static/data/documents/research_priorities.pdf 2015.01.23) alapján Csótó Mihály és Tamaskó Dávid készítette.

ciát MI-rendszerekkel támogatjuk, de a szegénység és a betegségek felszámolása sem elképzelhetetlen. Az MI nagy lehetőségeket tartogat, ezért érdemes megvizsgálni, hogyan lehet kihasználni az előnyeit, elkerülve a lehetséges buktatókat és csapdahelyzeteket.

Az MI kutatásában tapasztalt fejlődés alapján elérkezett az idő, hogy a kutatások ne csak mesterséges intelligencia alkalmazhatóbbá tételére, hanem annak társadalomra gyakorolt pozitív hatásainak maximalizálásra is kiterjedjenek, és meg is jelentek az első ilyen jellegű kezdeményezések (Pl. a Szövetség a Mesterséges Intelligencia Fejlesztéséért (Association for the Advancement of Artificial Intelligence (AAAI)) 2008-2009-es elnöki panelje az MI hosszú távú jövőjéről (Horvitz – Selman, 2009)). Ez a folyamat jelentős befolyással lehet az MI kutatások kiterjedésére, amelyek mindezidáig elsősorban a technikai megvalósíthatóságra összpontosítottak, azok hasznosítási területeire általában nem terjedtek ki. Jelen tanulmányra úgy lehet tekinteni, mint ezeknek az erőfeszítéseknek a természetes folytatására, ami azoknak a kutatási irányoknak a meghatározására összpontosít, amelyek célja a mesterséges intelligencia társadalmi előnyeinek kiaknázása. Ez a kutatás szükségszerűen interdiszciplináris, hiszen magában foglalja a társadalmat és a mesterséges intelligenciát is. Érinti a közgazdaságtant, a jogot, a filozófiát, a számítógépes biztonságot, a formális módszereket és természetesen az MI különböző ágait. A cél olyan MI fejlesztése, ami *hasznos* a társadalomra nézve és *robusztus* abban az értelemben, hogy jótéteményei garantáltak, azaz MI rendszereinknek azt kell tenniük, amit mi szeretnénk, hogy tegyenek.

2. RÖVID TÁVÚ KUTATÁSI PRIORITÁSOK

2.1 Az MI gazdasági hatásainak optimalizálása

Az MI ipari alkalmazásainak sikere a legkülönbözőbb területeken (a gyártási folyamatoktól kezdve az információs szolgáltatásokig) egyre nagyobb hatást gyakorol a gazdaság növekedésére, habár ennek a hatásnak a természetével kapcsolatban, és abban, hogyan lehet különbséget tenni a mesterséges intelligencia és egyéb információtechnológiák hatásai között, nincs egyetértés a szakemberek között. Sok közgazdász és informatikus egyetért abban, hogy érdemes kutatni azt, hogyan lehet maximalizálni az MI előnyeit a káros hatások – mint a növekvő társadalmi egyenlőtlenség és a munkanélküliség – csökkentése mellett (Mokyr, 2014; Brynjolfsson–McAfee, 2014; Frey–Osborne, 2013; Glaeser, 2014; Nilsson, 1984; Manyika, 2013). E megfontolások egy sor, a közgazdaságon és a pszichológián átívelő kutatási irányt ösztönöznek. A következőkben a teljesség igénye nélkül néhány példát mutatunk ezekre.

1. **Munkaerő-piaci előrejelzés**: mikor és milyen sorrendben várhatjuk a különböző munkák automatizálását (Frey–Osborne, 2013)? Hogyan fog ez hatni a kevésbé szakképzett munkások, a kreatív munkát végzők, valamint a különböző informatikai szakemberek munkabérére? Vannak, akik az állítják, hogy az MI nagyban növelné az emberiség egészének általános jólétét (Brynjolfsson–McAfee, 2014). Ugyanakkor az automatizáció növekedése még jobban eltolhatja a jövedelemelosztás mértékét a hatványtörvény irányába (Brynjolfsson–McAfee–Spance, 2014), és az ebből eredő egyenlőtlenségek aránytalanul jelenhetnek meg a faji, társadalmi és nemi vonalak mentén;

ezért a kutatások, melyek a gazdasági és társadalmi hatások aránytalanságának mértékével foglalkoznak, egyértelműen hasznosnak lehetnek.

2. **Egyéb piaci zavarok/diszruptív (radikális változással járó) hatások**: a gazdaság jelentős része, beleértve a pénzügyet, biztosításokat, biztosítás-statisztikát és egyéb fogyasztói piacokat könnyen tárgya lehet a kreatív rombolás jelenségének az MI rendszerek használatának köszönhetően, melyek képesek modellezni és megjósolni az ágensek cselekedeteit. Az ilyen piacokra jellemző a magas komplexitás és a komplexitás kezeléséből következő jelentős haszon kombinációja (Manyika, 2013).

3. **Politikák a káros hatások kezeléséhez**: milyen irányelvek segíthetik az egyre nagyobb mértékben automatizálódó társadalmak kiteljesedését? Brynjolfsson és McAfee (2014) például bemutat különböző megoldásokat a munkaerő-intenzív szektorok fejlődésének ösztönzésére, illetve az MI által létrehozott jólét felhasználására az alulfoglalkoztatott népesség támogatására. Mik az előnyei és hátrányai az olyan beavatkozásoknak, mint az oktatási reform, a gyakornoki programok, a munkaerőigényes infrastrukturális projektek, valamint a minimálbérre, az adózási struktúrára és a szociális hálóra vonatkozó szabályozások módosítása (Glaeser, 2014)? A történelemben számos példát találunk, amikor a népesség egy részének nem kellett dolgoznia az anyagi biztonságért, legyen szó az ókori arisztokráciától vagy a mai Katar számos állampolgáráról. Milyen társadalmi struktúrák és egyéb tényezők határozzák meg az ilyen társadalmak prosperitását? A munkanélküliség nem egyenlő a szabadidővel, szoros összefüggés mutatható ki a munkanélküliség és a boldogtalanság, az önbizalomhiány és az elszigeteltség között (Hetschko–Knabe–Schöb, 2014; Clark–Oswald, 1994); annak a megértése, hogy milyen politikai beavatkozások és normák szüntethetik meg ezt a kapcsolatot, jelentősen javíthatja az átlagos életszínvonalat. Az empirikus és elméleti kutatások az olyan témákkal kapcsolatban, mint a feltétel nélküli alapjövedelem, nagyban segítenék a lehetőségek feltárását (Van Parijs et al., 1992; Widerquist, 2013).

4. **Gazdasági mérések**: Elképzelhető, hogy a gazdasági mérőszámok, mint például az egy főre jutó reál GDP nem mutatja meg pontosan a mesterséges intelligencián és az automatizáción alapuló gazdaság előnyeit és hátrányait, ezért stratégiai tervezésre, politikai irányvonalak meghatározására e mutatók alkalmatlanok (Mokyr, 2014), így a döntéshozók, politikacsinálók számára hasznos lehet a mérőszámok továbbfejlesztése.

2.2 Jogi és etikai kutatás

A számottevő intelligenciát és az autonómiát tartalmazó rendszerek fejlesztése olyan fontos jogi és etikai kérdésekhez vezet, melyekre adandó válaszok hatással vannak az MI megoldások fejlesztőire és fogyasztói oldalra egyaránt. Az ezt érintő kérdések kiterjednek a jogra, a közpolitikára, a szakmai és filozófiai etikára is, megválaszolásukhoz az informatikusok, a jogi szakértők, a politológusok és etikával foglalkozó szakértők tudására is szükség van. Például:

1. **Kötelezettségek és törvényi szabályozás az autonóm járművek tekintetében**: ha az önműködő autók felére csökkentenék az évi mintegy 40 000 halálos közlekedési balesetet az Egyesült Államokban, elképzelhető, hogy az autógyártók nem 20 000 köszönőlevelet, hanem 20 000 pert kapnának a nyakukba. Milyen jogi keretek között tudjuk leginkább kihasználni az autonóm járművekben (mint például a drónok vagy az önvezető autók) rejlő biztonsági lehetőségeket (Vladeck, 2014)? Az MI-vel kapcsolatos jogi kérdéseket a létező (szoftver- és internetközpontú) „kiberjog" alapján vagy attól függetlenül kellene kezelnünk (Calo, 2014a)? A katonai és a kereskedelmi alkalmazásoknál is a kormányoknak kell döntést hozniuk arról, hogyan lehet leginkább a kérdést a megfelelő szakértelemmel kezelni; példa lehet egy olyan szakmai és akadémiai fórum vagy közösség létrehozása, mint amilyennek a felállítását Calo javasolta (Szövetségi Robotikai Bizottság (Federal Robotics Commission), Calo, 2014b).

2. **Gépi etika**: hogyan döntsön egy autonóm jármű olyan események között, amikor az egyik kimenet egy alacsony valószínűségű emberi sérülés, a másik pedig egy majdnem teljes bizonyossággal bekövetkező komoly anyagi kár? Hogyan kellene az ügyvédeknek, az etikával foglalkozó szakembereknek és döntéshozóknak a nyilvánosság elé tárnia e problémákat? Vajon szükséges az ilyen döntési helyzeteket nemzeti szabályozásban rögzíteni?

3. **Autonóm fegyverek**: lehetséges-e az emberi jogokkal összeegyeztethető autonóm fegyvereket készíteni (Churchill–Ulfstein, 2000)? Ha – mint azt már néhány szervezet javasolta – az autonóm fegyvereket betiltanák (Docherty, 2012; *The Scientists' Call To Ban Autonomous Lethal Robots*, 2015), definiálható lenne-e az autonómia ebben a környezetben? És egy ilyen tilalom egyáltalán érvényesíthető-e a gyakorlatban? Ha megengedhető vagy legális az élet kioltására alkalmas autonóm fegyverek használata, hogyan kellene ezeket a fegyvereket integrálni egy már kialakult ellenőrzési struktúrába úgy, hogy a felelősség és a kötelezettségek megfelelően kijelölésre kerüljenek, milyen technikai realitásokat és előrejelzéseket kell figyelembe venni ezeknél a kérdéseknél, és hogyan határozható meg ezekkel a fegyverekkel kapcsolatban az „érdemi emberi ellenőrzés" (Roff, 2014; Roff, 2013; Anderson– Reisner–Waxman, 2014)? Az autonóm fegyverek csökkenthetik a politika ingerküszöbét a fegyveres konfliktusokkal kapcsolatban, netalán ezek a fegyverek „véletlen" háborúkat is kirobbanthatnak (Asaro, 2008)? Végül, hogyan lehet az átláthatóságot és az aktív közbeszédet biztosítani a témával kapcsolatban?

4. **Magánélet/Privacy**: hogyan viszonyul az MI-rendszerek képessége a térfigyelő kamerákból, telefonhívásokból és e-mailekből stb. származó adatok összegyűjtése és feldolgozása terén a magánélethez való joghoz? Hogyan hatnak az adatvédelmi kockázatok a kiberbiztonságra és kiberhadviselésre (Singer–Friedman, 2014)? A mesterséges intelligencia és nagy adattömeg (big data) közötti szinergiák kihasználásának sikere részben attól függ, hogy képesek vagyunk-e fenntartani és megóvni a magánélet biztonságát (Manyika, 2011; Agrawal–Srikant, 2000).

5. **Szakmai etika**: milyen szerepet kellene játszani a számítógépes szakembereknek az MI fejlesztésének és használatának jogi és etikai kérdéseiben? A tanulmányban többször említett szakmai testületek és kutatások kiemelten foglalkoznak ezzel a kérdéssel (Boden et al., 2011; Horvitz – Selman, 2011).

Közpolitikai megközelítésből az MI (mint bármely feltörekvő új technológia) nagyszerű új előnyöket és elkerülendő csapdákat jelent, ahol a megfelelő politika biztosíthatja, hogy úgy élvezzük az előnyöket, hogy közben a kockázatok a minimális szintre csökkennek. Ez olyan szakpolitikai kérdéseket vet fel, mint például:

1. Milyen politikai és stratégiai megközelítéseket érdemes számba venni?

2. Milyen kritériumokat alapján értékelhetjük a különböző politikákat? A lehetőségek között szerepel az ellenőrizhetőség, a végrehajthatóság, a kockázatok csökkentésének képessége, a megfelelő irányú technológiai fejlesztések akadályainak mérséklése, az alkalmazkodóképesség, valamint a változó körülményekhez való alkalmazkodás képessége.

2.3 Számítástechnikai kutatás a robusztus, megbízható mesterséges intelligenciáért

Ahogy az autonóm rendszerek előfordulása egyre gyakoribb a társadalomban, úgy válik egyre fontosabbá, hogy ezek megbízhatóan, az elvárásoknak megfelelően működjenek. Az önvezető járművek, az automatikus kereskedelmi rendszerek, az autonóm fegyverek és hasonló megoldások fejlesztésének esetében éppen ezért kiemelt figyelem övezi a magas biztosítású rendszereket, ahol a megbízhatóság terén erős garanciák fogalmazhatók meg; Weld és Etzioni szerint „*társadalom elutasítja az autonóm ágenseket, kivéve ha van néhány hiteles eszközünk azok biztonságossá tételére.*" (Weld–Etzioni, 1994). A robusztusság számos területen szenvedhet csorbát, melyek külön kutatási területeket nyújthatnak a megbízhatóság tekintetében:

1. **Ellenőrzés (verification)**: hogyan lehet bizonyítani, hogy a rendszer kielégíti-e a kívánt formális jellemzőket? *(„Jól építettem meg a rendszert?")*

2. **Érvényesség (validity)**: hogyan biztosítható, hogy a rendszer, amely megfelel a formális követelményeknek, nem viselkedik az eredeti szándéktól különbözően, és ez nem jár nem kívánt következményekkel? *(„A megfelelő rendszert építettem meg?")*

3. **Biztonság (security)**: hogy lehet megelőzni a rendszer illetéktelenek általi manipulálását?

4. **Irányítás (control)**: hogyan lehet egy MI rendszert érdemben emberi irányítás alatt tartani, miután az már működésbe lépett? *(„Oké, rosszul építettem meg a rendszert, helyre tudom hozni?")*

2.3.1 Ellenőrzés

Verifikálás alatt azokat a módszereket értjük, amelyek nagy valószínűséggel biztosítják, hogy a rendszer meg fog felelni az előre meghatározott formai követelményeknek. Amennyiben lehetséges, fontos biztosítani azt, hogy a biztonsági szempontból kritikus helyzetekben a rendszerek verifikálhatóak legyenek.

A szoftverek formális ellenőrzése jelentősen fejlődött az elmúlt évek során (pl. a seL4 kernel (Klein et al., 2009), vagy a HACMS (Fisher, 2012)): nem csak azt lenne szükséges lehetővé tenni, hogy az MI rendszerek a különböző, ellenőrzött alapokra épüljenek; azt is biztosítani kell, hogy az MI rendszerek terveit, felépítését önmagában is ellenőrizni lehessen, különösen, ha azok „*komponens architektúrát*" követnek, így az egyéni összetevőkre vonatkozó garanciák a komponensek kapcsolatai alapján kombinálhatók a teljes rendszer tulajdonságainak javítása érdekében (ez a megközelítés tükröződik például Russel és Norvig (2010) munkájában).

Talán a legszembetűnőbb különbség a hagyományos szoftverek és a mesterséges intelligencia-alapú rendszerek ellenőrizhetőségében az, hogy míg az hagyományos szoftvereket egy állandó és ismert gépi modell definiálja, addig az MI rendszerek (főleg a robotok és más testet öltött rendszerek) olyan környezetben működnek, amelyről a rendszer tervezőjének legjobb esetben is csak részleges a tudása. Ennek alapján az MI rendszerek esetében célszerű lehet azok működésének helyességét a rendszer tudásának fényében ellenőrizni, elkerülve a valós környezet modellezésének problémáit (Dennis, 2013). A tervezési idő ismeretének hiánya a tanulási algoritmusok használatára ösztökél az ágens szoftverén belül, amely által az ellenőrzés még nehezebbé válik: a statisztikai tanulás elmélete megad un. ϵ-δ (valószínűleg közelítőleg helyes) határokat, leggyakrabban az olyan, nem valós körülményekre, mint a felügyelt tanulás azonos eloszlású adatokból, és az egyedi ágensek megerősítésen alapuló tanulása egyszerű architektúra és teljes megfigyelhetőség esetén, de még így is túlságosan nagy mintanagyság szükséges a megfelelő garanciák eléréséhez.

A kutatási módszerek, amelyek lehetővé teszik, hogy erős megállapításokat tehessünk a gépi tanulási algoritmusokkal kapcsolatban és a számítási költségek menedzselésére különböző numerikus feladat esetén, javíthatják lehetőségeinket ezen a területen, akár kiterjesztve a munkát a Bayes-tételre is (Henning–Kiefel, 2013; Gunter, 2014). Az adaptív szabályozás elméletére (Åström–Wittenmark, 2013), az úgynevezett *kiberfizikai rendszerekre* (Platzer, 2010) vagy a hibrid és robotrendszerek ellenőrzésére (Alur, 2011; Winfield–Blum–Liu, 2014) irányuló munka szintén rendkívül releváns, de ugyanezekkel a nehézségekkel kell szembenéznie. És természetesen mindezek a kérdések a sztenderd problémára rakódnak, miszerint be kell bizonyítani, hogy az adott szoftver-kezdemény megfelelően implementálható például a megerősítő tanulás algoritmusának egy kívánt típusára. Zajlottak kutatások a neurális hálózati alkalmazások ellenőrzése terén (Paulina–Tacchella, 2010; Taylor, 2006; Schumann–Liu, 2010) és a *részprogram* (partial programs, Andre–Russel, 2002; Spears, 2006) fogalma lehetővé teszi a tervezőknek, hogy tetszőleges „strukturális" megszorításokat tegyenek a viselkedésre vonatkozóan, de még így is sok a tennivaló, mielőtt lehetővé válik magas megbízhatósági szinten az, hogy egy tanuló ágens a tervezési kritériumokat kielégítő módon lesz képes tanulni életszerű kontextusokban.

2.3.2 Érvényesség

Egy ágens tervezésének verifikációs tétele a következő formát mutatja: „*ha a környezet eleget tesz φ feltételezéseknek, akkor viselkedés eleget tesz γ (ψ) követelményeknek.*" Két módja van annak, hogy egy ellenőrzött ágens elhibázza a helyes cselekvést: első esetben a környezetére vonatkozó φ feltételezés hamis a való életben, és olyan reakcióhoz vezet, ami megsérti a γ követelményeit. Második esetben a rendszer megfelel a γ formai követelményeknek, miközben továbbra is úgy viselkedik, amit a gyakorlatban erősen nem kívánatosnak értékelünk. Előfordulhat, hogy ez a nem kívánt esemény annak a következménye, hogy teljesül γ, miközben φ sérül, azaz φ megvalósulása esetén a nem kívánt esemény nem történt volna meg; előfordulhat az is, hogy γ önmagában hibás. Russel és Norvig (2010) egy egyszerű példával él: ha arra kérünk egy robot tisztítógépet, hogy annyi szennyeződést tisztítson föl, amennyit csak lehet, és rendelkezik a porgyűjtőjének kiürítésének képességével is, akkor ez azt eredményezi, hogy a gép ugyanazt a piszkot fogja újra és újra feltakarítani. A követelményeknek tehát nem a szennyeződés eltűntetésére, hanem a padló megtisztítására kell vonatkoznia. Az ilyen specifikációs hibák minden olyan szoftver ellenőrzése során előfordulnak, ahol általában megfigyelhető, hogy a helyes specifikáció megírása nehezebb, mint a helyes kódé. Sajnos a specifikáció ellenőrzése nem lehetséges: a „hasznos" és a „elvárt" fogalmak külön nem formalizálhatók, így nem lehet egyértelműen bizonyítani, hogy γ-nek való megfelelés szükségszerűen a célszerű viselkedéshez és egy hasznos ágenshez vezet.

Annak érdekében, hogy megbízhatóan működő rendszereket építsünk, természetesen minden alkalmazási területen el kell döntenünk, mit is jelent a „jó működés". Ez az etikai kérdés szorosan kapcsolódik ahhoz, hogy milyen mérnöki módszerek állnak rendelkezésünkre, hogy mennyire megbízhatóak ezek a technikák, és milyen kompromisszumokat köthetünk – minden területen, ahol a számítástechnika, a gépi tanulás és a tágabb MI szakértelem értékes. Például Wallach és Allan (2008) szerint kiemelten figyelembe veendők a különböző viselkedési sztenderdek (vagy etikai elméletek) számítási költségei: ha egy szabványt nem lehet elég kielégítően használni ahhoz, hogy helyes viselkedést eredményezzen a biztonsági szempontból kritikus helyzetekben, akkor olcsóbb megközelítésekre lehet szükség. Egyszerűsített szabályok kialakításához – például egy önvezető autó viselkedésének szabályozása kritikus helyzetekben – minden bizonnyal informatikusok és etikában jártas szakemberek is szükségesek. Az etikus gondolkodást leíró számítástechnikai modellek fényt deríthetnek a számítási költségekre és a megbízható érvelési módszerek életképességére egyaránt (Asaro, 2006; Sullins, 2011); az ilyen irányú kutatás például alkalmas lehet további területek alkalmazásának feltárására: szemantikus hálózatok használata az esetalapú következtetésben (McLaren, 2006), a hierarchikus korlátkielégítés (MacWorth, 2009), vagy a súlyozott prospektív abdukció (Pereira–Saptawijaya, 2007) alkalmazása a gépi etikához.

2.3.3 Biztonság

A biztonsággal kapcsolatos kutatások segíthetnek az MI robusztusabbá tételében. Mivel az MI rendszerek egyre növekvő számban kritikus szerepben kerülnek felhasználásra, egyre nagyobb felületet nyújtanak a kibertámadások számára is. Valószínű, hogy az MI és a gépi tanulási technológiákat is fel fogják használni kibertámadásokra. A robusztusság alacsony szinten szoros kapcsolatban van az ellenőrizhetőséggel és a hibáktól való mentességgel. A DARPA SAFE programjának célja például egy olyan, rugalmas metaadat szabályokat tartalmazó, integrált szoftver-hardver rendszer építése, amire olyan memóriabiztonsági, hiba-elkülönítési és egyéb protokollok építhetők, melyek javíthatják a biztonságot a kihasználható sérülékenységek megelőzésével (DeHon et al., 2011). Az ilyen programok nem szűntethetik meg az összes biztonsági hiányosságot (mivel az ellenőrzés csak olyan erős lehet, mint a követelmény, ami alapján az ellenőrzés történik), de jelentős mértékben csökkenthetik az olyan sérülékenységek okozta károkat, mint a közelmúltban terjedő „Heartbleed bug" vagy „Bash Bug". Az ilyen rendszerek használatát lehetőség szerint támogatni kell a biztonsági szempontból kritikus alkalmazásoknál, ahol a magasabb szintű biztonság költségei igazolhatók.

Magasabb szinten az MI és a gépi tanulás területén zajló specifikus kutatások egyre hasznosabbak lehetnek a biztonság számára. Ezek a technológiák alkalmazhatók a különböző jogtalan behatolások észlelésére (Lane, 2000), a rosszindulatú szoftverek (malware) azonosítására (Rieck et al., 2011), vagy esetleges sérülékenységek feltárására egyéb programok forráskódjának elemzése során (Brun – Ernst, 2004). Nem elképzelhetetlen, hogy az államok és a privát entitások között zajló kibertámadások is alkalmaznak majd a közeljövőben MI megoldásokat, ezáltal olyan rizikófaktort jelentenek, ami további kutatásokat motivál a káresemények elkerülése érdekében. Ahogy az MI rendszerek egyre komplexebbé válnak, és egymással is hálózati kapcsolatba kerülnek, intelligens módon kell kezelniük a bizalom kérdéskörét, ami szintén olyan újabb kutatási területek felé nyitja meg az utat, mint a statisztikai-viselkedési alapokon nyugvó bizalomépítés (Probst – Kasera, 2007), vagy a számítástechnikai reputációs modellezés (Sabater – Sierra, 2005).

2.3.4 Irányítás

Bizonyos típusú, biztonsági szempontból kritikus MI rendszerek – különösen a járművek és a fegyverrendszerek – esetében kívánatos lehet az emberi ellenőrzés valamilyen formájának megőrzése, jelentsen ez akár közvetlen emberi visszacsatolást (in the loop, on the loop - Hexmoor et al., 2009; Parasuraman et al., 2000) vagy valamilyen más protokollt. Sok esetben további technikai fejlesztés szükséges annak biztosításához, hogy az érdemi humán kontroll fennmaradhasson (UNIDIR, 2014).

Az önvezető járművek jó lehetőséget nyújtanak a hatékony kontroll-mechanizmusokkal való kísérletezésre. Az automata navigáció és az emberi ellenőrzés közötti váltást megvalósító rendszerek és eljárások tervezése szintén egy ígéretes kutatási terület. Az ilyen és ehhez hasonló problémák megtermékenyítően hatnak egyéb kutatásokra is, mint például az optimális feladat-elosztás meghatározása ember és gép alkotta egységek számára, illetve olyan helyzetek felismerése, ahol az ellenőrzést mindenképpen át kell adni, biztosítva a hatékony emberi döntést a legfontosabb döntéshozatali kérdésekben.

3. HOSSZÚ TÁVÚ KUTATÁSI PRIORITÁSOK

Számos mesterséges intelligenciával foglalkozó kutató által gyakran tárgyalt hosszú-távú cél olyan rendszerek fejlesztése, amelyek képesek az emberhez hasonlóan tág határok között tanulni tapasztalataikból, és meghaladni az emberi teljesítményt a legtöbb kognitív feladat esetében - ezáltal jelentős befolyást gyakorolva a társadalomra. Ha az elhanyagolhatónál nagyobb esélye van annak, hogy az ez irányú törekvéseket siker koronázza a belátható jövőben, akkor a korábban bemutatotthoz képest újabb, a továbbiakban részletezett kutatási területek nyílnak, melyek célja biztosítani, hogy az MI robusztus és jótékony hatású maradjon a jövőben is.

A kutatók véleménye jelentős mértékben eltérhet egy ilyen rendszer létrehozhatóságának valószínűségéről, de elenyésző azok aránya, akik nagy biztonsággal kijelentenék, hogy ez a valószínűség elhanyagolható – különösen a korábbi, hasonló jóslatok fényében. (Ernest Rutherford, korának vitathatatlanul legnagyobb atomfizikusa 1933-ban a nukleáris energiát a „fantazmagória" kategóriájába sorolta (Associated Press, 1933), míg Richard Wolley, a Királyi csillagász (Royal Astronomer, a legmagasabb csillagászi pozíció az Egyesült Királyságban) 1956-ban a bolygóközi űrutazást nevezte „teljes sületlenségnek" (Reuters, 1956)). Mindezeken túl, ahhoz hogy igazolhatók legyenek ezen a területen az MI robusztusságának kutatására fordított összegek, ennek a valószínűségnek nem kell magasnak lennie, csupán nem elhanyagolhatónak – épp annyira, mint amennyire a lakásbiztosításra fordított összegeket indokolja az otthon leégésének kicsi, de nem elhanyagolható valószínűsége.

3.1 Ellenőrzés

Visszaidézve a rövid távú prioritásokat, a kutatások, melyek ellenőrizhető alacsony szintű szoftvert és hardvert tesznek elérhetővé a különböző programhibák és problémák számos csoportját megszűntethetik az általános MI-rendszerekben; ahogy a rendszerek egyre erősebbeké és biztonságuk egyre kritikusabbá válik, a verifikálható biztonsági tulajdonságok is egyre értékesebbek lesznek. Ha a verifikálható összetevők teljes rendszerekre történő kiterjeszthetőségének elmélete elfogadott, akkor akár nagyon nagy rendszerek is részesülhetnek bizonyos típusú biztonsági garanciákból, potenciálisan akár olyan technológiák segítségével is, amelyeket kifejezetten tanuló ágensek és magas szintű összetevők irányítására terveztek. Az elméleti kutatások - kiváltképp a mindenre alkalmas mesterséges intelligencia-rendszernek terén - különösen hasznosak lehetnek.

A verifikáláshoz kapcsolódó, a hosszú távú aggodalmak esetében különösen releváns kutatási téma az olyan rendszerek ellenőrzése, amelyek képesek módosítani, bővíteni vagy fejleszteni saját magukat, akár többször egymás után (Irving, 1965, Vinge, 1993). A formalizált verifikációs megoldások egy-az-egyben történő alkalmazásának kísérlete erre a jóval általánosabb keretre új nehézségeket jelent, beleértve azt a kihívást, mely szerint a kellően erőteljes formális rendszerek kézenfekvő módon nem használhatnak formális módszereket azért, hogy megbizonyosodjanak funkcionálisan hasonló formális rendszerek pontosságáról - hacsak el nem tekintünk Gödel nemteljességi tételétől... (Fallenstein – Soares, 2014; Wallach – Allen, 2008). Egyelőre nem teljesen világos, hogy hogyan lehet meghaladni ezt a problémát, illetve hogy egyéb hasonló problémák felbukkannak-e a hasonló erősségű verifikációs módszerek kapcsán.

Végezetül az is elmondható, hogy sokszor nehéz valóban alkalmazni a formális verifikációs technológiákat fizikai rendszerekre, különösen olyan rendszerekre, amelyeknél a tervezés során nem tartották szem előtt a verifikálás problémáját. Ez a tény ösztönzőleg hathat az olyan, általános elméletre törekvő kutatásokra, melyek összekapcsolják a funkcionális specifikációkat a fizikailag megvalósuló eseményekkel. Ez a fajta elméletalkotás lehetővé tenné a formális eszközök használatát, hogy megjósolható és kontrollálható legyen az olyan rendszerek viselkedése, amelyek közelítenek a racionális ágensekhez, vagy az olyan eltérő szerkezetek, mint a kielégítő ágensek, illetve az olyan rendszereké, amelyeket nem lehet könnyen leírni a hagyományos ágens-formanyelvvel (erőteljes predikciós rendszerek, tételbizonyítók, korlátozott célú tudományos vagy mérnöki rendszerek stb.). Egy ilyen elmélet akár annak bizonyítását is lehetővé tenné, hogy a rendszerek korlátozhatók bizonyos tevékenységek elvégzésében, vagy bizonyos fajta érvelés alkalmazásában.

3.2 Érvényesség

Csakúgy, mint a rövid távú kutatási prioritások esetében, az érvényesség azokkal a nem kívánt viselkedésformákkal foglalkozik, melyek egy rendszer formai helyessége ellenére fordulnak elő. Hosszú távon az MI rendszerek még erősebbé és még autonómabbá válhatnak, így a validitásproblémák is jóval költségesebbek lehetnek. A rövid távú prioritásoknál már említett gépi tanulási módszerek számára létrehozott erős garanciák a hosszú távú biztonság szempontjából szintén fontosak lesznek. Az ezen a területen végzett munka hosszú távú hasznosítása érdekében a gépi tanulással kapcsolatos kutatások egyik fókuszpontja a nem várt általánosítások típusainak vizsgálata lehet, amely a mindenre alkalmas mesterséges intelligencia-rendszerek számára okozhatja a legtöbb problémát. Különösen fontos lehet megérteni mind elméleti, mind gyakorlati szempontból, hogy a magasabb szintű emberi fogalmak tanult értelmezése hogyan (nem) általánosítható gyökeresen eltérő kontextusokban (Tegmark, 2015). Ha bizonyos fogalmak megtanulása megbízhatóan történik, akkor ez lehetővé teheti azok használatát feladatok és kikötések definiálására, minimálisra csökkentve a nem szándékolt következményeket, akkor is, ha az autonóm MI rendszer mindenre alkalmas mesterséges intelligencia-rendszerré válik. Ez a témakör eddig kevéssé kutatott, így mind az elméleti, mind pedig az empirikus kísérletek hasznosak lehetnek a területen.

Az olyan matematikai módszerek, mint a formális logika, a valószínűség- és a döntéselmélet igazán gyümölcsözőnek bizonyultak a döntéshozás és a gondolkodás alapjainak megismerése során, ám továbbra is számos a megoldatlan probléma. A megoldások ezekre a problémákra sokkal megbízhatóbbá és kiszámíthatóbbá tehetik a mindenre alkalmas rendszerek viselkedését. Példák ezen a területen: érvelés és döntések korlátozott számítási erőforrások esetén (Horvitz, 1987; Russel – Subramanian, 1995), hogyan vehető figyelembe a kapcsolat az MI rendszerek viselkedése és környezetük, illetve egyéb ágensek viselkedése között (Halpern – Pass, 2013; Hintze, 2014; LaVictoire et al., 2014; Soares – Fallenstein, 2014; Tennenholtz, 2004), hogyan kell gondolkodnia a környezetébe ágyazott ágenseknek (Orseau – Ring, 2012; Soares, 2014), illetve hogyan érveljenek az olyan bizonytalan tényezők logikai következményeivel kapcsolatban, mint a különböző hiedelmek vagy egyéb determinisztikus számítások (Soares – Fallenstein, 2014, Probabilistic Numerics, 2014). Ezeket a témákat szoros kapcsolatuk miatt hasznos lehet egyben kezelni (Halpern – Pass, 2011, Halpern et al., 2014).

Kézenfekvő, hogy hosszú távon az autonóm és erőteljes ágenseket az élet minél több területén szeretnénk alkalmazni. Egyértelműen lefektetni a preferenciáinkat széles doménekben a közeljövő gépi etikájának stílusában nem feltétlenül célszerű, mivel nehézzé teheti az erőteljes MI rendszerek értékeinek „összehangolását" saját értékeinkkel és preferenciáinkkal (Soares, 2014; Soares – Fallenstein, 2014). Vegyük például egy olyan hasznossági függvény megalkotásának a nehézségét, amely a teljes joganyagot felöleli; még a jog szó szerinti interpretálása is messze meghaladja a jelenlegi lehetőségeinket, és igen gyenge eredményekkel járna a gyakorlatban (mivel a törvények írottak, feltételezzük, hogy interpretálásuk és alkalmazásuk rugalmasan, eseti elbírálás alapján történik). A megerősítéses tanulásnak is megvannak a saját problémái: ahogy a rendszerek egyre inkább mindenre alkalmas mesterséges intelligencia-rendszerekké válnak, egyre nagyobb az esélye egy, Goodhart törvényhez hasonló hatás bekövetkezésének: a szofisztikált ágensek megpróbálják manipulálni, vagy közvetlenül irányítani azokat a mechanizmusokat, amelyek az eredményességüket jelzik (Bostrom, 2014). Ez olyan kutatási területekhez vezethet, amelyek javíthatják az olyan rendszerek konstruálását, amelyek futásidőben képesek a tanulásra vagy értékek elsajátítására. Az inverz megerősítéses tanulás például egy érvényes megközelítést nyújthat, amiben egy rendszer következtet egy másik aktor preferenciáira, aki feltételezhetően szintén a megerősítéses tanulást alkalmazza (Russel, 1998; Ng – Russell, 2000). Más megközelítések eltérő feltételezéseket használhatnak az aktor alapvető kognitív modelljeiről, akinek a preferenciái a tanulás tárgyát képezik (preferenciák tanulása, Chu – Ghahramani, 2005), vagy kifejezetten az ember etikai értékelsajátítási folyamata inspirálja. Ahogy a rendszerek fejlettebbé válnak, episztémikusan nehezebb módszerek is használatosak lehetnek, így az ezen a területen végzett kutatások is hasznosak, mint például a Bostom (2014) által példaként említett előzetes módszertani áttekintés az indirekt célmeghatározás lehetőségeiről.

3.3 Biztonság

Egyelőre nem egyértelmű, hogy az MI hosszú távú fejlődése a biztonsággal kapcsolatos problémák megoldását könnyebbé vagy nehezebbé teszi; egyrészt a rendszerek egyre komplexebbé válnak mind szerkezetüket, mind viselkedésüket tekintve és az MI segítségével végrehajtott kibertámadások döbbenetesen hatékonyak lehetnek, ám másrészt az MI és a gépi tanulási technológiák használatával – az alacsony szintű rendszerek megbízhatóságának jelentős fejlődésével együtt – jóval ellenállóbb rendszerek jöhetnek létre, melyek sokkal kevésbé sebezhetők, mint a maiak. Titkosítási szemszögből úgy tűnik, hogy ebben a konfliktusban az MI inkább a védekezőket részesíti előnyben a támadókkal szemben – ez lehet az egyik oka annak, hogy a védelemmel kapcsolatos kutatásokat teljes szívvel támogassuk.

Habár a 2.3.3. pontban felsorolt kutatási témák egyre fontosabbak lesznek hosszú távon, a mindenre alkalmas mesterséges intelligencia-rendszerek egyedi biztonsági problémákat vetnek fel. Különösen igaz ez akkor, ha az érvényesség és az ellenőrzés problémái nem kerülnek megoldásra. Ebben az esetben hasznos lehet olyan „konténerek" kialakítása az MI rendszerek számára, amelyekben kevésbé ellenőrzött környezetben fordulhatnak elő a nem kívánt viselkedésmódok és következmények (Yampolskiy, 2012). Ennek a kérdésnek mind az elméleti, mind pedig a gyakorlati vizsgálata indokolt. Ha az MI általános

szabályozása korlátozásokkal nehéznek bizonyul, akkor egy MI rendszer és egy konténer párhuzamos fejlesztése megoldás lehet, mivel lehetővé teszi a tervezés gyengeségeinek és erősségeinek felhasználását a korlátozó stratégia kialakítása során (Bostrom, 2014). Szintén jelentős segítséget jelentenének az eltérés-felismerő és automatizált tevékenység-ellenőrző rendszerek. Összességében indokoltnak tűnik, hogy ez az újabb perspektíva – a támadásokkal szembeni védekezés a „belső", rendszeren belüli problémáktól és a külső aktoroktól is – érdekes és jövedelmező területté váljon az informatikai biztonság területén.

3.4 Irányítás

Említésre került már, hogy a különböző feladatokat önállóan végző mindenre alkalmas mesterséges intelligencia-rendszerek gyakran olyan hatások alanyaivá válnak, amelyek jelentősen megnehezítik az érdemi emberi ellenőrzés fenntartását (Bostrom, 2012; Bostrom, 2014; Omohundro, 2007; Shanahan, 2015). Az olyan rendszerek kutatása, amelyek nem alanyai ezeknek a hatásoknak, vagy minimálisra csökkenti a hatások befolyását, vagy lehetővé teszi a megbízható emberi ellenőrzést hasznosnak bizonyulhatnak a nem kívánt következmények megelőzésében, valamint egyfajta biztonságos tesztkörnyezetként szolgálhatnak különböző fejlettségű MI rendszerek számára.

Ha egy MI rendszer választhatja ki a legjobb módszereket egy adott feladat megoldására, akkor egyértelmű cél az olyan körülmények elkerülése, amelyek akadályozzák a rendszert a feladat elvégzésére való törekvésben (Bostrom, 2012; Omohundro, 2007), és igaz ez fordítva is, a nem előre alkotott helyzetek is hasznosak lehetnek a megismerés szempontjából (Wissner-Gross – Freer, 2013)). Ez ugyanakkor problémákkal is járhat, ha újra szeretnénk tervezni a rendszert, deaktiválni, vagy jelentősen megváltoztatni a döntéshozatali folyamatot; egy ilyen rendszer racionálisan elkerülné ezeket a változtatásokat. Azokat a rendszereket, amelyek nem mutatják ezt a viselkedést, *javítható* (corrigible) rendszereknek nevezték el (Soares et al., 2015), és mind az elméleti és a gyakorlati kutatások jól tervezhetők és hasznosak ezen a területen. Lehetséges például hasznosságfüggvények vagy döntési folyamatok tervezése, amelyek használatakor a rendszer nem fogja megpróbálni elkerülni saját lekapcsolását vagy átprogramozását (Soares et al., 2015). Elméleti keretrendszer fejlesztése is elképzelhető annak jobb megértése érdekében, hogy milyen potenciális rendszerek képzelhetők el, amelyek elkerülik a nem kívánt viselkedésmódokat (Hibbard, 2012; Hibbard; 2014, Hibbard, 2015).

Egy másik alapvető cél lehet a különböző helyettesítő erőforrások összegyűjtése, mint amilyenek például a környezeti információk, a diszruptív, radikálisan újat hozó megoldásoktól való védelem, és a nagyobb szabadságú cselekvés mind számos feladat megoldásához hozzájárulhat (Bostrom, 2012; Omohundro, 2007). Hammond és munkatársai (1995) stabilizációnak hívja azokat a helyzeteket, amikor „*az ágens beavatkozásának köszönhetően a környezet jobban illeszkedik az ágenshez az idő múlásával*". Az ilyen célok nem várt következményekhez vezethetnek, egyúttal azon körülmények és hatásaik jobb megértéséhez és mérsékléséhez is, amelyek esetén az erőforrás-gyűjtés vagy a radikális stabilizáció az optimális stratégia (vagy valószínűleg a rendszer által leginkább preferált). A lehetséges kutatási témák ezeken túl olyan, alkalmazási területükben valamilyen módon limitált (Bostrom, 2014), „domesztikált" célokra is kiterjedhetnek, mint az időpreferencia mérté-

kének hatása az erőforrás-gyűjtési stratégiákra, vagy a hasonló kutatási célokat bemutató, egyszerű rendszerek gyakorlati vizsgálata.

Végezetül szót kell ejteni a szuperintelligens gépekkel, vagy a gyors, fenntartható önfejlesztésre képes rendszerekkel („intelligencia robbanás") kapcsolatos kutatásokról, melyekket számos múltbéli és jelenlegi, az MI jövőjével foglalkozó projekt érintett, és amelyek az irányítás hosszú távú megőrzése szempontjából is értékesek lehetnek. Ilyen például az AAAI 2008-09-es, az MI hosszú távú jövőjével foglalkozó elnöki paneljének „Sebesség, aggályok és ellenőrzés" munkacsoportja, mely a következő megállapítást tette:

> *„Általános szkepszis veszi körül egy lehetséges intelligencia-robbanás következményeit… Ettől függetlenül közös meggyőződésünk, hogy a további kutatások a területen olyan módszerekhez vezethetnek, amelyek segítenek megérteni és ellenőrizni a komplex számítógépes rendszerek számos viselkedését annak érdekében, hogy elkerüljük a nem várt következményeket. A panel néhány tagja szerint további kutatás szükséges az „intelligencia-robbanás" fogalmának pontosítására, valamint a hasonló, gyorsan változó intelligenciák osztályozására. A szakmai munka elvezethet az ilyen jelenségek valószínűségének jobb megértéséhez, éppúgy, mint különböző változataik természetének, a hozzájuk kapcsolódó veszélyeknek és általános következményeknek a leírásához.*" (Horvitz – Selman, 2009)

> A Stanford Egyetem „A Mesterséges Intelligencia Száz Éve"-kutatása szintén beszél az „*MI-rendszerek feletti irányítás elvesztéséről*", mint a kutatás egyik területéről, különösen annak a lehetőségnek, hogy *„…egy nap elveszíthetjük az irányítást az MI rendszerek felett az olyan szuperintelligencia fejlődése miatt, amely nem az emberi elvárásoknak megfelelően cselekszik – és ezáltal egy ilyen erős rendszer veszélyeztethetné az emberiséget. Vajon lehetségesek-e ilyen disztópikus forgatókönyvek? Ha igen, hogyan következhetnek be ilyen helyzetek? … Milyen kutatásokat kellene támogatni annak érdekében, hogy jobban megértsük és felkészüljünk egy veszélyes szuperintelligencia létrejöttére, vagy az „intelligencia-robbanás" bekövetkeztére?*" (Horvitz, 2014)

A kutatások ezen a területen magukba foglalhatják bármelyik, a korábbiakban felsorolt kutatási prioritást éppúgy, mint az elméleti és előrejelző munkákat az intelligencia-robbanás és a szuperintelligencia (Bostrom, 2014; Chalmers, 2010) kérdéskörében, és kiterjeszthetik, vagy kritikus elemzés alá vehetik a már létező kutatói műhelyek (pl. Machine Intelligence Research Institute (Soares – Fallenstein, 2014)) megközelítéseit.

4. ÖSSZEFOGLALÁS

A mesterséges intelligencia megteremtésére irányuló törekvések példa nélkül álló szolgálatot tehetnek az emberiségnek, és ezért érdemes kutatásokat végezni a lehetséges hasznok maximalizálása, valamint a lehetséges buktatók elkerülése érdekében. Ez a tanulmány (a teljesség igénye nélkül) számos példát hozott arra, milyen kutatások segíthetnek biztosítani azt, hogy a mesterséges intelligencia robusztus, hasznos és az emberi érdekekkel összehangolható legyen.

Irodalom

Agrawal, R., Srikant, R. (2000): Privacy-preserving data mining. In: ACM Sigmod Record 29.2 (2000), pp. 439-450.
Alur, R. (2011): Formal verification of hybrid systems. In: Embedded Software (EMSOFT) Proceedings of the International Conference on. IEEE. 2011, pp. 273-278.
Anderson, K., Reisner, D., Waxman, M. (2014): Adapting the Law of Armed Conict to Autonomous Weapon Systems. In: International Law Studies 90
Andre, D., Russell, S. (2002): State abstraction for programmable reinforcement learning agents. In: Eighteenth national conference on Artificial intelligence. American Association for Artificial Intelligence. pp. 119-125.
Asaro, P. (2006): What should we want from a robot ethic? In: International Review of Information Ethics 6.12 (2006), pp. 9-16.
Asaro, P. (2008): How just could a robot war be? In: Current issues in computing and philosophy pp. 50-64.
Associated Press (1933): Atom-Powered World Absurd, Scientists Told". In: New York Herald Tribune (1933). September 12, p. 1.
Äström, Karl J.; Wittenmark, B. (2013): Adaptive control. Courier Dover Publications
Boden, M. et al. (2011): Principles of robotics. In: The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC)
Bostrom, N. (2012): The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. In: Minds and Machines 22.2, pp. 71-85.
Bostrom, N. (2014): Superintelligence: Paths, dangers, strategies. Oxford University Press
Brun, Y.; Ernst, M. D. (2004): Finding latent code errors via machine learning over program executions. In: Proceedings of the 26th International Conference on Software Engineering. IEEE, Computer Society. 2004, pp. 480-490.
Brynjolfsson, E.; McAfee, A. (2014): The second machine age: work, progress, and prosperity in a time of brilliant technologies. W.W. Norton & Company, 2014.
Brynjolfsson, E.; McAfee, A.; Spance, M. (2014): Labor, Capital, and Ideas in the Power Law Economy. In: Foreign Aff. 93 (2014), p. 44.
Calo, R. (2014a): Robotics and the New Cyberlaw. In: Available at SSRN 2402972
Calo, R. (2014b): The Case for a Federal Robotics Commission. In: Available at SSRN 2529151
Chalmers, D. (2010): The singularity: A philosophical analysis. In: Journal of Consciousness Studies 17.9-10 (2010), pp. 7-65.
Chu, W.; Ghahramani, Z. (2005): Preference learning with Gaussian processes. In: Proceedings of the 22nd international conference on Machine learning. ACM. 2005, pp. 137-144.
Churchill, R. R.; Geir Ulfstein, G. (2000): Autonomous institutional arrangements in multilateral environmental agreements: a little-noticed phenomenon in international law. In: American Journal of International Law (2000), pp. 623-659.
Clark, A. E.; Oswald, A. J.(1994): Unhappiness and unemployment. In: The Economic Journal (1994), pp. 648-659.
DeHon, A. et al (2011): Preliminary design of the SAFE platform. In: Proceedings of the 6th Workshop on Programming Languages and Operating Systems. ACM. 2011, p. 4.
Dennis, L. A. et al. (2013: Practical Veri_cation of Decision-Making in Agent-Based Autonomous Systems. In: arXiv preprint arXiv:1310.2431 (2013).
Docherty, B. L. (2012): Losing Humanity: The Case Against Killer Robots. Human Rights Watch, 2012.
Fallenstein, B.; Soares, N (2012): Vingean Reection: Reliable Reasoning for Self-Modifying Agents. Tech. rep. Machine Intelligence Research Institute, 2014. url: https://intelligence.org/files/VingeanReflection.pdf.

Fisher, K. (2012): HACMS: high assurance cyber military systems. In: Proceedings of the 2012 ACM conference on high integrity language technology. ACM. 2012, pp. 51-52.
Frey, C.; Osborne, M. (2013): The future of employment: how susceptible are jobs to computerisation? Working Paper. Oxford Martin School, 2013.
Glaeser, E. L. (2014): Secular joblessness. In: Secular Stagnation: Facts, Causes and Cures (2014), p. 69.
Good, I. J. (1965): Speculations concerning the first ultraintelligent machine. In: Advances in computers 6.31 (1965), p. 88.
Gunter, T. et al. (2014): Sampling for inference in probabilistic models with fast Bayesian quadrature. In: Advances in Neural Information Processing Systems. 2014, pp. 2789-2797.
Halpern, J. Y.; Pass, R. (2011): I don't want to think about it now: Decision theory with costly computation. In: arXiv preprint arXiv:1106.2657 (2011).
Halpern, J. Y.; Pass, R. (2013): Game theory with translucent players. In: arXiv preprint arXiv:1308.3778 (2013).
Halpern, J. Y.; Pass, R;, Seeman, L. (2014): Decision Theory with Resource-Bounded Agents. In: Topics in cognitive science 6.2 (2014), pp. 245-257.
Hammond, K. J.; Converse, T. M.; Grass, J. W. (1995): The stabilization of environments. In: Artificial Intelligence 72.1 (1995), pp. 305-327.
Hennig, P.; Kiefel, M. (2013): Quasi-Newton methods: A new direction. In: The Journal of Machine Learning Research 14.1 (2013), pp. 843-865.
Hetschko, C.; Knabe, A.; Schöb, R. (2014): Changing identity: Retiring from unemployment. In: The Economic Journal 124.575 (2014), pp. 149-166.
Hexmoor, H.; McLaughlan, B.; Tuli, G. (2009): Natural human role in supervising complex control systems. In: Journal of Experimental & Theoretical Artificial Intelligence 21.1 (2009), pp. 59-77.
Hibbard, B. (2012): Avoiding unintended AI behaviors. In: Artificial General Intelligence. Springer, 2012, pp. 107-116.
Hibbard, B. (2014): Ethical Artificial Intelligence. 2014. url: http://arxiv.org/abs/1411.1373.
Hibbard, B. (2015): Self-Modeling Agents and Reward Generator Corruption. In: AAAI-15 Workshop on AI and Ethics. 2015.
Hintze, D. (2014): Problem Class Dominance in Predictive Dilemmas". Honors Thesis. Arizona State University, 2014.
Horvitz, E (2014): One-Hundred Year Study of Artificial Intelligence: Reections and Framing. White paper. Stanford University, 2014. url: https://stanford.app.box.com/s/266hrhww2l3gjoy9euar.
Horvitz, E. J. (1987): Reasoning about beliefs and actions under computational resource constraints. In: Third AAAI Workshop on Uncertainty in Artificial Intelligence. 1987, pp. 429-444.
Horvitz, E.; Selman, B. (2009): Interim Report from the Panel Chairs. AAAI Presidential Panel on Long Term AI Futures. 2009. url: https://www.aaai.org/Organization/Panel/panel-note.pdf.
Klein, G. et al. (2009): seL4: Formal verification of an OS kernel. In: Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles. ACM. 2009, pp. 207-220.
Lane, T. D. (2000): Machine learning techniques for the computer security domain of anomaly detection. PhD thesis. Purdue University, 2000.
LaVictoire, P. et al. (2014): Program Equilibrium in the Prisoner's Dilemma via Löb's Theorem. In: AAAI Multiagent Interaction without Prior Coordination workshop. 2014.
Mackworth, A. K. (2009): Agents, bodies, constraints, dynamics, and evolution. In: AI Magazine 30.1 (2009), p. 7.
Manyika, J. et al. (2011): Big data: The next frontier for innovation, competition, and productivity. Report. McKinsey Global Institute, 2011.
Manyika, J. et al. (2013): Disruptive technologies: Advances that will transform life, business, and the global economy. Vol. 180. McKinsey Global Institute, San Francisco, CA, 2013.
McLaren, B. M. (2006): Computational models of ethical reasoning: Challenges, initial steps, and future directions. In: Intelligent Systems, IEEE 21.4 (2006), pp. 29{37.

Mokyr, J. (2014): Secular stagnation? Not in your life. In: Secular Stagnation: Facts, Causes and Cures (2014), p. 83.
Ng, A. Y.; Russell, S. (2000):Algorithms for Inverse Reinforcement Learning. In: in Proc. 17th International Conf. on Machine Learning. Citeseer. 2000.
Nilsson, N. J. (1984): Artificial intelligence, employment, and income". In: AI Magazine 5.2 (1984), p. 5.
Omohundro, S. M. (2007): The nature of self-improving Artificial intelligence. Presented at Singularity Summit 2007.
Orseau, L.; Ring, M. (2012): Space-Time embedded intelligence". In: Artificial General Intel-ligence. Springer, 2012, pp. 209-218.
Parasuraman, R.; Sheridan, T. B.; Wickens, C. D. (200): A model for types and levels of human interaction with automation. In: Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 30.3 (2000), pp. 286-297.
Pereira, L. M.; Saptawijaya, A. (2007): Modelling morality with prospective logic. In: Progress in Artificial Intelligence. Springer, 2007, pp. 99-111.
Platzer, A (2010): Logical analysis of hybrid systems: proving theorems for complex dynamics. Springer Publishing Company, Incorporated, 2010.
Probabilistic Numerics (2014): http://probabilistic-numerics.org. Accessed: 27 November 2014.
Probst, M. J.; Kasera, S. K. (2007): Statistical trust establishment in wireless sensor networks. In: Parallel and Distributed Systems, 2007 International Conference on. Vol. 2. IEEE. 2007, pp. 1-8.
Pulina, L.; Tacchella, A. (2010): An abstraction-re_nement approach to veri_cation of Artificial neural networks". In: Computer Aided Veri_cation. Springer. 2010, pp. 243-257.
Reuters (1956): Space Travel `Utter Bilge'. In: The Ottawa Citizen (1956). January 3, p. 1. url: http://news.google.com/newspapers?id=ddgxAAAAIBAJ&sjid=1eMFAAAAIBAJ&pg=3254%2C7126.
Rieck, K. et al. (2011): \Automatic analysis of malware behavior using machine learning". In: Journal of Computer Security 19.4 (2011), pp. 639-668.
Roff, H. M. (2013): Responsibility, liability, and lethal autonomous robots". In: Routledge Handbook of Ethics and War: Just War Theory in the 21st Century (2013), p. 352.
Roff, H. M. (2014): The Strategic Robot Problem: Lethal Autonomous Weapons in War. In: Journal of Military Ethics 13.3 (2014).
Russell, S. (1998): Learning agents for uncertain environments. In: Proceedings of the eleventh annual conference on Computational learning theory. ACM. 1998, pp. 101-103.
Russell, S. J. (1995); Subramanian, D.: Provably bounded-optimal agents. In: Journal of Artificial Intelligence Research (1995), pp. 1-36.
Russell, S.; Norvig P. (2010): Artificial Intelligence: A Modern Approach. 3rd. Pearson, 2010.
Sabater, J.; Sierra, C. (2005): Review on computational trust and reputation models. In: Artificial intelligence review 24.1 (2005), pp. 33-60.
Schumann, J. M.; Liu, Y. (2010): Applications of neural networks in high assurance systems. Springer, 2010.
Shanahan, M. (2015): The Technological Singularity. Forthcoming. MIT Press, 2015.
Singer, P. W.; Friedman, A. (2014): Cybersecurity: What Everyone Needs to Know. Oxford University Press, 2014.
Soares, N. (2014): Formalizing Two Problems of Realistic World-Models. Tech. rep. Machine Intelligence Research Institute, 2014. url: https://intelligence.org/files/RealisticWorldModels.pdf.
Soares, N. (2014): The Value Learning Problem. Tech. rep. Machine Intelligence Research Institute, 2014. url: https://intelligence.org/files/ValueLearningProblem.pdf.
Soares, N. et al. (2015): Corrigibility. In: AAAI-15 Workshop on AI and Ethics. 2015. url: http://intelligence.org/files/Corrigibility.pdf. 11.

Soares, N.; Fallenstein, B. (2014): Aligning Superintelligence with Human Interests: A Technical Research Agenda. Tech. rep. Machine Intelligence Research Institute, 2014. url: http://intelligence.org/files/TechnicalAgenda.pdf.

Soares, N.; Fallenstein, B. (2014): Questions of Reasoning Under Logical Uncertainty. Tech. rep. url: http://intelligence.org/files/QuestionsLogicalUncertainty.pdf. Machine Intelligence Research Institute, 2014.

Soares, N.; Fallenstein, B. (2014): Toward Idealized Decision Theory. Tech. rep. url: https://intelligence.org/files/TowardIdealizedDecisionTheory.pdf. Machine Intelligence Research Institute, 2014.

Spears, D. F. (2006): Assuring the behavior of adaptive agents". In: Agent technology from a formal perspective. Springer, 2006, pp. 227-257.

Sullins, J. P. (2011): Introduction: Open questions in roboethics". In: Philosophy & Technology 24.3 (2011), pp. 233-238.

Taylor, B. J. (2006): Methods and Procedures for the Veri_cation and Validation of Artificial Neural Networks. Springer, 2006.

Tegmark, M. (2015): Friendly Artificial Intelligence: the Physics Challenge". In: AAAI-15 Workshop on AI and Ethics. 2015. url: http://arxiv.org/pdf/1409.0813.pdf.

Tennenholtz, M (2004): Program equilibrium. In: Games and Economic Behavior 49.2 (2004), pp. 363-373.

The Scientists' Call To Ban Autonomous Lethal Robots (2015). International Committee for Robot Arms Control. Accessed January 2015. url: http://icrac.net/call/.

United Nations Institute for Disarmament Research (2014): The Weaponization of Increasingly Autonomous Technologies: Implications for Security and Arms Control. UNIDIR, 2014.

Van Parijs, P. et al. (1992): Arguing for Basic Income. Ethical foundations for a radical reform. Verso, 1992.

Vinge, V. (1993): The coming technological singularity. In: VISION-21 Symposium, NASA Lewis Research Center and the Ohio Aerospace Institute. NASA CP-10129. http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html

Vladeck, D. C. (2014): Machines without Principals: Liability Rules and Artificial Intelligence. In: Wash. L. Rev. 89 (2014), p. 117.

Wallach, W.; Allen, C. (2008): Moral machines: Teaching robots right from wrong. Oxford University Press, 2008.

Weaver, N.: Paradoxes of rational agency and formal systems that verify their own soundness. Preprint. url: http://arxiv.org/pdf/1312.3626.pdf.

Weld, D.; Etzioni, O. (1994): The first law of robotics (a call to arms). In: AAAI. Vol. 94. 1994, pp. 1042-1047.

Widerquist, K., et al. (2013):Basic income: an anthology of contemporary research. Wiley-Blackwell

Winfield, A. FT.; Blum, C.; Liu, W. (2014): Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In: Advances in Autonomous Robotics Systems. Springer, 2014, pp. 85-96.

Wissner-Gross, A.D., Freer C.E. (2013): Causal entropic forces. In: Physical review letters 110.16: 168702.

Yampolskiy, R. (2012): Leakproofing the Singularity: Artificial Intelligence Confinement Problem. In: Journal of Consciousness Studies 19.1-2, pp. 1-2.