Első lépések a robotok és a mesterséges intelligencia etikája felé

JOHN TASIOULAS

King's College London

ABSZTRAKT

Ez a cikk áttekintést nyújt a robotok és mesterséges intelligenciák által felvetett főbb elsőfajú etikai kérdésekről, öt nagy témakörben: funkcionalitás, eredendő jelentőség, jogok és felelősségek, mellékhatások és veszélyek. Az egyes rubrikák első betűi együttesen a FIRST rövidítést alkotják. Különös figyelmet szentelünk a funkcionalitás és az eredendő jelentőség rubrikáinak, mivel az előbbi központi szerepet játszik, az utóbbit pedig hajlamosak vagyunk elhanyagolni, mivel kissé ködös és vitatott jellegű. Az egyes rubrikák keretében felmerülő néhány illusztris kérdés feltárása mellett a cikk számos általánosabb témát is kiemel. Ezek a következők: az egymással kölcsönhatásban álló szintek sokasága, amelyeken az RRAI-kkal kapcsolatos etikai kérdések felmerülnek, annak felismerésének szükségessége, hogy az RRAI-k potenciálisan az emberi értékek teljes skáláját érintik (és nem kizárólag vagy elsősorban az etikai vagy jogi elvek valamely könnyen azonosítható részhalmazát), valamint annak szükségessége, hogy az RRAI-kkal kapcsolatos, gyakorlatilag kiemelkedő etikai megfontolásoknak a meglévő és előrelátható kapacitásaik reális értékelésén kell alapulniuk.

1. BEVEZETÉS

A robotok az emberi történelem szinte teljes ideje alatt csak mint Jekyll és Hyde-jelleggel felruházott képzeletbeli lények léteztek. Egyik alakjukban a betegségektől, szegénységtől és a munka fáradalmaitól mentes utópiát ígérnek, a másikban pedig az emberiség leigázására vagy elpusztítására törekszenek. A robotok azonban csak a múlt század közepén érték el, hogy

jelentős gyakorlati jelenlétet, amikor a General Motors az egyik üzemében "Unimate" nevű robotot telepített, hogy olyan kézi feladatokat - például hegesztést és permetezést - végezzen, amelyeket az emberi munkások számára túl veszélyesnek ítéltek.[1] Napjainkban a robotok olyannyira elterjedtek a gyártásban, hogy az ágazatban a munkanélküliség egyik fő okozói.[2] A robotok gyári alkalmazása azonban csak a kezdete a "robotforradalomnak" - amely maga is része a mesterséges intelligencia (AI) tudományának köszönhető szélesebb körű fejlődésnek -, amely életünk minden területére átalakító hatással volt, vagy annak ígérkezik.

A robotokat ma már a legkülönbözőbb területeken használják, vagy fejlesztik ki használatra. A vezető nélküli autókat már feltalálták, és várhatóan egy évtizeden belül megjelennek az utakon. Ezek az autók képesek akár 90%-kal csökkenteni a közlekedési balesetek számát, amelyek jelenleg világszerte évente több mint egymillió emberéletet követelnek, miközben a környezetszennyezés és a forgalmi torlódások is csökkennek (Bonnefon, Shariff, Rhawan 2006). A robotokat háztartási munkák elvégzésére is használják, beleértve a porszívózást, vasalást és a háziállatok sétáltatását. Az orvostudományban és a szociális ellátásban a robotok felülmúlják az orvosokat a rák bizonyos formáinak diagnosztizálásában vagy a műtétek elvégzésében, és alkalmazzák őket az autista gyermekek terápiájában vagy az idősek gondozásában. Már léteznek oktató robotok, valamint társas robotok, amelyek társaságot vagy akár szexet nyújtanak. Az üzleti világban a mesterséges intelligencia nagy szerepet játszik a tőzsdén, ahol a számítógépek a legtöbb döntést automatikusan hozzák meg, valamint a biztosítási és jelzáloghitel-iparban. Még az emberi munkaerő toborzása is nagyrészt automatizált folyamattá válik, sok elutasított álláspályázatot soha nem vizsgál meg emberi szem. A mesterséges intelligencián alapuló technológia, részben robotizált, a büntető igazságszolgáltatási rendszerben is szerepet játszik, segítve a rendőri munkát, valamint az óvadékkal, büntetésekkel és feltételes szabadlábra helyezéssel kapcsolatos döntéseket. A katonai célpontokat emberi beavatkozás nélkül kiválasztó és megtámadó autonóm fegyverrendszerek (AWS) kifejlesztése új korszakot ígér a katonai védelemben. És ez csak egy kis ízelítő a közelmúltbeli fejlesztésekből.

Ebben a cikkben a robotok és a mesterséges intelligencia (vagy ahogyan én nevezem őket: RAI-k) által felvetett néhány kulcsfontosságú etikai kérdést vizsgálok meg. Az általános kihívás természetesen az, hogy kiaknázzuk az RAI-k előnyeit, miközben megfelelően reagálunk az ezzel járó kockázatokra. Az előnyök és a kockázatok egyensúlyának szükségessége visszatérő kérdés a technológiai fejlődés történetében, de az RAI-k új és potenciálisan átfogó formában jelennek meg, ami nagyszabású következményekkel jár arra nézve, hogyan élünk mások között - a munka, a gondozás, a oktatás, a játék, a barátság, a szerelem terén -, és még arra nézve is, hogyan értjük meg, mit jelent az, hogy "embernek lenni".

1. http://my.ilstu.edu/~kldevin/Introduction_to_robotics2/Introduction_to_robotics6.html *eti*

2. 2012-ben például a gyártóiparban egy dolgozóra körülbelül 1563 robot jutott *kai jatekai 51*
(Németországban ez az adat körülbelül egy dolgozóra1,133 jutott), lásd Furman és Seamans (2018, 8).

az emberi lényt, és hogy ezeket az új technológiákat az "emberi fejlődés", vagy akár - mint a "transzhumanizmus" esetében - az emberi állapot meghaladása érdekében kell-e alkalmaznunk. Mielőtt ezekkel a kérdésekkel foglalkoznánk, először is tisztáznunk kell néhány kulcsfogalmat. [3]

2. MI AZ A ROBOT? MI A MESTERSÉGES INTELLIGENCIA?

Egy nemrégiben készült UNESCO-jelentés a robotokat négy jellemzővel rendelkező mesterséges lényekként írja le:

- *mobilitás, ami* fontos az emberi környezetben, például kórházakban és irodákban való működéshez;
- *interaktivitás, amelyet* érzékelők és működtetők tesznek lehetővé, amelyek releváns információkat gyűjtenek a környezetből, és lehetővé teszik a robot számára, hogy a környezetben cselekedjen;
- számítógépes interfészek vagy hangfelismerő és beszédszintetizáló rendszerek által lehetővé tett *kommunikáció*; és
- *autonómia*, abban az értelemben, hogy képesek saját maguk "gondolkodni" és saját döntéseiket meghozni a környezetükkel kapcsolatos cselekvés érdekében, közvetlen külső ellenőrzés nélkül (UNESCO 2017: 4).

Az ebben a cikkben tárgyalt kifinomult robotok a "mesterséges intelligencia" (AI) alapján működnek. Marvin Minsky, a mesterséges intelligencia úttörője szerint ez "annak tudománya, hogy a gépeket olyan dolgokra késztessük, amelyekhez intelligencia kellene, ha ember végezné őket, például arcfelismerés vagy nyelvi fordítás". A mesterséges intelligencia megértésében két megkülönböztetés fontos: a) általános és szűk értelemben vett mesterséges intelligencia, valamint b) felülről lefelé irányuló és alulról felfelé irányuló mesterséges intelligencia. Az első megkülönböztetés a mesterséges intelligencia képességeinek körére vonatkozik, a másik pedig a mesterséges intelligencia technikai működésére.

Az általános mesterséges intelligencia olyan intelligens gépekre utal, amelyek képesek az emberi szellemi képességek széles skáláját leutánozni, sőt, akár felül is múlni azokat. A mesterséges intelligenciának ezek a formái - bár a sci-fi szereplőiből, például a *Csillagok háborúja* C3PO-jából ismerősek - a legjobb esetben is csak a távoli jövőben léteznek. Amennyiben az elmúlt években jelentős előrelépés történt a mesterséges intelligencia terén, az a *szűk* értelemben *vett mesterséges intelligencia* területén történt. Ezek olyan gépek, amelyek az emberi képességeket reprodukálják vagy meghaladják egy korlátozott számú feladat tekintetében, például az autóvezetés, az orvosi diagnosztika vagy a nyelvi fordítás terén.

3. A robotika és a mesterséges intelligencia fejlődéséről és az általuk felvetett etikai kérdésekről

hasznos áttekintést nyújt Ed- monds (2017) és Tegmark (2017, különösen a 3. fejezet).

A mesterséges intelligencia algoritmusok segítségével működik, amelyek a különböző problémák megoldására vonatkozó szabályok vagy utasítások, amelyek általában egy számítógépbe vannak beépítve, és amelyeket a jelen célokra nagyjából két nagy csoportba lehet sorolni, amelyek kétféle robotnak felelnek meg. A *felülről lefelé irányuló* (vagy determinisztikus vagy zárt szabályokkal rendelkező) algoritmusok egy előre meghatározott program segítségével irányítják a robot viselkedését, aminek eredményeként a robot viselkedése nagymértékben kiszámítható. Ilyen algoritmusokat alkalmaztak a jövedelemadó-nyomtatványok elkészítésében és bizonyos típusú automatizált orvosi diagnózisokban. Az *alulról felfelé irányuló* (vagy sztochasztikus) algoritmusok ezzel szemben lehetővé teszik, hogy a robot "tanuljon" a múltbeli tapasztalatokból, és idővel felülvizsgálja algoritmusát (UNESCO 17-19)2017,4,. Ennek a "gépi tanulásnak" a példája a Google DeepMind algoritmusa, amely megtanulta magát játszani az Atari-játékokkal, például a Breakouttal, és olyan új, pontszámmaximalizáló stratégiákat talált ki, amelyek saját programozóit is meglepték. További példák találhatók a vezető nélküli autókban, a rendőrség által használt adatelemző rendszerekben és a vásárlási előzmények alapján vásárlási javaslatokat adó algoritmusokban. A "gépi tanulásnak" különböző fajtái vannak. Egyesek "neurális hálózatokat" alkalmaznak, amelyek az emberi agy működését mintázó, egymáshoz rétegesen kapcsolódó feldolgozó csomópontok. Az ilyen robotok nemcsak abban az értelemben élvezik az "autonómia" szintjét, hogy viselkedésük nem függ az emberi döntéshozataltól, vagy nem lehet emberi beavatkozás vagy ellenőrzés tárgya, hanem abban a radikálisabb értelemben is, hogy az ember nem tudja könnyen megjósolni.

Természetesen óvatosan kell bánnunk az olyan kifejezésekkel, mint "intelligencia", "érvelés", "döntés" és "autonómia", amikor a mesterséges intelligenciával kapcsolatban beszélünk. Ezek a kifejezések nem fedhetik el azt a tényt, hogy az AI-kat és az embereket még mindig hatalmas szakadék választja el egymástól. A mesterséges intelligencia rendszerek az információt úgy dolgozzák fel, hogy felismerik a szimbólumok közötti mintákat és kapcsolatokat, amelyek lehetővé teszik bizonyos problémák megoldását. De (még) nem képesek értelmes értelemben *megérteni*, hogy ezek a szimbólumok mit jelentenek a valós világban (Tegmark 3. 2017,fejezet). Sőt, még ha az RAI-k sikeresen el is érhetnek összetett célokat - például felismernek egy arcot a tömegben, vagy lefordítanak egy dokumentumot egy természetes nyelvről egy másikra -, akkor is hiányzik belőlük az emberi képesség, hogy elgondolkodjanak azon, hogy mi legyen a végső céljuk. Egyes filozófusok szerint a racionális autonómia ezen képessége a forrása annak a különleges méltóságnak, amely az embereket jellemzi, és megkülönbözteti őket a nem emberi állatoktól. Egyetlen általunk ismert vagy reálisan előrelátható RAI sem

rendelkezik közel sem ilyen racionális autonómiával. [4]

4. A világ egyik vezető informatikusának szkepticizmusát a mesterséges intelligencia körüli felhajtással kapcsolatban lásd Jordan (2018).

3. ETIKAI KÉRDÉSEK: KERETEK ÉS SZINTEK

A RAI-k számos etikai kérdést vetnek fel, amelyek legalább három, egymással összefüggő szinten merülnek fel. Az egyik szint a RAI-val kapcsolatos tevékenységek szabályozására meghozandó *jogszabályokra* vonatkozik. Ezek a törvények olyan közjogi normák, amelyek formális elfogadásuk révén minden polgárra nézve erkölcsileg kötelező érvényűek, és amelyeket általában intézményes végrehajtási mechanizmusok támogatnak, beleértve - a legvégső esetben - olyan büntetéseket, mint a pénzbírság és a szabadságvesztés. A kérdések egyik csoportja itt arra vonatkozik, hogy a RAI-k bizonyos aspektusait egyáltalán jogi szabályozás alá kell-e vonni; a kérdések másik csoportja pedig arra vonatkozik, hogy milyen mértékben kell konkrét törvényeket alkotnunk a RAI-k által felvetett problémák kezelésére, szemben az általánosabb jogi normákra való támaszkodással. Szükségünk van-e külön közlekedési törvényekre a vezető nélküli autókra? Hogyan kell rájuk alkalmazni a biztosítási és baleseti felelősségre vonatkozó jogszabályokat? Legyenek-e büntetőjogszabályok, amelyek tiltanak bizonyos robotokat vagy mesterséges intelligencia alkalmazásokat? Az ilyen kérdésekre vonatkozó hazai jogszabályok mellett a RAI-k olyan sürgető kérdéseket is felvetnek, amelyek regionális vagy nemzetközi jogi megoldásokat igényelnek, például az AWS-ek betiltására vagy a velük kapcsolatos fegyverkezési verseny kitörésének megakadályozására irányuló szerződések révén.

Második szinten arról van szó, hogy milyen *társadalmi erkölcsöt* kell kialakítanunk a RAI-kkal kapcsolatban. Ez annak a ténynek a felismerése, hogy nem minden olyan társadalmilag rögzült norma, amely megfelelően szabályozza az életünket, jogi norma, vagy annak kellene lennie. Nemcsak a törvényre támaszkodunk, hogy az embereket visszatartsuk a helytelen viselkedéstől, például a gyilkosságtól vagy a lopástól, hanem olyan erkölcsi normákra is, amelyeket gyermekkorunktól kezdve belénk nevelnek, és amelyeket a társadalom olyan informális mechanizmusok révén erősít meg, mint a kritika és más jogon kívüli szankciók. Vitatható, hogy maga a jogi szabályozás hatékonysága jelentősen csökkenne, ha nem támaszkodhatna a mögöttes etikai kultúra fenntartására. Ennek megfelelően el kell gondolkodnunk az erkölcsileg egészséges kultúra formájáról a RAI-k vonatkozásában.

Harmadik szinten az *egyének és a szövetségek* (pl. vállalkozások, egyetemek, szakmai testületek stb.) számára merülnek fel kérdések a RAI-kkal való kapcsolatukkal kapcsolatban. Bármilyen társadalmi szabályozási módok léteznek is ezekben a kérdésekben, az egyéneknek és egyesületeknek továbbra is saját erkölcsi ítélőképességükkel kell rendelkezniük. Ennek oka lehet az, hogy a

hatályos jog és a társadalmi erkölcs elmarad a technikai fejlődés mögött, vagy az, hogy valamilyen módon hiányos, vagy az, hogy bizonyos kérdésekben az egyéneknek saját döntési szabadságot biztosítanak. Azoknak a vállalatoknak, amelyek a fejlesztések élvonalában vannak, a gyorsan változó és átalakuló jellemzé-

a RAI-k szereplői indokolhatják saját etikai kódexük kidolgozását ezekben a témákban. Eközben mások "hippokratészi esküt" kértek az adattudósok számára, hogy az alkalmazandó jogi normáktól függetlenül etikai keretet hozzanak létre működésükre (Upchurch 2018).

Nehéz kérdések merülnek fel azzal kapcsolatban, hogy miként lehet a legjobban integrálni a RAI-k e három szabályozási módját, és komoly aggodalomra ad okot, hogy az iparági etikai kódexek hajlamosak lesznek háttérbe szorítani a demokratikusan elfogadott törvényeket ezen a területen, különös tekintettel arra, hogy a RAI-kkal kapcsolatos fejlesztéseket irányító kisszámú technológiai vállalat jelentős politikai befolyással rendelkezik. Ez a befolyás azonban azzal a mindig jelenlévő veszéllyel jár, hogy a nagyhatalmú vállalatok képesek lesznek úgy alakítani az ebből eredő törvényeket, hogy azok a közjó helyett inkább az érdekeiknek kedvezzenek (Nemitz 7)2018,. A nehézség részben abból a tényből fakad, hogy az etikai szabályozás három szintje bonyolult módon kapcsolódik egymáshoz. Lehetséges például, hogy erős erkölcsi okok szólnak az ellen, hogy felnőttek robotot hozzanak létre vagy használjanak szexuális partnerként (harmadik szint). De az egyéni autonómia tiszteletben tartása miatt jogilag szabadnak kell lenniük, hogy ezt megtehessék (első szint). Ugyanakkor jó okok szólhatnak egy olyan társadalmi erkölcsiség ápolása mellett is, amely általánosságban elítéli az ilyen tevékenységeket (második szint), így a szexrobotok eladását és nyilvános bemutatását különböző módon (területrendezési törvények, adózás, kor- és reklámkorlátozások stb. révén) jogilag korlátozzák, hasonlóan a cigarettára vagy a szerencsejátékra vonatkozó jogi korlátozásokhoz (ismét első szint). Tekintettel erre az összetettségre, nincs eleve biztosíték arra, hogy a szabályozás három szintjének integrálására egyetlen legjobb módszer létezik, bár az első és a második szinten mindenképpen szükséges lesz bizonyos egyetemes szabványok felé közelíteni, amennyiben a tárgyalt kérdés egységes megoldást igényel a különböző nemzeti joghatósági határok között.

Ezt az összetettséget tovább fokozza az a tény, hogy a mesterséges intelligencia és a robotika területe gyorsan változik, és jelentős hype-ok középpontjában áll, ami megnehezíti a reális jövőbeli forgatókönyvek és a puszta tudományos fantázia szétválasztását. Ennek fényében etikai gondolkodásunknak mindhárom szinten érzékenyen kell reagálnia a kérdéses időkeretre, néha a közvetlen aggodalomra okot adó kérdésekkel foglalkozva, máskor pedig a jövőbeli fejlemények számítva. Állandó veszélyt jelent, hogy olyan lehetséges fejlemények vonják el a figyelmünket, amelyek a legjobb esetben is csak a nagyon távoli jövőben fognak bekövetkezni, miközben elhanyagoljuk az itt és most fennálló sürgető problémákat. A következőkben arra teszünk kísérletet, hogy a hangsúlyt az itt és mostra, valamint a reális jövőbeli forgatókönyvekre helyezzük, bár elkerülhetetlenül szóba kerülnek a

spekulatívabb forgatókönyvek is.

4. ÖT NAGY ERKÖLCSI KÉRDÉS - EGY F*I*R*S*T ELEMZÉS

A RAI-k által felvetett erkölcsi kérdések közül sok, ha nem is az összes, öt fő címszó - funkcionalitás, eredendő jelentőség, jogok és felelősségek, mellékhatások és veszélyek - alá sorolható, és az egyes rubrikák első betűje kényelmesen létrehozza a "FIRST" rövidítést. Természetesen az öt különböző címszó közötti határok nem mindig élesek, és bár a RAI-kra általában összefoglalóan, csoportként fogok hivatkozni, a különböző típusú RAI-k mind az öt címszó alatt jelentősen eltérő aggályokat vetnek fel. A rövidítés azért helyénvaló, mert az alább tárgyalt kérdések a RAI-kkal való foglalkozásunk jogairól és helytelenségeiről szóló elsőrendű kérdések. Ezen túlmenően fontos másodrendű kérdések is felmerülnek, például az átláthatóság vagy a demokratikus elszámoltathatóság normáival kapcsolatos eljárásokkal kapcsolatban, amelyeket az elsőrendű kérdések kezelése során el kell fogadnunk. Ezek a másodrendű kérdések azonban nagyrészt túlmutatnak e cikk keretein. A következőkben elsősorban a funkcionalitásra és az in- herens jelentőségre összpontosítok, és csak nagyon tömörítve foglalkozom a másik három címszóval.

4.1 FUNKCIÓSÁG

Az első kérdés az, hogy egy javasolt RAI, például egy vezető nélküli autó, működőképes-e. A "funkcionalitást" itt tág értelemben értem, amely nem semleges az RAI által kitűzött célok erkölcsi minőségét vagy az ezek eléréséhez alkalmazott eszközöket illetően. A funkcionalitás az RAI azon képességére vonatkozik, hogy: (a) elérjen egy *értékelhető* célt, például az utasok elszállítását a kívánt célállomásra, és ezt meg is tegye: (b) *hatékonyan*, azaz megbízható sikerességgel, c) *hatékonyan*, azaz az erőforrások indokolatlan ráfordítása nélkül, és
(d) *erkölcsileg megfelelő módon, azaz* anélkül, hogy működésének szerves részeként megsértené az erkölcsi normákat, függetlenül a tervező szándékától, pl. az élethez vagy a magánélethez való jogokat vagy a környezetvédelmi normákat. Bár mindezek a dimenziók fontos kérdéseket vetnek fel, koncentráljunk az utóbbira, amely két nagy kérdést vet fel: (1) melyek azok az erkölcsi normák, amelyek az RAI-kra vonatkoznak, és (2) hogyan építhetők be az RAI-k működésébe?

Az első kérdés megválaszolására tett híres kísérlet Isaac Asimov "A robotika három törvénye" című műve:

1. A robot nem okozhat kárt az embernek, és nem engedheti meg, hogy az ember tétlenségével kárt okozzon.

kája felé 1

2. A robotnak engedelmeskednie kell az ember által adott utasításoknak, kivéve, ha ezek az utasítások ellentétesek az első törvénnyel.
3. A robotnak meg kell védenie saját létezését, amíg ez a védelem nem ütközik az első vagy a második törvénnyel. (Asimov 401950,).

Asimov törvényei azonban azonnal problémákba ütköznek. Az egyik a "sérülés" és a "kár" fogalmának tisztázatlansága az első törvényben. Ha egy robot testőr megsebesít egy potenciális bérgyilkost egy ártatlan személy védelme során, akkor "megsebesítette" vagy "megkárosította" őt? A bérgyilkos érdekei nyilvánvalóan sérültek, de vajon sérült-e? Különbséget kell tennünk a kár vagy sérülés nem moralizált és moralizált felfogása között (vagy a megkülönböztetés jogi változatában: *damnum* és *injuria*). Ha ezt megtesszük, valószínűtlennek tűnik, hogy a nem moralizált értelemben vett embereket károsító RAI-k teljes tilalma fenntartható lesz. Sőt, még az a követelmény is, hogy az RAI-k soha ne bántsanak emberi lényt, alábecsülheti az RAI-k által jogosan felmerülő dilemmák összetettségét. [5] Egy ismerős dilemma arra vonatkozik, hogy egy önvezető autónak hogyan kell reagálnia olyan helyzetekben, amikor választania kell az utasának okozott kár elkerülése - például egy szembejövő teherautó útjából való kitérés - és más emberek (sofőrök, utasok vagy gyalogosok) sérülésének elkerülése között, akiket halál vagy sérülés veszélye fenyeget, ha az autó kitér az utasa megmentése érdekében. Erre a "kocsi-problémára" ellentmondásos válaszokat kapunk, de úgy tűnik, hogy minden hihető válasz magában foglalja annak a mindenre kiterjedő lehetőségét, hogy egy RAI rosszat tesz egy embernek. Érdekes módon az empirikus vizsgálatok azt mutatják, hogy a legtöbb ember egyetért azzal, hogy az utasokat fel kell áldozni annak érdekében, hogy minél több járókelőt meg lehessen menteni, ugyanakkor a legtöbben szívesebben ülnének olyan kocsiba, amely mindig megmenti utasát (Bonnefon, Shariff, Rahwan 2006).[6] Ha ez így van, akkor a troliproblémára adott helyes válasz meghatározása gyakorlatilag irrelevánsnak bizonyulhat, mivel az első típusú autót nem vennék meg elég sokan ahhoz, hogy megérje azt gyártani. Asimov második és harmadik törvénye is megkérdőjeleződhet, ha meggyőződünk arról, hogy a fejlett, látszólag emberinek tűnő intellektuális és érzelmi tulajdonságokkal rendelkező RAI-k megszereznek valamit, ami megközelíti az emberi személyiséget és az ebből fakadó jogokat, beleértve az önrendelkezést is.

Asimov alapelvei korai és kezdetleges kísérletet jelentenek a RAI-k etikájának megalkotására. Hasonló elemi nehézségek azonban az újabb törekvéseket is sújtják, mint például a

5. Vö. még a Lordok Háza AI bizottságának 1252018,. cikkében felvázolt ötödik etikai keretelvvel: "Az emberi lényt bántani, elpusztítani vagy megtéveszteni képes autonóm hatalmat soha nem szabad a mesterséges intelligenciára ruházni".

6. Az önvezető autók balesetek kezelésével kapcsolatos etikai kérdésekről lásd: Nyholm 2018.

Gyakorlati etikai folyóirat

mint a 2017-ben megfogalmazott Asilomar AI-elvek. Ezen elvek némelyike már-már a közhelyszerűség határát súrolja, például az az elv, amely 6előírja, hogy "a mesterséges intelligenciával működő rendszereknek egész működési élettartamuk alatt biztonságosnak és védettnek kell lenniük, méghozzá ellenőrizhetően, ahol ez alkalmazható és megvalósítható". Más elvek nem jelentenek kivételt, de nem túlságosan homályosak, például az az15 elv, amely szerint "a mesterséges intelligencia által teremtett gazdasági jólétet széles körben meg kell osztani, hogy az egész emberiség javát szolgálja". Az elvek két másik tendenciáját is érdemes kiemelni, mivel ezek gyakran ismétlődnek az AI-k etikai elveiről szóló más nyilatkozatokban. Az első az a hallgatólagos feltételezés, hogy létezik egy megszámlálhatatlan értékelési szempontkatalógus, amely különösen fontos a RAI-k számára. Így a 11. elv megköveteli, hogy a mesterséges intelligencia rendszerek összeegyeztethetők legyenek "az emberi méltóság, a jogok, a szabadságjogok és a kulturális sokszínűség eszméivel". Kérdéses azonban, hogy a RAI-érzékeny értékek bármilyen értelmesen meghatározott listája rendben van-e. Miért ne lehetne további értékeket, például a jótékonyságot, a természeti környezet tiszteletben tartását vagy a közjó iránti aggodalmat felvenni a listára? Nincs okunk azt feltételezni, hogy a RAI-kra potenciálisan alkalmazható etikai értékek *eleve* nem érik el az emberi értékek teljes skáláját. Természetesen van némi felismerés erre vonatkozóan, amikor az elvek más értékekre, például a közjóra hivatkoznak. Itt azonban egy másik aggasztó vonás is felbukkan, mégpedig az a tendencia, hogy az enu- merált értékeket az értékről alkotott széles körben elfogadott *hiedelmekre* redukálják. Ezért a közjóról szóló23 elv kimondja: "A szuperintelligenciát csak széles körben osztott etikai eszmék szolgálatában szabad fejleszteni, és nem egy állam vagy szervezet, hanem az egész emberiség javára". Itt a közjó két különböző fogalma keveredik:

(1) olyan etikai értékek, amelyek valójában széles körben osztoznak az emberek között, és (2) amelyek objektíve minden ember javát szolgálnák. Az utóbbi normatív elképzelés, az előbbi empirikus elképzelés, amelynek normatív következményeit, ha vannak ilyenek, a valóban normatív elvekkel együtt kell kidolgozni. A probléma nem szűnik meg, ha a jogra hivatkozunk, nem pedig a széles körben elterjedt hiedelmekre. Az Európai Bizottság nemrégiben közzétett, *a megbízható mesterséges intelligenciára vonatkozó etikai iránymutatásai például az* emberi jogoknak alapvetően fontos szerepet tulajdonítanak. [7] Félretéve azt a tényt, hogy ez a jog nem tükrözi az összes etikai megfontolást (pl. környezeti értékek), amely a mesterséges intelligenciára vonatkozik, hogy nem közvetlenül kötelező érvényű a nem állami szereplőkre, és hogy nem minden rendelkezése kötelez minden államot (pl. mert nem ratifikálták a vonatkozó emberi jogi szerződéseket). A

7. "Hiszünk abban, hogy a mesterséges intelligencia etikájának megközelítése az uniós szerződésekben, az uniós chartában és a nemzetközi emberi jogi jogszabályokban rögzített alapvető jogokon alapul. Az alapvető jogok tiszteletben tartása ademokrácia és a jogállamiság keretein belül a legígéretesebb alapot nyújtja az absztrakt etikai elvek és értékek azonosításához, amelyek a mesterséges intelligencia kontextusában operacionalizálhatók". Európai Bizottság 2019,9.

Kötet kiadás 7,1

Az alapvetőbb az, hogy az ilyen törvények - az "emberi jogok" szavak által sugallt erőteljes erkölcsi töltet ellenére - nem alapvető etikai normák. Ehelyett, mint minden más törvényt, ezeket is meg kell fogalmazni és értékelni kell - és néhányszor komoly hiányosságokat kell megállapítani - az alapvető etikai normák, köztük az emberi jogok erkölcsisége szempontjából (lásd Tasioulas (készül)).

A RAI-kkal kapcsolatos helyes etikai megközelítésnek ezért túl kell lépnie a széles körben elterjedt hiedelmekre vagy a bevett jogra - beleértve az emberi jogi törvényeket is - való hivatkozáson, hogy a vonatkozó etikai értékek teljes skáláját figyelembe vegye. A normatív és az empirikus, hagyományos vagy jogi szempontok összemosására való hajlam talán várható a technológiai és "adatvezérelt" gondolkodásmóddal rendelkezők körében, akik érthető okokból a robotika és a mesterséges intelligencia közösségében túlsúlyban vannak. Ez ahhoz a végzetes következtetéshez vezethet, hogy az etikai normákat empirikus módszerek - például a "tömeges forrásszerzés" - alkalmazásával kell meghatározni a széles körben elterjedt etikai meggyőződések megállapítása érdekében. Ezen az úton azonban az etika a PR-ipar egyik ágává degenerálódik.

A második kérdés, hogy miként építhetjük be az etikai normákat a RAI-k működésébe, nem kisebb kihívás. Egyesek nagy ambíciókat táplálnak a robot erkölcsi bölcsek iránt, akik az átlagos embert messze felülmúló szakértői tudással rendelkeznek az erkölcsről. Julian Savulescu és Hannah Maslen azt állítják, hogy a "mesterséges etikai ágensek", köszönhetően emberfeletti sebességüknek, kiterjedt adatbázisuknak és a rájuk jellemző emberi vétkek, például az önzés hiányának, "ténylegesen segíthetik, sőt helyettesíthetik az embert a nehéz erkölcsi döntéshozatalban" (Savulescu és Maslen 2015). Hasonló irányvonalak mentén az RAI-kat javasolták a bűnözők elítélése során az emberi bírákat közismerten sújtó elfogultságok leküzdésére - például a napszaktól függően enyhébb vagy szigorúbb ítéletet hoznak, vagy ami még aggasztóbb, az elkövetők osztálya, etnikai hovatartozása vagy fajtája alapján. [8]

Mások, mint például az UNESCO nemrégiben közzétett jelentésének szerzői, szkeptikusak a RAI-k erkölcsi tökéletességét illetően. [9] E szkepticizmusnak két, egymással összefüggő forrása van. Először is, hogy az erkölcsi döntéshozatal a relevánsan különböző helyzetek végtelen számú potenciális szituációjával szembesül, amelyekkel egyetlen algoritmus vagy gépi tanulási folyamat sem elég érzékeny ahhoz, hogy foglalkozzon.

8. Lásd például Sunstein (hamarosan megjelenő) tanulmányát arról, hogy az algoritmusok hogyan segíthetnek korrigálni a jelenlegi elkövetői torzítást - azt a tendenciát, hogy az óvadékról szóló döntésekben túlzott hangsúlyt fektetnek a jelenlegi bűncselekmény tényére. Általánosabban, azzal az állítással kapcsolatban, hogy az emberi megismerés inkább egy "fekete doboz", mint az algoritmusok által a diszkrimináció felderítésével kapcsolatban potenciálisan elérhető átláthatóság, lásd Kleinberg, Ludwig, Mullainathan és Sunstein (megjelenés alatt).

9. "Nem tűnik valószínűnek, hogy bármilyen gép, amely nem rendelkezik olyan érzelmekkel, mint az empátia..., képes lenne kezelni az erkölcsileg releváns tények és preferenciák ilyen

változatosságát". UNESCO (2017, 44).

megfelelően. Másodszor pedig, hogy a helyes erkölcsi érvelés megköveteli, hogy az érvelő olyan érzelmi reakciókat fejlesszen ki, mint a bűntudat, a felháborodás és az empátia, amelyek megfelelően alkalmazkodnak a tárgyukhoz. Ezek a reakciók teszik lehetővé, hogy bizonyos helyzetek erkölcsi jelentőségét regisztráljuk, például azt, hogy sürgősen kell cselekednünk olyan helyzetekben, amelyek indokolják a közvetlen fenyegetettségtől való félelmet. De vitatható, hogy ezek természetüknél fogva meghaladják azon lények képességeit, amelyek nem osztoznak az emberi tudatosságban és életmódban. Mindkét gondolatmenetet az erkölcsfilozófia különböző irányzatai hangsúlyozták, a közelmúltban leginkább a neoarisztotelészi erényetika és a feminista elmélet. De némi támogatást kapnak a viselkedés- és agytudományoktól is, amelyek azt sugallják, hogy az érzelmekre való képesség nem egy különálló "modul", amely hozzáadódik kognitív gépezetünkhöz, hanem agyunk általános felépítésének szerves része (Pessoa 2018).

Az, hogy az RAI-k erkölcsi szakértőkké válhatnak-e, részben attól függ, hogy mi a helyes filozófiai elképzelés az erkölcsről. Ha a helyes nézet valami olyasmi, mint az utili- tarianizmus, amely egyetlen nagyon általános etikai elven nyugszik, és a jövőbeli következményekre vonatkozó ijesztő számításokat igényel, akkor az RAI-bölcsek kilátásai fényesnek tűnhetnek. Ha ezzel szemben a helyes erkölcsi megközelítés kontextus-specifikus ítélőképességet igényel, amely az értékek sokaságára hangolt erkölcsi érzelmek gazdag palettájából merít, és nem ad szerepet a mechanikusan alkalmazható általános elveknek, akkor a kilátások megfelelően borúsnak tűnnek.

Még ha félretesszük is az RAI-król mint erkölcsi szakértőkről szóló futurisztikus spekulációkat, és ehelyett arra összpontosítunk, hogy a különböző konkrét feladatok elvégzése során megfelelnek-e az alapvető erkölcsi normáknak, az erkölcsről szóló igazságnak fontos szerepe lesz abban, hogy az ilyen normákat hogyan lehet a legjobban integrálni a robotok működésébe. Természetesen sok múlik azon, hogy milyen feladatokat szeretnénk, ha a robotok elvégeznének, és milyen környezetben fognak működni, pl. hogy ez lehetséges emberi felügyelet vagy felülbírálat mellett vagy anélkül történik-e majd. Azonban nem lenne reális feltételezni, hogy az erkölcsfilozófiai nézeteltéréseket meg kell oldanunk, mielőtt etikai elveket programoznánk az RAI-k működésébe. Ennek az az oka, hogy a különböző erkölcsi filozófiák képviselőinek még akkor is jó okuk lehet arra, hogy az alapvető erkölcsi normák bizonyos magjában megegyezzenek, még akkor is, ha különböző indoklásokat és (sok esetben értelmezéseket) kínálnak rájuk vonatkozóan. Amint láttuk, kétféle megközelítést különböztethetünk meg az etikai normáknak a RAI-k működésébe való beillesztésére, bár mindkét megközelítés elemei keveredhetnek bármelyik adott RAI-ban. Az első egy felülről lefelé irányuló megközelítés, amely bizonyos elvek - mint

például a genfi egyezmények a hadviselésről, a hadseregben, a

az AWS-ek esetében, vagy a "chatbotok" esetében a beszélgetés ésszerűségének és nem sértő voltának normái - algoritmikus formába öntve. Ez egy ijesztő feladat, amely, ha sikeresen végrehajtják, lehetővé tenné egy RAI számára, hogy a halálos erő alkalmazásakor kifinomult arányossági ítéleteket hozzon, vagy különbséget tegyen a játékos humor és a sértő szidalmazás között. Egy másik megközelítés inkább alulról felfelé építkező jellegű. Ez egyfajta "gépi tanulás" alapján történne, amelynek során például az RAI-t az adott terület jogi szakértői által hozott rengeteg korábbi döntésnek tennék ki, majd a jövőbeni forgatókönyvek esetében a saját döntéseire extrapolálna. Az erényetikából kiindulva egyesek még azt is állítják, hogy a RAI-k etikai nevelésének helyes módja az, ha úgy neveljük őket, mint a gyerekeket, azon az alapon, hogy a jó jellem kialakulásához tisztességes nevelésre van szükség (Rini 2017).

Bár a RAI-k segítséget ígérnek bizonyos kihívások kezelésében, beleértve az emberi tökéletlenségek és korlátok leküzdését a fontos feladatok elvégzésében, gyakran hiányosak az egyébként értékes cél eléréséhez szükséges megfelelő normák betartásában. Amint fentebb megjegyeztük, a mesterséges intelligenciát hatalmas adathalmazok hajtják. Az adatok felhalmozásának eszközei azonban gyakran erkölcsileg kétes értékűek. Az egyik leghírhedtebb példa erre az olyan online platformok, mint a Facebook és a Google által végzett célzott hirdetések, amelyekből bevételeik mintegy 90%-át szerzik. Kényelmes, de egyben idegesítő is lehet, ha az online hírfolyamunkban olyan hirdetések jelennek meg, amelyek az ízlésünkre és érdeklődési körünkre vannak szabva. Ez a folyamat azonban olyan algoritmusokat foglal magában, amelyek az Ön által látogatott weboldalakról, az Ön által küldött e-mailekből és a mobilos nyomkövetésből származó adatokon alapulnak. És komoly kérdés, hogy az ezeket a platformokat használó emberek tisztában vannak-e azzal, hogy ők "adattehenek", akiket könyörtelenül fejnek a kereskedelmi szempontból értékes információkért, nem is beszélve arról, hogy ehhez ténylegesen hozzájárultak-e. Ennek eredményeképpen ezek az üzleti modellek joggal váltottak ki aggodalmakat azzal kapcsolatban, hogy sértik a magánélethez fűződő jogokat, vagy a gazdasági kizsákmányolás formáit jelentik. Hasonló aggályok számos más tevékenységre is kiterjednek, beleértve az olyan platformokat, mint például az Axon vállalat által kifejlesztett platform, amely több mint 20 millió gigabájt közbiztonsággal kapcsolatos, a rendőrségi testkamerákból származó adatot tárol (Goode 2018).

Az, hogy az egyéneknek joguk van ellenőrizni, hogy mi történik a személyes adataikkal, nem jelent mindenre kiterjedő megoldást az ilyen adatok gyűjtése és felhasználása által okozott etikai problémákra. Ennek egyik oka az, hogy az adatokra szükség lehet egy létfontosságú társadalmi jó előmozdításához - például egy fertőző betegség kitörésének megelőzéséhez vagy egy terrortámadás

megelőzéséhez. Ilyen esetekben az egyéneknek vétójogot adni arra vonatkozóan, hogy hozzáférjenek-e az adataikhoz, vagy azokat bizonyos módon használják-e fel, aránytalannak tűnik a jó értékéhez képest.

Gyakorlati etikai folyóirat

ami elfelejtődik. A társadalmi javak ilyen módon történő elérése azonban az egyéni beleegyezés hiányában más szigorú feltételek teljesítését is megkövetelheti, például az adatfelhasználók céljainak és módszereinek átláthatóságát, a számonkérés mechanizmusait stb.

Egy másik hiba, amelynek kiküszöbölésééért a területen dolgozók küzdenek, az algoritmikus elfogultság, amely még akkor is felmerülhet, ha az adatgyűjtés eszközei nem sértik az erkölcsi normákat, például a magánélet védelmét és a kizsákmányolás tilalmát. Az RAI-kat olyan algoritmusok vezérlik, amelyeket adathalmazokon képeznek ki, és ezekből általánosítva működnek a jövőbeli forgatókönyvekre. Az egyik probléma abból adódik, hogy a képzési adatok önmagukban is hibásak lehetnek a pontos ítéletek vagy döntések alapjaként. Ez nem meglepő, mivel az adatokat éppen azok a hibás lények (emberek) generálják, akiknek a hiányosságait - amelyek közül sok káros és igazságtalan társadalmi viselkedésminták formájában gyűlik össze - az RAI-kat elsősorban azért fejlesztették ki, hogy leküzdjék. Az adatok különösen statisztikailag torzak lehetnek, pl. nem foglalják magukban a kisebbségi csoportokat, vagy előítéleteket és történelmi diszkriminációs mintákat tartalmaznak. A valós életben tapasztalható algoritmikus torzításokra a közelmúltban olyan példákat hoztak fel, mint egy angol rendőrség által használt algoritmus, amely diszkriminálta a szegényebb területekről származó embereket annak eldöntése során, hogy őrizetben tartsák-e a bűnözőket, olyan álláskereső eszközök, amelyek a magas jövedelmű állások esetében a férfiakat részesítették előnyben a nőkkel szemben, valamint az internetes képkereséstől a rendőrségi végrehajtásig számos alkalmazásban használt arcfelismerő algoritmusok, amelyek sokkal nagyobb hibaarányt mutatnak a nők és a nem fehér bőrűek esetében (Oswald, Grace, Unwin and Barnes 2018; Burgess 2017; O'Neil 2016; valamint Buolamwini and Gebru 2018). Különösen kirívó példa erre a COMPAS kockázatértékelési algoritmus, amelyet egyes amerikai bíróságok az ítélethozatal során használnak. Célja, hogy megjósolja annak valószínűségét, hogy egy elkövető a következő két évben ismét elkövet egy bűncselekményt, és ezt a célt 70%-os pontossággal teljesíti. Hibái azonban kifejezett faji előítéletet mutattak: A COMPAS kétszer nagyobb valószínűséggel állapította meg tévesen a magas kockázatot fekete elkövetők esetében, mint fehér elkövetők esetében ("hamis pozitív"), ugyanakkor a fehér elkövetők kétszer nagyobb valószínűséggel jelölték meg tévesen alacsony kockázatúnak a fekete elkövetőkhöz képest ("hamis negatív") (Larson, Mattu, Kirchner és Angwin 2016). [10]

Még ha mondjuk egy kockázatértékelő algoritmus hatékonyan képes is pontosan megjósolni annak valószínűségét, hogy a bizonyos jellemzőkkel rendelkező gyanúsítottak elkövetnek majd bűncselekményt, ez nem jelenti azt, hogy

10. A COMPAS-algoritmus vitáját a méltányosságról és a diszkriminációról szóló szélesebb körű filozófiai vitába helyező újabb vitát lásd Binns (2018). Lásd még, általánosabban, Barocas és Selbst, (2016).

nem vet véget etikai aggályainknak. Először is, van valami problémás abban, ha egy személyről az alapján ítélkeznek, hogy *más* emberek, akiknek különböző tulajdonságai megegyeznek vele, hogyan viselkedtek a múltban. Nem követeli meg a személyes autonómiájának tiszteletben tartása, hogy őt az alapján értékeljük, amilyen egyéniség ő, a korábbi szavai és tettei alapján, nem pedig másoké alapján, akik osztoznak az ő (esetleg nem kiválasztott) tulajdonságaiban? A probléma súlyosbodik azokban az esetekben, amikor a releváns jellemzők, mint például a faji hovatartozás, a nem vagy a szegénység, maguk is súlyos igazságtalanságok hosszú történelmének középpontjában állnak, vagy azok termékei. Az eredmény egy "ördögi kör", amelyben a RAI állandósítja és súlyosbítja azt az igazságtalanságot, amely az adatállományát létrehozta, például azáltal, hogy a letartóztatások során a kisebbségi vagy szegény gyanúsítottakra összpontosít, ami a kisebbségi elítélések számának növekedéséhez vezet, ami viszont a kisebbségi gyanúsítottakkal szembeni nagyobb diszkrimációhoz vezet, egy végtelen lefelé tartó spirálban.

Az algoritmikus elfogultság problémája nem oldható meg az olyan adatok egyszerű kiszűrésével, amelyek kifejezetten érzékeny kategóriákra, például életkorra, osztályra, nemre vagy fajra vonatkoznak, mivel más, látszólag ártalmatlan adattípusok korrelálhatnak ezekkel a kategóriákkal (vagy ezek helyettesítői lehetnek). Az ilyen kihívásokra adott válasz egyik fontos része az, hogy alaposan megvizsgáljuk, milyen célokra használják az algoritmust (Fry 622018,). Például előfordulhat, hogy bizonyos kontextusokban valóban az eredmények előrejelzésének elsődlegessége számít, míg más kontextusokban ez kevésbé fontos. Vitatható, hogy például annak meghatározása, hogy mely gyanúsítottakat kell óvadék ellenében szabadlábra helyezni, nagyobb teret enged a gyanúsított bűncselekmény elkövetésének előre jelzett valószínűségén alapuló döntésnek, mint az ítélethozatal, mivel az előbbi nem jár ugyanolyan mértékű morális ellenszenvvel, mint az utóbbi. Az algoritmusok igazságos működésével kapcsolatos, már említett kihívásokat súlyosbítja az a tény, hogy kereskedelmi okokból gyakran sem az algoritmust, sem az adatokat, amelyeken az algoritmus kiképzésre került, nem hozzák nyilvánosságra. Sőt, az alulról felfelé irányuló algoritmusok esetében még az RAI-t működtető személyek számára is átláthatatlan lehet, hogy pontosan milyen algoritmus irányítja a tevékenységét.

Ha egy RAI a fent említett módon funkcionális, felmerül a kérdés, hogy milyen szinten kell a funkcionalitás követelményét meghatározni, különösen az emberi lények által teljesíthető képességekkel összehasonlítva. A válasz a kitűzött cél értékétől függ, valamint attól, hogy milyen kockázatokat vagyunk hajlandóak vállalni a cél elérése érdekében. Egyes feladatok esetében, például a szobatisztító vagy társas robotok által végzett feladatok esetében elegendő lehet, ha egy RAI

megfelelően teljesít, még ha nem is olyan jól, mint az ember. Más feladatok, mint például a gépjárművezetés, az orvosi diagnosztika vagy a bűnözők elítélése esetében a tét olyan fontosnak ítélhető - az élet, a szabadság és az igazságszolgáltatás -, hogy

hogy az RAI funkcionalitásának legalább olyan jónak, vagy talán lényegesen jobbnak kell lennie, mint amit az emberek általában elérnek. A RAI-k kiváló teljesítménye olyasvalami, amire azért is szükség lehet, hogy ellensúlyozza a rájuk való támaszkodással járó nemkívánatos mellékhatásokat és veszélyeket, például a munkahelyek elvesztését vagy a rosszindulatú ügynökök szabotázsát (lásd 4.4alább4.5,). A funkcionalitás minimálisan elfogadható szintjének meghatározását megnehezítő tényező az összehasonlítási alap: azt kell-e összehasonlítani, amire az emberek elvileg képesek, vagy azt, amire ténylegesen képesek? Például egy olyan világban, ahol emberek milliói nem férnek hozzá az alapvető oktatáshoz és egészségügyi ellátáshoz, egy robotoktató leckéi vagy egy robotorvos diagnózisai akkor is jelentős értéket képviselhetnek, ha határozottan alacsonyabb szintűek, mint az emberi oktatók és orvosok szolgáltatásai, amelyekhez az egyének - akár anyagi, akár más okokból - nem férnek hozzá. Az embereket megfosztani az RAI-k által nyújtott létfontosságú előnyöktől egy olyan idealizált emberi szolgáltatási szint nevében, amely valószínűleg elérhetetlen marad számukra, nem lenne helyes.

Egy további szempont, amit szem előtt kell tartanunk, az a RAI-kra vonatkozó erkölcsi normák dinamikus minősége; különösen az, hogy ezek a normák idővel hogyan fejlődhetnek a technológiai fejlődés, valamint a lehetőségek és költségek új profiljának eredményeként. Itt nem az erkölcsi normáinknak a technológiai fejlődés által történő egyirányú meghatározására gondolok, arra a fajta nézetre, amely egyeseket tévesen arra késztet, hogy azt állítsák, hogy "a magánéletnek vége", csak azért, mert nincs bolondbiztos módszer a magánélet megsértésének megakadályozására. A gondolat inkább az, hogy az olyan erkölcsi normákat, amelyeket az RAI-kkal kapcsolatos lehetőségek megjelenése előtti korszakban dolgoztak ki, újra kell értékelni a megjelenésük és az általuk ígért előnyök vagy az általuk jelentett kockázatok fényében. Ha például a RAI-k fejlődése jelentősen növeli a járványok kitörésének digitális felismerése révén történő előrejelzésének lehetőségét, akkor lehet, hogy kevésbé korlátozó normákat kell kialakítanunk a RAI-k által végezhető digitális felügyelet típusaira vonatkozóan, hogy megvédjenek minket az ilyen fenyegetésektől. A megfigyelés olyan formái, amelyeket a múltban jogosan ítéltek a magánélethez való jog megsértésének, ebben az új technológiai környezetben már nem feltétlenül azok (Vayena és Tasioulas 2016).

4.2 EREDENDŐ JELENTŐSÉGE

Még ha az RAI-k egy adott feladatban vagy szerepkörben megfelelő funkcionalitást is képesek elérni, felmerülhetnek kérdések azzal kapcsolatban, hogy miért fontos az adott feladat vagy szerepkör RAI-kra és nem emberekre bízása. Néha az "emberi tényező" kiküszöbölése előnyös lehet.

Kötet kiadás 7,1

Egy olyan idős személy esetében, akinek például fürdés közben segítségre van szüksége, a robotgondozó minimálisra csökkentheti a zavarodottság kockázatát.[11] Az emberi tényező kiiktatása azonban aggasztó is lehet. Ez már most is megmutatkozik abban a széles körben elterjedt aggodalomban, hogy az embereknek tudniuk kell, hogy a "másik", akivel online vagy telefonon érintkeznek, ember-e vagy gép. Az egyik olyan kontextus, ahol az emberi tényező jelentősége különösen hangsúlyosnak tűnik, az a RAI-k esetében, amelyek olyan döntéseket hoznak, amelyek súlyos következményekkel járnak az emberekre nézve. Itt elsősorban a következőkről van szó
(1) az alulról felfelé építkező vagy sztochasztikus algoritmusokon alapuló döntések, amelyek ezért nem kiszámíthatóak, és (2) az olyan döntések, amelyek az érintett ember valamilyen személyes értékelésén alapulnak, pl. az érdemei, a sivatagi helyzete vagy a jogosultságai.

Ezen aggodalomról való elmélkedés során hasznos lesz, ha szem előtt tartunk egy élénk, bár tömörített és vitatható emlékeztetőt az emberi lét jelentőségéről, különösen az olyan műtárgyakkal, mint a gépek, szemben. Ebben a tekintetben aligha tehetünk jobbat David Wiggins következő kijelentésénél:

> *[A]z adott sajátos állati természetben és a törvény által fenntartott életi módban való részesedésünk szerves része a személy és a személy közötti szoros nyelvi összhangnak, és szerves része az emberi érzékenységnek, amely lehetővé teszi az értelmezést. Másodszor, ez a sajátos állati természetben és tevékenységi módban való részesedés előfeltétele annak az emberi szolidaritásnak (ahol van), amely elítéli az emberi lénynek - egyikünknek - puszta dologként vagy puszta eszközként való kezelését. Harmadszor pedig, hogy a természetben és tevékenységben való rokonság szerves része annak a képnek - egy nem determinisztikus képnek -, amelyet arról alkotunk, hogy képesek vagyunk - egyesével és közösen - egy nem általunk választott és nálunk nagyobb jelentésekkel teli kereten belül meghatározni közvetlen és közvetett céljainkat. Ebben a keretben találhatjuk meg helyünket és gyakorolhatjuk képességeinket. Nem úgy tekintünk magunkra, mint valamilyen funkcióval rendelkező dolgokra - mi a fenére lehet egy ember, mint ember? -, hanem mint autonóm, önmozgó, eleven lényekre, olyan lényekre, akik megtalálják magukat a világban, és arra törekszenek, hogy saját nyomot hagyjanak rajta, a legjobbat hozzák ki abból, amit ott találnak, és (ha szerencsések) keresnek valamit, amit mindegyikünk a saját megfelelő munkájának vagy hivatásának tekinthet (Wiggins 91).2016,*

Ebben a szövegben Wiggins hármas jelentőséget tulajdonít a közös emberi természetnek: *hermeneutikai* szempontból lehetővé teszi egy bizonyos fajta

kölcsönös megértést egy

11. Az RAI-k egészségügyi és szociális ellátásban való használatának átgondolt megvitatását lásd Coeckelbergh (2015), annak fényében, hogy a modernitás kontextusában a technikai-logikai fejlődés hogyan hat az emberi kapcsolatokra.

etikai szempontból ez egy bizonyos fajta szolidaritás alapja, amely egy olyan közös eredendő értéken alapul, amely nem egyeztethető össze többek között azzal, hogy embertársainkat pusztán eszközként kezeljük egyéni céljaink eléréséhez; *metafizikai szempontból* pedig az embereknek megvan az a racionális önállóságuk, hogy meghatározzák életüket alakító céljaikat, szemben a műtárgyakkal, amelyek természetét és tevékenységét egy mások által meghatározott és meghatározott cél határozza meg.

Most térjünk vissza a RAI-k döntéshozatali funkcióinak a RAI-khoz való hozzárendelésével kapcsolatos aggályainkhoz. Gondoljunk például a tartósan munkanélküli emberek helyzetére, akiknek az álláspályázatát rendszeresen elutasítják az automatizált rendszerek, amelyek ma a munkaerő-toborzást uralják. Miután hónapokig vagy akár évekig sikertelenül pályáznak állásokra, előfordulhat, hogy ezeknek az embereknek egyszer sem olvassa el és értékeli a pályázatukat egy másik ember. Még ha feltételezzük is, hogy az adott algoritmus megfelel a megfelelő funkcionalitásnak, azaz ugyanolyan hatékony, eredményes és megfelel a megfelelőségi normáknak, mint az átlagos emberi munkaerő-felvételi szakember, az a tény, hogy ez egy nem emberi döntéshozatali mód, aggodalomra ad okot. Nehéz pontosan meghatározni az aggodalmat, de a gondolat nagyjából az, hogy az álláskereső egy rideg, elidegenítő és végső soron potenciálisan tiszteletlen folyamatnak van kitéve, mert a jelentkezése soha nem jut el egy embertársához. Ezt sugallja a *Guardian* egyik nemrégiben megjelent cikkéből vett részlet:

> *"Kicsit embertelen, hogy soha nem tudok kapcsolatba lépni egy munkáltatóval" - mondja Robert, a negyvenes éveiben járó vízvezeték-szerelő, aki állásbörzéket és toborzókat használ ideiglenes munkák keresésére. Harry négy hónapja24, keres munkát. A kiskereskedelemben, ahol ő keres, "majdnem minden munkakörben" van valamilyen teszt vagy játék, a személyiségtől a matematikáig, hogy kiszűrjék a jelentkezőket. Hetente négy-öt tesztet tölt ki, amikor az állásokat közzéteszik. Az elutasítások gyakran azonnaliak, bár egyes szolgáltatók időbeli késleltetéssel küldött elutasító e-maileket kínálnak, feltehetően azért, hogy fenntartsák azt az illúziót, hogy az ember időt töltött egy olyan pályázat elbírálásával, amely már megbukott egy automatizált szűrésen (Buryani 2018).*

Vagy nézzünk meg két másik példát, amelyek még fontosabb érdekeket érintik: a bűnözők RAI-k általi megfigyelése és az autonóm fegyverrendszerek (AWS) használata. Természetesen komoly kérdés, hogy az RAI-k képesek-e az emberi bírákhoz és katonákhoz hasonló, vagy azoknál magasabb szintű funkcionalitást elérni. [12] De még ha

12. Az itt rejlő lehetőségek viszonylag derűlátó megítélését lásd Turner (2018).

Kötet kiadás 7,1

tudnak, akkor zavarhat bennünket az a tény, hogy egy RAI döntései ilyen drasztikusan befolyásolják az emberi életet és szabadságot. Egy szkeptikus megkérdezhetné: milyen jogos panasza lehet egy vádlottnak vagy ellenséges katonának *pusztán az* ellen, hogy egy nem emberi lény ítélte el vagy ölte meg? Az aggodalom itt nem más, mint etikai aggályoknak álcázott nos- talgia? Talán az aggodalom megfogalmazásának egyik módja az a gondolat, hogy a mások életéről és szabadságáról szóló döntések olyan jelentősek, hogy valami értéket veszítünk, ha azokat nem egy olyan szereplő hozza meg, aki felelősséget tud vállalni értük. És paradigmatikusan számunkra ez a szereplő egy emberi lény - valaki, aki képes megérteni és együttérezni embertársunk helyzetével, és a saját autonóm értékelésének fényében döntést hozni arról, hogy milyen okok szólnak amellett, hogy egy bizonyos módon kezeljenek minket, és nem másként. Itt az emberi lény Wiggins által említett mindhárom dimenziója szerepet játszik. Ebből következik, hogy nincs olyan döntéshozó, *akitől* számon kérhetnénk, hogy miért döntött úgy, ahogyan döntött, és akit ezen okok értékelése alapján felelősségre vonhatnánk. Létezik egy értékes emberi szolidaritás és kölcsönösség - az emberek embertársként ismerik el egymást, és ennek alapján alakítják ki egymáshoz való viszonyulásukat és döntéseiket -, amely a dehumanizált, teljesen automatizált döntéshozatal kontextusában elvész. Ehhez hozzátehetünk egy további megfigyelést. Gyakran előfordul, hogy a másik emberrel való bánásmódot befolyásoló relatív okok többféleképpen is kiegyensúlyozhatók, és a döntések egy sora racionálisan támogatható, de egyetlen döntés sem lehet az egyetlen helyes válasz. A bűnözők elítélése során például a bíró előtt nyitva álló büntetési tételek jellemzően bizonyos durva határok között változnak, és nincs olyan pontos büntetési tétel, amelyet egyértelműen ki lehetne jelölni, mint amelyik egyedüliként indokolt. Erre a helyzetre válaszul egyesek mindenekelőtt a hasonló esetek kezelésének következetességét tartják fontosnak, azt állítva, hogy az azonos hibát elkövető bűnözőknek pontosan ugyanolyan mértékű büntetést kell kapniuk. A RAI-k - ragaszkodhatnak hozzá - különösen alkalmasak arra, hogy a "hasonló eseteket" ilyen módon kezeljék. Mások azonban úgy látják, hogy a bírónak helye van a mérlegelési jogkörében - és az ebből következően az ítéletek különbözőségében -, amit ez generál. Azok számára, akik a második megközelítést fogadják el, értéket jelent, ha egy kegyes bíró kifejezheti értékeit és jellemét például azzal, hogy a szóba jöhető lehetőségek közül egy enyhébb büntetést választ. Értéket képvisel egy olyan büntető igazságszolgáltatási rendszer, amely a bűnelkövetőknek lehetőséget biztosít a mérlegelési jogkörben történő kegyelemre. Az irgalom megadása itt egyfajta ajándékozás, egyik személy részéről a másiknak, ami talán az előbbi azon reményét tükrözi, hogy az utóbbi valóban megbánta korábbi rossz cselekedeteit. Az automatizált büntetéskiszabási rendszer esetében ez az érték a következő lenne

Gyakorlati etikai folyóirat

súlyosan korlátozni vagy megszüntetni. A robot nem egy értékekkel és saját karakterrel rendelkező egyén, aki úgy reagálhat az elkövető kegyelmi kérésére, mint egy ember a másikra, és az enyhébb büntetést választja, ha a szigorúbb büntetés is aránylag nyitva áll.

Most, még ha általános aggodalom is van a RAI-kkal kapcsolatban, amelyek az emberi érdekeket érintő fontos döntéseket hoznak, még mindig van két további kérdés, amellyel foglalkozni kell. Az első az emberi tényező súlya egy adott esetben. Vajon ez csupán egy ok a többi között, vagy pedig kötelezettség szintjére emelkedik, sőt, talán még emberi joggal is társul? A Rathenau Intézet egyik jelentése a tartalmas emberi kapcsolatokhoz való emberi jogot (valamint a nem méréshez, elemzéshez vagy edzéshez való jogot) javasolta (UNESCO 392017,), amely feltehetően veszélyben van a fent tárgyalt háromféle esetben. Eközben a korábban említett UNESCO-jelentés elutasítja az AWS-eket azon "vezérelv alapján, hogy a gépek nem hozhatnak élet-halál kérdésében döntést az emberről", mivel az ember megöléséről szóló döntés gépre való delegálása az emberi méltóság megsértése (UNESCO 2017, 54). [13] Természetesen, még ha egy súlyos értéket fel is áldozunk, lehet, hogy mindent összevetve a robotbírák vagy AWS-ek előnyei egyes esetekben igazolják ezt az áldozatot. Még radikálisabban, egyesek szerint üdvözölnünk kellene az "ember-gép szakadék" esetleges eltörlését az ember-gép hibridek vagy kiborgok megjelenése révén (lásd alább a 4.3. pontot). Reálisabban nézve azonban, még ha eleve jelentőséget is tulajdonítunk az emberi tényezőnek, annak súlya az egyes területeken eltérő lehet: a rákdiagnosztikában talán nem vagy elhanyagolható értéket képvisel, és mindent annak a kérdésnek rendelünk alá, hogy melyik a legmegfelelőbb diagnosztikai módszer, míg a büntetőjogi ítélethozatalban érthetőnek tűnik, hogy jelentőséget tulajdonítunk az emberi tényezőnek, még akkor is, ha ez az ítélethozatal megalapozottságának általános csökkenését eredményezi.

A második kérdés az emberi tényező elvesztésével kapcsolatos aggodalmak enyhítésének módjaival kapcsolatos, miközben az RAI-k továbbra is fontos szerepet kapnak. Minimális esetben úgy gondolnánk, hogy a robotbírák döntései által érintettek számára lehetővé kell tenni, hogy kérhessék a döntések alapjául szolgáló információk nyilvánosságra hozatalát. [14] Ez utóbbi követelmény...

13. A gyilkolásra tervezett robotok hasznos tárgyalását lásd: R. Sparrow, "Killer Robots", *Journal of Ap- plied Philosophy* (242007); 62-77.

14. Ez összhangban van azzal a "magyarázathoz való joggal", amely egyesek szerint az EU általános adatvédelmi rendeletében (2016) szereplő "automatizált döntéshozatal tilalmához való jogból" fakad; de azzal kapcsolatban, hogy szkeptikusan állítják, hogy létezik-e ilyen (megvalósítható) jog, lásd Wachter, Mittelstadt és Floridi (2017).

A RAI-k megalkotása nagyobb kihívást jelent, mint amilyennek látszik, és nem csak azért, mert a kereskedelmi érdekek miatt az algoritmusokat titokba burkolják, hanem azért is, mert a RAI-k alkotói gyakran őszintén elismerik, hogy nem teljesen értik, hogyan képesek az alulról felfelé építkező algoritmusok alapján működő alkotásaik olyan sikeres teljesítményt nyújtani, amilyet nyújtanak. Ezen túlmenően a robotbírák által hozott ítéleteket emberi felügyeletnek és felülbírálatnak is helye lehet. Az álláspályázatok esetében pedig előfordulhat, hogy a pályázatok egy véletlenszerű mintáját emberi munkaerő-felvevők értékelik, így a rendszer nem teljesen mentes a közvetlen emberi beavatkozástól. Mindezek a javaslatok arra az általános következtetésre vezetnek, hogy a robotrepülőgépek értékes szerepet játszhatnak az embereknek *a* fontos döntések meghozatalában, ahelyett, hogy teljesen helyettesítenék őket az ilyen feladatok végrehajtásában. Kétséges azonban, hogy ez mindenre alkalmas megoldást szolgálhat-e, mivel lehetnek olyan esetek, például az AWS-ek által hozott harctéri döntések, ahol az átfogó emberi felügyelet egyszerűen összeegyeztethetetlen az RAI-k által ígért előnyök biztosításával.

Eddig az esetek egy széles kategóriájára összpontosítottam, amelyekben az a tény, hogy egy RAI funkciót lát el, potenciális etikai jelentőséggel bír: azokra az esetekre, amikor olyan döntést hoznak, amely súlyosan érinti az emberi érdekeket. Az emberi tényezővel kapcsolatos aggodalom azonban másfajta esetekben is felmerül, nem utolsósorban azokban, amelyekben az RI-ket partnerként kezelik baráti vagy intim kapcsolatokban. Az emberi tényező hiánya itt a hátborzongás és az undor érzését keltheti. Ennek forrása abban rejlik, hogy egy robotot, amely végső soron egy gép vagy eszköz, bevezetünk egy kapcsolati formába, amelyet eddig az emberek számára tartottunk fenn. Néhányan azt fogják válaszolni, hogy az "emberi tényező" itt zsigeri "undorító tényezővé" fajult, és nem bír nagyobb normatív súllyal, mint a homofóbok vagy a rasszisták ellenszenve az azonos nemű vagy vegyes fajú párokkal szemben. Az ilyen elutasítás azonban túl gyors lenne. Ideális esetben egy romantikus kapcsolatban a felek nem egyszerűen eszközként, hanem öncélként értékelik egymást, részben azáltal, hogy képesek szabadon viszonozni az olyan érzéseket és érzelmeket, mint a szeretet és a ragaszkodás. Ebben az összefüggésben egy emberi alakú robottal való randevú talán nem olyan abszurd vagy furcsa, mint egy iPhone-hoz vagy egy intelligens hűtőszekrényhez való romantikus kötődés, de vitathatatlanul egy kontinuumon van velük.

Képzeljünk el egy olyan társadalmi világot, amelyben az emberek nagy száma az RAI-kat az emberekkel szemben előnyben részesíti barátként és szerelmi partnerként, talán azért, mert a felhasználó saját igényei szerint programozhatók, és hiányzik belőlük az a függetlenség, amely az emberi barátok és szerelmesek esetében súrlódások, csalódások és árulások forrása lehet. De maga a

"felhasználó" szó is rávilágít arra, hogy mi hiányzik egy ilyen kapcsolatból. *etikája felé*[85]

az autonóm emberi lények közötti személyes kapcsolatban potenciálisan elérhető értékhez képest. Egy eszköz olyan szerepet játszik, amely egy személyhez tartozik. Az, hogy az eszköznek bizonyos jó hatásai lehetnek a felhasználó boldogságára vagy lelkiállapotára, nem törli el azt a tényt, hogy az eszköz *pusztán* eszköz marad, és nem személy. Ennek a pontnak a másik oldala az az érzés, hogy aki a robotbarátokat részesíti előnyben az emberi barátokkal szemben, annak nem sikerült igazán felnőnie. Kicsit olyan, mintha egy középkorú férfi még mindig a gyerekkori plüssmackóját kezelné "legjobb barátjaként".

Persze, még ha egyet is értünk azzal, hogy itt valami értékes dolog veszik el, ebből semmi sem következik arról, hogy az embereknek a robotokkal való kapcsolatokat kellene-e választaniuk, vagy hogy az ilyen kapcsolatokat társadalmi erkölcsünknek el kellene-e ítélnie, vagy törvényileg tiltania kellene-e. Végül is lehet, hogy a robottal való kapcsolat az egyetlen reálisan elérhető kapcsolat egyes emberek számára, vagy hogy az emberi tényező elvesztését a kapcsolat egyéb előnyei kompenzálják. De az is lehet, hogy egyszerűen tiszteletben kell tartanunk az emberek szabad döntését, hogy ilyen kapcsolatokba lépnek, még akkor is, ha ezeket a kapcsolatokat mélységes hiányosságoknak ítéljük. De e kérdések bármelyikének megválaszolásához a kiindulópontot az jelenti, ha mélyebben elgondolkodunk valaminek az értékén, amit nagyrészt természetesnek vettünk: az emberi tényező jelenlétén a mindennapi életünkben. [15]

4.3 JOGOK ÉS FELELŐSSÉGEK

Ha egy RAI működőképes, és nincsenek kényszerítő okok a használaton kívül helyezése ellen, akkor felmerül a kérdés, hogy rendelkezik-e olyan erkölcsi státusszal, amely jogokat és felelősségeket ruház rá. Mint a céljaink előmozdítására létrehozott műtárgy, kétségesnek tűnik, hogy egy RAI rendelkezhet-e az ilyen státusz megalapozásához szükséges eredendő értékkel. Ha az RAI-k közel járnának a racionális autonómiára való általános képességünk megismétléséhez, akkor indokolt lenne az emberi lényekhez hasonló erkölcsi státuszt biztosítani számukra, megfelelő jogokkal és felelősségekkel. Valószínű, hogy az elkövetkező években az ember és a gép közötti különbség fokozatosan elmosódik, sőt, a különböző hibridek vagy kiborgok (az emberrel integrált RAI-k) megjelenésével el is tűnik. A "transzhumanizmus" egyik támogatója, Ray Kurzweil szavaival élve:

> *A számítógépek kezdetben nagy, távoli gépek voltak légkondicionált helyiségekben, amelyeket fehér köpenyes technikusok kezeltek. Később átkerültek az asztalunkra, majd a számítógépek alá.*

15. A gyökeres barátkozás lehetősége által felvetett néhány kérdés feltárását lásd Danaher és McArthur szerk. (2017).

> *a karunkban, és most a zsebünkben. Hamarosan rutinszerűen a testünkbe és az agyunkba is beletesszük őket. Végső soron inkább nem-biológiai, mint biológiai emberek leszünk (Kurzweil A teljesebb kifejtésért2002. lásd: Kurzweil 2005).*

Az ilyen fejlemények radikális hatással lehetnek az RAI-kra vonatkozó erkölcsi normák tartalmára, például az önvédelemhez való jogra, amely igazolhatja a fenyegetést jelentő emberek megölését vagy bántalmazását. Amint azonban már láttuk, az általános mesterséges intelligencia nagyon távoli lehetőségnek tűnik, még ha nem is zárható ki teljesen, mint logikus lehetőség.

Sürgetőbb kérdés az, hogy van-e létjogosultsága annak, hogy a RAI-knak jogi személyiséget tulajdonítsanak, a megfelelő jogokkal és felelősségekkel együtt, más, a jog által elismert "mesterséges személyekkel", például a társaságokkal analóg módon.[16] A kérdés nem az, hogy a RAI-kat minden jogi szempontból az emberekkel azonos módon kezeljük-e, mivel a RAI-k jogi személyiségének nem kell pontosan megegyeznie a közönséges emberek vagy "természetes személyek" jogi személyiségével. A jogokat és kötelezettségeket más, és valószínűleg kisebb mértékűnek ítélt jogok és kötelezettségek is alkothatják. És a vonatkozó köteg tartalmilag is változhat a RAI-k egyik fajtájánál (önvezető autók) a másiknál (egészségügyi RAI-k).

Az alulról felfelé irányuló algoritmusokkal működő RAI-k esetében hihetőbbnek tűnik, hogy jogi személyiséget tulajdonítsunk a RAI-knak, mivel viselkedésük nem teljesen kiszámítható. A felülről lefelé irányuló algoritmusok alapján működő RAI-k esetében, amelyek viselkedésüket nagymértékben kiszámíthatóvá teszik, meggyőző érvnek tűnik a gyártók, a tulajdonosok vagy a felhasználók jogi felelősségének tulajdonítása mellett. Ennek megfelelően az Európai Parlament felvetette annak lehetőségét, hogy "hosszú távon különleges jogi státuszt teremtsen a robotok számára, hogy legalább a kifinomultabb autonóm robotok elektronikus személy státuszával rendelkezzenek, akik felelősek az általuk okozott károk megtérítéséért, és esetleg az elektronikus személyiséget alkalmazzák azokra az esetekre, amikor a robotok autonóm döntéseket hoznak vagy más módon önállóan lépnek kapcsolatba harmadik felekkel" (Európai Parlament, 2017,59. bekezdés f) pont). Ez a javaslat a régi angol common law "deodand" fogalmára emlékeztet, amely szerint az állatok vagy dolgok - beleértve az (akkoriban) új technológiákat, például a vonatokat is - elveszíthették a tulajdonjogukat, ha kárt okoztak egy másik személynek, anélkül, hogy a tulajdonosuknak vétkességet tulajdonítottak volna. Az elkobzás eredményeként a szóban forgó dolog értékét arra lehetett felhasználni, hogy kártalanítsanak mindenkit, akinek kárt okoztak. A doktrínát végül

16. Néhány friss vita: Schwitzgebel and Garza (2015); Gunkel (2018); Bryson, Diamantis, and

Grant, (2017); és Turner (2018).

a tizenkilencedik században eltörölték, és helyükbe az újonnan megjelenő technológiák által károsultak kártalanítására szolgáló, hiba nélküli felelősség- és biztosítási rendszerek léptek.

Tekintettel arra a feltételezésre, hogy a RAI-k jogi személyiséggel való felruházása nem az önálló erkölcsi személyiségükön alapulna, azt a kérdést, hogy a RAI-kat jogi személynek kell-e nyilvánítani, az emberi lények és más erkölcsileg megfontolandó lények számára e jogi újítás elfogadásából származó előnyök és terhek általános egyensúlya alapján kell meghatározni (Bryson, Diamantis és Grant, 2017). A kritikusok felhívták a figyelmet a RAI-k jogi személyiséggel való felruházásának különböző buktatóira. Az egyik buktató a visszaélés lehetősége (a 4.5. pontban tárgyalt veszélyek kategóriájába tartozó megfontolás): a RAI-k jogi személyiségét más gátlástalan szereplők, például gyártóik és üzemeltetőik felhasználhatják arra, hogy kikerüljenek a részükről fennálló felelősség alól. Egy másik lehetőség akkor merül fel, ha nincs megfelelő emberi szereplő, aki "állna" az RAI cselekedetei mögött: hogyan fog az RAI kártalanítani másokat a kötelezettségszegéséért? (Bryson, Diamantis és Grant, 2017). A sikeres peres fél kapna-e tulajdonjogot az RAI felett, vagy az RAI-k rendelkezhetnek-e olyan vagyontárgyakkal, pl. bankszámlákkal, amelyeket kártérítésként követelhetnek? Ezek azonban nem perdöntő ellenvetések, és semmiképpen sem zárják ki véglegesen a korlátozott jogi személyiségű formák célszerűségét egyes RAI-k esetében.

Még ha komoly akadályok is állnak a RAI-k jogi személyiséggel való felruházásának útjában, vitatható, hogy az alulról építkező RAI-k kellőképpen különböznek a legtöbb géptől, hogy a hatályos jogot felül kell vizsgálni, hogy figyelembe vegye autonóm viselkedési képességüket. Az UNESCO-jelentés ennek megfelelően egy hármas rendszert javasol: (1) a meglévő jog alkalmazása a felelősség kijelölésében a felülről lefelé irányuló robotokkal kapcsolatban; (2) az alulról felfelé irányuló robotok esetében a jog mellett a gyakorlati kódexek és etikai iránymutatások alkalmazása; és (3) az olyan alulról felfelé irányuló robotok esetében, amelyek kárt okozhatnak az embereknek, pl. az AWS-ek vagy az önvezető autók, annak mérlegelése, hogy milyen mértékben kell a robotra bízni az autonóm döntéseket, és hol van szükség értelmes emberi ellenőrzésre (UNESCO 2017, 48-9). Kétséges azonban, hogy egy ilyen rendezett séma végső soron védhető. A (3) felszólítás plauzibilisnek tűnik, bár meglehetősen homályos, de az (1) téves. Nincs okunk feltételezni, hogy a felülről lefelé irányuló RAI-k esetében csak jogi normákra van szükség, mint ahogyan egy szabványos autó használói számára sem csak jogi felelősségek merülnek fel.

Mindezen erkölcsi és jogi kérdések mögött egy nehéz technikai probléma megoldásának szükségessége áll: hogyan lehet biztosítani a RAI-k "nyomon

követhetőségét" annak érdekében, hogy erkölcsi vagy jogi felelősséget lehessen rájuk hárítani. A nyomon követhetőség magában foglalja, hogy meg tudjuk határozni azokat az okokat, amelyek egy RAI-t arra késztettek, hogy úgy viselkedjen, ahogyan viselkedett, és biztosítani tudjuk a következő feltételeket

különösen nehéznek tűnik az alulról felfelé irányuló algoritmusokat alkalmazó RAI-k esetében. Tekintettel arra, hogy a RAI-k potenciális veszélyt jelentenek az emberi érdekekre, nehéznek tűnik indokolni, hogy létrehozásukat és használatukat egy adott esetben engedélyezzük, ha nincs megfelelő válasz erre a kihívásra.

4.4 MELLÉKHATÁSOK

A RAI-k alkalmazása elkerülhetetlenül jó és rossz mellékhatásokkal jár. Az egyik pozitív mellékhatás például az, hogy az emberek számára nagyobb lehetőséget biztosítanak személyes kapcsolataik fejlesztésére vagy szabadidős tevékenységek végzésére. Másrészről viszont az RAI-k jelentős mértékű munkanélküliséget és végső soron súlyos társadalmi nélkülözést, egyenlőtlenséget és nyugtalanságot okozhatnak. Ezek - helyesen szólva - mellékhatások, mivel még az emberi munkaerő helyettesítésére szolgáló RAI-k működési céljai között sem szerepel a munkanélküliség vagy az ezzel járó társadalmi problémák előidézése. Ezek legfeljebb előre látható, de nem szándékolt következmények.

A közgazdászok nem értenek egyet a RAI-k széles körű használatának a humán foglalkoztatásra gyakorolt lehetséges hatásával kapcsolatban. Egy tanulmány szerint az amerikai munkahelyek akár 50%-át - beleértve az ügyvédek, orvosok és könyvelők által végzett munkákat - is veszélyezteti az automátrix (Frey és Osborne 2013). Az Egyesült Királyságban a következő húsz évben több mint minden harmadik munkahelyet átvehetik a RAI-k, és a hatás aránytalanul nagy mértékben érinti az ismétlődő, alacsony fizetésű munkakörökben dolgozókat (Tovey, 2014). Ezek a fejlemények azonban nem feltétlenül okoznak jelentős munkanélküliséget, ahogyan azt a technológiai innováció története is mutatja: új munkahelyek, amelyekről jelenleg sokszor még csak sejtésünk sincs, keletkezhetnek, részben az új technológiák által elért termelékenységnövekedés eredményeként, és gyakran olyan új igényekre reagálva, amelyeket a technológiai fejlődés maga is segített generálni (Milanovic 2016; Autor 2015). Mások pesszimistább álláspontot képviselnek, különösen, ha arra számítanak, hogy a RAI-k olyan emberfeletti képességekre tesznek szert, amelyek nagyjából minden emberi munkaerőt feleslegessé tesznek (Brynjolfson and McAfee 2014; Drury 2018; Aeon 2018). Bármi is legyen az igazság, valószínű, hogy rövidebb távon sok ember fogja elveszíteni a munkáját a RAI-k miatt, és nagy nehézségekkel kell majd szembenézniük az átképzés során, hogy végül bármilyen új munkakörök - amelyek középpontjában talán az ítélőképességet, kreativitást vagy érzelmi intelligenciát igénylő képességek állnak - létrejöjjenek.

Ezek a lehetőségek arra kényszerítenek bennünket, hogy újraértékeljük a munka értékét az emberi életben. Ez az érték részben a jövedelemszerzéshez

kapcsolódik, ezért az egyik népszerű válasz az RAI által okozott munkanélküliség veszélyére az egyetemes alapjövedelem (UBI) politikája.[93]

a RAI-k alkalmazása révén elért termelékenységnövekedésből finanszírozza (Van Parijs és Vanderborght 2017).[17] Az UBI rendszeres és feltétel nélküli készpénzkifizetést biztosítana mindenki számára, függetlenül attól, hogy dolgozik-e, vagy teljesít-e további feltételeket, például hogy fogyatékossággal él-e vagy aktívan munkát keres-e. A munka azonban potenciálisan a jövedelemtermelésen kívül más értékeket is szolgál: fontos forrása a megelégedettségnek és az önértékelésnek, a felelősség és az önfegyelem erényeit erősíti, és az értékes társadalmi szerepvállalás középpontjába állítja az embereket. Szükségünk lesz-e arra, hogy valamilyen módon korlátozzuk a RAI-k gazdasági életünkbe való behatolását annak érdekében, hogy megőrizzük az emberek megfelelő hozzáférését ezekhez az értékekhez? Talán a RAI-kat elsősorban olyan eszközként kellene használnunk, amelyek inkább segítik, mintsem helyettesítik az emberi munkaerőt. Önpusztító lenne-e a RAI-k használatának korlátozása az emberi teljesítmények szférájának megőrzése érdekében, amennyiben ezeket az "eredményeket" beárnyékolná az a tudat, hogy a RAI-k egyenértékű, vagy akár sokkal jobb munkát is végezhetnének? Vagy vannak-e megvalósítható, sőt talán még előnyösebb módok arra, hogy ezeket az értékeket más tevékenységeken keresztül valósítsuk meg, például a családi élet, a művészet, a vallásgyakorlás vagy a sport révén? Vagy egy olyan világban, amely megszabadult az emberi munka szükségességétől, felfedezhetnénk egy teljesen új értékrendet, amely értelmet adna az életünknek?

A munkanélküliség csak az egyik aggasztó lehetséges mellékhatása a RAI-nak. Egy diffúzabb, nagyjából hasonló jellegű aggodalom az, hogy a RAI-kra való túlzott támaszkodás értékes készségek elsorvadásához és a saját életünkért és döntéseinkért való felelősségérzetünk csökkenéséhez vezethet. Az orvosok, sofőrök és pilóták például elveszíthetik azokat a készségeket, amelyekre szükségük van ahhoz, hogy vészhelyzetekben jól teljesítsenek; az átlagemberek pedig túlzottan rászorulhatnak a RAI-kra, amikor olyan mindennapi döntéseket hoznak, mint az étel, amit esznek, az újság, amit olvasnak, vagy a politikai pártok, amelyekre szavaznak a választásokon. Ráadásul minél jobban támaszkodunk a RAI-kra a döntések meghozatalában, annál inkább felfedezhetjük, hogy életünket egyre inkább olyan szempontok alakítják, amelyekre az automatizált döntéshozatal a legérzékenyebb. Ez nem feltétlenül jelenti azt, hogy az elért "eredmények" rosszabbak lennének azoknál, amelyek akkor születtek volna, ha a megfontolások szélesebb körét vettük volna figyelembe, de azt jelenti, hogy életünk körvonalai egyre inkább az RAI-k képességeinek függvényei lesznek, ahelyett, hogy a kiemelkedő értékmegfontolások által lehetővé tett összes útvonalat átgondolnánk.[18]

Vannak más komoly aggodalmak is, beleértve a lehetséges maró hatásokat a

17. További lehetséges válaszok közé tartozik a bértámogatás és a garantált kormányzati foglalkoztatás; e három stratégia megvitatását lásd: Furman és Seamans 21-252018,.

18. Az ilyen jellegű mellékhatások megvitatását a jogalkalmazás "egyenjogúbb" és "kodifikáltabb"

formái közötti elmozdulásra ösztönző mesterséges intelligenciával kapcsolatban lásd Re and Solow-Niederman (2019).

a többi emberrel való kapcsolataink minősége. Minél inkább a vágyaink kiszolgálására kialakított - gyakran antropomorf formákkal és hangokkal felruházott - gépekkel való interakciók körül forog az életünk, annál inkább fennáll a veszélye annak, hogy engedünk a kísértésnek, hogy ugyanezt az instrumentalizáló hozzáállást embertársainkra is kiterjesszük. Ez az aggodalom különösen élesen jelentkezik a társas kapcsolatokra vagy szexre használt RAI-kkal kapcsolatban. Mások, mint láttuk, azt állítják, hogy üdvözölnünk kellene az "ember-gép szakadék" végleges eltörlését, egy olyan folyamatot, amelyet az ember-gép hibridek vagy kiborgok megjelenése felgyorsít.

4.5 FENYEGETÉSEK

A kifejezetten rosszindulatú célok - például a magánéletet sértő megfigyelés, pénzügyi csalás vagy terrortámadás - megvalósítására létrehozott RAI-k komoly veszélyt jelenthetnek érdekeinkre és értékeinkre. Az ilyen RAI-k nem funkcionálisak a fentiekben meghatározott értelemben (4.1.). Veszélyt jelenthet azonban az is, ha a hasznos feladatok elvégzésére tervezett RAI-kat szabotálják vagy aláássák működésüket - például ha algoritmusaikat szándékosan hamis vagy korrupt adatokkal táplálják, vagy ha rosszindulatú ügynökök feltörik őket. Egy olyan világ, amelyben az Ön okostelefonja kémkedik Ön után, vagy az AWS-ek terroristák kezébe kerülnek, aligha távoli lehetőség. A fenyegetések pedig nemcsak bűnözőktől, terrorista csoportoktól vagy vállalatoktól származnak, hanem talán mindenekelőtt a kormányoktól, amelyek gyakran más jellegű csoportokkal együttműködve dolgoznak. A RAI-k önkényuralmi eszközként való felhasználásának legfrissebb, szemléletes példája a kínai kormány által létrehozott Social Credit System, amelynek keretében az egyének a róluk gyűjtött adatok alapján "állampolgári pontszámot" kapnak, amely alapján meghatározzák, hogy jogosultak-e munkahelyre, külföldi utazásra és egyéb juttatásokra (*The Economist* 2016).

Az a válasz, hogy ez a fajta kormányzati fenyegetés nem igazán vonatkozik a de- mokráciákra, nem meggyőző. Ez durván alábecsüli azt, hogy a demokráciákban a politikusok és a bürokraták milyen módon használhatják ki a széles körben elterjedt félelmeket, például a terrorizmussal vagy a bevándorlással kapcsolatban, hogy elnyomó intézkedéseket hozzanak, és meghosszabbítsák hatalmukat. Egy másik nagy aggodalomra ad okot a "megelőző" vagy "biztosításmatematikai" megközelítések térnyerése a rendőri tevékenységben, amelyek során a mesterséges intelligenciát a jövőbeli bűncselekmények előrejelzésére használják, és a leendő bűnözőket a valószínűsíthetően elkövetett bűncselekmények alapján tartóztatják le, nem pedig a ténylegesen elkövetett bűncselekmények alapján, a polgári

szabadságjogokat fenyegető kockázatokkal együtt, amelyeket a *Minority Report* című filmben szemléletesen ábrázoltak (lásd például Ferguson 2017). Sőt, különösen alattomos jelenség ebben az összefüggésben az a mód, ahogyan egyes gyanúsítottak

a kormányzati megfigyelés és a vállalati adatgyűjtés formái összefonódnak, amint azt Edward Snowden drámai módon megmutatta, amikor felfedte, hogy az NSA hozzáfér a Google, a Facebook és a Microsoft által gyűjtött adatokhoz.[19]

Az egyik legsúlyosabb fenyegetés, amelyet a RAI-k jelentenek, magának a demokráciának a megfelelő működése. [20] A demokrácia nem csak azt követeli meg, hogy minden polgárnak legyen szavazati joga, hanem azt is, hogy a szavazatát szabad és tájékozott mérlegelés és vita után gyakorolhassa a szóban forgó kérdésekről. Szabad információáramlásra van szükség ahhoz, hogy a demokratikus mérlegelés lehetővé tegye a politikaformálás alakítását, és biztosítsa a tisztségviselők elszámoltathatóságát. Az elmúlt években aggályok merültek fel azzal kapcsolatban, hogy a RAI-ket e demokratikus folyamatok megvalósítására használják. Az e célból alkalmazott módszerek közé tartozott az egyének mikrocélú megcélzása személyre szabott politikai hirdetésekkel, amelyek a közösségi médiaplatformokról, például a Facebookról illegálisan begyűjtött adatokon alapulnak, vagy a Twitter és más platformok propagandával való telítése érdekében embernek álcázott robotok - "botok" - használata, illetve gyakorlatilag észrevehetetlen audiovizuális hamisítványok készítése. A veszélyt itt nem egyszerűen az adatok tiltott származása jelenti, és még csak nem is az, hogy az elkészített üzenetek és képek megtévesztőek vagy manipulatívak lehetnek, hanem az, ahogyan az ilyen tevékenységek hozzájárulnak a választók különböző kategóriái számára különböző információs "univerzumok" kialakításához, és ezáltal a demokratikus mérlegeléshez nélkülözhetetlen közös nyilvánosságot erodálják (Bartlett 2018). Ami a demokráciát fenyegető veszély nagyságát illeti, Onora O'Neill, éles figyelmeztetést adott ki:

> *A nem megtévesztés az egyik alapvető kötelesség. Amikor a technológiára gondolok, azon tűnődöm, hogy vajon lesz-e demokrácia évek20 múlva, mert ha nem találunk megoldást erre a problémára, akkor nem lesz. Az emberek olyan üzeneteket és tartalmakat kapnak, amelyeket robotok terjesztenek, nem pedig más emberek, nemhogy más polgártársak. Ez ijesztő (O'Neill 2018; lásd még Helbing et al 2017).*

A demokráciát fenyegető kockázatok jelentősnek tűnnek, ha az RAI lehetséges mellékhatásaival együtt - amelyek aláássák a polgártársaink iránti személyes felelősségérzetünket és az emberiség különlegességének érzését -, a demokráciát fenyegető kockázatok is jelentősnek tűnnek.

A demokráciát fenyegető veszély kezelésének módjai közé tartozik a technológia-specifikus mea...

19. A vitát, beleértve azt is, hogy a kormányzati szervek gyakran megkerülik az adatgyűjtésre

vonatkozó jogi korlátozásokat azáltal, hogy megvásárolják, követelik vagy feltörik a vállalati ügynökök birtokában lévő adatokat, lásd Pasquale 48-512015,. o. *eti*

kája felé99

20. A digitális technológiáknak a demokráciára gyakorolt pozitív és negatív hatásairól lásd Susskind (2018).

Kötet kiadás 7,1

a személyes adatok fokozott védelmét, az online platformszolgáltatók általi adatfelhasználás nagyobb átláthatóságát, valamint a közösségi médiafiókok szigorúbb regisztrációs eljárásait. O'Neill még a kínai kormány által gyakorolt internetcenzúrához hasonló internetes cenzúra lehetőségét is felveti, ami rosszabb gyógymód lehet, mint a betegség, amelyet kezelni kíván. Fontos azonban, hogy foglalkozzunk politikai rendszereink olyan strukturális jellemzőivel is, amelyek a RAI-kkal kölcsönhatásban a demokrácia aláásása érdekében. Például az Egyesült Államok laza kampányfinanszírozási törvényei megkönnyítik, hogy a forrásokat az "álhírek" széles körű terjesztésére fordítsák. Általánosabban fogalmazva, egy rejtélyes problémával állunk szemben. A RAI-k által jelentett legtöbb társadalmi fenyegetés elhárításához vitathatatlanul a fokozott demokratikus elszámoltathatóságra van szükség; ez azonban versenyt jelenthet számunkra az idővel, mivel az egyik legsúlyosabb fenyegetés éppen a demokratikus folyamatok működését fenyegeti. A RAI-k által felvetett problémákra demokratikus megoldásokra van szükségünk, mielőtt még magát a demokráciát is tönkretennék.

Természetesen egyesek szerint a RAI-k által jelentett legnagyobb veszély az, amiből ez a cikk kiindult: hogy olyannyira intelligensebbek lesznek az embereknél, hogy végül leigáznak vagy kiirtanak minket saját céljaik érdekében. Ezt a végveszélyes forgatókönyvet hangsúlyozták a RAI-k olyan prominens személyiségei, mint Bill Gates és Elon Musk, valamint vezető tudósok, köztük a néhai Stephen Hawking, aki élete vége felé megfigyelte:

> *Egy szuperintelligens mesterséges intelligencia rendkívül jól fogja elérni a céljait, és ha ezek a célok nem egyeznek a miénkkel, akkor bajban vagyunk... Valószínűleg nem vagy gonosz hangyagyűlölő, aki rosszindulatból lép a hangyákra, de ha te vagy a felelős egy vízerőműves zöld energia projektért, és a térségben van egy hangyaboly, amit el kell árasztani, akkor kár a hangyákért. Ne hozzuk az emberiséget azoknak a hangyáknak a helyzetébe (Griffin 2015).*

Néhányan felelőtlenül spekulatívnak minősítették az ilyen figyelmeztetéseket, arra hivatkozva, hogy az általános mesterséges intelligencia, nem is beszélve az emberfeletti mesterséges intelligenciáról, nem reális lehetőség a belátható jövőben. Ahogy Daniel Dennett (2019) fogalmazott: "eszközöket készítünk, nem kollégákat". E nézet szerint a világvége-forgatókönyvek olyan fantáziák, amelyek elterelik a figyelmünket más, sürgős, az AI-val kapcsolatos problémákról, amelyekkel szembe kell néznünk. Egy másik válasz azonban arra a feltételezésre összpontosít, amely Hawking figyelmeztetésének alapjául szolgál. Miért feltételezzük, hogy a szuperintelligens RAI-k céljai aggasztó módon nem lesznek összhangban a mi céljainkkal? Ha a RAI-k valóban emberfeletti

képességeket fejlesztenek ki, akkor ezek közé nem fog tartozni az a képesség is, hogy gondolkodjanak, és

az erkölcsnek? [21] Más szóval, az "intelligencia" fogalmának elszegényedése, ha azt az összetett célok elérésének képességére korlátozzuk, függetlenül azok erkölcsi értékétől. Képzeljünk el tehát egy olyan világot, amelyben igazságos és jóindulatú RAI-k irányítanak bennünket, amelyek intelligenciájukban és jóságukban messze felülmúlnak minden embert. Vajon ez a forgatókönyv a RÁI-k emberiséget szolgáló ígéretének végső beteljesülése, vagy érdekeink és értékeink mélységes elárulása? Nyilvánvaló, hogy az utóbbi következtetés mellett szól. Részben arról van szó, hogy olyan elveken és megfontolásokon alapuló kormányzási formáknak vagyunk alávetve, amelyek potenciálisan meghaladják a legtöbb, vagy talán minden emberi lény felfogóképességét, ami jelentős eltérést jelent a felvilágosodás felfogásától, amely szerint az uralom olyan normák szerint történik, amelyeket az alávetettek racionálisan felfoghatnak és elfogadhatnak.[22] Végül is, tekintve, hogy az emberi erkölcs az emberi helyzetünkben rejlő lehetőségekre és korlátokra van hangolva, miért kellene feltételezni, hogy a szuperintelligens RAI-k, akik nem osztoznak az emberi természetben, hajlandóak lennének nagy súlyt fektetni bármire, amit mi erkölcsi megfontolásokként ismerhetünk el? De ezt az aggodalmat félretéve, az egyik legfontosabb érték, amelyet az állítólagos igazságos és jóindulatú RAI-kormányzóknak feltehetően el kellene ismerniük, az emberi szabadság, nemcsak az egyes emberek személyes életválasztásainak szintjén, hanem a közösségi önrendelkezésüket gyakorló emberi lények csoportjainak szintjén is. Nehéz elképzelni, hogy uralmukkal hogyan kerülhetnék el, hogy ne ássák alá súlyosan ezt a szabadságot.[23] E megfontolások fényében talán kiderülne, hogy a szuper-RAI-k egy jóindulatú fajának meglenne a kegye, hogy az emberekre bízza a saját útjukat, olyan minimális normák betartása mellett, amelyek elkerülik az emberi hibák pusztítóbb megnyilvánulásait.

5. KÖVETKEZTETÉS

A RAI-k etikájára vonatkozó szilárd megközelítés kidolgozásának feladata több szinten is működik, beleértve a jogi szabályozást, a társadalmi erkölcsöt és a személyes erkölcsi normákat, amelyek összetett kölcsönhatásban állnak egymással. Az e három területen felmerülő legfontosabb elsőrendű kérdések

21. A RAI-k értékeit illetően az ellenkező, a nem-alignációs tézis részletesebb védelmét lásd Bostrom (2014).

22. Ezt hangsúlyozza Kissinger (2018): "A legnehezebb, de legfontosabb kérdés a világgal kapcsolatban, amely felé tartunk, ez: Mi lesz az emberi tudattal, ha saját magyarázó erejét felülmúlja a mesterséges intelligencia, és a társadalmak már nem képesek értelmezni az általuk lakott világot a számukra értelmes módon?".

23. Egy további, inkább spekulatív megfontolás, hogy a nem-emberek általi uralom támadás lenne a fajhoz való hűség vagy a fajjal való azonosulás ellen, lásd Williams 149-1522006,.

Első lépések a robotok és a mesterséges intelligencia

eti

kája felé103 *Kötet kiadás 7,1*

Az általam javasolt szintek nagyrészt a FIRST séma szerint gyűjthetők össze, a funkcionalitás, az eredendő jelentőség, a jogok és felelősségek, a mellékhatások és a veszélyek rubrikáinak megfelelően. Ugyanakkor kétségbe vontam azt az elképzelést, hogy a létező erkölcsi vagy jogi elveknek van olyan hasznos szegmense, amely elsősorban vagy kizárólag az RAI-k etikájára vonatkozik. Ehelyett az RAI-kat úgy kell tekinteni, mint amelyek potenciálisan az emberi értékek teljes skáláját érintik a FIRST séma által azonosított öt rubrikában. Bár már mind az öt rubrika alapján végeztek munkát, az RAI-k etikájának átgondolása még korai szakaszban van. Ráadásul az emberi tényező eredendő jelentősége olyan kérdés, amely érthetetlen okokból még nem kapott olyan mértékű figyelmet, mint amilyet érdemelne. Meg kell küzdenünk magának a jelentőségnek az elképzelésével, és azzal, hogy az erkölcsi érték különböző formáit nyeri el, olyan jellemzőktől függően, mint a döntéshozatal területe (pl. rákdiagnózis vagy büntetőítélet) vagy a kapcsolat formája (pl. ügyvéd vagy szerető), amelyről szó van. Az ezen öt rubrika alatt felmerülő kérdések megválaszolásakor alapvető fontosságú, hogy a RAI-k meglévő és előrelátható jövőbeli képességeinek reális értékelése vezéreljen és korlátozzon, és ne engedjük, hogy etikai gondolkodásunkat utópisztikus (vagy disztópikus) spekulációkkal eltérítsék, amelyek olyan lehetőségeken alapulnak, amelyek a legjobb esetben is a távoli jövőben vannak, még ha nem is esnek szigorúan a tudományos és technológiai lehetőségek körén kívül.

KÖSZÖNETNYILVÁNÍTÁS

Hálás vagyok Roger Brownswordnak, Rebecca Lowe-nak, Claudia Chwalisznak, Francesca Rossinak, Jose Suchnak, David Nelken-nek, a KCL/Queen's Ontario jogfilozófiai kollokvium résztvevőinek, és - különösen - Hannah Maslen-nek, Annette Zimmermann-nak és egy névtelen bírálónak a korábbi tervezetekhez fűzött hasznos megjegyzéseikért. A tanulmány elkészítését az Élet Jövője Intézet ösztöndíja tette lehetővé. Hálás vagyok továbbá Napoleon Xanthoulisnak a kiváló kutatási segítségért. A cikk rövidített változata Tasioulas (megjelenés előtt) címen jelenik meg.

HIVATKOZÁSOK

Aeon. (2018). "Az embereknek nem kell jelentkezniük." (videó) *Aeon*, <https://aeon.co/videos/the-robots-are- coming-for-our-jobs-why-the-human-workforce-is-at-risk> [Hozzáférés: 132019. június].

Asilomar AI alapelvek. (2017). <https://futureoflife.org/ai-principles/> [Hozzáférés: június13 2019].

Asimov, I. (1950). "Runaround.". I, *Robot*. New York City: Doubleday.

Autor, D. (2015). "Miért van még mindig olyan sok munkahely? A munkahelyek története és jövője

Automatizálás." *Journal of Economic Perspectives (*293): 30.

Barocas, S. és Selbst, A. (2016). "A Big Data egyenlőtlen hatása". *California Law Review* 104: 670-732.

Bartlett, J. (2018). *Az emberek a technika ellen: Hogyan öli meg az internet a demokráciát (és hogyan menthetjük meg)*.

London: Penguin.

Binns, R. (2018). "Méltányosság a gépi tanulásban: Lessons from Political Philosophy." *Journal of Machine Learning Research* 81: 1-11 <https://arxiv.org/abs/1712.03586> [Hozzáférés: 132019. június].

Bostrom, N. (2014). *Szuperintelligencia: Útvonalak, veszélyek, stratégiák*. Oxford: Oxford University

Press. Bonnefon, J.-F., Shariff, A. és Rahwan, I. (2006). "The Social Dilemma of Autonomous Járművek." *Science* 352: 1573-1576. <http://science.sciencemag.org/content/352/6293/1573> [Accessed on 13 2019. június].

Bryson, J, Diamantis, M. és Grant, T. (2017). "A népből, a népért és a nép által: The Legal Lacuna of Synthetic Persons". *Mesterséges intelligencia és jog* 25: 273-29.

Buolamwini, J. és Gebru, T. (2018). "Nemi árnyalatok: Interszekcionális pontossági egyenlőtlenségek a kereskedelmi nemek osztályozásában". *Proceedings of Machine Learning Research* 81: 1-15. <http://proceed- ings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf> [Hozzáférés: 132019. június].

Buranyi, S. (2018). "Dehumanizáló, áthatolhatatlan és frusztráló: a munkakeresés zord valósága a mesterséges intelligencia korában". *Guardian* március 4. <https://www.theguardian.com/inequality/2018/mar/04/ dehumanising-impenetrable-frustrating-the-grim-reality-of-job-hunting-in-the-age-of-ai> [Hozzáférés: 132019. június].

Brynjolfson, E. és McAfee, A. (2014). *A második gépkorszak: Munka, haladás és jólét a briliáns technológiák korában*. New York City, NY: WW Norton & Co.

Burgess, M. (2017). "Az emberekhez hasonlóan a mesterséges intelligencia is lehet szexista és rasszista". *Wired* április 13. <http://www.wired.co.uk/article/machine-learning-bias-prejudice> [Hozzáférés: 2019. június 13.].

Coeckelbergh, M. (2015). "Mesterséges ágensek, jó gondoskodás és modernitás". *Theoretical Medicine and Bioethics* 36: 265-277.

Danaher, J. és McArthur, N. szerk. (2017), *Robotszex: Social and Ethical Implications*.

Cambridge MA: MIT Press.

Dennett, D.C. (2010). "Vajon az AI eléri a tudatosságot? Téves kérdés." *Wired* (Febr. 9.) <https:// www.wired.com/story/will-ai-achieve-consciousness-wrong-question/> [Hozzáférés: 132019. június]

Első lépések a robotok és a mesterséges intelligencia

eti

kája felé107 *Kötet kiadás 7,1*

Drury, C. (2018). "Mark Carney arra figyelmeztet, hogy a robotok elveszik a munkahelyeket, ami a marxizmus felemelkedéséhez vezethet". *The Independent* április <https://www.independent.co.uk/news/uk/home-news/mark-carney-marxism- 14automation-bank-of-england-governor-job-losses-capitalism-a8304706.html> [Hozzáférés: 132019. június].

Edmonds, D. (2017). "Taníthatunk-e etikát a robotoknak?" *BBC News*, 2017. október 15. <http://www. bbc.co.uk/news/magazine-41504285> [Hozzáférés: 132019. június 15.].

Európai Bizottság (2019). *Etikai iránymutatások a megbízható mesterséges intelligenciához.* Európai Bizottság: Brüsszel. <https://ec.europa.eu/futurium/en/ai-alliance-consultation> [Hozzáférés: 2019. június 13.].

Európai Parlament (2017). Jelentés a Bizottságnak szóló ajánlásokkal a robotikára vonatkozó polgári jogi szabályokról, Jan 27. <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP// NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN> [Hozzáférés: 132019. június].

Ferguson, A.G. (2017). *A nagy adatmennyiségű rendfenntartás felemelkedése*. New York City, NY: New York University Press.

Frey, C.B. és Osborne, M.A. (2013). "A foglalkoztatás jövője: Mennyire érzékenyek a munkahelyek a számítógépesítésre. " < https://www. oxfordmartin. ox. ac. uk/downloads/academic/The_Future_of_ Employment.pdf> [Hozzáférés: 132019. június].

Fry, H. (2018). *Hello World: Hogyan legyünk emberek a gépek korában*. New York: Doubleday.

Furman, J. és Seamans, R. (2018). "AI and the Economy." <https://papers.ssrn.com/sol3/papers. cfm?abstract_id=3186591> [Hozzáférés: 132019. június].

Goode, L. (2018). "Az arcfelismerő szoftver elfogult a fehér férfiakkal szemben, állapította meg egy kutató". *The Verge* (Feb 11). <https://www.theverge.com/2018/2/11/17001218/facial-recognition-software- accuracy-technology-mit-white-men-black-women-error> [Hozzáférés: 132019. június].

Griffin, A. (2015). "Stephen Hawking: Hawking: A mesterséges intelligencia eltörölheti az emberiséget, ha túl okos lesz, mivel az emberek olyanok lesznek, mint a hangyák." *The Independent* (október 8.) <https://www.indepen- dent.co.uk/life-style/gadgets-and-tech/news/stephen-hawking-artificial-intelligence-could-wipe-out- humanity-when-it-gets-too-clever-as-humans-a6686496.html> [Hozzáférés: 132019. június].

Gunkel, D. (2018). *Robotjogok*. Cambridge, MA: MIT Press.

Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R.V., and Zwitter, A. (2017). "Túléli-e a demokrácia a Big Data-t és a mesterséges intelligenciát?" (Will Democracy Survive Big Data and Artificial Intelligence?) *Scientific American* (Feb 25 <https://www.scientificamerican.com/article/will-democracy-survive-big- data-and-artificial-intelligence/?redirect=1> [Hozzáférés: 132019. június].

Lordok Háza AI bizottsága (2018). A *mesterséges intelligencia az Egyesült Királyságban; készen, hajlandó és képes?* House of Lords Paper

100. <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm> [Hozzáférés: 132019. június].

Első lépések a robotok és a mesterséges intelligencia etikája felé109

Gyakorlati etikai folyóirat

Jordan, M. (2018). "Mesterséges intelligencia - A forradalom még nem történt meg". *Medium* április 18,<https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-.
5e1d5812e1e7> [Hozzáférés: 132019. június].

Kissinger, H.A. (2018). "Hogyan ér véget a felvilágosodás". *The Atlantic* (június) <https://www.the- atlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-histo- ry/559124/> [Hozzáférés: 132019. június].

Kleinberg, J., Ludwig, J., Mullainathan, S., és Sunstein, C. (megjelenés előtt). "Discrimination in the Age of Algorithms." <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329669> [Hozzáférés: 132019. június].

Kurzweil, R. (2002). "We Are Becoming Cyborgs." <http://www.kurzweilai.net/we-are-becoming-cyborgs> [Hozzáférés: 132019. június].

--- (2005). *A szingularitás közel van: Amikor az ember túllép a biológián*. London: Duckworth.

Larson, J., Mattu, S., Kirchner, L. és Angwin, J. (2016). "Hogyan elemeztük a COMPAS Visszaesési algoritmus." *ProPublica* May <https://www.propublica.org/article/how-we-analyzed-23the-compas-recidivism-algorithm> [Hozzáférés: 132019. június].

Milanovic, B. (2016). "Három tévhit, ami miatt félni kell a robotgazdaságtól". *Evonomics*, Sept. 1. <http://evonomics.com/three-fallacies-robot-economy-branko/> [Hozzáférés: 132019. június].

Nemitz, P. (2018). "Alkotmányos demokrácia és technológia a mesterséges intelligencia korában". *Philosophical Transactions Royal Society A (*3762133).

Nyholm, S. (2018). "Az önvezető autókkal bekövetkező balesetek etikája: A Road Map, I." *Filozófiai iránytű* 17.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin.

---(2018). "Vajon lesz-e demokrácia 20 év múlva?". Interjú Elena Cuéval. *Huffington Post* május 1. <https://www.huffingtonpost.com/entry/onora-oneill-i-wonder-wheth- er-we-will-have-democracy_us_5a4f8a12e4b0cd114bdb324f> [Hozzáférés: 2019. június 13.].

Oswald M., Grace, J., Urwin, S., and Barnes, G.C. (2018). "Algoritmikus kockázatértékelési rendészeti modellek: Lessons from the Durham HART Model and 'Experimental' Proportionality (A durhami HART-modell és a "kísérleti" arányosság tanulságai)". *Információs és kommunikációs technológiai jog (*272018): 223.

Pasquale, F. (2015). *A fekete doboz társadalom: A pénzt és az információt irányító titkos algoritmusok*. Cambridge MA: Harvard University Press, 2015.

Pessoa, L. (2018). "A robotok megismeréséhez olyan gépekre van szükség, amelyek egyszerre gondolkodnak és éreznek". *Aeon* április 13.
<https://aeon.co/ideas/robot-cognition-requires-machines-that-both-think-and-feel> [Hozzáférés: 2019. június 13.].

Re, R.M. és Solow-Niederman, A. (megjelenés előtt). "A mesterségesen intelligens igazságszolgáltatás fejlesztése".

Stanford Technology Law Review 22.

Rini, R. (2017). "Jó robotok nevelése". *Aeon* April <https://aeon.co/essays/creating-robots-18capable-of-moral-reasoning-is-like-parenting> [Hozzáférés: 132019. június].

Savulescu, J. és Maslen, H. (2015). "Az erkölcsfejlesztés és a mesterséges intelligencia: erkölcsi mesterséges intelligencia?" *Beyond Artificial Intelligence (A mesterséges intelligencián túl): The Disappearing Human-Machine Divide*. Eds. Romportl, J., Zackova, E. és Kelemen, J. Svájc: Springer International Publishing.

Schwitzgebel, E. és M. Garza, M. (2015). "A mesterséges intelligenciák jogainak védelme".

Középnyugati filozófiai tanulmányok 39.

Sparrow, R. (2007). "Gyilkos robotok." *Journal of Applied Philosophy* 24: 62-77.

Sunstein, C. (megjelenés előtt). "Algoritmusok, az előítéletek korrigálása".

Social Research.

Susskind, J. (2018). A *jövő politikája: Együtt élni a technika által átalakított világban*. Oxford: Oxford University Press.

Tasioulas, J. (2019). "AI and Robot Ethics." *Ethics and the Contemporary World*. Ed. Edmonds, D. London: Routledge.

---(hamarosan). "Az emberi jogok megmentése az emberi jogi jogtól". *Vanderbilt Journal of Transnational Law*.

Tegmark, M. (2017). *Élet 3.0: Embernek lenni a mesterséges intelligencia korában*. London: Penguin.

The Economist. (2016). "Big Data, Meet Big Brother: Kína feltalálja a digitális totalitárius államot". *The Economist* december 17. <https://www.economist.com/briefing/2016/12/17/china-invents-the-digital-totalitarian-state> [Hozzáférés: 132019. június].

Tovey, A. (2014). "Tízmillió munkahelyet veszélyeztet a fejlődő technológia". *Telegraph* (november 10.)

<https://www.telegraph.co.uk/finance/newsbysector/industry/11219688/Ten-million-jobs-at-risk-from-advancing-technology.html> [Hozzáférés: 132019. június].

Turner, J. (2018). *Robotszabályok*. London: Palgrave MacMillan.

UNESCO (2017). A COMEST jelentése a robotika etikájáról. Párizs, UNESCO. <http://unesdoc. unesco.org/images/0025/002539/253952E.pdf> [Hozzáférés: 132019. június].

Upchurch, T. (2018). "A társadalomért való munkához az adattudósoknak hippokratészi esküre van szükségük fogakkal". *Wired*, április 8. <http://www.wired.co.uk/article/data-ai-ethics-hippocratic-oath-cathy-o- neil-weapons-of-math-destruction> [Hozzáférés: 132019. június].

Van Parijs, P. és Vanderborght, Y. (2017). *Alapjövedelem: Radikális javaslat a szabad társadalomért és a józan gazdaságért*. Cambridge, MA: Harvard University Press.

Vayena, E. és Tasioulas, J. (2016). "A nagy adatok és az emberi jogok dinamikája: A tudományos kutatás esete". *Philosophical Transactions of the Royal Society A* 28: 374.

Wachter, S. Mittelstadt, B. és Floridi, L. (2017). "Miért nem létezik az általános adatvédelmi rendeletben az automatizált döntéshozatal magyarázatához való jog". *International Data Privacy Law* 7: 76-99.

Wiggins, D. (2016). "Azonosság, szubsztancia és az emberi személy". *Continuants: Tevékenységük, létük és identitásuk - Tizenkét esszé*. Oxford: Oxford Unversity Press.

Williams, B. (2006). "Az emberi előítélet". A *filozófia mint humanista diszciplína*. Princeton: Princeton University Press.