

# STIFTUNG FÜR EFFEKTIVEN ALTRUISMUS

## Mesterséges intelligencia: lehetőségek és kockázatok

---

### Vitairat

A mesterséges intelligencia (AI) és az egyre összetettebb algoritmusok minden eddigénél jobban befolyásolják életünket és civilizációnkat. A mesterséges intelligencia alkalmazási területei sokrétűek, a lehetőségek pedig messzemenőek: különösen a számítógépes hardverek fejlődésének köszönhetően egyes mesterséges intelligencia-algoritmusok már ma is felülmúlják az emberi szakértők teljesítményét. Alkalmazási területük a jövőben tovább fog növekedni, és az AI teljesítménye javulni fog. Konkrétan az várható, hogy a megfelelő algoritmusok egyre inkább optimalizálják magukat - emberfeletti szintre. Ez a technológiai fejlődés valószínűleg történelmileg példátlan etikai kihívások elé állít bennünket. Jó néhány szakértő úgy véli, hogy a mesterséges intelligencia a globális lehetőségek mellett globális kockázatokat is rejt magában, amelyek meghaladják majd például a nanotechnológia lehetőségeit - amelyet történelmileg szintén sokáig alábecsültek. A tudományos kockázatelemzés azt is sugallja, hogy a magas potenciális kárszinteket akkor is nagyon komolyan kell venni, ha a bekövetkezés valószínűsége alacsony.

2015. november 12.

## A Hatékony Altruizmusért Alapítvány vitairata.

Előnyben részesített hivatkozás: Mannino, A., Althaus, D., Erhardt, J., Gloor, L., Hutter, A. és Metzinger, T. (2015). Mesterséges intelligencia: lehetőségek és kockázatok. A Hatékony Altruizmusért Alapítvány vitairatai. (2): 1-17.

Első közzététel: 2015. november 12.

[www.ea-stiftung.org](http://www.ea-stiftung.org)

# Tartalomjegyzék

Összefoglaló .....	1
Bevezetés.....	3
A jelenlegi mesterséges intelligenciák előnyei és kockázatai	3
Automatizálás és munkanélküliség .....	5
Általános intelligencia és szuperintelligencia.....	7
Mesterséges tudat .....	10
Összefoglaló .....	11
Visszaigazolás .....	12
Támogatók .....	12
Irodalom .....	13

**Adriano Mannino**, filozófus és társelnök, Foundation for Effective Altruism (Alapítvány a Hatékony Altruizmusért)  
**David Althaus**, tudományos munkatárs, Foundation for Effective Altruism (Alapítvány a Hatékony Altruizmusért)  
**Dr. Jonathan Erhardt**, tudományos munkatárs, Foundation for Effective Altruism (Alapítvány a Hatékony Altruizmusért)  
**Lukas Gloor**, tudományos munkatárs, Foundation for Effective Altruism (Alapítvány a Hatékony Altruizmusért)  
**Dr. Adrian Hutter**, Fizika Tanszék, University of Basel  
**Prof. Thomas Metzinger**, filozófiaprofesszor, Mainzi Egyetem





# Mesterséges intelligencia: lehetőségek és kockázatok

## Összefoglaló

A mesterséges intelligencia (AI) és az egyre összetettebb algoritmusok minden eddiginél jobban befolyásolják életünket és civilizációnkat. A mesterséges intelligencia alkalmazási területei sokrétűek, a lehetőségek pedig messzemenőek: különösen a számítógépes hardverek fejlődésének köszönhetően egyes mesterséges intelligencia-algoritmusok már ma is felülmúlják az emberi szakértők teljesítményét. Alkalmazási területük a jövőben tovább fog növekedni, és az AI teljesítménye javulni fog. Konkrétan az várható, hogy a megfelelő algoritmusok egyre inkább optimalizálják magukat - emberfeletti szintre. Ez a technológiai fejlődés valószínűleg történelmileg példátlan etikai kihívások elé állít bennünket. Jó néhány szakértő úgy véli, hogy a mesterséges intelligencia a globális lehetőségek mellett globális kockázatokat is rejt magában, amelyek hasonlóak például a nukleáris technológiához - amelyet történelmileg szintén sokáig alábecsültek.

- meghaladja. A tudományos kockázatelemzés azt is sugallja, hogy a nagy potenciális károkat még akkor is nagyon komolyan kell venni, ha a bekövetkezés valószínűsége alacsony.

## Jelenlegi

A mesterséges intelligenciák emberekkel szembeni fölénye már bebizonyosodott szűk, jól tesztelt alkalmazási területeken (pl. önvezető autók és az orvosi diagnosztika részterületei). E technológiák fokozott alkalmazása nagy lehetőségeket rejt magában (pl. jelentősen kevesebb baleset a közúti közlekedésben, kevesebb hiba a betegek orvosi kezelése során, vagy számos újfajta terápia feltalálása). Az összetettebb rendszerekben, ahol több algoritmus nagy sebességgel lép kölcsönhatásba egymással (pl. a pénzügyi piacon vagy az előrelátható katonai alkalmazásokban), megnő a kockázata annak, hogy az új mesterséges intelligencia technológiák váratlanul rendszerszintű hibát okoznak vagy visszaélnak velük. Fennáll a veszélye egy olyan mesterséges intelligencia-fegyverkezési versenynek, amely feláldozza a technológiai fejlődés biztonságát annak ütemének. Mindenesetre a lényeges kérdés az, hogy milyen célokat vagy etikai értékeket kell beprogramozni egy mesterséges intelligencia algoritmusba, és hogyan lehet technikailag garantálni, hogy a célok stabilak maradjanak és ne lehessen őket manipulálni. Az önvezető autók esetében például felmerül a kérdés, hogyan döntse el az algoritmus, ha egy több gyalogossal való ütközés csak úgy kerülhető el, hogy az autó egyetlen utasa kerüljön veszélybe - és hogyan biztosítható, hogy az önvezető autók algoritmusai ne hibázzanak rendszerszinten.

**intézkedés** A mesterséges intelligencia témájáról szóló tényszerű és racionális diskurzus előmozdítása szükséges ahhoz, hogy az előítéletek csökkenjenek, és a hangsúlyt a legfontosabb és legsürgetőbb biztonsági kérdésekre lehessen helyezni.

A mesterséges intelligencia kutatásának fejlődése lehetővé teszi, hogy egyre több emberi munkát végezzenek el a gépek. **intézkedés** A jogi keretet az új technológiákhoz kell igazítani. A mesterséges intelligencia gyártóit kötelezni kell arra, hogy többet fektessenek be a technológiák biztonságába és megbízhatóságába, és tartsák be az olyan elveket, mint a kiszámíthatóság, az átláthatóság és a nem manipulálhatóság, hogy a váratlan katasztrófák kockázata minimálisra csökkenjen.





## Bevezetés

A tudás keresése végigvonul az emberiség történelmében. Amikor a társadalmak szerkezetükben és dinamikájukban jelentős változásokon mentek keresztül, ez a legtöbb esetben az új technológiai találmányoknak is köszönhető volt. A kőszerszámok első használata és a fejlődési "nagy lépés" között, amikor a Homo sapiens feltalálta a művészetet és elkezdte festeni a barlangfalakat, körülbelül kétmillió év telt el. Néhány tízezer évbe telt a földművelés és a letelepedés. Az első szimbólumok néhány ezer évvel később jelentek meg, és később alakultak ki az első írások. A mikroszkópot a 17. században találták fel. A 19. századi iparosodás tette lehetővé az első több millió lakosú városok létrejöttét. Alig egy évszázaddal később az atomot felbontották, és az emberek elrepültek a Holdra. Feltalálták a számítógépet, és azóta a számítógépek számítási teljesítménye és energiahatékonyasága rendszeres időközönként megduplázódott [1]. A technológiai fejlődés gyakran exponenciálisan fejlődik. Az emberi szellemi képességek esetében ez nem így van.

Az elmúlt év során számos neves tudós és vállalat hangsúlyozta a mesterséges intelligencia kérdésének sürgető fontosságát, és azt, hogy mennyire fontos, hogy a döntéshozók foglalkozzanak a mesterséges intelligencia kutatásának előrejelzéseivel [2]. A mesterséges intelligencia biztonsági mozgalmának képviselői közé tartozik Stuart Russell [3], Nick Bostrom [4], Stephen Hawking [5], Sam Harris [6], Max Tegmark [7], Elon Musk [8], Jann Tallinn [9] és Bill Gates [10].

Bizonyos tárgykörökben (azaz *szakterület-specifikusan*) a mesterséges intelligenciák már többször elérték vagy akár meg is haladták az emberi szintet.

## A jelenlegi mesterséges intelligenciák előnyei és kockázatai

Életünket és civilizációnkat egyre inkább az algoritmusok és a terület-specifikus mesterséges intelligenciák (AI) befolyásolják és uralják.

[19]: Gondoljunk csak az okostelefonokra, a légi közlekedésre [20] vagy az internetes keresőmotorokra [21]. A pénzügyi piacok is egyre bonyolultabb algoritmusokra támaszkodnak, amelyeket egyre kevésbé értünk [22, 23]. A legtöbbször az ilyen algoritmusok használata incidensek nélkül zajlik, de mindig fennáll annak a lehetősége, hogy egy

2011-ben Watson legyőzte a két legjobb emberi játékost a Jeopardy! [12], 2015-ben pedig az első pókerváltozatot - *fix limit holdem heads-up* - teljesen játékelméletileg oldották meg a *Cepheus* [13] segítségével. A mesterséges neurális hálózatok ma már képesek felvenni a versenyt az emberi szakértőkkel a rákos sejtek diagnosztizálásában [14], és a kézzel írt kínai karakterek felismerésében is megközelítik az emberi szintet [15]. Már 1994-ben egy öntanuló backgammon program a világ legjobb játékosainak szintjét érte el azzal, hogy olyan stratégiákat talált, amelyeket ember még soha nem használt [16]. Időközben már olyan algoritmusok is léteznek, amelyek képesek a semmiből megtanulni különböző számítógépes játékokat, és elérik az (emberfeletti) emberi szintet [17, 18]. Ez azt jelenti, hogy lassan közeledünk egy olyan *általános intelligenciához*, amely - legalábbis elvileg - képes bármilyen jellegű probléma önálló megoldására.

A nagyobb hatalom nagyobb felelősséggel jár. A technológia csupán egy eszköz; az számít, hogyan használjuk. Még a meglévő mesterséges intelligenciák alkalmazása is jelentős etikai kihívások elé állít bennünket, amelyeket e vitairat következő részében ismertetünk. Az ezt követő fejezet olyan fejleményeket tárgyal, amelyek arra utalnak, hogy a mesterséges intelligencia kutatásának előrehaladása középtávon olyan mértékben fogja ösztönözni a gazdasági automatizálást, hogy az a munkaerőpiac jelentős átrendeződését fogja eredményezni. Az utolsó két fejezet a mesterséges intelligencia kutatásának hosszú távú és lehetséges kockázatait tárgyalja az (emberfeletti) intelligencia és a mesterséges tudat lehetséges létrehozásával kapcsolatban.

valószínűtlen *fekete hattyú esemény* [24] következik be, amely azzal fenyeget, hogy az egész rendszert káoszba taszítja. 2010-ben például sokkoló tőzsdei összeomlás következett be az Egyesült Államokban [25], mivel a számítógépes algoritmusok előre nem látható módon léptek kölcsönhatásba a pénzügyi piaccal [26]. Perceken belül a jelentős részvények értékük több mint 90%-át elvesztették, majd visszaemelkedtek a kiindulási értékükre. Katonai alkalmazásokban a valószínűsége, hogy egy

Nagyobb a valószínűsége annak, hogy ez a "visszatérés az alaphelyzethez" nem következik be[27]. Az ilyen jellegű pusztító hibák megelőzése érdekében általában véve tanácsosnak tűnik, hogy jelentősen többet fektessünk a mesterséges intelligenciák biztonságába és megbízhatóságába. Sajnos jelenleg gazdasági ösztönzők vannak arra, hogy a mesterséges intelligencia teljesítményének javítását előnyben részesítsék a mesterséges intelligencia biztonságával szemben.

#### A mesterséges intelligencia létrehozásának négy kritériuma

A biztonság minden gép esetében alapvető fontosságú, de a terület-specifikus mesterséges intelligenciák megalkotása új etikai kihívásokkal jár, amint átveszik a korábban emberek által végzett, szociális dimenzióval rendelkező kognitív munkát. Vegyünk például egy olyan algoritmust, amely értékeli a banki ügyfelek hitelképességét, és (anélkül, hogy kifejezetten erre lenne programozva) diszkriminatív döntéseket hoz bizonyos népcsoportokkal szemben. Még azok a technológiák is érdekes kihívások elé állíthatják a gépi etikát, amelyek alapvetően csak a meglévő tevékenységeket helyettesítik [28]: Az önvezető járművek például felvetik a kérdést, hogy milyen kritériumok alapján kellene döntenie egy közlekedési baleset esetén. Például a járműveknek a legnagyobb prioritást az utasok túlélésének kell-e adniuk, vagy elkerülhetetlen baleset esetén az áldozatok számát a lehető legalacsonyabb szinten kell tartani [29]?

Ezért Eliezer Yudkowsky, a mesterséges intelligencia teoretikusa és Nick Bostrom filozófus négy alapvető javasolt, amelyeknek az új mesterséges intelligenciák megalkotásakor irányadónak kell lenniük [30]: A mesterséges intelligencia működésének 1) *érthetőnek* kell lennie, és 2) cselekedeteinek *elvileg kiszámíthatónak* kell lennie; mindkettőnek olyan időablakon belül, amely elegendő mozgásteret biztosít a felelős szakértők számára a reagálásra és a vétőellenőrzésre egy esetleges meghibásodás esetén. Ezenkívül a mesterséges intelligenciát 3) nem szabad könnyen *manipulálni*, és ha baleset történik, 4) *a felelősséget* egyértelműen meg kell határozni.

#### A (szakterület-specifikus) mesterséges intelligencia előnyei

Az algoritmusok és a terület-specifikus mesterséges intelligenciák alapvetően számos előnnyel járnak. Pozitívan befolyásolták az életünket, és a jövőben is így fognak tenni, feltéve, hogy megtesszük a szükséges

ővintézkedéseket. Az alábbiakban két tanulságos példát tárgyalunk.

Az önvezető autók már régen nem tudományos fantasztikum többé [31, 32], és belátható időn belül az önvezető autók is

kereskedelmi forgalomban kapható legyen: A *Google* által kifejlesztett, mesterséges intelligencia algoritmusok által teljesen autonóm módon irányított *Google Driverless Car* már 2011-ben megkezdte első tesztvezetéseit az Egyesült Államokban [33, 34]. A munkára vagy pihenésre fordítható időmegtakarítás mellett az önvezető autók másik előnye a megnövekedett biztonság. 2010-ben például világszerte 1,24 millió ember halt meg közúti balesetben, szinte kizárólag emberi hiba miatt [35]. Évente sok életet lehetne tehát megmenteni, mivel az önvezető autók már most is bizonyíthatóan biztonságosabbak, mint az emberek által vezetett járművek [36, 37].

Az emberek túlságosan nagy része azonban még mindig szkeptikus az önvezető autókkal kapcsolatban, valószínűleg azért, mert túlbecsülik a kockázatokat és saját vezetési képességeiket. Egy tanulmány például arra a következtetésre jutott, hogy az amerikai sofőrök 93%-a úgy véli, hogy általában jobb vezetési képességekkel rendelkezik, mint az átlag [38] - ami statisztikailag lehetetlen. Irreális optimizmus [39] és a kontroll illúziója [40] valószínűleg azt is eredményezik, hogy az emberek alábecsülik a vezetés kockázatát [41, 42].

Az orvosok is túlbecsülik képességeiket [43], ami végzetes hibákhoz vezethet. Csak az Egyesült Államokban becslések szerint évente 44 000 és 98 000 ember hal meg kórházakban kezelési hibák miatt [44]. Ebben az összefüggésben üdvözlendő az IBM által kifejlesztett Watson mesterséges intelligencia [45], amely 2011-ben legyőzte a legjobb emberi játékosokat a *Jeopardy!* kvízműsorban, és ezzel hírnevet szerzett [12]. Watson nem csak a kvízműsorokban jobb az embernél: 2013 óta a kórházak is alkalmazhatják Watsont például a rák diagnosztizálására. Mivel "Doktor Watson" nagyon rövid idő alatt hatalmas mennyiségű információt képes befogadni és kombinálni, a diagnosztika terén részben felülmúlja az emberi kollégákat [46, 47].

Meglepőnek tűnhet, hogy a jelenlegi mesterséges intelligencia pontosabb betegségdiagnózisokat tud felállítani, mint az emberi orvosok. Régóta ismert azonban, hogy a *statisztikai érvelés* általában jobb, mint a klinikai érvelés, azaz az emberi szakértők ítéletei [48, 49]. És persze az olyan mesterséges intelligenciák, mint Watson, statisztikai következtetésekre készültek. Következésképpen a számítógépek diagnózishoz való (nem) igénybe vétele emberi életek között dönthet.

### Kognitív torzulások - Tévedni emberi dolog

Az egyik ok, amiért az emberi szakértők kevésbé kompetensek a statisztikai megítélésben, mint a mesterséges intelligenciák, az a fent említett, túlságosan is emberi tendencia, hogy túlbecsülik saját képességeiket. Ezt a tendenciát nevezzük *túlzott bizalmi torzításnak* [50]. A *túlzott magabiztossági torzítás* csak egy a számos kognitív torzítás közül [51, 52], amelyek szisztematikusan félrevezethetik az emberi gondolkodást. A mesterséges intelligenciákat viszont úgy lehet megtervezni, hogy ne legyenek kognitív elfogultságaik. Elvileg a mesterséges intelligenciák előrejelzéseibe vetett bizalom növekedése - feltéve, hogy azok megbízhatóak és érthető kritériumok alapján készültek - számos társadalmi és politikai kihívás esetében is a racionalitás jelentős növekedéséhez vezethet. A probléma itt az lenne, hogy a mesterséges intelligencia erősségeit úgy használjuk ki, hogy közben ne adjuk fel az emberi cselekvési autonómiát a megfelelő rendszereknek.

### Összefoglaló és kilátások

Az új, alapvetően előnyös technológiákkal szembeni irracionális félelmek még mindig széles körben elterjedtek [53].

**ajánlás - Felelős használat:** Mint minden más technológia esetében, a mesterséges intelligencia kutatása során is nagy figyelmet kell fordítani arra, hogy a (lehetséges) előnyök egyértelműen meghaladják a (lehetséges) hátrányokat. A tényszerű-rationális diskurzus előmozdítása szükséges ahhoz, hogy az irracionális előítéletek és félelmek csökkenjenek, és az elavult jogi kereteket az új technológiáknak megfelelően meg lehessen reformálni. A mesterséges intelligenciák nagyszabású alkalmazása során a fentiekben vázolt négy alapelvet kell betartani [30]. ■

### Automatizálás és munkanélküliség

A gépi tanulás és a robotika területén az elmúlt években elért sikereket tekintve úgy tűnik, csak idő kérdése, hogy a magas intelligenciát igénylő, összetett munkákat is átfogóan átvegyék a gépek [56].

Ha a gépek számos feladatot gyorsabban, megbízhatóbban és olcsóbban tudnak elvégezni, mint az emberi munkaerő, az messzemenő hatással lesz a munkaerőpiacra. Olyan közgazdászok, mint Cowen [57], McAfee és Brynjolfsson [58] azt jósolják, hogy a technológiai fejlődés még jobban ki fogja nyitni a jövedelmi különbségeket, és nagymértékű bércsökkenés és a munkanélküliség tömeges növekedése várható.

Egy 2013-ban közzétett elemzés szerint az Egyesült Államokban az összes munkahely 47%-a nagy valószínűséggel automatizálható lesz 10-20 éven belül [59]. A címen.

Ez a technofóbia lehet az egyik oka annak is, hogy a Watson vagy az önvezető autók iránt szkeptikusan viszonyulnak. Az új technológiákkal kapcsolatos aggodalmak azonban nem mindig irracionálisak. A legtöbb technológia az emberiség javára használható, de veszélyes is lehet, ha rossz kezekbe kerül, vagy ha nem fordítanak kellő figyelmet a biztonságra és a nem kívánt mellékhatásokra.

Hasonló a helyzet a mesterséges intelligenciával: az önvezérelt autók megkönnyíthetik az életünket és életeteket menthetnek, de az összetett számítógépes algoritmusok összeomolhatnak a tőzsdén is. Bár a közeljövőben a legtöbb terület-specifikus mesterséges intelligencia viszonylag egyszerű és biztonságos kialakítású lehet, a hosszú távú fejlesztéseket figyelembe kell venni: A nem túl távoli jövőben a mesterséges intelligencia akár egzisztenciális fenyegetést is jelenthet, hasonlóan a biotechnológiához (pl. új típusú vírusok lehetséges szintézise révén) [54, 55, 4].

A legnehezebben automatizálható tevékenységek azok, amelyek magas szociális intelligenciát (pl. PR-tanácsadás), kreativitást (pl. divattervezés) vagy érzékeny és rugalmas mozdulatokat (pl. sebészet) igényelnek. Ezek a területeken a mesterséges intelligencia kutatásának állása még mindig messze van az emberi szakértők szintjétől.

### A számítógépes automatizálás előnyei és hátrányai

Különösen azok az emberek és országok fognak profitálni a technológiai fejlődésből, amelyek tudják, hogyan használják ki az új technológiai lehetőségeket és az ezzel járó adatárdatot (*big data*) [60]. Ezek különösen a jól képzett számítógépes szakemberekkel rendelkező országok. Emellett a jövőben egyre fontosabbá válik, hogy az emberek pontos képet kapjanak a különböző számítógépes algoritmusok előnyeiről és hátrányairól a



döntéshozatal és munkateljesítmény, amelyhez a jó oktatás központi szerepet játszik [61].

A szórakoztatóiparban is lesznek mélyreható újítások: A jobb grafika, az új szórakoztató technológiák és az egyre olcsóbbá váló mobileszközök új képességei révén a videojátékok és az internet-hozzáférés addiktív hatása is növekszik [62]. Ennek a fejlődésnek a társadalmi és pszichológiai következményei még kevéssé ismertek, de vannak arra utaló jelek, hogy ezek a tendenciák tartósan megváltoztatják a szociális viselkedésünket [63], a figyelmünket és a gyermekek felnövekedésének módját [64]. A belátható jövőben, amikor a kifinomult virtuális valóságokat a nem tudósok is megtapasztalják majd, és egyre mélyebben behatolnak az életünkbe, ez a hatás sokkal hangsúlyosabbá válhat. A virtuális valóságokban való gyakori elmerülés, vagy az olyan eljárások, mint a teljes test illúziók, amelyek során a szubjektív érzetet átmenetileg egy virtuális avatárra vetítik [65], valószínűleg jelentős hatást gyakorolnak.

Végül az oktatás területén a szórakoztatóipar nagy lehetőségeket kínál a tanulási tartalmak játékosítása révén [66]; ugyanakkor fennáll a veszélye annak, hogy nő azon serdülők aránya, akik a kóros videojáték- vagy internethasználat miatt kirekesztődnek az oktatásból. [67] nehézségekbe ütközik az oktatás befejezése.

### Utópiák és disztópiák

A technológiai fejlődés növeli a társadalom termelékenységét [68], ami növeli az átlagos életszínvonalat [69]. Amikor egyre több munkát végeznek el a gépek, ez teret enged az emberek számára a szabadidő eltöltésének és az önmegvalósításnak - legalábbis azok számára, akik képesek ezt kihasználni. A növekvő automatizálás hátránya azonban az lehet, hogy a termelékenység növekedése a társadalmi egyenlőtlenségek növekedésével jár együtt, így az *átlagos* életszínvonal növekedése nem esik egybe a *medián* életminőség növekedésével. Az olyan szakértők, mint például Erik Brynjolfsson, az MIT közgazdászprofesszora, különböző okokból [70] attól tartanak, hogy a technológiai fejlődés azzal fenyeget, hogy az emberek többségének helyzete romlik.

Egy olyan globális versenygazdaságban, ahol a mesterséges intelligencia technológiája olyannyira előrehaladott, hogy számos tevékenységet már a

Ha a munkát gépek is el tudják végezni, akkor az automatizálható emberi munkaerő bére csökken [58]. Szabályozás nélkül a bérszint sok ember számára a létminimum alá süllyedhet. A társadalmi egyenlőtlenségek meredeken növekedhetnek, ha a gazdasági teljesítmény nő, de bérfizetés nélkül nincs újraelosztás. Ennek a fejleménynek az ellensúlyozására McAfee és Brynjolfsson azt javasolja, hogy az emberek által végzett bizonyos tevékenységeket támogatni lehetne. A technológiai fejlődés előnyeinek a lakosság egésze számára történő elosztásának további módjai a feltétel nélküli alapjövedelem és a negatív jövedelemadó [71, 72].

Egyes szakértők olyan jövőbeli forgatókönyvekre is figyelmeztetnek, amelyekben a változások még súlyosabbak lesznek. Robin Hanson közgazdász például úgy véli, hogy még ebben az évszázadban lehetővé válik az emberi agy szimulációjának, az úgynevezett *teljes agyi emulációnak (WBE)* [73] a virtuális valóságban történő digitális futtatása. A WBE-ket meg lehetne sokszorozni, és ha elegendő hardver áll rendelkezésre, akkor a biológiai agynál sokszor gyorsabban működhetnének - ami a munka hatékonyságának óriási növekedését eredményezné [74]. Hanson azt jósolja, hogy ilyen esetben "népességrobbanás" következne be a WBE-k körében, mivel számos területen rendkívül költséghatékony munkavállalóként lehetne őket alkalmazni [75]. Hanson feltételezései ellentmondásokkal [61], és nem szabad azt feltételezni, hogy a *legvalószínűbb* jövőt vázolják fel. Jelenleg a kutatások - például a lausanne-i EPF *Blue Brain Project* - még messze vannak az első agyszimulációtól, nemhogy valós idejű (vagy akár felgyorsított) virtuális valóság-inputokat tudnának biztosítani számukra. Mindazonáltal fontos, hogy figyelemmel kísérjük a hardverek fejlődését a WBE-k lehetőségét illetően. Ha a Hanson által felvázolt forgatókönyv megvalósulna, akkor ez etikai szempontból nagy jelentőséggel bírna: Egyrészt a komplex szimulációk által helyettesített emberek közül sokan munkanélkülivé válhatnak. Másrészt felmerül a kérdés, hogy az alkalmazott WBE-k milyen körülmények között rendelkeznének fenomenális tudattal és szubjektív preferenciákkal, azaz, hogy szenvedést is éreznének-e (esetleg kényszerű) munkavégzésük során.

**2. ajánlás - előrelátó cselekvés:** Az éghajlatváltozás kérdéséhez hasonlóan a kutatók és a döntéshozók számára is ösztönzőket kell teremteni a mesterséges intelligencia jövőbeli forgatókönyveinek kezelésére. Ez megteremtheti az elővigyázatossági intézkedések alapját. Különösen a mesterséges intelligencia hatásvizsgálata és a biztonság területén kellene megfelelő szakértői konferenciákat tartani, szakértői bizottságokat létrehozni és kutatási projekteket finanszírozni. ■

**ajánlás - Oktatás:** Az oktatás tartalmának célzott kiigazítása segíthetne jobban felkészíteni az embereket az új kihívásokra. A számítógépes és programozási ismeretek például erősen felértékelődnek, míg a memorizált tudás veszít értékéből. A tanulási tartalmak játékosítása nagy lehetőségeket rejt magában, amelyeket támogatni kell. Az internet társadalmi és pszichológiai hatásait tovább kell vizsgálni, és meg kell akadályozni a videojátékok és az online média káros fogyasztását. ■

**ajánlás - Nyitottság az új intézkedésekre:** A növekvő automatizálás negatív hatásainak társadalmi tompítására a humán munkaerő támogatása, a feltétel nélküli alapjövedelem és a negatív jövedelemadó mint lehetséges intézkedések kerültek felvetésre. Tisztázni kell, hogy milyen egyéb lehetőségek léteznek, és melyik intézkedéscsomag a leghatékonyabb. Ennek érdekében az előnyöket és hátrányokat szisztematikusan elemezni és politikai szinten megvitatni kell. A folyamat során felmerülő empirikus kérdések megválaszolásához finanszírozást kell fordítani. ■

## Általános intelligencia és szuperintelligencia

Az "általános intelligencia" egy szereplő azon képességét méri, hogy ismeretlen környezetek széles körében képes-e elérni céljait [76, 77]. Az ilyen típusú hírszerzés (katasztrófa)kockázatot jelenthet, ha a szereplő céljai nem egyeznek a miénkkel. Amikor az általános intelligencia eléri az emberfeletti szintet, *szuperintelligenciának* nevezzük: a szuperintelligencia minden tekintetben felülmúlja az emberi intelligenciát, beleértve a tudományos kreativitást, a józan ész és a szociális kompetenciát. A szuperintelligencia e definíciója nyitva hagyja, hogy a szuperintelligencia rendelkezik-e tudattal vagy sem [78, 79].

### Az általános mesterséges intelligencia összehasonlító előnyei az emberrel szemben

Az emberek intelligens kétlábú "biorobotok", amelyek tudatos önmodellel rendelkeznek, és amelyeket az evolúció több milliárd év alatt hozott létre. Ezt a tényt érvként használták [80, 81, 82], hogy a mesterséges intelligencia létrehozása nem lehet túl nehéz, mivel a mesterséges intelligencia kutatása sokkal gyorsabban és céltudatosabban haladhat, szemben az evolúcióval, amely csak lassú és céltalan, pazarló generációs lépésekben halad előre. Amellett, hogy az evolúció biztosítja az *AI koncepciójának igazolását*, természetesen lehetővé teszi a célzott humán kutatást is, hogy a biológiai

tervezni és ennek megfelelően gyorsabban haladni előre.

Az ember biológiai agyához képest a számítógépes hardver számos előnyt kínál [4, 60. o.]: Az alapelemek (a modern mikroprocesszorok) több milliószor gyorsabban "tüzelnek", mint a neuronok; a jelek több milliószor gyorsabban továbbítódnak; és egy számítógép lényegesen több alapelemből állhat - a szuperszámítógépek akár egy gyárcsarnok méretűek is lehetnek. A jövő digitális intelligenciája a szoftverkomponensek tekintetében is nagy előnyökkel rendelkezik a biológiai agyhoz képest [4, 60-61. o.]: A szoftverek például könnyen szerkeszthetők vagy megsokszorozhatók, így a terv előnyei többféleképpen is felhasználhatók. A mesterséges intelligenciát nagy adatbázisokkal lehet ellátni, így a potenciálisan releváns információk bármikor előhívhatók.

Néhány fontos területen, mint például az energiahatékonyság, a tisztán fizikai sérülésekkel szembeni ellenálló képesség és a *kíméletes degradáció* [83], a mesterséges hardver még mindig elmarad az emberi agytól. Különösen a termodinamikai hatékonyság és a komplexitás csökkentése között nincs közvetlen összefüggés az információfeldolgozás szintjén [84, 85]. Az elkövetkező évtizedekben azonban a számítógépes hardverek folyamatosan fejlődni fognak. Tekintettel a fent említett komparatív előnyökre és a

A hardver [86] és a szoftverek előre jelzett gyors fejlődésével valószínűnek tűnik, hogy az emberi intelligenciát egy napon a gépek megelőzik. Pontosan ki kell deríteni vagy meg kell becsülni, hogy ez hogyan és mikor következhet be, és milyen következményei lehetnek egy ilyen forgatókönyvnek.

### Időhorizontok

A mesterséges intelligencia különböző szakértői foglalkoztak azzal a kérdéssel, hogy mikor éri el az első gép az emberi intelligencia szintjét. A száz legsikeresebb mesterséges intelligencia-szakértő körében végzett, idézettségi indexszel mért felmérés szerint e szakértők többsége valószínűnek tartja, hogy ez már az évszázad első felében bekövetkezik [4, 19. o.]. A szakértők többsége azt is feltételezi, hogy az ember egy napon szuperintelligenciát hoz létre, ha a technológiai fejlődés (globális katasztrófák következtében) nem szenved komoly visszaesést [4, 20. o.]. Az időbecslések között nagy a szórás: egyes szakértők nagyon biztosak abban, hogy legkésőbb 2040-re legalább emberi intelligenciával rendelkező gépek lesznek, míg (néhány) másik szerint ezt a szintet soha nem éri el. Még ha némileg konzervatívabb feltételezésekkel is élünk, mert figyelembe akarjuk venni, hogy az emberi szakértők hajlamosak túlságosan biztosak lenni becsléseikben [87, 88], akkor is teljesen helytelen lenne a szuperintelligencia témáját a "science fiction" kategóriájába sorolni: Még a konzervatív feltételezések is azt sugallják, hogy nem elhanyagolható annak a valószínűsége, hogy ebben az évszázadban emberi intelligenciaszintű mesterséges intelligenciát fejlesztenek ki.

### Az általános mesterséges intelligencia céljai

Racionális szereplőként egy mesterséges intelligencia pontosan arra törekszik, amit a céljai/célfüggvénye mond [89]. Az, hogy egy mesterséges intelligencia *etikusan* fog-cselekedni, azaz lesznek-e olyan céljai, amelyek nem ütköznek az emberek és más szenvedésre képes lények érdekeivel, teljesen nyitott kérdés: A mesterséges intelligencia minden lehetséges célt követhet [90]. Hibás antropomorfizmus lenne azt feltételezni, hogy bármilyen szuperintelligencia úgy törődne az etikai kérdésekkel, mint a (tipikus) emberek. Amikor mesterséges intelligenciát építünk, explicit vagy implicit módon meghatározzuk a célját is.

Néha ezeket az igényeket olyan kritikával illetik, hogy minden olyan kísérlet, amely a mesterséges intelligencia célját az emberi értékek szerint próbálja meghatározni, egyenlő a "rabszolgasorba taszítással", mivel a mi emberi értékeinket *kényszerítenénk rá a* mesterséges intelligenciára [91]. Ez a kritika azonban félreértéseken alapul. Az "előírja" kifejezés azt sugallja, hogy egy bizonyos, "valódi" cél már létezik, amelyet egy mesterséges intelligencia már a létrehozása előtt elérhet. Ez az elképzelés azonban képtelenség: nincs "szellem a gépezetben", nincs olyan cél, amely független lenne a szereplőt létrehozó folyamatoktól. Az intelligenciát létrehozó folyamat szükségszerűen meghatározza az intelligencia működését és céljait. Ha szuperintelligenciát akarunk építeni, akkor annak (fő) céljaiért mi, és semmi/senki más nem felelős. Továbbá az sem igaz, hogy a mesterséges intelligenciának bármilyen módon ártani kellene a célok miatt, amelyeket elkerülhetetlenül adunk neki. (Az etikailag releváns értelemben vett sérülés lehetősége is feltételezi a tudatosság meglétét - ez az előfeltétel egy szuperintelligencia esetében sem kell, hogy teljesüljön.) Egészen analógikusan, *volens nolens* alakítjuk ki az általunk létrehozott biológiai gyermekek - azaz biológiai intelligenciák - értékeit vagy céljait. Természetesen ez nem jelenti azt, hogy a gyermekek ezáltal etikátlan módon "rabszolgasorba" kerülnek. Éppen ellenkezőleg, erős etikai kötelességünk, hogy alapvető etikai értékeket adjunk át biológiai gyermekeinknek. Ugyanez vonatkozik minden általunk létrehozott mesterséges intelligenciára.

Stuart Russell informatikus professzor hangsúlyozza [3], hogy az etikai célok programozása nagy kihívás, mind technikai szinten (Hogyan lehet az összetett célokat úgy megragadni egy programozási nyelvben, hogy ne szülessenek nem kívánt eredmények?), mind etikai, morálfilozófiai szinten (Mely célok valójában?). A Russell által említett első problémát a szakirodalomban *értékterhelési problémának* is nevezik [92].

Bár egy szuperintelligencia lehetséges céljainak tárháza hatalmas, néhány megbízható kijelentést tehetünk a cselekvéseiről. Számos instrumentálisan racionális köztes cél létezik, amelyek a legkülönbözőbb végső célokkal rendelkező szereplők számára hasznosak. Ezek közé tartozik a cél- és önfenntartás, az intelligenciafejlesztés, a tudásfejlesztés és a fizikai erőforrások felhalmozása [93]. Ha egy mesterséges intelligencia célja megváltozik, az ugyanolyan negatív (vagy még negatívabb) hatással lehet az eredeti cél elérésére, mintha felbomlana.

megzavarták volna. Az intelligencia növelése azért fontos, mert nem jelent mást, mint a célok elérésére való képesség növelését változó környezetben - ezért van lehetőség az úgynevezett *intelligencia-robbanásra*, amikor egy mesterséges intelligencia rekurzív önfejlesztés révén rövid idő alatt nagyfokú intelligenciára tesz szert [94, 95]. (A rekurzív önfejlesztés alapötletét először I. J. Good fogalmazta meg [96]; időközben konkrét algoritmusokat is kifejlesztettek [97]). Az erőforrások felhalmozása és az új technológiák feltalálása nagyobb hatalmat ad a mesterséges intelligenciának, ami szintén a célok jobb elérését szolgálja. Ha egy újonnan keletkezett szuperintelligencia célfunkciója nem tulajdonít értéket a szenvedésre képes lények jólétének, akkor ott, ahol az (inter-) intelligenciája számára szükséges, a szenvedésre képes lények jólétét is figyelembe venné.

)cél elérése hasznos lenne, kíméletlenül halált és szenvedést okozna.

Hajlamosak lehetnénk azt feltételezni, hogy egy szuperintelligencia nem jelent veszélyt, mert ez csak egy számítógép, amely szó szerint kihúzhatja a dugót. A *definíció szerint* azonban egy szuperintelligencia nem lenne ostoba: ha fennáll a veszélye annak, hogy kihúzzák a dugót, akkor először a teremtők által kívánt módon viselkedne, amíg ki nem találja, hogyan minimalizálhatja az önkéntelen kikapcsolás kockázatát [4, 117. o.]. Egy szuperintelligencia képes lehet arra is, hogy eddig ismeretlen biztonsági rések (úgynevezett *nulladik napi exploitok*) segítségével kijátssza a nagy bankok és a kézfegyver-arszenálok biztonsági rendszereit, és így zsarolja és együttműködésre kényszerítse a világ lakosságát. Amint azt már az elején említettük, a "visszatérés az alapszintre" már nem biztos, hogy lehetséges.

### Mi a tét

A legjobb esetben egy szuperintelligencia megoldhatná az emberiség számos problémáját, azaz segítene megoldani a nagy tudományos, etikai, ökológiai és gazdasági problémákat.

**ajánlás - Tájékoztató:** A mesterséges intelligencia biztonságának hatékony javítása a mesterséges intelligencia szakértőinek, a befektetőknek és a döntéshozóknak az oktatásával kezdődik. A mesterséges intelligencia fejlődésével kapcsolatos kockázatokról szóló információkat könnyen hozzáférhetővé kell tenni. Az ügyet támogató szervezetek közé tartozik az Oxfordi Egyetemen működő Future of Humanity Institute (FHI), a Berkeley-i Machine Intelligence Research Institute (MIRI), a bostoni Future of Life Institute (FLI) és a német nyelvű országokban a Foundation for Effective Altruism (EAS).

mikai kihívások a jövőben. Ha azonban egy szuperintelligencia céljai nem esnek egybe a mi preferenciáinkkal vagy az összes érző lény preferenciáival, akkor egzisztenciális fenyegetéssé válik, és esetleg több szenvedést okozhat, mint ami nélküle valaha is létezett volna [98].

### Racionális kockázatkezelés

Azokban a döntési helyzetekben, ahol a tét potenciálisan nagyon nagy, a következő elvek fontosak:

1. A drága óvintézkedések meghozatala még alacsony kockázati valószínűség esetén is megéri, ha van elég nyerhető/veszteség [89].
2. Ha egy területen kevés a konszenzus a szakértők között, akkor tanácsos az episztemikus szerénység, azaz nem szabad túlságosan bízni saját véleményünk megbízhatóságában.

A mesterséges intelligencia kutatásának kockázatai globálisak. Ha a mesterséges intelligencia kutatói kudarcot vallanak a szuperintelligencia etikussá tételére tett első kísérletükben, lehet, hogy nem lesz második esély. Teljesen indokolt, hogy a mesterséges intelligencia kutatásának hosszú távú kockázatait még a globális felmelegedés kockázatainál is nagyobbaként értékeljük. Ehhez képest azonban a téma alig kapott figyelmet. Ezzel a vitaanyaggal rámutatunk, hogy ezért még inkább érdemes jelentős forrásokat fektetni a mesterséges intelligencia kutatásának biztonságába.

Ha az itt tárgyalt forgatókönyvek bekövetkezésének (talán csekély, de) több mint elenyésző valószínűsége van, akkor a mesterséges intelligenciának és a hozzá kapcsolódó lehetőségeknek és kockázatoknak globális prioritást kell élvezniük. A mesterséges intelligencia kutatás jó eredményének valószínűsége többek között a következő intézkedésekkel maximalizálható:

**ajánlás - Mesterséges intelligencia biztonsága:** Az elmúlt években látványosan megnőtték a mesterséges intelligencia kutatásába történő befektetések [86]. A mesterséges intelligencia biztonságának kutatása viszont viszonylag messze elmaradt. A világon az egyetlen olyan szervezet, amely kiemelten foglalkozik a mesterséges intelligencia biztonságának elméleti és technikai problémáival kapcsolatos kutatásokkal, a Machine Intelligence Research Institute (MIRI). A mesterséges intelligencia területén a kutatási támogatások odaítélésekor meg kell követelni, hogy a kutatási projektek biztonsági szempontból releváns szempontjait azonosítsák, és megfelelő óvintézkedéseket tegyenek. A magas kockázatú mesterséges intelligenciával kapcsolatos valamennyi kutatás betiltása nem lenne célszerű, és a kutatás gyors és veszélyes áthelyezéséhez vezetne az alacsonyabb biztonsági normákkal rendelkező országokba. ■

**7. ajánlás - Globális együttműködés és koordináció:** A gazdasági és katonai ösztönzők olyan kom- petitív légkört teremtenek, amelyben szinte biztos, hogy veszélyes fegyverkezési verseny alakul ki. A folyamat során a mesterséges intelligencia kutatásának biztonsága a gyorsabb fejlődés és a költségcsökkentés javára csökkenne. A fokozott nemzetközi együttműködés ellensúlyozhatja ezt a dinamikát. Ha a nemzetközi koordináció sikeres, akkor a biztonsági szabványok terén (a tudományos és az ipari mesterséges intelligencia kutatásának migrációja vagy annak veszélye révén) nagyobb valószínűséggel kerülhető el a "versenyfutás az alulról lefelé". ■

## Mesterséges tudat

lehetne kifejleszteni?

Az embereknek és sok nem emberi állatnak van fenomenális tudata - szubjektíven és belsőleg bizonyos módon érzi magát embernek vagy nem emberi állatnak [99]. Vannak érzékszervi benyomásaik, (kezdetleges vagy kifejezett) én-érzetük, fájdalmat éreznek, ha fizikailag sérülnek, és képesek pszichés szenvedést vagy örömet átélni (lásd például az egereken végzett depressziós vizsgálatokat [100]). Röviden: ők *érző* lények. Ez azt jelenti, hogy a saját magukra vonatkozó értelemben is károsodhatnak. A mesterséges intelligencia kapcsán felmerül a kérdés: létezhetnek-e olyan gépek is, amelyek anyagi-funkcionális struktúrája képes megvalósítani egy szenvedő "belső életet"? Thomas Metzinger filozófus és kognitív tudós a szenvedés fogalmának négy kritériumát adja meg, amelyeknek a gépeknek is meg kellene felelniük:

1. Tudatosság.
2. egy fenomenális önmodell.
3. a negatív értékek (azaz a sérült szubjektív preferenciák) reprezentálásának képessége az önmodellben belül.
4. Átláthatóság (azaz amit érzékelünk, azt visszavonhatatlanul "valóságosnak" érezzük - a rendszer így kénytelen azonosulni tudatos önmodelljének tartalmával) [101, 102].

Pontosabban meg kell különböztetni két kapcsolódó kérdést: Először is, hogy a gépek valaha is kifejleszthetik-e a tudatosságot és a szenvedés képességét; másodsor, ha az első kérdésre igenlő a válasz, milyen *típusú* gépeket

A gépeknek (lesz) tudata.

Ezt a két kérdést filozófusok és AI-szakértők egyaránt vizsgálják. A kutatás jelenlegi állását vizsgálva az első kérdésre könnyebb válaszolni, mint a másodikra. A szakértők között viszonylag szilárd konszenzus van abban, hogy a gépeknek elvileg lehet tudatuk, és hogy a gépi tudatosság legalábbis a *neuromorfikus* számítógépekben lehetséges [103, 104, 105, 106],

107, 108, 109]. Az ilyen számítógépek hardvere ugyanolyan funkcionális szerveződésű, mint a biológiai agyé [110]. A második kérdésre nehezebb válaszolni: A neuromorfikus számítógépeken kívül milyen típusú gépeknek lehet tudata? Ezen a területen a tudományos konszenzus kevésbé kifejezett [111]. Vitatott például, hogy a tiszta szimulációknak - mint például a *Blue Brain Project* szimulált agyának - lehet-e tudata. A kérdésre különböző szakértők pozitív választ adnak [109, 105], de egyesek negatívan is válaszolnak [111, 112].

Tekintettel a szakértők bizonytalanságára, helyénvalónak tűnik, hogy *óvatos* álláspontot foglaljunk el: A jelenlegi ismeretek szerint legalábbis elképzelhető, hogy sok kellően összetett számítógép, köztük a nem neuromorf számítógépek is képesek lesznek szenvedni.

Ezeknek a megfontolásoknak messzemenő etikai következményei vannak. Ha a gépek képtelenek lennének a tudatosságra, etikailag nem lenne kifogásolható, ha munkásként hasznosítanánk őket, és olyan kockázatos feladatokat bíznanék rájuk, mint az aknák hatástalanítása vagy veszélyes anyagok kezelése [4, 167. o.]. Ha van egy

Ha azonban az összetett mesterséges intelligenciák valószínűleg rendelkeznek tudattal és szubjektív preferenciákkal, akkor hasonló etikai-jogi biztosítékokat kell alkalmazni, mint az emberek és számos nem emberi állat esetében [113]. Ha például a *Blue Brain Project* virtuális agya tudattal fog rendelkezni, akkor etikai szempontból rendkívül problematikus lenne például depressziós állapotba helyezni azt (és vele együtt számos másolatot vagy "klónt") a depresszió szisztematikus kutatása érdekében. Metzinger arra figyelmeztet, hogy a tudatos gépekkel visszaélhetnek kutatási célokra, és "másodrendű állampolgároként" nemcsak hogy nincsenek jogaik, és felcserélhető kísérleti eszközként használják őket, hanem ez a tény a belső élményük szintjén is negatívan tükröződhet [106]. Ez a kilátás azért különösen aggasztó, mert elképzelhető, hogy egy nap hatalmas számban fognak mesterséges intelligenciákat létrehozni [4, 75]. Így egy

A legrosszabb forgatókönyv csillagászati, történelmileg elképzelhetetlen számú áldozatot és szenvedést eredményezne.

Ezek a disztópikus forgatókönyvek rámutatnak a technológiai fejlődés egy fontos következményére: még ha csak "kisebb" etikai hibákat követünk is el, például bizonyos számítógépeket tévesen öntudatlannak vagy erkölcsileg jelentéktelennek minősítünk, ez történelmileg példátlan katasztrófákhoz vezethet a történelmileg példátlan technológiai hatalom miatt. Ha az érző lények teljes száma nagymértékben megnő, akkor etikai értékeink és empirikus értékeléseink csekély mértékű javulása nem elegendő.

- mindkettőnek *jelentősen* javulnia kell ahhoz, hogy megbirkózzon a jelentősen megnövekedett felelősséggel. Ezért, tekintettel a gépi tudatossággal kapcsolatos bizonytalanságunkra, különös óvatossággal kell eljárunk a mesterséges intelligencia területén. Ez az egyetlen módja annak, hogy elkerüljük a leírtakhoz hasonló lehetséges katasztrófális forgatókönyveket.

**8. ajánlás - Kutatás:** Ahhoz, hogy etikus döntéseket tudjunk hozni, elengedhetetlen annak ismerete, hogy mely természetes és mesterséges rendszerek rendelkeznek tudattal, és különösen a szenvedés képességével. Ugyanakkor még mindig sok a bizonytalanság, különösen a gépi tudatosság területén. Ezért ésszerűnek tűnik a megfelelő interdiszciplináris kutatások (filozófia, idegtudomány, informatika) támogatása. ■

**ajánlás - Szabályozás:** Ma már bevett gyakorlat, hogy az élő kísérleti alanyokon végzett kísérleteket etikai bizottságok vizsgálják felül [114, 115]. Mivel fennáll annak a lehetősége, hogy a neuromorfikus számítógépek és a szimulált élőlények is tudatosságot vagy szubjektív belső perspektívát fejlesszenek, a velük kapcsolatos kutatásokat is az etikai bizottságok szigorú felügyelete mellett kell végezni. A szenvedésre képes mesterséges lények (váratlan) létrehozását el kell kerülni vagy el kell halasztani, különösen azért, mert nagyon nagy számban jelenhetnének meg, és kezdetben - társadalmi-politikai érdekképviselet hiányában - valószínűleg jogtalanul maradnának. ■

## Összefoglaló

Az új, meglepő potenciállal rendelkező mesterséges intelligencia-technológiák kezdeti változatai már ma is léteznek, legyen szó akár az önvezető járművekről, Watsonról mint az orvosi diagnosztika segítőjéről vagy az amerikai hadsereg által tesztelt legújabb drónokról. Belátható időn belül ezek az alkalmazások készen állnak majd a széles körű piaci bevezetésre. Legkésőbb addigra jól átgondolt jogi keretre lesz szükség ahhoz, hogy e technológiai lehetőségek potenciálját úgy valósítsuk meg, hogy a negatív általános fejlődés kockázata a lehető legalacsonyabb maradjon.

Minél nagyobb a fejlődés a központi területeken

A mesterséges intelligencia technológiával egyre fontosabbá és sürgetőbbé válik, hogy a felmerülő kihívásokat racionális és előremutató módon kezeljük. Az új technológiák kutatói és fejlesztői is felelősek azért, hogy hozzájárulásuk hogyan változtatja meg a világot. A politikával és a jogalkotással ellentétben, amelyek általában lemaradnak a legújabb fejleményekhez képest, a mesterséges intelligencia kutatói és fejlesztői közvetlenül részt vesznek az eseményekben; ők azok, akik a legjobban ismerik a dolgokat.

Sajnos erős gazdasági ösztönzők vannak az új technológiák minél gyorsabb fejlesztésére.

<sup>1</sup> Az olyan egyesületek, mint a People for the ethical treatment of reinforcement learners (PETRL), amellett érvelnek, hogy a mesterséges intelligenciáknak, ha érzőek, ugyanolyan erkölcsi megfontolást kell kapniuk, mint a "biológiai intelligenciáknak": <http://petrl.org/>.

anélkül, hogy időt "veszítene" a drága kockázatelemzésekre. Ezek a kedvezőtlen feltételek növelik annak kockázatát, hogy egyre inkább elveszítjük a mesterséges intelligencia technológiák és azok használata feletti ellenőrzést. Ezt a lehető legtöbb szinten ellensúlyozni kell: politikailag, magában a kutatásban és általában minden olyan személy körében, aki releváns módon foglalkozhat a témával. A mesterséges intelligencia fejlesztésének a lehető legelőnyösebb irányba való terelésének egyik legfontosabb előfeltétele, hogy a

Ezt nem csak néhány szakértő ismeri el, hanem a széles körű közbeszéd is az előttünk álló nagy (talán a legnagyobb) kihívásként tartja számon.

A fent említett konkrétabb követeléseken túlmenően ezért ezt a vitairatot arra is szeretnénk felhasználni, hogy jelentős lendületet adjunk és kérést fogalmazzunk meg annak érdekében, hogy a "mesterséges intelligencia kockázatai és lehetőségei" témáját, mint például az éghajlatváltozás vagy a fegyveres konfliktusok megelőzése, mielőbb globális prioritásként ismerjék el.

## Visszaigazolás

Szeretnénk köszönetet mondani mindazoknak, akik segítettek nekünk a kutatásban vagy a vitairat megírásában. Külön meg kell említeni Kaspar Etert és Massimo Manninót a dolgozat szerkezetével kapcsolatos tanácsaikért; Prof. Oliver Bendelt a "A jelenlegi mesterséges intelligenciák előnyei és kockázatai" című fejezethez adott impulzusaiért; valamint Prof. Jürgen Schmidhuber professzort az "Általános intelligencia és szuperintelligencia" és a "Mesterséges tudat" című fejezetekhez, valamint a mesterséges intelligencia különböző területein folyó kutatások jelenlegi állásáról szóló írásaiért. .

## Támogatók

A vitaanyag főbb pontjait a következők támasztják alá:

- **Prof. Dr. Fred Hamker**, a Mesterséges Intelligencia professzora, Chemnitzi Műszaki Egyetem
- **Prof. Dr. Dirk Helbing**, a Számítógépes Társadalomtudományok professzora, ETH Zürich
- **Prof. Dr. Malte Helmert**, a Mesterséges Intelligencia professzora, Bázeli Egyetem
- **Prof. Dr. Manfred Hild**, Digitális rendszerek professzora, Beuth University of Applied Sciences (Berlin)
- **Prof. Dr. Dr. Eric Hilgendorf**, a Robotjogi Kutatóközpont vezetője, Würzburgi Egyetem
- **Prof. Dr. Marius Kloft**, a gépi tanulás professzora, Humboldt Egyetem, Berlin
- **Prof. Dr. Stefan Kopp**, a Bielefeldi Egyetem Társadalmi Kognitív Rendszerek professzora
- **Prof. Dr. Dr. Franz Josef Radermacher**, az Ulmi Egyetem Adatbázisok és Mesterséges Intelligencia professzora.

- [1] Koomey, J. G., Berard, S., Sanchez, M. & Wong, H. (2011). A számítástechnika elektromos hatékonyságának történelmi tendenciáinak következményei. *IEEE Annals of the History of Computing*, 33(3), 46-54.
- [2] Brockman, J. (2015). *Mit gondoljunk a gondolkodó gépekről: Napjaink vezető gondolkodói a gépi intelligencia koráról*. Harper Perennial.
- [3] Russell, S. (2015). Jobb emberekké tesznek minket? (<http://edge.org/response-detail/26157>)
- [4] Bostrom, N. (2014). *Szuperintelligencia: utak, veszélyek, stratégiák*. Oxford University Press.
- [5] BBC. (2015a). Stephen Hawking arra figyelmeztet, hogy a mesterséges intelligencia véget vethet az emberiségnek. (<http://www.bbc.com/news/technology-30290540>)
- [6] Harris, S. (2015). Elkerülhetjük a digitális apokalipszist? (<https://edge.org/response-detail/26177>)
- [7] The Independent. (2014). Stephen Hawking: "A transzcendencia a mesterséges intelligencia következményeit vizsgálja - de elég komolyan vesszük-e az AI-t?" (<http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence--but-are-we-taking-ai-seriously-enough-9313474.html>)
- [8] The Guardian. (2014). Elon Musk 10 millió dollárt adományoz, hogy a mesterséges intelligencia jót tegyen az emberiségnek. (<http://www.theguardian.com/technology/2015/jan/16/elon-musk-donates-10m-to-artificial-intelligence-research>).
- [9] SBS. (2013). Mesterséges irrelevancia: Jönnek a robotok. (<http://www.sbs.com.au/news/article/2012/07/18/artificial-irrelevance-robots-are-coming>).
- [10] BBC. (2015b). A Microsoft-os Bill Gates kitart amellett, hogy a mesterséges intelligencia fenyegetést jelent. (<http://www.bbc.com/news/31047780>)
- [11] Silver, N. (2012). *A jel és a zaj: Miért bukik meg oly sok előrejelzés - de néhány nem*. Pingvin.
- [12] PCWorld. (2011). Az IBM Watson legyőzte az emberi Jeopardy ellenfeleket. ([http://www.pcworld.com/article/219893/ibm\\_watson\\_vanquishes\\_human\\_jeopardy\\_foes.html](http://www.pcworld.com/article/219893/ibm_watson_vanquishes_human_jeopardy_foes.html))
- [13] Bowling, M., Burch, N., Johanson, M. & Tammelin, O. (2015). A heads-up limit hold'em póker megoldott. *Science*, 347(6218), 145-149.
- [14] Cireşan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. (2013). Mitózis észlelése emlőrák szövettani képeken mély neurális hálózatok segítségével. MICCAI 2013. (<http://people.idsia.ch/~juergen/deeplearningwinsMICCAIgrandchallenge.html>)
- [15] Cireşan, D., Meier, U. & Schmidhuber, J. (2012). Több oszlopos mély neurális hálózatok képosztályozáshoz. *Computer Vision and Pattern Recognition 2012*, 3642-3649.
- [16] Tesauro, G. (1994). A TD-Gammon, egy önképző backgammon program, mester szintű játékot ér el. *Neural Computation*, 6(2), 215-219.
- [17] Koutník, J., Cuccu, G., Schmidhuber, J. & Gomez, F. (2013). Nagyméretű neurális hálózatok fejlesztése látásalapú megerősítő tanuláshoz. In *Proceedings of the 75th Annual Conference on Genetic and Evolutionary Computation* (pp. 1061-1068). ACM.
- [18] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . Ostrovski, G. et al. (2015). Emberi szintű irányítás mély megerősítő tanulással. *Nature*, 578(7540), 529-533.
- [19] Slavin, K. (2012). Hogyan alakítják világunkat az algoritmusok. (<http://ed.ted.com/lessons/kevin-slavin-how-algorithms-shape-our-world>)



- [20] Tagesanzeiger. (2008). Számítógépes hiba bénítja az amerikai légi forgalmat. (<http://www.tagesanzeiger.ch/külföld/amerika/ComputerPannelegt-USFlugverkehr-lahm/story/13800972>)
- [21] Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). A PageRank Citation Ranking: Rendet teremt a világhálón. (<http://ilpubs.stanford.edu:8090/422/>)
- [22] Be van drótozva. (2010). Az algoritmusok átveszik az irányítást a Wall Street felett. ([http://www.wired.com/2010/12/ff\\_ai\\_flashtrading/all/](http://www.wired.com/2010/12/ff_ai_flashtrading/all/))
- [23] Lin, T. C. (2012). Az új befektető. *UCLA L. Rev.* 60, 678-735.
- [24] Taleb, N. N. (2010). *A fekete hattyú: A rendkívül valószínűtlen törekénység hatása*. Random House.
- [25] Lauricella, T. & McKay, P. (2010). A Dow megrázó, 1.010,14 pontos utat tesz meg. *Wall Street Journal* (2010. május 7.).
- [26] Értékpapírok, U., Bizottság, E. és a határidős árutőzsdei kereskedési bizottság. (2010). A 2010. május 6-i piaci eseményekkel kapcsolatos megállapítások. *A CFTC és a SEC munkatársainak jelentése a felmerülő szabályozási kérdésekkel foglalkozó vegyes tanácsadó bizottságnak*.
- [27] Spiegel. (2015). Gondolkodó fegyverek: A mesterséges intelligencia kutatói figyelmeztetnek a mesterséges intelligenciára. (<http://www.spiegel.de/netzwelt/netzpolitik/elon-musk-and-stephen-hawking-warn-of-autonomous-weapons-a-1045615.html>)
- [28] Bendel, O. (2013). A gépi etika felé. In: *Technológiaértékelés és a nagy átmenetek politikai területei* (pp. 343-347). A prágai PACITA 2013 konferencia jegyzőkönyvei.
- [29] Goodall, N. J. (2014). Gépi etika és automatizált járművek. In *Road Vehicle Automation: Lecture Notes in Mobility*. (S. 93-102). Springer International Publishing.
- [30] Bostrom, N. & Yudkowsky, E. (2013). A mesterséges intelligencia etikája. In *Cambridge Handbook of Artificial Intelligence*. Cambridge University Press.
- [31] Dickmanns, E. D., Behringer, R., Dickmanns, D., Hildebrandt, T., Maurer, M., Thomanek, F. & Schiehlen, J. (1994). A látó személygépkocsi "VaMoRs-P". In *International Symposium on Intelligent Vehicles 94 (Intelligens járművek 94. nemzetközi szimpózium)* (68-73. o.).
- [32] Dickmanns, E. (2011). Esti előadás: A dinamikus jövőkép mint az AGI kulcseleme. 4th Conference on Artificial General Intelligence, Mountain View, CA. (<https://www.youtube.com/watch?v=YZ6nPhUG2i0>)
- [33] Thrun, S. (2011). A Google vezető nélküli autója. ([http://www.ted.com/talks/sebastian\\_thrun\\_google\\_s\\_driverless\\_car](http://www.ted.com/talks/sebastian_thrun_google_s_driverless_car))
- [34] Forbes. (2012). Nevada elfogadja a vezető nélküli autókra vonatkozó szabályokat. (<http://www.forbes.com/sites/alexknapp/2012/02/17/nevada-passes-regulations-for-driverless-cars/>).
- [35] Organization, W. H. et al. (2013). *A WHO 2073. évi globális helyzetjelentése a közúti közlekedésbiztonságról: A cselekvés évtizedének támogatása*. Egészségügyi Világszervezet.
- [36] Simonite, T. (2013). Offline kézírásfelismerés többdimenziós rekurrens neurális hálózatokkal. *MIT Technology Review*, okt. 25.
- [37] CNBC. (2014). Az önvezető autók biztonságosabbak, mint az emberek által vezetett autók: Bob Lutz. (<http://www.cnbc.com/id/101981455>)
- [38] Svenson, O. (1981). Mindannyian kevésbé vagyunk kockázatosak és ügyesebbek, mint vezetőtársaink? *Acta Psychologica*, 9(6), 143-148.
- [39] Weinstein, N. D. (1980). Irreális optimizmus a jövőbeli életeseményekkel kapcsolatban. *Journal of Personality and Social Psychology*, 39(5), 806.
- [40] Langer, E. J. (1975). Az ellenőrzés illúziója. *Journal of Personality and Social Psychology*, 32(2), 311.
- [41] Von Hippel, W. & Trivers, R. (2011). Az önbecsapás evolúciója és pszichológiája. *Behavioral and Brain Sciences*, 34(1), 1-56.
- [42] Trivers, R. (2011). *A bolondok bolondsága: A család és az önbecsapás logikája az emberi életben*. Basic Books.
- [43] Berner, E. S. & Graber, M. L. (2008). A túlzott magabiztosság mint a diagnosztikai hiba oka az orvostudományban.



- [44] Kohn, L. T., Corrigan, J. M., Donaldson, M. S. et al. (2000). *To Err Is Human: Egy biztonságosabb egészségügyi rendszer kiépítése*. National Academies Press.
- [45] The New York Times. (2010). Mi az IBM Watson? (<http://www.nytimes.com/2010/06/20/magazine/20Computer-t.html>)
- [46] Be van drótozva. (2013). Az IBM Watson jobb a rák diagnosztizálásában, mint az emberi orvosok. (<http://www.wired.co.uk/news/archive/2013-02/11/ibm-watson-orvos-orvos>).
- [47] Forbes. (2013). Az IBM Watson megkapja az első üzletrészét az egészségügyben. (<http://www.forbes.com/sites/bruceupbin/2013/02/08/ibms-watson-gets-its-first-piece-of-business-in-healthcare/>)
- [48] Dawes, R. M., Faust, D. & Meehl, P. E. (1989). Klinikai kontra biztosításmatematikai megítélés. *Science*, 243(4899), 1668-1674.
- [49] Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. (2000). Klinikai kontra mechanikai előrejelzés: metaanalízis. *Pszichológiai értékelés*, 72(1), 19.
- [50] West, R. F. & Stanovich, K. E. (1997). A túlzott magabiztosság területspecifikussága és általánossága: egyéni különbségek a teljesítménybecslés torzításában. *Psychonomic Bulletin & Review*, 4(3), 387-392.
- [51] Tversky, A. & Kahneman, D. (1974). Bizonytalanság alatti ítélezés: Heurisztikák és torzítások. *Science*, 785(4157), 1124-1131.
- [52] Pohl, R. (szerk.). (2004). *Kognitív illúziók: kézikönyv a gondolkodás, az ítélőképesség és az emlékezet tévedéseiről és torzításairól*. Psychology Press.
- [53] Brosnan, M. J. (2002). *Technofóbia: Az információs technológia pszichológiai hatása*. Routledge.
- [54] Yudkowsky, E. (2008). A mesterséges intelligencia mint a globális kockázat pozitív és negatív tényezője. *Globális katasztrófakockázatok*, 7, 303.
- [55] Bostrom, N. (2002). Egzisztenciális kockázatok. *Journal of Evolution and Technology*, 9(1).
- [56] Smith, A. & Anderson, J. (2014). A mesterséges intelligencia, a robotika és a munkahelyek jövője. Pew Research Center.
- [57] Cowen, T. (2013a). *Az átlagnak vége: Amerikát a nagy stagnálás korszakán túlra juttatni*. Pingvin.
- [58] Brynjolfsson, E. & McAfee, A. (2014). *A második gépkorszak: Munka, haladás és jólét a briliáns technológiák korában*. WW Norton & Company.
- [59] Frey, C. B. & Osborne, M. A. (2013). A foglalkoztatás jövője: Mennyire érzékenyek a munkahelyek a számítógépesítésre? *Oxford Martin Programme on Technology and Employment*. ([https://web.archive.org/web/20150109185039/http://www.oxfordmartin.ox.ac.uk/downloads/academic/The\\_Future\\_of\\_Employment.pdf](https://web.archive.org/web/20150109185039/http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf))
- [60] Helbing, D. (2015). *Thinking Ahead - Essays on Big Data, Digital Revolution, and Participatory Market Society*. Springer.
- [61] Cowen, T. (2013b). EconTalk Episode with Tyler Cowen: Tyler Cowen on Inequality, the Future, and Average is Over. ([http://www.econtalk.org/archives/2013/09/tyler\\_cowen\\_on.html](http://www.econtalk.org/archives/2013/09/tyler_cowen_on.html))
- [62] Griffiths, M., Kuss, D. & King, D. (2012). Videójáték-függőség: múlt, jelen és jövő. *Current Psychiatry Reviews*, 8(4), 308-318.
- [63] Srivastava, L. (2010). A mobiltelefonok és a társadalmi viselkedés fejlődése. *Behavior & Information Technology*, 24(2), 111-129.
- [64] Prensky, M. (2001). Tényleg másképp gondolkodnak? *On the Horizon*, 47(2).
- [65] Metzinger, T. (2015a). Virtuális megtestesülés robotokban. *SPEKTRUM*, 2, 48-55.
- [66] Kapp, K. M. (2012). *A tanulás és az oktatás játékosítása: Játékalapú módszerek és stratégiák a képzésben és az oktatásban*. Pfeiffer.
- [67] Bavelier, D., Green, S., Hyun Han, D., Renshaw, P., Merzenich, M. & Gentile, D. (2011). Nézőpont: agyak a videójátékokról. *Nature Reviews Neuroscience*, 72, 763-768.

- [68] Fagerberg, J. (2000). Technológiai fejlődés, strukturális változás és termelékenységnövekedés: összehasonlító tanulmány. *Structural Change and Economic Dynamics*, 77(4), 393-411.
- [69] Galor, O. & Weil, D. N. (1999). A malthusi stagnálástól a modern növekedésig. *American Economic Review*, 150-154.
- [70] Brynjolfsson, E. (2014). EconTalk epizód Erik Brynjolfssonnal: Brynjolfsson a második gépkorszakról. ([http://www.econtalk.org/archives/2014/02/brynjolfsson\\_on.html](http://www.econtalk.org/archives/2014/02/brynjolfsson_on.html))
- [71] Hughes, J. J. (2014). A technológiai munkanélküliség és az alapjövedelem-garancia elkerülhetetlen vagy kívánatos? *Journal of Evolution and Technology*, 24(1), 1-4.
- [72] Krugman, P. (2013). Szimpátia a ludditák iránt. *New York Times*, 73. (<http://www.nytimes.com/2013/06/14/opinion/krugman-sympathy-for-the-luddites.html>)
- [73] Bostrom, N. & Sandberg, A. (2008). Teljes agyi emuláció: útiter. Oxford: Az emberiség jövője Intézet.
- [74] Hanson, R. (2012). Az Emulált Elmék Rendkívüli Társasága. ([http://library.fora.tv/2012/10/14/Robin\\_Hanson\\_Extraordinary\\_Society\\_of\\_Emulated\\_Minds](http://library.fora.tv/2012/10/14/Robin_Hanson_Extraordinary_Society_of_Emulated_Minds)).
- [75] Hanson, R. (1994). Ha a feltöltések az elsők. *Extropia*, 6(2), 10-15.
- [76] Legg, S. & Hutter, M. (2005). Az intelligencia univerzális mércéje a mesterséges ügynökök számára. In *International Joint Conference on Artificial Intelligence* (19. kötet, pp. 1509). Lawrence Erlbaum Associates Ltd.
- [77] Hutter, M. (2007). Univerzális algoritmikus intelligencia: matematikai felülről lefelé történő megközelítés. In *Artificial General Intelligence* (6. kötet, 2. rész, 227-290. o.). Springer.
- [78] Bostrom, N. (1998). Meddig tart a szuperintelligencia? *International Journal of Future Studies*, 2.
- [79] Schmidhuber, J. (2012). Filozófusok és futuristák, felzárkózni! Válasz a Szingularitásra. *Journal of Consciousness Studies*, 79(1-2), 173-182.
- [80] Moravec, H. (1998). Mikor lesz a számítógépes hardver az emberi agyhoz hasonló. *Journal of Evolution and Technology*, 7(1), 10.
- [81] Moravec, H. (2000). *Robot: a puszta géptől a transzcendens elméig*. Oxford University Press.
- [82] Shulman, C. & Bostrom, N. (2012). Mennyire nehéz a mesterséges intelligencia? Evolúciós érvek és szelekciós hatások. *Journal of Consciousness Studies*, 79(7-8), 103-130.
- [83] Sengupta, B. & Stemmler, M. (2014). Energiafogyasztás a neurális számítás során. *Proceedings of the IEEE*, 702(5), 738-750.
- [84] Friston, K. (2010). A szabad energia elve: egységes agyelmélet? *Nature Reviews Neuroscience*, 77, 127-138.
- [85] Sengupta, B., Stemmler, M. & Friston, K. (2013). Információ és hatékonyság az idegrendszerben - Összegzés. *PLoS Comput Biol*, 9(7).
- [86] Eliasmith, C. (2015). A mesterséges elmék előestéjén. In T. Metzinger & J. M. Windt (Eds.), *Open mind*. MIND Group. (<http://open-mind.net/papers/@@chapters?nr=12>)
- [87] Armstrong, S., Sotala, K. & ÓhÉigeartaigh, S. S. (2014). A híres mesterséges intelligencia-előrejelzések tévedései, meglátásai és tanulságai - és mit jelentenek a jövőre nézve. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3), 317-342.
- [88] Brenner, L. A., Koehler, D. J., Liberman, V. & Tversky, A. (1996). Túlzott magabiztosság a valószínűségi és gyakorisági ítéleteknél: kritikai vizsgálat. *Organizational Behavior and Human Decision Processes*, 65(3), 212-219.
- [89] Peterson, M. (2009). *Bevezetés a döntéselméletbe*. Cambridge University Press.
- [90] Armstrong, S. (2013). Általános célú intelligencia: az ortogonalitás tézisének érvelése. *Analysis and Metaphysics*, (12), 68-84.
- [91] Noë, A. (2015). A "szingularitás" etikája. (<http://www.npr.org/sections/13.7/2015/01/23/379322864/the-etika-a-szingularitas-etikaja>).
- [92] Bostrom, N. (2012). A szuperintelligens akarat: Motiváció és instrumentális racionalitás a fejlett mesterséges ügynökökben. *Minds and Machines*, 22(2), 71-85.

- [93] Omohundro, S. M. (2008). Az alapvető mesterséges intelligencia meghajtók. In *Proceedings of the First AGI Conference, 777, Frontiers in Artificial Intelligence and Applications* (Vol. 171, pp. 483-492).
- [94] Solomonoff, R. (1985). A mesterséges intelligencia időskálája: gondolatok a társadalmi hatásokról. *Human Systems Management, 5*, 149-153.
- [95] Chalmers, D. (2010). A szingularitás: filozófiai elemzés. *Journal of Consciousness Studies, 77*(9-10), 7-65.
- [96] Good, I. J. (1965). Spekulációk az első ultraintelligens géppel kapcsolatban. In *Advances in Computers* (pp. 31-88). Academic Press.
- [97] Schmidhuber, J. (2006). Gödel-gépek: Teljesen önreferenciális optimális univerzális önfejlesztők. In *Mesterséges általános intelligencia* (pp. 119-226).
- [98] Tomasik, B. (2011). A csillagászati jövőbeli szenvedés kockázatai. Alapítványi Kutatóintézet. (<http://foundational-research.org/publications/risks-of-astronomical-future-suffering/>)
- [99] Nagel, T. (1974). Milyen érzés denevérenek lenni? *The Philosophical Review, 435-450*.
- [100] Durgam, R. (2001). A depresszió rágcsálómodelljei: tanult tehetetlenség triádossal elrendezéssel a Tats-ban. *Curr Protoc Neurosci, 8*.
- [101] Metzinger, T. (2012). Két alapelv a robotetika számára. In H. E & G. J-P (Eds.), *Robotika és jogalkotás* (pp. 263-302). NOMOS. ([http://www.blogs.uni-mainz.de/fb05philosophie/files/2013/04/Metzinger\\_RG\\_2013\\_penultimate.pdf](http://www.blogs.uni-mainz.de/fb05philosophie/files/2013/04/Metzinger_RG_2013_penultimate.pdf))
- [102] Metzinger, T. (2015b). *Empirikus perspektívák a szubjektivitás önmodell-elméletének szemszögéből: Rövid bemutatás példákkal*. Önkiadás. (<http://www.amazon.de/Empirische-perspectives-view-self-model-theory-of-subjectivity-ebook/dp/B01674W53W>).
- [103] Moravec, H. P. (1988). *Mind Children: A robotok és az emberi intelligencia jövője*. Harvard University Press.
- [104] Chalmers, D. J. (1995). Hiányzó qualia, elhalványuló qualia, táncoló qualia. *Tudatos tapasztalat, 309-328*.
- [105] Chalmers, D. J. (1996). *A tudatos elme: Egy alapvető elmélet keresése*. Oxford University Press.
- [106] Metzinger, T. (2014). *Az Ego alagút. Az én új filozófiája: Az agykutatástól a tudat etikájáig*. Piper.
- [107] Metzinger, T. (2015c). Mi van, ha szenvedniük kell? (<https://edge.org/response-detail/26091>)
- [108] Dennett, D. C. (1993). *A tudatosság magyarázata*. Penguin UK.
- [109] Bostrom, N. (2003). Egy számítógépes szimulációban élünk? *The Philosophical Quarterly, 53*(211), 243-255.
- [110] Hasler, J. & Marr, B. (2013). Útitervezés a nagyméretű neuromorfikus hardverrendszerek megvalósításához. *Frontiers in Neuroscience, 7*(118).
- [111] Koch, C. (2014). What it Will Take for Computers to Be Conscious, MIT Technology Review. (<http://www.technologyreview.com/news/531146/what-it-will-take-for-computers-to-be-conscious/>)
- [112] Tononi, G. (2015). Integrált információelmélet. *Scholarpedia, 70*(1), 4164. ([http://www.scholarpedia.org/article/Integrated\\_Information\\_Theory](http://www.scholarpedia.org/article/Integrated_Information_Theory)).
- [113] Singer, P. (1988). Hozzászólás Frey "Erkölcsi helytállás, az életek értéke és a fajiság" című írásához. *Between the Species: A Journal of Ethics, 4*, 202-203.
- [114] Swissethics, a svájci elismert etikai bizottságok szövetsége. (o.d.). (<http://www.swissethics.ch/>)
- [115] A Szenátus Állatkísérleti Kutatási Bizottsága. (2004). Állatkísérletek a kutatásban. ([http://www.dfg.de/download/pdf/dfg\\_im\\_profil/geschaeftsstelle/publikationen/dfg\\_terversuche\\_0300304.pdf](http://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/dfg_terversuche_0300304.pdf), publisher=Deutsche Forschungsgemeinschaft)





Az Alapítvány a Hatékony Altruizmusért (EAS) egy független agytröszt és projektszervezet az etika és a tudomány metszéspontjában. Munkájának eredményeit vitaanyagok formájában a társadalom és a politika rendelkezésére bocsátja. Adományozási és karrier-tanácsadást is kínál. A hatékony altruizmus (EA) az alapítvány vezérgondolata: Forrásaink

- Az idő és a pénz - korlátozott. Hogyan használhatjuk őket úgy, hogy a legtöbb szenvedést megelőzzük és a legtöbb életet megmentjük? És milyen racionális okok szólnak amellett, hogy erőforrásokat fektessünk be a fenntartható és hatékony szenvedéscsökkentésbe? Ezeket a kérdéseket filozófiai, gazdasági, kognitív és szociálpszichológiai szempontból vizsgáljuk.