Jogalkotási folyóirat

Kötet | 45Kiadvány 2 Cikk 1

6-7-2019

A mesterséges általános intelligencia veszélyei és ígéretei

Brian S. Haney

A MESTERSÉGES ÁLTALÁNOS INTELLIGENCIA VESZÉLYEI ÉS ÍGÉRETEI

Brian S. Haney[†]

BEVEZETÉS

A legtöbb ember úgy gondolja, hogy az ember és a technológia összeolvadása a távoli jövőben történhet meg; az igazság az, hogy az emberek már most is kiborgok. Egy okostelefonnal az ember gyakorlatilag bármilyen kérdésre gyorsan válaszolhat, korlátlan mennyiségű információt tárolhat a memóriájában, és bármilyen számítást elvégezhet.[1] A modern technológiai vállalatok az okostelefonokról adatokat gyűjtenek az emberekről, és közvetlenül a fejlett mesterséges intelligencia ("AI") rendszereken keresztül táplálják azokat.[2] A mesterséges intelligencia-rendszerek eleve maximalizálják a fogyasztók limbikus rendszerébe, az agy jutalomközpontjába érkező elektromos impulzusokat a gazdasági növekedés és fejlődés ösztönzése érdekében.[3] Elon Musk a Nemzeti Kormányzók Szövetségének 2017-es nyári ülésén kijelentette: "[a] legnagyobb kockázat, amellyel civilizációnknak szembe kell néznie, a mesterséges intelligencia".[4]

Musk nincs egyedül, sőt, egyre több tudós és iparági vezető hívja fel a figyelmet arra, hogy a mesterséges intelligencia milyen egzisztenciális veszélyeket jelent az emberre nézve.[5] A mesterséges intelligencia témájával foglalkozó jogi kutatások azonban vagy tagadták, vagy viszonylag figyelmen kívül hagyták a mesterséges intelligencia gyorsuló ütemű fejlődését.[6] Ehelyett a mesterséges intelligencia szabályozásának szentelt jelenlegi jogi tudományosság arra ösztönözte a szabályozókat, hogy ne hagyják magukat megzavarni az "AI apokalipszisre" vonatkozó állításoktól, és erőfeszítéseiket a "közvetlenebb károkra" összpontosítsák.[7] Összefoglalva, az AI-szabályozással foglalkozó jogi ösztöndíj messze elmaradt, és téves tanácsokat ad a szabályozóknak és a tudósoknak.[8] Valójában minden olyan feladat, amelynek elvégzésére az emberek intelligenciát használnak, az AI automatizálásának célpontja.[9] Továbbá gyakran előfordul, hogy amint egy mesterséges intelligenciával működő rendszer eléri az emberi szintű teljesítményt egy adott feladatban, röviddel ezután ugyanez a mesterséges intelligenciával működő rendszer meghaladja a legképzettebbek teljesítményét.

---

† Brian S. Haney, J.D. Notre Dame Law School 2018, a Martian Tech- nologies vezérigazgatója. Külön köszönet illeti LaDarien Harrist, Mike Gallaghert, Ryan Claudeanost, Delaney Foremant, Maria Munoz-Roblest, Brian Wongchaowart, Bill Greent, Ned Rooneyt és a Jogászújságot.

1 *The Joe Rogan Experience #1169 - Elon Musk*, THE JOE ROGAN EXPERIENCE (2018. szeptember 6.), https://www.youtube.com/watch?v=ycPr5-27vSI.

2 *Id*.

3 *Id*.

4 *Elon Musk at the National Governors Association 2017 Summer Meeting*, C-SPAN, 50:00 (July 15, 2017), https://www.c-span.org/video/?c4676772/elon-musk-national-governors-association-2017-summer-meeting.

5 *Lásd:* MILES BRUNDAGE ET AL., A MŰVES INTELLIGENCIA MÉLTALÁNOS HASZNÁLATA: FORECASTING, PRE- VENTION, AND MITIGATION (122018); *lásd még* MAX TEGMARK, LIFE BEING3.0 HUMAN IN THE AGE OF ARTI- FICIAL INTELLIGENCE (2017).

6 *Lásd* Ryan Calo, *Mesterséges intelligencia politika:* DAVIS L. REV. 399,432 (2017).

7 *Lásd id*. 431.

8 *Lásd id*.

9 BRUNDAGE ET AL., *Supra* note. 5.

embereket a feladat elvégzésében.[10] Sok AI-kutató arra számít, hogy az AI-rendszerek idővel minden feladatban elérik, majd meghaladják az emberi szintű teljesítményt.[11]

A mesterséges intelligencia technológiája olyan jövőt alakít ki, ahol a hamis képek és videók olcsók, széles körben elérhetők és megkülönböztethetetlenek a valóditól, ami teljesen átformálja azt a módot, ahogyan az emberek az igazságot a bizonyítékokkal társítják.[12] Még azok is, akik kételkednek abban, hogy a jövőben létre fog jönni a mesterséges általános intelligencia ("AGI"), azaz a bármilyen cél elérésére képes mesterséges intelligencia[13], egyetértenek abban, hogy a mesterséges intelligenciának mélyreható következményei lesznek minden területen, többek között az egészségügyben, a jogban és a nemzetbiztonságban.[14] E cikk célja kettős. Először is, ez a cikk meghatározza és elmagyarázza a mesterséges intelligencia csúcstechnológiáját, különös tekintettel a mély megerősítő tanulásra, a Google által 2013-ban kifejlesztett, áttörést jelentő gépi tanulás egyik típusára.[15] Másodszor, ez a cikk három akadályt határoz meg, amelyeket a szabályozóknak le kell küzdeniük a mesterséges intelligencia szabályozása során.

Ez a cikk három fő módon járul hozzá a jelenlegi jogi és mesterséges intelligencia tudományhoz. Ez az első, amely a mély megerősítő tanulásra összpontosít, különösen a mesterséges intelligencia által jelentett egzisztenciális fenyegetésekre, és ez az első, amely a mesterséges intelligencia alapjául szolgáló formális modellekkel foglalkozik. A cikk három részből áll. Az I. rész elmagyarázza a mesterséges intelligencia alapvető kifejezéseit és fogalmait, és a mesterséges intelligencia számos gyakorlati alkalmazását vizsgálja a modern iparban. A II. rész a mély megerősítő tanulást ismerteti, amely a mesterséges intelligencia viszonylag új keletű áttörése, és amely sok tudós szerint az AGI felé vezető utat jelent. A III. rész a mesterséges intelligencia szabályozásának témájával foglalkozó jogi tudományosságot vizsgálja, és három olyan kérdést tárgyal, amelyekkel a szabályozóknak foglalkozniuk kell ahhoz, hogy a mesterséges intelligencia erős szabályozási keretét kialakítsák.

I. MESTERSÉGES INTELLIGENCIA

A kortárs tudósok a mesterséges intelligencia számos különböző definícióját mutatták be. Max Tegmark, az MIT professzora például tömören úgy határozza meg az AI-t, mint "nem biológiai intelligencia".[16] Ray Kurzweil, a Google munkatársa a mesterséges intelligenciát "olyan gépek létrehozásának művészeteként írta le, amelyek olyan funkciókat látnak el, amelyek elvégzéséhez embereknek intelligenciára van szükségük".[17] Nils Nilsson, a Stanford professzora szerint az AI "a műtárgyak intelligens viselkedésével foglalkozik".[18] Általánosságban és e cikk alkalmazásában az AI olyan intelligens gépek tanulmányozására és fejlesztésére utal, amelyek képesek az emberi kognitív funkciók - például jóslatok készítése, beszédfolyamatok vagy játék - gondolkodási folyamatait leutánozni.

Bár a mesterséges intelligencia különböző kategóriákat foglal magában, a mesterséges intelligencia szabályozásának összefüggésében két típusa a legfontosabb. Az első a szűk mesterséges intelligencia, más néven gyenge mesterséges intelligencia.[19] A szűk értelemben vett mesterséges intelligencia korlátozott célok[20] elérésére képes, és az emberi intelligencia fejlesztésére irányuló kísérletekkel függ össze, szemben az emberi intelligencia megkettőzésével.

10 *Id*. 16.
11 *Id*.
12 GREG ALLEN & TANIEL CHAN, ARTIFICIAL INTELLIGENCE AND NATIONAL SECURITY (312017).
13 TEGMARK, 5,31. o. (2017).
14 ALLEN & CHAN, *fenti* megjegyzés 12.

15 Methods and Apparatus for Reinforcement Learning, U.S. Patent Application No. 14/097,862 (benyújtott 2013. december 5-én) (elérhető a https://patents.google.com/patent/US20150100530A1/en oldalon); *lásd még* TEGMARK, *su- pra* note at 5,84.

16 TEGMARK, *Supra* note5 .

17 RAY KURZWEIL, AZ INTELLIGENS GÉPEK KORA (141992).

18 NILS J. NILSSON, MESTERSÉGES INTELLIGENCIA: (11998).

19 NILS J. NILSSON, A MESTERSÉGES INTELLIGENCIA KERESÉSE (3882010).

20 *Lásd* TEGMARK, *fenti* megjegyzés 5.

emberi intelligencia.[21] A mesterséges intelligencia második típusa a mesterséges általános intelligencia ("AGI"), más néven erős mesterséges intelligencia.[22] Az AGI bizonyításához egy mesterséges intelligencia-ügynöknek képesnek kell lennie bármilyen cél elérésére.[23] Az AGI-hez kapcsolódik az az állítás, hogy egy programozott számítógép lehet elme, és legalább olyan jól tud gondolkodni, mint az emberek.[24] Végső soron az AGI számos AI-kutató jelenlegi célja.[25] Például az OpenAI, egy nonprofit szervezet, amely a terület úttörő kutatásait finanszírozza, a honlapján azt állítja, hogy küldetése "[d]árnyékolja és megvalósítja a biztonságos mesterséges általános intelligenciához vezető utat".[26] Mégis, egyelőre úgy tűnik, hogy csak szűk körű mesterséges intelligenciát fejlesztettek ki és alkalmaznak sikeresen.[27]

A. AI A MODERN SZAKMAI IPARÁGAKBAN

A szűk értelemben vett mesterséges intelligencia alkalmazása világszerte felforgatja a modern iparágakat.[28] Még a jogi iparág sem mentesül ettől a maró erőtől.[29] A technológiával támogatott felülvizsgálat ("TAR") forradalmasítja a felfedési folyamatot, és az AI az innováció élvonalába tartozik.[30] Az ügyfelek ma már gyakran kérik fel a peres ügyvédeket, hogy állítsanak fel e-discovery relevancia-hipotéziseket, és alkalmazzanak prediktív kódolási modelleket (a TAR egy fajtája) az elektronikus információk feltárásához.[31] Ebben a folyamatban a peres ügyészek először meghatározzák a keresendő kulcsszavakat, és azonosítják az áttekintendő dokumentumok kezdeti csoportját.[32] Ezután a dokumentumokat vizsgáló ügyvédek felülvizsgálják, kódolják és pontozzák a dokumentumok kezdeti halmazát bizonyos kulcsszavak előfordulása alapján a dokumentum relevanciájával kapcsolatban.[33] A felülvizsgálat során az e-discovery ügyvédek felügyelt tanulási algoritmusokat képeznek és modelleznek a dokumentumok osztályozására, amelyek a dokumentumokat felülvizsgáló ügyvédek döntései alapján osztályozzák a dokumentumokat a kezdeti dokumentumhalmazban.[34] Más szóval, az algoritmus a valódi ügyvédek döntéseinek elemzésével és megismétlésével tanulja meg, hogy mely dokumentumok relevánsak.[35] Ezen túlmenően a prediktív kódolási modellek több millió felderíthető dokumentumot képesek relevancia alapján osztályozni.[36]

21 *Lásd* NILSSON, *Supra* note19 , 388-89. o.

22 *Lásd* NICK BOSTROM, SUPERINTELLIGENCE: (232017).

23 TEGMARK, *fenti* megjegyzés 5.

24 NILSSON, *fenti* megjegyzés 19.

25 *Id*.

26 *Az OpenAI-ról*, OPENAI https://openai.com/about/ (utolsó látogatás 201910,. május).

27 *Lásd általában* Nick Bostrom, *Are You Living in A Computer Simulation?* , FILOZÓFIA 53Q. (2003)211,.243

28 ALLEN & CHAN, *Supra* note 12; *lásd még* HEMANT TANEJA, UNSCALED: HOW AI AND NEW GENERA- TION OF UPSTARTS AREATING THE ECONOMY OF THE FUTURE (12018).

29 RICHARD SUSSKIND, TOMORROW'S LAWYERS (112d ed. 2017).

30 Scott D. Cessar, Christopher R. Opalinski, & Brian E. Calla, *Controlling Electronic Discovery Costs: Cutting "Big Data" Down to Size*, ECKERT SEAMANS (2013. március 5.), https://www.eckertseamans.com/publi- cations/controlling-electronic-discovery-costs-cutting-big-data-down-to-size; *lásd még* Nicholas Barry, *Man Versus Machine Review: The Showdown Between Hordes of Discovery Lawyers and a Computer-Utilizing Pre- dictive-Coding Technology*, VAND15. J. ENT. & TECH. L. (2013)343,.344

31 KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS 240-42 (2017).

32 Barry, *Supra* note at 30,351.

33 GORDON V. CORMACK & MAURA R. GROSSMAN, EVALUATION OF MACHINE-LEARNING PROTOCOLS FOR TECHNOLOGY-ASSISTED REVIEW IN ELECTRONIC DISCOVERY (1542014), http://plg2.cs.uwaterloo.ca/~gvcormac/calstudy/study/sigir2014-cormackgrossman.pdf.

34 Barry, *Supra* note at 30,354.

35 *Id*.

36 *Lásd pl.* ASHLEY, *Supra* note at 31,250.

A másik példa a mesterséges intelligencia miatt gyorsan fejlődő iparágra az egészségügy.[37] Egy évtized múlva az egészségügyi ágazat a mesterséges intelligenciának köszönhetően egészen másképp fog kinézni, mint ma.[38] Jelenleg a nagy adatok által vezérelt mesterséges intelligencia érezhető elmozdulást hoz létre az orvosi gyakorlatban a tömeges ellátástól a személyre szabott ellátás felé.[39] A modern kórházakban praktizáló egészségügyi szakemberek ugyanis ma már elektronikus adatbázisokban, elektronikus egészségügyi nyilvántartásokban ("EHR") tárolják a betegek adatait.[40] Ez lehetővé teszi, hogy a gépi tanuló algoritmusok elemezzék a betegek egészségügyi adatait, és drasztikusan javítsák a betegellátást.[41] Ezek az adatvezérelt erőforrások nemcsak azt teszik lehetővé, hogy az orvos gyakorlatilag mindent tudjon a beteg kórtörténetéről anélkül, hogy valaha is találkozna a beteggel, hanem az orvosi munka segítésével drasztikusan csökkentik az egészségügyi ellátással kapcsolatos költségeket is.[42] Például 2016-ban a Stanford kutatói olyan mesterséges intelligenciát fejlesztettek ki, amely pontosabban tudta diagnosztizálni a tüdőrákot, mint az emberi patológusok.[43] Egy másik példa a D-Wave adiabatikus kvantumszámítógépe, amely képes a rákdiagnosztikához szükséges gépi tanulási algoritmusok futtatására.[44] Röviden, az EHR-ek, a nagyméretű adatok és a mesterséges intelligencia átalakítják az egészségügyi ellátást.[45]

A mesterséges intelligencia okozta zavarok harmadik példája a védelmi iparban zajlik. A mesterséges intelligencia már most is alapvető eszköz a kiberbiztonságban.[46] Mike Rogers admirális, a Nemzetbiztonsági Hivatal igazgatója szerint a mesterséges intelligencia és a gépi tanulás a kiberbiztonság jövőjének alapját képezi.[47] Márciusban 2,2017,jelentést tettek közzé a Fehér Házban, amely szerint orosz programozók mesterséges intelligenciával végrehajtott kibertámadást indítottak a Védelmi Minisztérium több mint 10 000 alkalmazottjának személyes közösségi média fiókjai ellen.[48] Emellett a modern hadviselésben a mesterséges intelligenciát a csatatéren is használják.[49] Például a haditengerészeti hajókra telepített amerikai Phalanx rakétavédelmi rendszer mesterséges intelligenciát használ az ellenséges rakéták és repülőgépek által jelentett fenyegetések észlelésére, követésére és támadására.[50] A kereskedelmi AI-rendszerekkel való terrorista visszaélés azonban komoly problémát jelent.[51] Terrorista szervezetek már most is használják a drónokban lévő mesterséges intelligencia-rendszereket robbanóanyagok szállítására és balesetek okozására.[52]

A szűk körű mesterséges intelligencia továbbra is megváltoztatja az olyan professzionális iparágak működését, mint a jog, az egészségügy és a védelem.[53] Számos AI-kutató hivatkozott megfigyelhető

37 TEGMARK, *Supra* note at 5,102.

38 TANEJA, *Supra* note at 28,73.

39 *Id*.

40 Kate Monica, *Apple EHR Patient Data Viewer Now in Use at Health39 Systems*, EHRINTELLIGENCE (2018. április 2.), https://ehrintelligence.com/news/apple-ehr-patient-data-viewer-now-in-use-at-39-health-sys- tems.

41 *Lásd* XIAOQIAN JIANG ET AL., A PATIENT-DRIVEN ADAPTIVE PREDICATION TECHNIQUE TO IMPROVE PERSONALIZED RISK ESTIMATION FOR CLINICAL DECISION SUPPORT 137 (2012).

42 *Lásd* Alvin Rajkomar et al., *Scalable and Accurate Deep Learning with Electronic Health Records*, NATURE PARTNER J. (2018), https://www.nature.com/articles/s41746-018-0029-1.pdf.

43 *Lásd* Lloyd Minor, *Crunching the Image Data Using Artificial Intelligence to Look at Biopsies*, STAN- FORD MED. (2017), https://stanmed.stanford.edu/2017summer/artificial-intelligence-could-help-diagnose-can- cer-predict-survival.html.

44 *Lásd* Brian S. Haney, *Quantum_Machine_Learning_Cancer_Diagnostics*, GITHUB (2019. február 24.), https://github.com/Bhaney44/Leap/blob/master/Quantum_Machine_Learning_Cancer_Diagnostics.py.

45 *Id*.

46 *Lásd általánosságban* BRUNDAGE ET AL .,*supra* note 5.

47 ALLEN & CHAN, *Supra* note at 12,18.

48 *Lásd* Massimo Calbresi, *Inside Russia's Social Media War on America*, TIME (201718,. május), http://time.com/4783932/inside-russia-social-media-war-america/.

49 *Lásd United States Navy Fact File: MK15 - Phalanx Close-In Weapons System (CIWS)*, U.S. DEP'T

NAVY (utolsó látogatás 201913,. május),
http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=487&ct=2.

50 TEGMARK, *Supra* note at (5,1112017).

51 *Lásd általánosságban* BRUNDAGE ET AL .,*supra* note 5.

52 *Lásd id*.

53 *Lásd általában* TEGMARK, *fenti* megjegyzés. 5.

az információtechnológiai ár- és teljesítménykinetika történelmi mintázatai, amelyek alátámasztják azt az érvet, hogy a mesterséges intelligencia technológiák fejlődési üteme a vártnál sokkal gyorsabban fog végbemenni.[54] Sőt, ezek a kutatók feltételezik, hogy a mesterséges intelligencia technológiák továbbra is gyorsuló ütemben fognak fejlődni.[55]

B. A GYORSULÓ MEGTÉRÜLÉS TÖRVÉNYE

A gyorsuló megtérülés törvénye ("LOAR") kimondja, hogy az információs technológia alapvető mérőszámai általában kiszámítható és exponenciális pályát követnek.[56] Az információs technológiák valóban exponenciális módon épülnek egymásra; ezt a jelenséget Moore törvényének nevezték el, és könnyen mérhető a legtöbb olyan folyamatban, ahol az információ mintái fejlődnek.[57] A LOAR a számítástechnika[58] árára és teljesítményére való alkalmazását írja le, és Gordon Moore, az Intel alapítója javasolta 1965-ben.[59] A Moore-törvény azt jósolja, hogy másfélévente a számítógépek feldolgozási teljesítménye megduplázódik, miközben a költségek a felére csökkennek.[60] Általánosságban azt jelenti, hogy az informatika teljesítménye másfél évente megduplázódik.[61] Az elmúlt ötvenhárom év igazolta Gordon Moore jóslatát; [62]egy okostelefon ma nagyobb számítási teljesítményt nyújt, mint az egész NASA 1969-ben - amikor az Apollo-11 leszállt a Holdra.[63] A Moore-törvényt a mesterséges intelligenciára alkalmazva sok mesterséges intelligencia kutató úgy véli, hogy jelenleg a szuperintelligens mesterséges intelligencia kifejlesztésének küszöbén állunk.[64]

Irving J. Good 1965-ben mutatta be először a szuperintelligencia fogalmát.[65] Good kijelentette: "[l]etezzük az ultraintelligens gépet úgy, mint olyan gépet, amely messze felülmúlja bármely ember minden szellemi tevékenységét, legyen az bármilyen okos is".[66] Good szerint "[a]lévén, hogy a gép tervezése az egyik ilyen intellektuális tevékenység, egy ultraintelligens gép még jobb gépeket tudna tervezni; ekkor kétségtelenül intelligencia-robbanás következne be, és az ember intelligenciája messze lemaradna".[67] Good valóban azt jósolta, hogy az első ultraintelligens gép lesz "az utolsó találmány, amelyet az embernek valaha is meg kell tennie".[68] A közelmúlt tudósai átvették Good elemzését, és hasonlóan definiálták a szuperintelligenciát. Nick Bostrom oxfordi professzor például úgy definiálja a szuperintelligenciát, mint "minden olyan értelmet, amely gyakorlatilag minden érdeklődési területen jelentősen meghaladja az emberek kognitív teljesítményét".[69] Max Tegmark szerint a szuperintelligencia "[g]enerális intelligencia, amely messze meghaladja az emberi szintet".[70]

54 *Lásd* BOSTROM, *Supra* 22. lábjegyzet, a következő címen 85.
55 *Lásd* RAY KURZWEIL, HOGYAN KÉSZÜLJÜK AZ ELMET (2502012).
56 *Lásd id*.
57 *Lásd id*. 256.
58 *Lásd* MARTINE ROTHBLATT, VIRTUÁLIS EMBER (482014).
59 *Lásd* KURZWEIL, *Supra* note at 55,251.
60 *Lásd* SUSSKIND, *Supra* note at 29,11.
61 *Lásd* ROTHBLATT, *Supra* note at 58,28.
62 *Lásd* KURZWEIL, *Supra* note at 55,251.
63 *Lásd* MICHIO KAKU, A JÖVŐ FIZIKÁJA (232011).
64 *Lásd általában* BOSTROM, 22. lábjegyzet; *lásd még* KURZWEIL, 55. lábjegyzet; *lásd még* TEGMARK, 55. *lábjegyzet.* megjegyzés 5.
65 *Lásd általában* Irving J. Good, *Speculations Concerning the First Ultraintelligent Machine*, AD-6 VANCES IN COMPUTERS (311966).
66 *Id*. 33.
67 *Id*.
68 *Id*.
69 BOSTROM, *Supra* note at 22,22.
70 *Lásd* TEGMARK, *fenti* jegyzet, lásd:5, TEGMARK, *fenti* jegyzet. 39.

A LOAR alkalmazása a mesterséges intelligenciára azt bizonyítja, hogy a szűk értelemben vett mesterséges intelligenciáról az AGI-ra és a szuperintelligenciára való áttérés sokkal közelebb lehet, mint azt általában gondolják.[71] Egyelőre az AGI legkorábbi becslések szerint 2029-re várható.[72] Ray Kurzweil szerint a XXI. század a LOAR-nak köszönhetően olyan technológiai fejlődést és innovációt hoz, amely ma még éveknek20,000 tűnhet.[73] Emellett Bostrom és Eliezer Yudkowsky, a mesterséges intelligencia teoretikusa azt jósolta, hogy a közvélemény a mesterséges intelligencia antropomorfizmusa miatt a mesterséges intelligencia fejlődésének gyors kinetikáját fogja érzékelni.[74] A mesterséges intelligencia antropomorfizmusa az emberi intelligenciaszintek nem emberi entitásoknak való tulajdonítására utal.[75] Az emberek az intelligenciaspektrum szélsőséges végpontjainak tekinthetik a falusi bolondot és Albert Einsteint,[76] de a kettő közötti különbség nagyobb relatív skálán valójában *csekély*.[77] Így egy mesterséges intelligencia-rendszer fejlődése a falusi idióta intelligenciájától Einstein intelligenciájáig, az AGI intelligenciájáig és végül a szuperintelligenciáig gyorsabb lehet a vártnál.[78]

Érdekes módon ezeket az előrejelzéseket alátámasztja az a hatalmas mennyiségű információ, amelyet az emberek a digitális korszak hajnalán kezdtek gyűjteni.[79] Valójában az emberek által gyűjtött információk mennyisége is egyre gyorsul.[80] Az adatok, amelyeket a világról szóló információk digitális reprezentációjaként határoznak meg,[81] elképesztő sebességgel keletkeznek. Kétnaponta az emberek több mint öt kvintillió bájtnyi adatot hoznak létre, vagyis annyi adatot, mint a civilizáció hajnalától kezdve egészen a 2003.[82]

Michael Kremer harvardi professzor és közgazdász szerint "az emberi fejlődés alapvető mozgatórugója nem a nyersanyagok, hanem a problémák technológiai megoldása".[83] A mesterséges intelligencia kontextusában az adatok a technológiai fejlődés hajtóereje az emberi programozók helyett.[84] A technológiai megoldások hajtóereje pedig az a felismerés, hogy minden információ számokként ábrázolható.[85] Az egy adott problémához rendelkezésre álló adatok mennyisége és típusa nagyban meghatározza a fejleszthető AI-rendszerek erejét.[86] A LOAR tehát mélyreható hatással lesz az AI fejlődésére az AGI és a szuperintelligencia irányába.

Egyesek azonban azt állítják, hogy az AGI talán soha nem fog megvalósulni.[87] A Microsoft néhai társalapítója, Paul Allen például azt állítja, hogy a tudományos fejlődés szabálytalan, és feltételezi, hogy a XXI. század végére az embereknek még

71 *Lásd id*. 157.

72 *Lásd* KURZWEIL, *Supra* note at 55,261.

73 *Lásd* Ray Kurzweil, *The Law of Accelerating Returns, in* KURZWEIL NETWORK (2001), http://www.kur- zweilai.net/the-law-of-accelerating-returns.

74 *Lásd* BOSTROM, *Supra* 22. lábjegyzet, a következő címen 85.

75 *Lásd* Eliezer Yudkowsky, *A mesterséges intelligencia mint a globális kockázat pozitív és negatív tényezője*, MACHINE INTELLIGENCE RES. INST., (212008), https://intelligence.org/files/AIPosNegFactor.pdf.

76 *Lásd* BOSTROM, *Supra* 22. lábjegyzet, a következő címen 85.

77 *Lásd* Yudkowsky, *Supra* note. 75.

78 *Lásd* BOSTROM, *Supra* 22. lábjegyzet, a következő címen 86.

79 *Lásd* ETHEM ALPAYDIN, MACHINE LEARNING (112016).

80 *Lásd* SUSSKIND, *Supra* note at 2911.

81 ALPAYDIN, *Supra* note at 79,3.

82 SUSSKIND, *Supra* note at 29,11.

83 Michael Kremer, "*Népességnövekedés és technológiai változás: Kr.e. egymillió évtől1990 a* ," 108Q. J. OF ECON. 3 (1993). (*idézi* SAIFEDEAN AMMOUS, THE BITCOIN STANDARD: THE DECENTRALIZED ALTER-NATIVE TO CENTRAL BANKING (2018).

84 ALPAYDIN, *Supra* note at 79,12.

85 *Id*.

86 SEBASTIAN RASCHKA & VAHID MIRJALILI, PYTHON MACHINE LEARNING, (22d. szerk. 2017).

87 Paul G. Allen & Mark Greaves, *The Singularity Isn't Near*, MIT TECH. REV. (201112,. október), https://www.technologyreview.com/s/425733/paul-allen-the-singularity-isnt-near/.

AGI elérése.[88] Másrészt Max Tegmark szerint a vita alapvető igazsága - hogy az emberiség valaha is létrehozza-e az AGI-t - továbbra is bizonytalan.[89] Tegmark azonban azt is kifejti, hogy a legtöbb AI-szakértő szerint az AGI 2047 körül fog megvalósulni.[90] Egy tudós szerint az AI hatásával kapcsolatos kérdések csak még sürgetőbbé válnak, ahogy közeledünk az exponenciális inflexiós ponthoz, és a növekedés hirtelen és drámai függőleges pályára áll.[91] Egyelőre az a kérdés, hogy a társadalom közeledik-e ehhez a fordulóponthoz, vagy még mindig a lassabb, fokozatos fejlődési szakaszban van.[92] Ma az emberiség legegyértelműbb útja az AGI létrehozása felé a mély megerősítő tanulás.

II. AGI FEJLESZTÉS

A gépi tanulás a mesterséges intelligencia egyik részterülete, amely a gépek azon képességére összpontosít, hogy megtanulják és lemásolják az emberi elmével kapcsolatos kognitív viselkedést.[93] A gépi tanulás általában adatbányászatot, mintafelismerést és természetes nyelvi feldolgozást foglal magában.[94] Ezek a technikák az elmúlt években egyre népszerűbbé váltak, mivel az internet megjelenése óta az emberek által előállított és összegyűjtött adatok mennyisége robbanásszerűen megnőtt.[95] A gépi tanulás legújabb áttörése a mély megerősítéses tanulás.[96] A mély megerősítéses tanulás a gépi tanulás két hagyományos modelljét - a felügyelt tanulást és a megerősítéses tanulást - ötvözi, hogy az algoritmusok az embertől függetlenül tanulhassanak.[97]

A mesterséges intelligencia szabályozásával kapcsolatos legtöbb tanulmány a gépi tanulás felügyelt vagy nem felügyelt módszereire összpontosít, mivel 2014-ig ez volt a gépi tanulás egyetlen két típusa, amelyet széles körben használtak.[98] Valójában a legtöbb jogi tanulmány a mély neurális hálózatokra, a felügyelt tanulási algoritmusok egyik típusára összpontosít.[99] A Google azonban 2013-ban kifejlesztette a tanulás egy új típusát, a "mély megerősítéses tanulást", amelyet később szabadalmaztatott.[100] Az 1980-as években úttörő szerepet játszó megerősítéses tanulás egy olyan gépi tanulási technika, amelyet a behaviorista pszichológia ihletett, ahol egy intelligens ágens hajlamát, hogy bizonyos módon cselekedjen, egy jutalmazási struktúra befolyásolja.[101] Az intelligens ágens olyan entitás, amely szenzorok segítségével információkat gyűjt a környezetéről, majd ezeket az információkat feldolgozva dönt arról, hogyan reagáljon a környezetére.[102] A mély megerősítéses tanulás a megerősítést kombinálja

88 *Id*.

89 TEGMARK, 5. lábjegyzet, 54.

pont. 90 *Id*. 157.

91 Michael Guihot et al., *Nudging Robots: Innovatív megoldások a mesterséges intelligencia szabályozására*, VAND20. J. ENT. & TECH. L. (2017)385,.400

92 *Lásd id*.

93 *Lásd általánosságban* ALPAYDIN, *fenti* megjegyzés. 79.

94 *Lásd* Michael Simon et al., *Lola v. Skadden and the Automation of the Legal Profession*, 20 YALE J.L. & TECH (234,2018253) (*idézi* Bernard Marr, *What Everyone Should Know About Cognitive Computing*, FORBES (2016. március 23., 3:28), https://www.forbes.com/sites/bernardmarr/2016/03/23/what-every-one-should-know-about-cognitive-computing/#5630f9005088.

95 *Lásd id*. 252.

96 *Lásd* TEGMARK, *fenti* jegyzet, lásd:5, TEGMARK, *fenti* jegyzet. 85.

97 *Lásd id*.

98 *Lásd id*. 83.

99 *Lásd általában* Calo, 6. lábjegyzet; *lásd még* John O. McGinnis, *Accelerating AI*, NW104. U. L. REV. 1253 (2010).

100 '862 Application, *fenti* megjegyzés 15.

101 RICHARD S. SUTTON & ANDREW G. BARTO, REINFORCEMENT LEARNING: AN INTRODUCTION (552017); *lásd még* TEGMARK, *supra* note at5, (2017). 85.

102 TEGMARK, *Supra* note at 5,84.

tanulás mély neurális hálózatok használatával.[103] A mély megerősítéses tanulás olyan megerősítéses tanulási algoritmusra utal, amely mély neurális hálózatot használ függvényközelítőként, és amelyet e rész későbbi részében ismertetünk.[104] Ez a rész először a mély neurális hálózatokat ismerteti. Másodszor, ez a rész a megerősítő tanulást ismerteti. Harmadszor, ez a rész elmagyarázza a mély megerősítő tanulást.

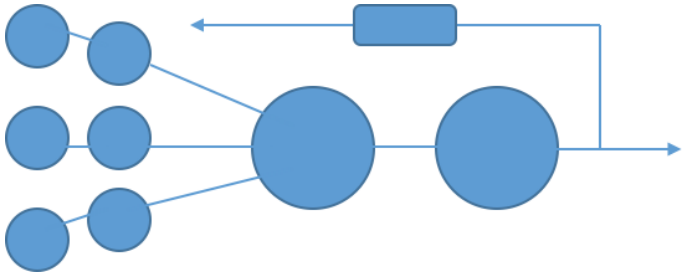
A. MÉLY NEURÁLIS HÁLÓZATOK

Az emberi agy "neuronoknak" nevezett feldolgozó egységekből áll.[105] Az agy minden egyes neuronja más neuronokkal szinapszisoknak nevezett struktúrákon keresztül kapcsolódik.[106] Egy biológiai neuron dendritekből áll - más neuronoktól érkező különböző elektromos impulzusok vevői -, amelyek a neuron sejttestében gyűlnek össze.[107] Amint a neuron sejtteste elegendő elektromos energiát gyűjtött össze, hogy meghaladjon egy küszöbértéket, a neuron elektromos töltést továbbít az agy más neuronjainak a szinapszisokon keresztül.[108] A biológiai agyban ez az információátvitel adja az alapját a modern ideghálózatok működésének.[109]

A mesterséges neuronok lényegében a biológiai neuronról mintázott logikai kapuk.[110] Mind a mesterséges, mind a biológiai neuronok különböző forrásokból kapnak bemenetet, és a bemeneti információt egyetlen kimeneti értékre képezik le.[111] A mesterséges neuronhálózat egymással összekapcsolt mesterséges neuronok csoportja, amelyek képesek egymás viselkedését befolyásolni.[112] A mesterséges neurális hálózatban a neuronok a biológiai agyban lévő szinapszisok erősségét modellező súlykoefficiensekkel vannak összekapcsolva.[113] A neurális hálózatokat nagy adathalmazok felhasználásával képzik.[114] A képzési folyamat lehetővé teszi a súlykoefficiensek beállítását, hogy a neurális hálózat kimenete vagy előrejelzése pontos legyen.[115] Miután a neurális hálózatot betanították, új adatokat táplálnak a hálózaton keresztül, hogy előrejelzéseket készítsenek.[116]

1957-ben Frank Rosenblatt közzétett egy algoritmust - a perceptront -, amely automatikusan megtanulja az optimális súlykoefficienseket egy mesterséges neurális hálózat számára.[117] A perceptron modellt az alábbiakban szemléltetjük:[118]

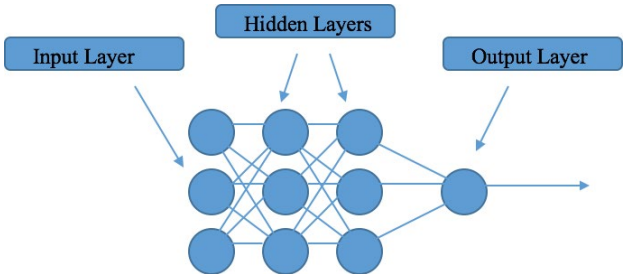
103 Fei-Fei Li, Justin Johnson, & Serena Yeung, *14. előadás: Mély megerősítéses tanulás*, STANFORD U. SCH. OF ENG'G (2017), https://www.youtube.com/watch?v=lvoHnicueoE (utolsó hozzáférés: 201913,. május).
104 *Id*.
105 ALPAYDIN, *Supra* note at 79,86.
106 *Id*.
107 RASCHKA & MIRJALILI, *uo.* 86. lábjegyzet, at 18.
108 *Id*.
109 ALPAYDIN, *Supra* note at 79,86.
110 KURZWEIL, *Supra* note at 55,38.
111 JOHN D. KELLEHER & BRENDAN TIERNEY, ADATTUDOMÁNY (1312018).
112 TEGMARK, *Supra* note at 5,72.
113 ALPAYDIN, *Supra* note at 79,88.
114 *Id*. 89.
115 *Id*.
116 KELLEHER, *Supra* note111 , at 127.
117 RASCHKA & MIRJALILI, *uo.* 86. lábjegyzet, at 18.
118 *Id*.



A perceptronban a bal szélső három kör a bemeneti értékeket $xj...m$, a hozzá tartozó súlyértékek pedig $_{j...m}w$ a bemeneti értékektől jobbra lévő három kör.[119] A bemeneti értékek és a súlyértékek aggregálásra kerülnek, jellemzően egy összegző egyenlet segítségével, amelyet az első nagy kör (balról jobbra) képvisel.[120] A második nagy kör a küszöbfüggvényt jelképezi, egy előre meghatározott értéket, amelynek túllépése esetén a modell 1 kimenetet jelez.[121] Ha a küszöbfüggvényt nem lépi túl, a modell 0-t ad ki.[122] A kimenetet a jobbra mutató nyíl jelzi.[123] A modell tetején lévő doboz a hibafüggvényt ábrázolja.[124] Abban az esetben, ha a modell kimenete hibás, akkor a hibafüggvény lép működésbe.[125] Ha a hibafüggvény működésbe lép, a súlyértékek a perceptron tanulási szabály szerint frissülnek.[126] A perceptron tanulási szabály formális ábrázolása a következő mint: $\Delta wj = \eta(^{(i)}y\ ^{i}- \hat{y})^{(i)}x$, ahol η a tanulási ráta, $y^{(i)}$ a valós osztálycímke a az i-edik gyakorló minta, és $y^{\wedge}$ a megjósolt osztálycímke.[127] A valódi osztálycímke a kimeneti címke, a megjósolt osztálycímke pedig a perceptron kimenete.[128]

Minden neurális hálózatnak van egy bemeneti és egy kimeneti rétege.[129] A bemeneti és kimeneti réteg között azonban a neurális hálózatok több rejtett réteget tartalmaznak.[130] A rejtett rétegek száma változhat, és az adott modelltől függ.[131] Fontos megjegyezni, hogy míg a perceptron modellek általában lineáris osztályozási feladatokra korlátozódnak, ez a korlátozás nem vonatkozik a többrétegű hálózatokra.[132] A többrétegű perceptron modell ugyanis egy univerzális approximátor, azaz olyan algoritmus, amely elegendő neuron esetén bármilyen függvényt képes a kívánt pontossággal közelíteni.[133] A mély neurális hálózat olyan hálózat, amely több rejtett réteggel rendelkezik.[134] Ez lehetővé teszi, hogy a neurális hálózat több absztrakciós réteget vegyen figyelembe.[135] Az alábbi ábra egy mély neurális hálózat egyszerű modellje.[136]

119 *Id*. 19.
120 *Lásd* ALPAYDIN, *fenti* jegyzet, lásd:79, ALPAYDIN, fenti jegyzet. 89.
121 *Lásd Id*.
122 *Lásd* RASCHKA & MIRJALILI, 86. lábjegyzet, a következő címen 18.
123 *Lásd* KURZWEIL, *Supra* note at 55,132.
124 *Lásd* RASCHKA & MIRJALILI, 86. lábjegyzet, a következő címen 18.
125 *Lásd* ALPAYDIN, *fenti* jegyzet, lásd:79, ALPAYDIN, fenti jegyzet. 90.
126 *Lásd id*.
127 *Lásd* RASCHKA & MIRJALILI, 86. lábjegyzet, a következő címen 21.
128 *Lásd id*. 22.
129 *Lásd* KURZWEIL, *Supra* note at 55,132.
130 *Lásd* ALPAYDIN, *fenti* jegyzet, lásd:79, ALPAYDIN, fenti jegyzet. 100.
131 *Lásd* KELLEHER & TIERNEY, 111,132. lábjegyzet.
132 *Lásd* ALPAYDIN, *fenti* jegyzet, lásd:79, ALPAYDIN, fenti jegyzet. 99.
133 *Lásd id*.
134 *Lásd* TEGMARK, *fenti* jegyzet, lásd:5, TEGMARK, *fenti* jegyzet. 76.
135 ALPAYDIN, *Supra* note at 79,88.
136 KELLEHER & TIERNEY, *Supra* note at111, (a következő idézetben szereplő illusztráción alapuló 132modell).



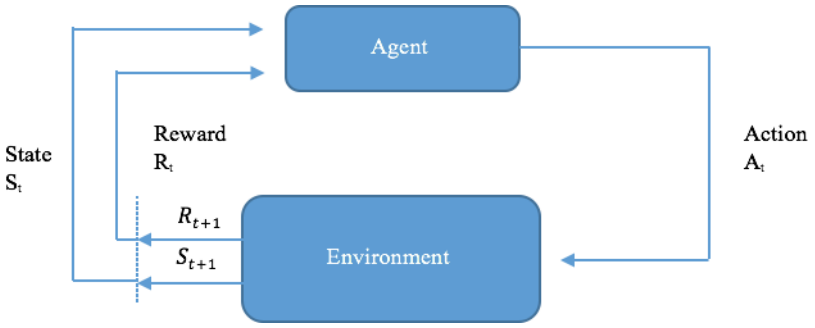
Minden neuron egy rejtett egységet képvisel egy rétegben, és a modell egy komplex tulajdonságát határozza meg.[137] A rejtett egységek megfelelnek a megfigyelt, de közvetlenül nem megfigyelt rejtett tulajdonságoknak.[138] A rejtett egységek egymást követő rétegei pedig a jellemzők absztrakciójának növekvő rétegeinek felelnek meg.[139]

Valójában a rejtett egységek minden egyes rétege egy-egy jellemző-kivonó funkcióként működik, mivel kissé bonyolultabb jellemzők elemzését biztosítja.[140] A jellemző-kivonás a dimenziócsökkentés módszere - a bemeneti jellemzők csökkentésének módszere -, amely lehetővé teszi a nyers bemenet átalakítását kimenetté oly módon, hogy az adattudósok megfigyelhessék az adatok rejtett jellemzőit.[141] A későbbi rejtett egységek a rejtett jellemzőket a korábbi jellemzők kombinálásával vonják ki a bemeneti tér egy kissé nagyobb részében.[142] A kimeneti réteg a teljes bemenetet figyeli, hogy végső előrejelzést készítsen.[143] Más szóval, a mély neurális hálózatok a kezdeti bemenetük bonyolultabb függvényeit tanulják meg, amikor minden rejtett réteg kombinálja az előző réteg értékeit.[144] Emellett a mély neurális hálózatok számos kontextusban kiválóan alkalmasnak bizonyultak előrejelzések készítésére.[145] Ezek a modellek azonban a tanuláshoz adatokra és legalább minimális emberi beavatkozásra van szükség a tanulási folyamat felügyeletéhez.[146] A megerősítéses tanulás egy újabb gépi tanulási technika, amely egyiket sem igényli.[147]

B. MÉLY MEGERŐSÍTÉSES TANULÁS

A megerősítéses tanulás egyfajta gépi tanulási technika, amelyet a behaviorista pszichológia ihletett.[148] Formálisan a megerősítéses tanulás egy ágens-környezet kölcsönhatáson keresztül írható le, a Markov-döntési folyamattal ("MDP").[149] Az alábbi modell az ágens-környezet kölcsönhatást írja le egy MDP-ben.[150]

137 ALPAYDIN, *Supra* note at 79,100.
138 *Id*.
139 *Id*.
140 KELLEHER & TIERNEY, *Supra* note at 111,135.
141 ALPAYDIN, *Supra* note at 79,102.
142 *Id*.
143 KURZWEIL, *Supra* note at 55,132.
144 ALPAYDIN, *Supra* note at 79,104.
145 *Lásd általában* ASHLEY, *fenti* megjegyzés. 31.
146 ALPAYDIN, *Supra* note at 79,106.
147 *Id*. 127.
148 SUTTON & BARTO, *Supra* note at 101,38.
149 Alex Kendall et al., *Learning to Drive in a Day*, (1201811,. szeptember) (nem publikált tanulmány) (elérés: https://arxiv.org/abs/1807.00412).
150 SUTTON & BARTO, *Supra* note 101, 38. o. (az előző idézetben szereplő illusztráción alapuló modell).



A környezet minden egyes időpontban, amikor a környezet létezik, állapotokból áll.[151] Az ágens cselekedetei az egyes állapotokban meghatározzák a környezet valószínűségi fejlődését.[152]

Kezdetben az ágens elé kerül a környezet állapota, amely több lehetséges cselekvést is tartalmaz.[153] Ezután az ágens végrehajt egy akciót az adott állapotban, és továbblép a környezet következő állapotába, ahol jutalmat kap.[154] Az ágens az ágens politikája alapján választja ki, hogy melyik akciót hajtja végre, amikor egy állapotot mutatunk neki.[155] A politika az a mód, ahogyan az ágens döntéseket hoz vagy cselekvéseket választ egy állapoton belül. [156] Például egy nagyfokú integritással rendelkező személynek van egy politikája, amely rutinszerűen irányítja a döntéshozatalát, hogy a helyes dolgot válassza, amikor etikai dilemmákkal szembesül. Hasonlóképpen, egy kapzsi személynek olyan politikája van, amely rutinszerűen irányítja a döntéshozatalt, hogy a legmagasabb dollárértéket hozó cselekvést válassza. A politika célja, hogy az ágens úgy haladjon előre a környezetben, hogy maximalizálja a jutalmat.[157]

Egy érték-funkció határozza meg az *s* állapotban való tartózkodás és a politika követésének értékét.

π a környezet végső állapotáig, amelyet végállapotnak nevezünk.[158] A végállapot zárja le az epizódot, amely egy környezet összes állapotából áll.[159] Az π politika π végrehajtásának várható értékét az *s* állapot mellett $V^{\pi}(s)$ jelöli. [160] Egy MDP kontextusában az V^{π} értékfüggvény egyenlő az π politika végrehajtásáért járó diszkontált jutalmak várható összegével:[161]

$$V^{\pi}(s) = \mathrm{E}[R(s_0) + \gamma R(s_1) + \cdots |s_0 = s, \pi(s)]$$

A várható jövőbeni jutalmakat γ diszkontfaktorral diszkontáljuk.[162] A diszkontfaktor jellemzően a következő: $0 < \gamma < 1$.[163] Ez lehetővé teszi az értékfüggvény meghatározását

151 Li, Johnson, & Yeung, *Supra* note. 103.
152 MYKEL J. KOCHENDERFER, DÖNTÉSHOZATAL BIZONYTALANSÁG MELLETT (772015).
153 SUTTON & BARTO, *fenti* 101,39. lábjegyzet.
154 KOCHENDERFER, *Supra* note, 152,77. o.; *lásd még* SUTTON & BARTO, *Supra* note,101, 77. o. 39.
155 SUTTON & BARTO, *fenti* 101,39. lábjegyzet.
156 KOCHENDERFER, *supra* note at 152,77.
157 SUTTON & BARTO, *fenti* 101,50. lábjegyzet.
158 Li, Johnson, & Yeung, *Supra* note. 103.
159 *Id*.
160 KOCHENDERFER, *supra* note at 152,80.
161 Ahmad El Sallab et al., *Deep Reinforcement Learning Framework for Autonomous Driving*, (Apr. 20178,), (nem publikált tanulmány) (elérhető a https://arxiv.org/pdf/1704.02532.pdf oldalon. *Lásd az A. függelékben a jelölések összefoglalóját*).
162 SUTTON & BARTO, 101,92. lábjegyzet.
163 *Id*.

véges időben, és lehetővé teszi, hogy a jelenbeli jutalmak értéke értékesebb legyen, mint a jövőbeli jutalmaké.[164] Az $\pi^*(s)$ optimális politikát úgy definiáljuk, hogy az a politika, amely maximalizálja a várható értéket a többi politikához képest.[165] Az MDP modell célja az optimális politika megtalálása:[166]

$$\pi^*(s) = arg \max_{\pi} V^{\pi}(s)$$

Egy adott MDP optimális politikájának megtalálását általában Q-tanulással oldják meg.[167] A Q-tanulás ezt a problémát egy Q-értékfüggvény maximalizálásával oldja meg: $\mathrm{Q}(s,\ a)$.[168] A Q-értékfüggvény egy állapot-cselekvés pár értékét írja le.[169] Valójában a Q-tanuló algoritmus célja az optimális Q-értékfüggvény felfedezése.

Q^* bármely állapot-cselekvés párosra.[170] A Bellman-egyenlet kifejezi egy állapot értéke és az azt követő állapotok értékei közötti kapcsolatot.[171] Az algoritmus a Q-értékfüggvény konvergenciájáig örökké folytatódik.[172] A Q-értékfüggvény konvergenciája Q-t* jelenti, és kielégíti a Bellman-egyenletet, amelyet a következőképpen határozunk meg:[173]

$$Q^*(s,a) = \mathbb{E}_{s' \sim \varepsilon}[r + \gamma \max_{a'} Q^*(s', a)|s, a]$$

Az π^* optimális π ágens politikája megfelel az Q^* által meghatározott cselekvésnek minden egyes állapotban.[174] Felmerül azonban az a probléma, hogy $Q(s,\ a)$ értékét minden állapot-akció párra ki kell számítani, ami számítási szempontból kivitelezhetetlen lehet.[175] Például minden olyan állapot-akció pár értékének kiszámítása, ahol a nyers bemenet pixelek egy Atari játékban, óriási számítási teljesítményt igényelne.[176] Az egyik megoldás a Q-érték függvény becslésére egy függvényapproximátor használata:[177]

$$Q(s,a;\emptyset) \approx (s,a)$$

Itt $\emptyset$ a függvény paramétereit jelenti.[178] Ha pedig a $\emptyset$-t egy mély neurális hálózat határozza meg, akkor az algoritmus egy mély megerősítő tanulási algoritmus, amelyet mély Q-hálózatnak ("DQN") nevezünk.[179]

A DQN egy mély tanulási modell, amely egy mély neurális hálózatot ("DNN") kombinál egy Q-tanulási algoritmussal.[180] A DQN a tapasztalatok visszajátszását használja

164 *Lásd* KOCHENDERFER, *Supra* note at 152,78.
165 *Lásd id*. 79.
166 Sallab et al., 71-72. o. (*lásd a* 161,71-72. pontot).
167 SUTTON & BARTO, *fenti* 101,108. lábjegyzet.
168 *Id*. 107.
169 862. sz. kérelem, a15, *fenti* megjegyzés 1.
170 Li, Johnson, & Yeung *fenti* jegyzet. 103.
171 SUTTON & BARTO, *fenti* 101,47. lábjegyzet.
172 Sallab et al., *Supra* note at 161,72.
173 862. sz. kérelem, a15, *fenti* megjegyzés 5.
174 MAXIM LAPAN, DEEP REINFORCEMENT LEARNING HANDS-ON, (1022018).
175 Li, Johnson, & Yeung, *Supra* note. 103.
176 LAPAN, *Supra* note at 174,125.
177 862. sz. kérelem, a15, *fenti* megjegyzés 5.
178 *Id*.
179 Li, Johnson, & Yeung, *Supra* note. 103.
180 Manon Legrand, *Deep Reinforcement Learning for Autonomous Vehicle Control Among Human Drivers*, at (2016-17-es 26tanév) (kiadatlan doktori értekezés, Université Libre de Bruxelles) https://ai.vub.ac.be/sites/default/files/thesis_legrand.pdf.

az algoritmus régi tapasztalatainak tárolása a neurális hálózat képzéséhez.[181] Egy tapasztalat egy megfigyelt állapot-akció párosból, a kapott azonnali jutalomból és a következő megfigyelt állapotból áll.[182] "Egy ágens tapasztalatát egy *t* időlépésnél $_{et-vel}$ jelöljük, és egy olyan tuple ($_{st}$, $_{at}$, $_{rt}$, $_{st+1}$), amely az aktuális $_{st}$ állapotból, az $_{at}$ választott akcióból, az $_{rt}$ jutalomból és a következő $_{st+1}$ állapotból áll".[183] Az összes időlépés tapasztalatait egy visszajátszási memóriában tároljuk, sok epizódon keresztül, és a DNN betanításához használjuk.[184] A DNN kimenete egy érvényes akciónak felel meg, mivel a DNN a Q-érték függvény közelítőjeként szolgál.[185] Így a hálózat egy előrecsatolásos áthaladása után a kimenetek az állapot-akció pár becsült Q-értékei.[186] Ez lehetővé teszi, hogy az algoritmus általánosíthasson a múltbeli tapasztalatok összegyűjtött adataiból.[187] Max Tegmark, az MIT professzora szerint ugyanis "a mély megerősítéses tanulás egy teljesen általános technika".[188]

III. A MESTERSÉGES INTELLIGENCIA SZABÁLYOZÁSA

A mesterséges intelligencia fenyegetésével foglalkozó jogi tudomány két különálló táborra oszlik.[189] Az egyik tábor elismeri a mesterséges intelligencia rosszindulatú és meggondolatlan használata által jelentett potenciális veszélyeket, míg a másik szerint a mesterséges intelligencia apokalipszise csupán a tudományos fantasztikum szüleménye.[190] E táborok egyike sem foglalkozik igazán a mesterséges intelligencia által jelentett egzisztenciális fenyegetésekkel a katasztrófa megelőzéséhez szükséges azonnalisággal.[191] Ez a rész először azokat az érveket tárgyalja, amelyek azzal az elképzeléssel kapcsolatosak, hogy a mesterséges intelligencia nem jelent veszélyt az emberiségre. Ezután ez a rész azokat az érveket tárgyalja, amelyek az AI-szabályozásban az iparági vezetők aggályait tükrözve fejlesztették a tudományt. Végül ez a rész a mesterséges intelligencia szabályozásának három folyamatban lévő és megválaszolatlan kérdésével foglalkozik.

Azok a tudósok, akik azt állítják, hogy a mesterséges intelligencia apokalipszise csupán tudományos fikció, tévednek.[192] Ezeket a tudósokat különösen John McGinnis képviseli, aki megjegyzi, hogy "az emberek által irányíthatatlanná váló gépektől való egzisztenciális rettegés és a politikai aggodalom a gépek pusztító ereje miatt a forradalmasított csatatéren" túlzó.[193] McGinnis az AGI-vel kapcsolatos problémákat valójában egy gondolkodási hibának tulajdonítja, amikor az emberek antropomorfizálják a mesterséges intelligenciát, és tévesen attól tartanak, hogy az AGI szükségszerűen az emberi rosszindulatot fogja tükrözni.[194] McGinnis tehát a barátságos mesterséges intelligencia lehetőségét sugallja, és arra ösztönöz, hogy hagyjunk fel azzal a feltételezéssel, hogy a mesterséges intelligencia

181 HADO VAN HASSELT, ARTHUR GUEZ, & DAVID SILVER, ASS'N FOR THE ADVANCEMENT OF ARTIFI- CIAL INTELLIGENCE, PROCEEDINGS OF THE THIRTIETH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE: DEEP REINFORCEMENT LEARNING WITH DOUBLE Q-LEARNING (2094,20162095).

182 Volodymyr Mnih et al., *Human-Level Control Through Deep Reinforcement Learning*, NATURE518 INT'L J. SCI. 529,529 (2015).

183 Legrand, *Supra* note at180, A 72."tuple" egy lista vagy tömbhöz hasonló adattárolási formátum. *Id*. 9.

184 VAN HASSELT, GUEZ, & SILVER, *fenti* megjegyzés. 181.

185 Legrand, *Supra* note at 180,27.

186 Mnih et al., *Supra* note. 182.

187 KOCHENDERFER, *supra* note at 152,124.

188 TEGMARK, *Supra* note at 5,85.

189 *Lásd* Calo, *Supra* note, 6,432. o.; *lásd még* Matthew U. Scherer, *Regulating Artificial Intelligence Systems:* 29 HARV. J.L. & TECH. 353 (2016).

190 *Lásd általában* McGinnis, 99. lábjegyzet; *lásd még* Scherer, 99. lábjegyzet. 189,394.

191 *Lásd általában* BOSTROM, *Supra* 22. lábjegyzet, at 85.

192 *Lásd id.*; lásd *még* TEGMARK, *fenti* megjegyzés. 5.
193 *Lásd* McGinnis, *Supra* note at 99,1254.
194 *Id*.

akaraterővel kell rendelkeznie, mint egy embernek.[195] Feltételezi, hogy az akaraterő hiánya semmissé teszi a gonosz mesterséges intelligenciát övező félelmet.[196]

Érdekes módon az antropomorf érv mindkét irányba hat. Nick Bostrom és Eliezer Yudkowsky ugyanis meggyőzően hozta fel az antropomorf érvet annak magyarázatára, hogy az emberek miért fogják drasztikusan alábecsülni a mesterséges intelligencia fejlődését.[197] Bostrom és Yudkowsky azzal érvel, hogy az emberi antropomorfizmus miatt a közvélemény úgy fogja érzékelni, hogy a mesterséges intelligencia fejlődése gyors ütemű lesz.[198] A mesterséges intelligencia emberi antropomorfizmusa itt is arra utal, hogy nem emberi entitásoknak emberi szintű intelligenciát tulajdonítanak.[199] Amint azt az I. részben Einstein és a falu bolondja összehasonlítása is mutatja, az intelligenciaszintek közötti különbség nagyobb relatív léptékben *jelentéktelen*.[200] Így az AI fejlődése az AGI és a szuperintelligencia felé a vártnál gyorsabb lesz, mivel a két intelligenciaszint közötti különbség szélesebb léptékben sokkal kisebb, mint azt az emberek felismerik.[201]

Egy másik jogtudós, Ryan Calo még nyersebben érvel amellett, hogy a mesterséges intelligencia nem jelent egzisztenciális fenyegetést az emberiségre, és hogy az AGI csupán a "képregények anyagát" jelenti.[202] Calo továbbá azt állítja, hogy "az AI apokalipszisre fordított aránytalan figyelem és erőforrások elvonhatják a döntéshozók figyelmét az AI közvetlenebb ártalmainak kezelésétől.". . ."[203] Azt állítja, hogy a gépi tanulás területén semmi sem utal arra, hogy az emberiség hamarosan képes lesz modellezni az emlősök, nem is beszélve az emberi intelligenciáról.[204] Ez az állítás nyilvánvalóan téves. A megerősítő tanulás és a Markov-döntési folyamatok ugyanis szó szerint modellezik az emberi kognitív funkciókat, a döntéshozatalt, a racionális cselekvőképességet és az intelligenciát.[205] Ráadásul az adattermelés, a számítási teljesítmény és a globális GDP exponenciális növekedése mind azt a következtetést támasztja alá, hogy az AGI hamarabb érkezik, mint az emberek gondolnák.[206]

Ezért az AGI által az emberiségre jelentett egzisztenciális fenyegetés azonnali kárt jelent. Ez a fenyegetés nem analóg egy terminátorszerű robot világuralomra törésével. Ehelyett ez a fenyegetés a mély megerősítéses tanulással tanuló ágensekből kifejlesztett AGI terméke.[207] Amint az AGI szintű ágensek létrejönnek, gyorsan képesek lesznek arra, hogy szoftverarchitektúrájukat hatékonyabban fejlesszék, mint bármelyik ember. Ezek az ágensek képesek lesznek bármilyen, jutalmazási rendszerrel korrelált cél elérésére egy virtuális környezetben. A mély megerősítő tanulási rendszerek már képesek rakéták, rakéták, autók és repülőgépek irányítására.[208] És az ezekhez szükséges szoftverek

195 *Id*. 1263-64.
196 *Id*.
197 BOSTROM, *Supra* note, 22,85. o.; *lásd* YUDKOWSKY, *Supra* note,75, 85. o. 21.
198 BOSTROM, *Supra* note at 22,85.
199 YUDKOWSKY, *Supra* note at 75,21.
200 *Id*.
201 BOSTROM, *Supra* note at 22,86.
202 Calo, *Supra* note at 6,432.
203 *Id*. 431.
204 *Id*. 432.
205 *Lásd még:* TEGMARK, *fentebb,* 5,85. o.; *lásd általában:* BOSTROM, *fentebb*, 85. o. 22,239.
206 BOSTROM, *Supra* 22. lábjegyzet, 1-4. pont.
207 C-SPAN, *supra* note 4.
208 Mnih et al., *Supra* note, 182,529. o.; *lásd még* Legrand, *Supra* note, 180,27. o.; *lásd még* Kendall et al, *supra* note ;149 *lásd még:* U.S. Patent No. to 8,678,321Bezos et al. (20145,. március).

alkalmazások nyílt forráskódúak.[209] Így ma már mindenki, aki rendelkezik internet-hozzáféréssel, potenciálisan hozzáférhet a világ legfejlettebb fegyverzetirányítási rendszereihez is.[210]

A jogi tudomány mégis teljesen figyelmen kívül hagyja ezt a megkerülhetetlen igazságot. Néhány jogtudós azonban tett lépéseket a helyes irányba anélkül, hogy kifejezetten foglalkozott volna az AGI szabályozásának kérdésével. Matthew Scherer például azt állítja, hogy a mesterséges intelligencia szabályozásának kiindulópontja egy olyan jogszabály kellene, hogy legyen, amely meghatározza a mesterséges intelligencia szabályozásának általános elveit.[211] Javasolja a mesterséges intelligencia fejlesztési törvényt ("AIDA"), amely létrehozna egy ügynökséget, amelynek feladata a mesterséges intelligencia rendszerek biztonságának tanúsítása lenne.[212] Az ügynökségnek szabályokat kellene alkotnia a mesterséges intelligencia meghatározására.[213] Az AIDA lényege, hogy a mesterséges intelligencia rendszerek biztonságának értékelésével kapcsolatos érdemi feladatot egy független, szakemberekből álló ügynökségre bízná, így a konkrét mesterséges intelligencia rendszerek biztonságáról szóló döntéseket elszigetelné a választási politika által gyakorolt nyomástól.[214]

Más tanulmányok a különböző szabályozási kereteket tárgyalják, amelyeket a mesterséges intelligenciával kapcsolatos kérdések elemzésére lehet alkalmazni, valamint néhány konkrét példát mutatnak be a mesterséges intelligencia szabályozásának problémáira.[215] A cikk meggyőzően érvel amellett, hogy a mesterséges intelligenciát, "függetlenül a benne rejlő lehetőségektől, óvatosan kell kezelni".[216] A szerzők az innováció előmozdítása érdekében árnyalt, érzékeny és alkalmazkodó szabályozási keretet szorgalmaznak.[217] Miközben az AI-szabályozás területén korlátozott előrelépés történt, az intelligens gépek etikájával és biztonságával kapcsolatos kutatások gyors növekedése nem hozott valódi előrelépést.[218] Ahogy az egyik írás megjegyzi, "[a] publikált tanulmányok nagy többsége nem tesz mást, mint arról vitatkozik, hogy a meglévő etikai iskolák közül, amelyek évszázadokon át az emberi társadalom igényeinek kielégítésére épültek, melyik lenne a helyes, ha mesterséges utódainkra alkalmaznánk".[219] Továbbá, még a terület progresszívebb tudományos munkássága is kvázi kizárólag a szűk értelemben vett mesterséges intelligenciára összpontosít, nem pedig az AGI-re.[220] Így a tudósok által javasolt szabályozási keretek egyike sem foglalkozott megfelelően az AGI fejlődésének számos fontos kérdésével.

209 *Üdvözöljük a Spinning Up in Deep RL!: User Documentation*, OPENAI (utolsó látogatás: 2019. március 20.) spinningup.openai.com.

210 *TensorFlow 2.0 Alpha is Available*, TENSORFLOW, (2019), https://www.tensorflow.org/install; *lásd még* RICHARD WU ET AL., AAAI 2017 FALL SYMPOSIUM SERIES, A FRAMEWORK USING MACHINE VISION AND DEEP REINFORCEMENT LEARNING FOR SELF-LEARNING MOVING OBJECTS IN A VIRTUAL ENVIRONMENT (2017), https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/16003/15319.

211 Scherer, *supra* at 189,394.

212 *Id*. 393.

213 *Id*. 394.

214 *Id*. 393.

215 *Lásd általában* Guihot et al., *Supra* note. 91.

216 *Id*. 454.

217 *Id*.

218 *Lásd* Roman Yampolskiy & Joshua Fox, *Safety Engineering for Artificial General Intelligence*, TOPOI (322172012).

219 *Id*.

220 *Id*.

A. AZ AGRÁR- ÉS VIDÉKFEJLESZTÉS KORTÁRS KÉRDÉSEI

A szakértők gyanúja szerint a kibertámadók hamarosan olyan stratégiákat fognak alkalmazni, amelyek mély megerősítő tanulási ágenseket használnak olyan támadások kidolgozására, amelyeket a jelenlegi technikai védelmi rendszerek nem képesek megakadályozni.[221] Egy tudós konkrétan részletezi a rosszindulatú mesterséges intelligencia szoftverek fejlesztésére vonatkozó irányelveket.[222] Az ösztöndíj azért íródott, hogy bemutassa, hogy gyakorlatilag lehetséges olyan gépi tanulási algoritmusokat fejleszteni, amelyek képesek ártani az embereknek.[223] Ráadásul az embereknek már ma is megvan a hatalmuk, hogy nukleáris fegyverek használatával elpusztítsák az életet a Föld bolygón, és egy AGI minden bizonnyal ugyanezzel a képességgel rendelkezne.[224] A modern mesterséges intelligenciával foglalkozó tudósok az AGI-k, különösen a mély megerősítő ágensek létrehozásának folyamatát a nukleáris fegyverek építéséhez hasonlítják.[225] Ez a rész három olyan konkrét kérdéssel foglalkozik, amelyeket a mesterséges intelligenciára vonatkozó megfelelő szabályozási keretnek figyelembe kellene vennie.

Az első kérdés a verseny problémája. Ha a szabályozó hatóságok megpróbálják felügyelni az AGI-t fejlesztő vállalatokat, akkor ez a felügyelet elfojtja az innovációt, és lehetővé teszi, hogy olyan országok, mint Kína és Oroszország, az Egyesült Államok előtt fejlesszenek AGI-t.[226] Valójában nagy a valószínűsége annak, hogy bármelyik AGI-t létrehozó szervezet döntő előnyre tesz szert a világ többi részével szemben.[227]

A DQN algoritmusokat például gyakran használják részvények kereskedelmére, ahol egy ügynök minden egyes állapotban képes egy részvény megvásárlására, eladására vagy tartására.[228] Az ágens célja a portfólió értékének maximalizálása.[229] A DQN-algoritmusok portfóliókezelésre való alkalmazása sikeresnek bizonyult.[230] Ha egy entitás képes lenne AGI-t létrehozni, akkor ez felhasználható lenne egy olyan ágens létrehozására, amely képes lenne a piacokat úgy manipulálni, hogy egyetlen szereplő minimális idő alatt rendkívüli vagyonra tegyen szert.[231] Ez lehetővé tenné egy ilyen entitás számára, hogy a tömegek által nem ismert, egységes központi hatalmi erővé fejlődjön.[232]

A verseny problémája még ijesztőbbé válik, ha figyelembe vesszük, hogy a mesterséges intelligencia területén ma a legnagyobb szereplők tőzsdén jegyzett vállalatok. Az olyan vállalatok, mint a Google, a Facebook, az Apple és a Microsoft a mesterséges intelligencia fejlesztésének legnagyobb szereplői közé tartoznak, és technológiájuk teljesítménye és skálázhatósága gyorsan növekszik.[233] Az e vállalati szereplők, a külföldi kormányok és az Egyesült Államok közötti hatalmi egyenlőtlenség további problémákat vet fel a szabályozók számára.[234] Ha a szövetségi kormány elkezdi szabályozni a mesterséges intelligenciát, óvakodnia kell attól, hogy a fejlődés ütemének belföldi lassítása minden bizonnyal hátrányos helyzetbe hozza az Egyesült Államokat a külföldi szereplőkkel szemben. A végső kérdés a

221 BRUNDAGE ET AL., *supra* note 5, at 34; *see also* HYRUM S. ANDERSON, ET AL., LEARNING TO EVADE STATIC PE MACHINE LEARNING MALWARE MODELS VIA REINFORCEMENT LEARNING, (22018) (ac- cessed at https://arxiv.org/abs/1801.08917).

222 Federico Pistono & Roman Yampolskiy, *Etikátlan kutatás: (*2016), (publikálatlan cikk) (elérhető: https://arxiv.org/abs/1605.02817).

223 *Id*.

224 Yampolskiy & Fox, *fenti* megjegyzés. 218.

225 BOSTROM, *Supra* note at 22,104.

226 TEGMARK, *Supra* note at 5,9.

227 BOSTROM, *Supra* note at 22,103.

228 LAPAN, *Supra* note at 174,217.

229 *Id*. 220.

230 *Lásd általában* Zhipeng Liang et al., *Deep Reinforcement Learning in Portfolio Management,* (Aug. 201829,), (publikálatlan tanulmány) (elérhető a https://arxiv.org/abs/1808.09940 oldalon).

231 TEGMARK, *fenti* megjegyzés, 5,15-16. o.

232 *Id*.

233 Guihot et al., *supra* note at 91,437.

234 *Id*.

A verseny problémája az, hogy a szabályozóknak a biztonság és a szabadság közötti egyensúlyozással kell szembenézniük. Ha a szabályozók nagyobb hangsúlyt fektetnek a biztonságra, azt a szabadság rovására teszik, amely lehetővé tette a hazai technológiai iparágak vezetőinek az innovációt. Másrészt, ha a szabályozók nagyobb hangsúlyt fektetnek a szabadságra, azt a választópolgárok biztonságának rovására teszik. Ezért a szabályozóknak olyan keretet kell kialakítaniuk, amely érzékeny a vállalatok, a külföldi kormányok és a nemzetbiztonsági ügynökségek közötti versenyre.

A második probléma, amellyel a szabályozó hatóságok szembesülnek, a "magányos farkas" koncepció, amelyben a fenyegetést elszigetelt incidensnek tekintik, nem pedig széles körű társadalmi problémának. A mesterséges intelligencia szabályozása sok szempontból a matematika vagy a számítástechnika szabályozásához hasonló. A mesterséges intelligencia kutatásához ugyanis csak egy személyi számítógépre van szükség.[235] Érdekes módon a tudósok megosztottak az AGI kifejlesztésére irányuló potenciális projekt méretét illetően.[236] Az egyik tudós megjegyzi, hogy az AGI-hez vezető út egy hatalmas kormányzati projekt részeként egy kis csoport vagy akár egyetlen egyén munkájából is megvalósulhat.[237] Az AGI-hez vezető út nagyságrendje nagymértékben függ a mesterséges intelligencia eléréséhez használt módszerektől.[238] Például, ha a teljes agy emulációjának jelenlegi módszereit alkalmazzák, akkor valószínű, hogy az AGI kifejlesztéséhez hatalmas mennyiségű kódot kell kidolgozniuk a szakértő informatikusoknak és mérnököknek.[239] Fontos megjegyezni, hogy míg maga a projekt hatalmas léptékű lehet, addig a mesterséges intelligenciától az AGI-ig való áttörés megvalósításával megbízott egyéni csoport nagyon kicsi lehet.[240] A Manhattan-projekt például csúcspontján nagyjából 130 000 embert foglalkoztatott.[241] Az atombombát azonban tudósok és mérnökök egy kisebb csoportja hozta létre, J. Robert Oppenheimer és Leslie Groves tábornok vezetésével a Los Alamos-i Tudományos Laboratóriumban.[242]

A magányos farkas problémájának egy másik kérdése akkor fog jelentkezni, ha a mesterséges intelligencia területén egyetlen ember fog áttörést elérni. Ebben az esetben lehetséges, hogy minden, amit jelenleg a mesterséges intelligenciáról tudunk, a süllyesztőbe kerül. A tudománytól nem idegenek az egyszerű, de forradalmi áttörések, amelyek gyökeresen megváltoztatják az emberiség természeti világról alkotott képét.[243] Egy tudós szerint azonban valószínű, hogy a szabályozó szervek a legtöbb, AGI kifejlesztésére potenciálisan képes emberről tudnának.[244] Bár meg kell jegyezni, hogy egy olyan megvilágosodás az AI terén, mint amilyen Einstein *Annus* Mirabilis-dokumentumaiban kifejezett fizikai megvilágosodása volt, nem zárható ki a lehetőségek köréből. Ezért lehetséges, hogy egyetlen egyén lesz az első, aki létrehozza az AGI-t, és röviddel ezután példátlan mértékű hatalomra tehet szert.[245] A szabályozóknak olyan keretrendszert kell kialakítaniuk, amely lehetővé teszi a magányos AGI támadások és fenyegetések azonosítására és megelőzésére szolgáló technológia alkalmazását.

A harmadik kérdés, amellyel a technológiai szabályozóknak szembe kell nézniük, az ellenőrzési probléma. Az ellenőrzési probléma két különböző módon elemezhető a megbízóügynöki keretrendszerben.[246]

235 BOSTROM, *Supra* note at 22,103.
236 *Id*.
237 *Id*. 101.
238 *Id*.
239 KURZWEIL, *Supra* note at 55,124.
240 BOSTROM, *Supra* note at 22,101.
241 F.G. GOSLING, A MANHATTAN PROJEKT: AZ ATOMBOMBA KÉSZÍTÉSE (541999).
242 *Id*. 35.
243 *Lásd általában* Albert Einstein, *On the Electrodynamics of Moving Bodies (*1905) *in* THE 2COLLECTED PAPERS OF ALBERT EINSTEIN: THE SWISS YEARS: WRITINGS, 1900-1909 (140Anna Beck trans., 1990),

https://einsteinpapers.press.princeton.edu/vol2-trans/154; *lásd még* BRIAN GREENE, FABRIC OF THE COSMOS (1282005).

244 BOSTROM, *Supra* note at 22,103.

245 *Id*.

246 *Id*. 155.

Az első keretrendszer akkor létezik, ha a projekt szponzora megbízóként, a tudósok és mérnökök egy csoportja pedig a projekt szponzorának megbízottjaként jár el.[247] Ebben a keretben az ellenőrzési probléma akkor jelentkezik, ha az AGI-t fejlesztő tudósok és mérnökök rosszindulatú célokra használják fel a munkájuk során szerzett ismereteket és információkat.[248] Például az Apple, a Google és a Facebook kutatói saját cégeik mesterséges intelligenciafejlesztésének eredményeképpen hatalmas hatalomra249 és arra a képességre tettek szert, hogy fejlett mesterséges intelligencia-rendszereket fejlesszenek ki vagy módosítsanak saját személyes hasznukra vagy mások kárára.

A második keretrendszerben a megbízó az emberi alkotó, az ügynök pedig a mesterséges intelligencia rendszer.[250] Ebben a keretrendszerben az irányítási probléma akkor jelentkezik, ha egy AGI-rendszert fejlesztenek ki, és annak cselekvései az alkotója által nem ellenőrizhetők.[251] Az AGI megfékezésére több különböző módszert mutattak be. Nick Bostrom például bokszolási módszereket javasolt az AGI információhoz való hozzáférésének felosztására és korlátozására.[252] Max Tegmark továbbá felvetette, hogy lehetséges egy "kapuőr mesterséges intelligencia" létrehozása, egy olyan szuperintelligencia, amelynek célja, hogy a lehető legkisebb mértékben avatkozzon be, hogy megakadályozza egy másik szuperintelligencia létrejöttét.[253] Ezért a szabályozóknak olyan keretet kell kialakítaniuk, amely ellenőrzi, hogy az AI-kutatók hogyan használják fel hatalmukat, és olyan keretet, amely lehetővé teszi az AGI-rendszerek szabályozását, hogy azokat emberi szereplők irányíthassák.

Összefoglalva, a mesterséges intelligenciát szabályozó hatóságok három fő problémája a verseny, a magányos farkas koncepció és az ellenőrzés. E problémák gyakorlati megoldásának egyik gyakorlati módja az önszabályozó AGI-technológia.[254] Ehhez lényegében az AGI értékeinek az AGI létrehozójának értékeivel összhangban lévő programozására lesz szükség.[255] Ennek a megoldásnak az egyik fő előnye, hogy lehetővé teszi, hogy az állami szabályozók viszonylag sötétben maradjanak az AI-technológia működésével kapcsolatban.[256] Ezzel a megoldással kapcsolatban azonban két jelentős probléma van. Először is, ha létezik egy olyan AGI-rendszer, amely képes szabályozni az összes többi AI-rendszert, akkor szükség lesz egy olyan szabályozási mechanizmusra, amely megfékezi a szabályozó AGI-t, hogy az ne váljon az emberek felett ellenőrzést gyakorló egységes hatalommá. Másodszor, az emberiségnek biztosítania kell, hogy a szabályozó AGI-t ne győzhesse le és ne győzhesse le semmilyen más mesterséges intelligencia vagy AGI. Valóban, ha létezne egy AGI, amely képes fejleszteni önmagát, akkor bármely emberi programozó képességei gyorsan messze lemaradnának, miközben Irving J. Good hírhedt szavai: "...az utolsó találmány, amit az embernek valaha is meg kell tennie..." prófétai jelleggel fognak visszhangozni.[257]

KÖVETKEZTETÉS

A tudományban a lehetőségek spektruma van lefektetve. Az egyik oldalon azok állnak, akik szerint a mesterséges intelligencia a közeljövőben örökre meg fogja változtatni az emberi életet, a másik oldalon pedig azok, akik szerint a mesterséges intelligencia apokalipszise csupán tudományos fikció. Az igazság az, hogy

247 *Id*. 155-56.
248 *Id*.
249 Guihot et al., *supra* note at 91,455.
250 BOSTROM, *Supra* note at 22,156.
251 TEGMARK, *Supra* note at 5,187.
252 BOSTROM, *Supra* note at 22,156.
253 TEGMARK, *Supra* note at 5,176.
254 Guihot et al., *Supra* note, 91,433-37. o.

255 TEGMARK, *Supra* note at 5,261.
256 BOSTROM, *Supra* note at 22,156.
257 Good, *fenti* megjegyzés 65.

hogy egyik tábor sem érti teljesen az AGI-t vagy annak a világunkra gyakorolt hatását.[258] 1988-ban bekövetkezett halálakor Richard Feynman Nobel-díjas fizikus tábláján a következő szavak szerepeltek: "Amit nem tudok létrehozni, azt nem értem".[259] Ebből következik, hogy amíg az emberiség nem hozza létre az AGI-t, addig az emberiség számára felfoghatatlan. Ez a valóság ironikus sorsot jelent az emberiség számára. Az emberiségnek ugyanis először meg kell értenie az AGI-t ahhoz, hogy irányítani tudja, de az emberiség nem értheti meg az AGI-t, amíg nem hozza létre. Továbbá Max Tegmark szerint "fogalmunk sincs, mi fog történni, ha az emberiségnek sikerül emberi szintű AGI-t létrehoznia".[260] Így nem vehetjük biztosra, hogy az AGI létrehozása pozitív kimenetelű lesz.[261] A mesterséges intelligencia területén ugyanis általános konszenzus van abban, hogy egyetlen szabályrendszer sem képes kontrollálni mindazt, amit az AGI szabályozása megkövetel.[262] De bármennyi mintát is lehet felismerni és trendeket követni, a mesterséges intelligencia jövője nem fog magától megtörténni.[263]

A legtöbb ember azt gondolja, hogy a múlt determinisztikus kapcsolatban áll a jövővel, de az igazság az, hogy a jövő alapvetően bizonytalan.[264] A huszadik század eleje óta az emberiség meggyőző bizonyítékokkal rendelkezik arról, hogy az emberek által a mindennapi tapasztalataikban érzékelt téridő egésze csak az egyén szubjektív megfigyeléséhez képest létezik.[265] A kvantumfizikában, valamint a fekete lyukak középpontjában pedig a klasszikus fizika és a téridő törvényei teljesen összeomlanak.[266] Ez azért fontos, mert a létezés alapvető szintjén a téridő nélkül a masszív részecskék függetlensége elpárolog, és az ember által érzékelt idő előrehaladása megszűnik létezni.[267] Ezt bizonyítják a szuperpozíció, a nem-lokalitás, az időszimmetria és a kvantum összefonódás elvei.[268] Valójában a jövő, csakúgy, mint a múlt, a kvantumbizonytalanság alapvető állapotában létezik. A mesterséges intelligencia a kulcsa annak, hogy az ilyen bizonytalanság mellett is maximalizáljuk a jólét valószínűségét. Ezért azonnal dolgoznunk kell a biztonságos és virágzó jövő megteremtésén.[269]

258 *Lásd* BOSTROM, *fenti* megjegyzés (azzal 22érvelve, hogy már egy szimulációban élünk; ez egy valós lehetőség, és erős érv, hogy egy entitás, esetleg a téridőben, megérti az AGI-t).

259 *Lásd* Michael Way, "*Amit nem tudok létrehozni, azt nem értem*", J. OF CELL SCI. (2017), http://jcs.biologists.org/content/joces/130/18/2941.full.pdf.

260 *Lásd* TEGMARK, *fenti* jegyzet, lásd:5, TEGMARK, *fenti* jegyzet. 156.

261 *Lásd* PETER THEIL, ZERO TO ONE (1952014).

262 *Lásd* Yampolskiy & Fox, *Supra* note. 218.

263 *Lásd* THEIL, *fenti* megjegyzés 261.

264 *Lásd általában* Einstein, *Supra* note 243.

265 *Lásd általában id*.

266 *Lásd* STEPHEN HAWKING, A BRIEF HISTORY OF TIME 111 (1996); lásd *még* ROBERT J. SPITZER, NEW PROOFS FOR THE EXISTENCE OF GOD (1232010).

267 *Lásd* GREENE, *fenti* jegyzet, 243,192-93.

o. 268 *Lásd id*. 199-208.

269 *Lásd* THEIL, *fenti* megjegyzés 261.

A függelék

A jelölés összefoglalása	
Jelölés	**Jelentése**[270]
$Q^*(s,a)$	Az *a* cselekvés értéke az optimális politika szerint
γ	Kedvezménytényező
$\mathbb{E}[x]$	Véletlen változó várakozása
$arg\max_a f(a)$	Az a olyan értéke, amelynél $f(a)$ maximális értéket vesz fel.
r	Jutalom
st	Állapot t időpontban
π	Politika
$\pi*$	Optimális politika
$V^{\pi}(s)$	A politika végrehajtásának várható értéke egy adott államból.

270 *Lásd általában* SUTTON & BARTO, *fenti* megjegyzés. 101.