Közpolitika és szuperintelligens mesterséges intelligencia: egy vektormező-megközelítés[1]

(2018) 4.3 verzió (első verzió: 2016).

Nick Bostrom†, Allan Dafoe†, Carrick Flynn†

[in Liao, S. M. M. (szerk.): *A mesterséges intelligencia etikája* (Oxford University Press, 2020)]

[www.nickbostrom.com/papers/aipolicy.pdf]

ABSZTRAKT

Megvizsgáljuk a szuperintelligens mesterséges intelligencia spekulatív kilátásait és annak normatív következményeit a kormányzásra és a globális politikára nézve. A gépi szuperintelligencia olyan átalakító fejlemény lenne, amely számos politikai kihívást és lehetőséget jelentene. Ez a tanulmány azonosítja ennek a feltételezett politikai kontextusnak egy sor jellegzetes vonását, amelyekből levezetjük a politikai desideráták korrelatív csoportját - olyan megfontolásokat, amelyeknek különös súlyt kell adni a következő esetekben a hosszú távú mesterséges intelligencia politika, mint más politikai kontextusokban. Hozzájárulásunk egy olyan "vektormezőt" ír le, amely megmutatja a lehetséges normatív alaphelyzetektől vagy szakpolitikai álláspontoktól való *irányváltozást.* A normatív változások irányára való összpontosításnak köszönhetően megállapításaink a szereplők széles köre számára relevánsak lehetnek, bár az elvontan megfogalmazott kívánalmaknak megfelelő konkrét szakpolitikai lehetőségek kidolgozása további munkát igényel.

Kulcsszavak : mesterséges intelligencia, etika, politika, technológia, globális kormányzás, mesterséges intelligencia, szuperintelligencia

A radikálisan átalakító mesterséges intelligencia kilátásai

Mára széles körben elterjedt nézetté vált, hogy a mesterséges intelligencia (AI) egy általános célú, átalakító potenciállal rendelkező technológia.[2] Ebben a tanulmányban arra a még mindig ellentmondásosabbnak és spekulatívabbnak tekintett lehetőségre összpontosítunk: a gépi szuperintelligenciára - az emberek kognitív képességeit messze meghaladó általános mesterséges intelligenciára -, amely képes forradalmi technológiai és gazdasági fejlődést hozni az ágazatok széles körében.

[1] Előző cím: "Policy Desiderata for Superintelligent AI" (Politikai kívánalmak a szuperintelligens mesterséges intelligenciához)

†Governance of AI Program, Future of Humanity Institute, Oxfordi Egyetem.

A hozzászólásokért és vitáért köszönettel tartozunk Stuart Armstrongnak, Michael Barnettnek, Seth

Baumnak, Dominic Beckernek, Nick Becksteadnek, Devi Borgnak, Miles Brundage-nek, Paul Christianónak, Jack Clarknak, Rebecca Crootofnak, Richard Danzignak, Daniel Dewey-nak, Eric Drexlernek, Sebastian Farquharnak, Sophie Fischernek, Ben Garfinkelnek, Katja Grace-nek és Tom Grantnek, Hilary Greaves, Rose Hadshar, John Halstead, Robin Hanson, Verity Harding, Sean Legassick, Wendy Lin, Jelena Luketina, Matthijs Maas, Luke Muehlhauser, Toby Ord, Mahendra Prasad, Anders Sandberg, Carl Shulman, Andrew Snyder-Beattie, Nate Soares, Mojmir Stehlik, Jaan Tallinn, Alex Tymchenko és több névtelen bíráló. Ezt a munkát részben a H2020 Európai Kutatási Tanács és az Élet Jövője Intézet támogatásával végeztük.

[2] Az ilyen jellegű jelentések számos példája között lásd West & Allen 2018.

a mai civilizációra jellemzőnél sokkal rövidebb időskálán. Ebben a tanulmányban nem azt állítjuk, hogy ez egy hihető vagy valószínű fejlemény; [3]inkább annak néhány aspektusát elemezzük*, hogy mi következne,* ha a radikális gépi szuperintelligencia ebben az évszázadban bekövetkezne.

Különösen a gépi intelligencia forradalmának a kormányzásra és a globális politikára gyakorolt hatásaival foglalkozunk. Mi lenne a kívánatos közpolitikai megközelítés abban a feltételezésben, hogy közeledünk a gépi szuperintelligencia korszakába való átmenethez? Milyen általános tulajdonságokat kellene keresni azokban a javaslatokban, amelyek arra irányulnak, hogy a világ hogyan kezelje az ilyen átmenetből adódó kormányzási kihívásokat?

Ezeket a kérdéseket tágan értelmezzük. Így "kormányzás" alatt nem csak az államok tevékenységét értjük, hanem a transznacionális kormányzást is, amely [4]magában foglalja a mesterséges intelligenciával foglalkozó technológiai cégek, befektetők, nem kormányzati szervezetek és más érintett szereplők által kialakított normákat és megállapodásokat; valamint a sokféle globális hatalom, amely az eredményeket alakítja.[5] És bár az etikai megfontolások relevánsak, nem merítik ki a
a vizsgálat hatókörét - a fontos választói csoportok prudenciális érdekeire, valamint a technikai és politikai megvalósíthatóság szempontjaira összpontosító kívánatos szempontokat kívánunk figyelembe venni. Úgy véljük, hogy a radikális kontextusban az általunk vizsgált irányítási kihívások sok tekintetben eltérnek azoktól a kérdésektől, amelyek a közeljövőben megvalósuló mesterséges intelligencia fejlesztésekkel kapcsolatos vitákat uralják.

Talán hasznos lenne röviden szólni arról, hogy milyen képességeket képzelünk el a szuperintelligencia korszakába való átmenet során. Ahogyan elképzeljük a forgatókönyvet, olcsó, általánosan intelligens gépeket fejlesztenek ki, amelyek majdnem az egész világot helyettesíthetik.
emberi munka, beleértve a tudományos kutatást és egyéb feltalálói tevékenységet.[6] A gép korai változatai
szuperintelligencia gyorsan fejlettebb változatokat építhet, ami hihetően egy "intelligencia" kialakulásához vezethet.
robbanás".[7] A gépi intelligencia felgyorsulása a technológiai fejlődés más formáit is elősegítheti. a fejlődés, amely innovációk sokaságát eredményezi, például az orvostudomány és az egészségügy, a közlekedés, az energia, az oktatás és a környezeti fenntarthatóság területén. A gazdasági növekedés mértéke drámai mértékben, [8]valószínűleg több nagyságrenddel növekedne.[9]

Ezek a fejlemények kihívást jelentenek majd annak biztosítása szempontjából, hogy a mesterséges intelligencia fejlesztése, alkalmazása és szabályozása felelősségteljes és általánosan előnyös módon történjen. Néhány, a mesterséges intelligenciával kapcsolatos kormányzási kérdés már elkezdődött, mint például a halálos autonóm fegyverek etikája, a mesterséges intelligenciával [10]kiegészített

[3] Mindazonáltal érdemes megjegyezni, hogy sok AI-kutató komolyan veszi a szuperintelligencia lehetőségét ebben az évszázadban. Sőt, a gépi tanulással foglalkozó közösségen belül a többségi vélemény szerint valószínűbb, hogy az emberi szintű gépi intelligencia kifejlesztése (2050Müller & Bostrom 2016) vagy (2060Grace et al. 2018), és valószínű (75%), hogy a szuperintelligencia évek30 múlva fog kialakulni.
[4] Hale & Held 2011.

[5] Barnett & Duvall 2005.
[6] Kivételt képezne, ha kifejezetten az emberi munkaerő iránt lenne kereslet, például ha a fogyasztók a "kézzel" készült termékeket részesítenék előnyben.
[7] Jó 1965.
[8] Nordhaus 2015.
[9] Hanson ch2016,. 16.
[10] Nehal et al. 2016.

felügyelet, [11]méltányosság, elszámoltathatóság és átláthatóság a következetes algoritmikus döntésekben,[12]
és a hazai szabályozási keretek kialakítása.[13] A gépi szuperintelligenciára való áttérés, különösen az Európai Unió, jelentős, sőt egzisztenciális kockázatot jelent.[14] Az elmúlt években több kormányzati
a szervek jelentéseket készítettek és nemzeti stratégiákat jelentettek be a mesterséges intelligenciáról, beleértve a kapcsolódó kormányzási kihívásokat is.[15]

E tanulmány céljaira a szuperintelligencia lehetséges megjelenése ebben az évszázadban, valamint az ezzel kapcsolatos egyéb kiegészítő állítások *feltételezéseknek* tekinthetők - nem állítjuk, hogy elegendő bizonyítékot kínálunk arra, hogy ezek hihetőek, de segítenek meghatározni azokat a hipotetikus kormányzási forgatókönyveket, amelyeket elemezni kívánunk. Az az olvasó, aki meg van győződve arról, hogy valamelyik állítás téves, elemzésünket tekintheti (esetleg elgondolkodtató) intellektuális gyakorlatnak. Azok az olvasók, akik valamilyen pozitív valószínűséget tulajdonítanak ezeknek a kilátásoknak, úgy tekinthetik hozzájárulásunkat, mint erőfeszítést arra, hogy beszélgetést kezdjünk arról, ami a század későbbi szakaszában a legfontosabb politikai kérdéssé válhat: hogyan nézhet ki a kormányzás kívánatos megközelítése a gépi szuperintelligencia korszakában.

A normatív elemzés "vektormező" megközelítése

Tegyük fel, hogy optimista módon, a legáltalánosabban fogalmazva, átfogó célunk egy széles körben vonzó és befogadó közeli és hosszú távú jövő megvalósításának biztosítása, amely végül is megvalósítja az emberiség kívánatos fejlődési lehetőségeit, miközben tekintettel vagyunk minden olyan lényre, akiknek érdekeit döntéseink érinthetik. A gépi szuperintelligencia világára vonatkozó ideális kormányzási javaslat tehát olyan lenne, amely elősegíti ezt a célt.

De mit jelentene ez a homályos törekvés a gyakorlatban? Természetesen sokféle nézet létezik a különböző értékek és etikai normák relatív fontosságáról, és sokféle szereplő (államok, cégek, pártok, egyének, nem kormányzati szervezetek stb.) van, akik különböző ideológiai elkötelezettséggel és különböző preferenciákkal rendelkeznek a jövőbeli társadalom megszervezésének, valamint az előnyök és felelősségek elosztásának módját illetően. E sokrétűség fényében az egyik mód az lenne, ha egy bizonyos normatív norma mellett érvelnénk, és megpróbálnánk megmutatni, hogy ez mennyire vonzóbb vagy racionálisan védhetőbb, mint az alternatívák. Mind a normatív etikában, mind a tágabb értelemben vett politikai diskurzusban gazdag irodalom létezik, amely erre tesz kísérletet. Ebben a tanulmányban azonban nem célunk, hogy egy bizonyos alapvető etikai elmélet, normatív perspektíva, társadalmi választási eljárás vagy politikai preferencia mellett érveljünk.

Egy másik módszer az lenne, ha egyszerűen feltételeznénk egy bizonyos normatív standardot, érvek nélkül, majd megvizsgálnánk, hogy mi következik ebből az adott kérdéssel kapcsolatban; majd esetleg megismételnénk ezt az eljárást különböző lehetséges normatív standardok esetében. Itt szintén nem ezt fogjuk tenni.

[11] Calo 2010.
[12] FAT/ML 2018.

[13] Scherer 2016.
[14] Yudkowsky 2008; Bostrom 2014; Russell et al. 2016.
[15] Lásd például: House of Lords Select Committee on Artificial Intelligence (Lordok Házának mesterséges intelligenciával foglalkozó bizottsága). 2018.

Ehelyett ebben a dokumentumban azt a megközelítést alkalmazzuk, hogy megpróbálunk némileg semlegesek lenni a befolyásos szereplők között általánosan elfogadott normatív nézetek, ideológiák és magánérdekek között. Ezt úgy tesszük, hogy a *politikai irányváltásra* összpontosítunk - számos lehetséges értékelési szempontból -, amelyet a bevezetőben vázolt, radikálisan átalakuló gépi szuperintelligencia forgatókönyvében várhatóan bekövetkező különleges körülmények összessége von maga után.

Más szóval, arra törekszünk, hogy (metaforikusan vagy minőségileg) felvázoljuk a politikai következmények "vektormezőjét", amely a lehetséges normatív álláspontok széles skálája szempontjából releváns. Egyes politikai ideológiák például azt vallják, hogy a gazdasági egyenlőség a közpolitika egyik központi fontosságú célkitűzése, míg más ideológiák szerint a gazdasági egyenlőség nem különösebben fontos, vagy az államoknak csak nagyon korlátozott felelősségük van e tekintetben (pl. a szegénység legszélsőségesebb formáinak enyhítése). A vektorméret-megközelítés ezután megkísérelhetne olyan irányú politikai változtatási következtetéseket levezetni, amelyeket sematikusan a következőképpen ábrázolhatnánk: "Bármilyen nagy hangsúlyt is fektessenek *X-re* az államok a jelenlegi körülmények között a gazdasági egyenlőség céljára, vannak bizonyos különleges körülmények *Y*, amelyek várhatóan fennállnak a fent leírt radikális mesterséges intelligencia kontextusában, és amelyek arra késztetik Önt, hogy úgy gondolja, hogy az államoknak *e* körülmények között inkább a gazdasági egyenlőség céljára kellene fY(*X*) hangsúlyt fektetniük.
Az ötlet az, hogy *f* itt egy viszonylag egyszerű függvény, amely a lehetséges függvények egy terében van definiálva.
értékelési normák vagy ideológiai álláspontok. Például *f* egyszerűen hozzáadhatna egy kifejezést *X-hez*, amely megfelelne annak az állításnak, hogy a gazdasági egyenlőségnek adott hangsúlyt *Y* körülmények között (az összes figyelembe vett ideológiai álláspont szerint) bizonyos mértékben növelni kell. Vagy *f* egy bonyolultabb történet elmondását is megkövetelhetné, talán a következők szerint: "Bármilyen nagy hangsúlyt is fektetünk a gazdasági egyenlőségre mint politikai célkitűzésre a jelenlegi körülmények között, *Y* körülmények között másképp kell elképzelni a gazdasági egyenlőséget - a gazdasági egyenlőtlenség bizonyos dimenziói valószínűleg irrelevánssá válnak, más dimenziók pedig valószínűleg a jelenleginél fontosabbá vagy politikailag relevánsabbá válnak". (Az egyenlőséggel kapcsolatos kérdéseket az alábbi "elosztás" című részben tárgyaljuk.)

Ez a vektoros megközelítés csak annyiban eredményes, amennyiben vannak bizonyos minták abban, hogy a különleges körülmények hogyan hatnak a szakpolitikai értékelésekre a különböző értékelési pozíciókból. Ha a radikális mesterséges intelligencia kilátásai teljesen eltérő és sajátos következményekkel járnának minden egyes ideológia vagy érdekplatform számára, akkor az f függvény nem lenne több egy keresőtáblánál. A politikai elemzésnek ekkor vissza kellene térnie a fent említett eljárásmódokhoz, azaz vagy megpróbálna meghatározni (vagy egyszerűen csak feltételezne) egy egyedileg helyes vagy megfelelő normatív standardot, vagy pedig a lehetséges standardok egy sorát vizsgálná meg, és külön-külön vizsgálná meg azok politikai következményeit.

Mi azonban azt állítjuk, hogy legalább néhány érdekes mintát találhatunk az f-ben, és a következőkben igyekszünk ezek közül néhányat jellemezni. Ehhez először is azonosítunk néhány olyan szempontot, amelyekben a szuperintelligens mesterséges intelligencia kilátásai *különleges körülményeket* jelentenek - olyan *kihívásokat* vagy lehetőségeket, amelyek vagy egyediek az ilyen mesterséges intelligencia kontextusában, vagy várhatóan szokatlan módon vagy szokatlan mértékben jelennek meg. Ezután elmagyarázzuk, hogy ezek a különleges

körülmények milyen viszonylag egyértelmű következményekkel járnak a politikára nézve abban az értelemben, hogy vannak bizonyos politikai tulajdonságok, amelyek sokkal fontosabbak ezekben a különleges körülmények között (mint az ismertebb körülmények között) számos széles körben osztott prudenciális és morális preferencia kielégítése szempontjából. Ezeket a különösen releváns és fontos politikai tulajdonságokat a *desideráták*, vagyis a kívánatos tulajdonságok halmazaként fejezzük ki. A kívánatos tulajdonságokat négy címszó alá soroljuk

(hatékonyság, allokáció, népesség és folyamat), tehát a politika bizonyos irányokba való elmozdításának indoklására szolgálnak (ahhoz képest, hogy mi lenne a preferált politikai pont, ha a különleges körülményeken kívül működnénk).

A fejlett mesterséges intelligencia irányítására vonatkozó határozott javaslat ideális esetben nagymértékben megfelelne mindezen kívánalmaknak. Lehetnek olyan további kívánalmak is, amelyeket itt nem határoztunk meg; nem állítjuk, hogy listánk teljes. Továbbá, egy erős politikai javaslatnak feltehetően számos más normatív, prudenciális és gyakorlati megfontolást is integrálnia kellene, amelyek vagy sajátosak az egyes értékelési álláspontokhoz, vagy nem jellemzőek a radikális mesterséges intelligencia kontextusára. Hozzájárulásunk célja, hogy kiemeljünk néhány olyan témát, amelyet érdemes szem előtt tartani annak további vizsgálatakor, hogy miként kell megközelítenünk a kormányzási és globális politikai kihívásokat a szuperintelligens mesterséges intelligencia kilátásainak fényében.[16]

Hatékonyság

E címszó alatt azokat a kívánalmakat csoportosítjuk, amelyek a rendelkezésre álló tortaszelet méretének védelmével vagy növelésével kapcsolatosak. Egy eredmény akkor lenne nem hatékony, ha Pareto-hátrányban lenne valamely más lehetséges eredménnyel szemben - például ha erőforrás-pazarlással, a fejlődési lehetőségek elpazarlásával, a kölcsönösen előnyös együttműködésből származó elérhető nyereség elvesztésével stb. járna. A nagyobb hatékonyság kívánatos volta általában magától értetődőnek tekinthető; a hatékonyságnak azonban vannak olyan dimenziói, amelyek a radikális mesterséges intelligencia átalakításával összefüggésben különös jelentőséget kapnak. Ezek közé tartozik a technikai lehetőség, az AI-kockázat, a katasztrofális globális koordinációs hibák lehetősége és a turbulencia csökkentése, amelyeket az alábbiakban sorra veszünk.

Technológiai lehetőség

A gépi szuperintelligencia (az ebben a tanulmányban elképzelt típus) képes lenne a termelési lehetőségek határát sokkal messzebbre és sokkal gyorsabban kiterjeszteni, mint ami normális körülmények között lehetséges. A szuperintelligens mesterséges intelligencia rendkívül általános célú technológiai előrelépést jelentene, amely megszüntethetné az emberi munkaerő iránti igény nagy részét, és tömegesen növelné a teljes tényezőtermelékenységet. Az ilyen mesterséges intelligencia különösen a kutatás-fejlesztés terén tudna gyors előrelépést elérni, és felgyorsítaná a a technológiai érettség megközelítése.[17] Ez lehetővé tenné a gyors külső birodalom felhasználását csillagászati erőforrások, beleértve a települést is, amelyek hozzáférhetővé válnának az automatizált önreprodukáló űrhajó.[18] Ez egy hatalmas belső fejlődési területet is megnyitna, így
az egészség, az élettartam és a szubjektív jólét nagymértékű javulása, gazdagabb élettapasztalatok, önmagunk és mások mélyebb megértése, és szinte minden szempontból történő finomítás.
a létezésnek az a formája, amelyet mi választunk.[19] Így mind a kifelé irányuló kiterjedt növekedés,

[16] Elemzésünket olyan kívánatos adatokra korlátozzuk, amelyek megfelelnek egy alapvető univerzalizálhatósági kritériumnak. Például, ha van olyan szempont, amelyben a különleges körülmények a szokásosnál erősebb okot adnának *A* szereplőnek arra, hogy kárt okozzon *B szereplőnek*, és *B* szereplőnek a szokásosnál erősebb okot arra, hogy kárt okozzon *A szereplő*nek, akkor bizonyos értelemben lenne egy

általános minta, amelyet fel lehetne ismerni és a következő politikai ajánlássá lehetne desztillálni: "helyezzünk nagyobb hangsúlyt egymás megtámadására". De ebben az általánosított formában a szakpolitikai változás senki számára nem lenne kívánatos; így mivel nem válik általánosíthatóvá, nem vennénk fel kívánatosnak.

[17] A "technológiai érettség" alatt olyan képességek elérését értjük, amelyek a gazdasági termelékenység és a természet feletti ellenőrzés olyan szintjét teszik lehetővé, amely megközelíti a megvalósítható maximumot (Bostrom 2013). [18] Tipler 1980; Armstrong & Sandberg 2013.

[19] Pearce 1995; Bostrom 2005; Bostrom 2008.

és az intenzív növekedés befelé irányuló iránya, drámai fejlődés követheti a szuperintelligencia fejlődését.

A meglepően magas növekedési plafon (és a gyors felemelkedés kilátása) miatt különösen fontosnak kell tartanunk, hogy ezt a potenciált ne pazaroljuk el. Ennek a desiderátumnak két aspektusa van: (a) a belső és külső termelési lehetőségek határát kifelé kell tolni, hogy a Földről származó élet *végül* elérje a teljes értékmegvalósítási potenciálját, és (b) ennek a fejlődésnek lehetőleg *elég hamar* meg kell történnie ahhoz, hogy mi (pl. a jelenleg létező emberek vagy bármely olyan szereplő, aki ezeket a kritériumokat használja a javasolt mesterséges intelligenciapályák értékelésére) élvezhessük az előnyök egy részét. E két szempont relatív súlya a szereplő értékrendjétől függ.[20]

Külön kiemelendő, hogy létezhet egy olyan technológiai szint, amely lehetővé tenné, hogy az emberi élettartamot a biológiai öregedés és a helyi balesetek ténylegesen ne korlátozzák - egy olyan szint, amelyet
nem sokkal a szuperintelligencia megjelenése után érhető el.[21] Következésképpen, a szereplők, akiknek fontos a saját hosszú távú túlélésük (vagy a családjuk vagy más létező emberek túlélése), a szuperintelligens mesterséges intelligencia kifejlesztése felé vezető út kívánatos volta meglehetősen érzékenyen függhet attól, hogy az elég gyors lesz-e ahhoz, hogy ezeknek az embereknek esélyt adjon arra, hogy az AI átmenet megmentse az életüket.[22]

Még ha eltekintünk is az élet meghosszabbításának lehetőségétől, az, hogy a létező emberek élete összességében mennyire jól alakul, meglehetősen érzékenyen függhet attól, hogy életüknek van-e olyan utolsó szakasza, amelyben megtapasztalhatják azt a jobb életszínvonalat, amelyet egy pozitív mesterséges intelligencia átmenet után érnének el.

Mesterséges intelligencia kockázata

A mesterséges intelligencia okozta pusztítás elkerülése politikai célként különös jelentőséget kap a jelen kontextusban, mivel valószínűsíthető, hogy az ilyen pusztítás kockázata - beleértve a különösen szélsőséges kimeneteleket, például az emberiség kihalását - nem lenne a gépek fejlődésével, hanem az emberiség pusztulásának kockázatával járna.
szuperintelligencia, nagyon kicsi.[23] A javasolt politika értékelésének egyik fontos kritériuma a
A hosszú távú mesterséges intelligencia-fejlesztés tehát azt jelenti, hogy mennyi minőséggel kiigazított erőfeszítést fordítanának a mesterséges intelligencia biztonságára és az ezen az úton történő támogató tevékenységekre. A megfelelő kockázatcsökkentő erőfeszítések közé tartozhat például a mesterséges intelligencia vezérlésére szolgáló skálázható módszerek alapkutatása, a mesterséges intelligencia fejlesztőinek ösztönzése a megfelelő technikák alkalmazására, és ~~általában véve olyan feltételek~~ elősegítése, amelyek biztosítják, hogy a szuperintelligens mesterséges intelligencia fejlesztése körültekintően és óvatosan történjen.

[20] Beckstead 4-52013,. fejezet; Bostrom 2003b.
[21] Talán digitális formában (Sandberg & Bostrom 2008) vagy biológiai formában, fejlett biotechnológiai vagy nanotechnológiai eszközökkel (Drexler 7. 1986,fejezet; Freitas 1999). Bizonyos értelemben már most is

lehetséges lehet, hogy egyes jelenleg létező egyének csillagászati élettartamot érjenek el, nevezetesen úgy, hogy hétköznapi eszközökkel életben maradnak, amíg egy intelligencia-robbanás vagy más technológiai áttörés meg nem történik. Szintén a krionika (Bostrom 2003c; Merkle 1994).

[22] Azok a szereplők azonban, akik nagyon magas diszkontrátával számolnak a jövőbeli életük időtartamára, inkább elhalasztják a szuperintelligenciát, amíg a halál küszöbén állnak, mivel a gépi szuperintelligencia megjelenése pillanatnyilag megnövekedett kockázati szintet jelenthet (lásd alább a "Mesterséges intelligencia kockázata" című részt).

[23] Bostrom 2014; Russell & Norvig, 1036-1040. o.2010,

A globális koordináció katasztrofális meghibásodásának lehetősége

A katasztrofális globális koordinációs kudarcok elkerülése szintén különös jelentőséggel bír a jelenlegi kontextusban, mivel az ilyen kudarcok viszonylag valószínűnek tűnnek. A katasztrofális koordinációs kudarc többféleképpen is bekövetkezhet.

A gépi szuperintelligencia lehetővé teheti olyan technológiák felfedezését, amelyek megkönnyítik az emberiség elpusztítását - például valamilyen biotechnológiai vagy nanotechnológiai alapú "világvége-eszköz" megépítésével, amelyet, ha egyszer már feltaláltak, olcsó és könnyen megépíthető. Egy ilyen könnyen hozzáférhető világvége-eszköz kifejlesztésének *előzetes* megállítása vagy *utólagos* megfékezése a globális megállapodás, felügyelet, korlátozás és együttműködés szélsőséges és újszerű formáit igényelné.

A koordinációs problémák a kockázatot növelő mesterséges intelligencia-technológiai verseny dinamikájához vezethetnek, amelyben a fejlesztők a szélnek vetik az óvatosságot, miközben azért versengenek, hogy elsőként érjék el a szuperintelligenciát.[24] A
A verseny dinamikája a biztonsági kutatásba történő befektetések csökkenéséhez vezethet, a késedelmes ellenőrzési módszerek telepítése és tesztelése iránti hajlandóság csökkenéséhez, valamint a jelentős számítási költséggel járó vagy a teljesítményt más módon akadályozó ellenőrzési módszerekre való támaszkodás lehetőségeinek csökkenéséhez.

Általánosabban, a koordinációs hibák a fejlett mesterséges intelligencia fejlesztése és alkalmazása során különböző típusú "versenyfutásokhoz" vezethetnek. Például a mesterséges elmék érdekeit védő jóléti rendelkezések erodálódhatnak egy olyan hiper-versenyképes globális gazdaságban, amelyben a digitális munkavállalókkal szembeni rossz bánásmód és kizsákmányolás elleni szabályozást bevezető joghatóságok versenyhátrányba kerülnek és marginalizálódnak. Az evolúciós dinamika nem kívánatos irányokba és olyan módon is alakíthatja a fejlődést, amelyet hatékony globális koordináció nélkül lehetetlen elkerülni.[25]

Ha a technológiai fejlődés növeli a globális koordináció katasztrofális kudarcának kockázatát, akkor még fontosabbá válik a koordinációs problémák megoldására szolgáló lehetőségek és mechanizmusok kidolgozása. Ez magában foglalhatja a meglévő globális kormányzás javítására irányuló fokozatos munkát.
mechanizmusok és az együttműködési normák megerősítése.[26] Ez magában foglalhatja a következők előnyben részesítését is
olyan fejlődési utak, amelyek döntő stratégiai előnyt biztosítanak valamelyik szereplőnek, amelyet szükség esetén a világ stabilizálására lehet felhasználni, ha a koordináció egzisztenciális kudarcának jelentős kockázata merül fel.[27]

[24] Armstrong et al. 2016.
[25] Bostrom 2004; Alexander 2014.
[26] Például a tudósoknak és a filantrópoknak többet kellene befektetniük a globális kormányzás és a

világkormányzás lehetőségeinek megértésébe; a politikai döntéshozóknak többet kellene befektetniük a meglévő globális koordinációs problémák megoldásába, hogy gyakorlatot és intézményi tapasztalatot szerezzenek a nagyobb kihívásokhoz; a globális kormányzással foglalkozó fórumoknak pedig többet kellene befektetniük a hipotetikus koordinációs kihívások mérlegelésébe.

[27] A stabilizálás magában foglalhatja a veszélyes technológia ellenőrzésének központosítását, vagy egy olyan megfigyelési rendszer bevezetését, amely lehetővé teszi a technológia pusztító célú alkalmazására irányuló bármely lépés időben történő felderítését és elfogását; vö. Bostrom 2018.

A turbulencia csökkentése

A gépi intelligencia forradalmában bekövetkező változások sebessége és nagyságrendje kihívások elé állítaná a meglévő intézményeket. A rendkívül turbulens körülmények között a már meglévő megállapodások megroppanhatnak, és a hosszú távú tervezés megnehezülhet. Ez megnehezítheti az egyébként lehetséges koordinációból származó előnyök megvalósítását - mind nemzetközi szinten, mind a nemzeteken belül. Hazai szinten veszteséget okozhat a rosszul átgondolt, elsietett szabályozás, vagy a jól átgondolt szabályozás, amely nem tud lépést tartani a gyorsan változó technológiai és társadalmi körülményekkel. Nemzetközi szinten az alkalmazkodási hibák kockázata valószínűleg még nagyobb, mivel ott gyengébbek a kormányzati intézmények és kisebb a kulturális kohézió, és általában évekig vagy évtizedekig tart a jól átgondolt normák, politikák és intézmények kidolgozása és végrehajtása. Az ebből eredő hatékonyságveszteségek a jólét átmeneti csökkenése vagy a hosszú távon rosszabb eredmények megnövekedett kockázata formájában jelentkezhetnek. Ezért kívánatos, hogy az ilyen turbulenciákat minimalizálják vagy jól kezeljék.

A hatékonysággal kapcsolatos kívánalmak

Az előző megfigyelésekből a következő kívánalmakat vonjuk le:

- *Gyors előrehaladás.* Ez két összetevőre oszlik: (a) olyan politikák, amelyek nagy valószínűséggel vezetnek a biztonságos szuperintelligencia végleges kifejlesztéséhez és annak alkalmazásához a jólét új forrásainak megcsapolásához; és (b) a gyors mesterséges intelligencia fejlődés, hogy a társadalmilag hasznos termékek és alkalmazások időben széles körben elérhetővé váljanak.
- *Mesterséges intelligencia biztonsága.* Olyan technikákat dolgoznak ki, amelyek lehetővé teszik (túlzott költségek, késedelem nélkül,
  vagy teljesítménybüntetés) annak biztosítására, hogy a szuperintelligens mesterséges intelligencia a kívánt módon viselkedjen.[28] Szintén,
  a szuperintelligencia kialakulása és korai alkalmazása során a körülmények olyanok, amelyek a rendelkezésre álló legjobb biztonsági technikák és egy általánosan óvatos megközelítés alkalmazására ösztönöznek.
- *Feltételes stabilizáció.* A fejlődési pálya és a tágabb politikai kontextus olyan, hogy *ha* drasztikus stabilizációs intézkedések hiányában katasztrofális globális koordinációs kudarc következne be, *akkor* a szükséges stabilizációt időben végrehajtják a katasztrófa elkerülése érdekében. Ez azt jelentheti, hogy (valamelyik szereplő vagy szereplők számára) megvalósítható lehetőségnek kell lennie egy egyeduralom létrehozására, vagy az intenzív globális felügyelet rendszerének bevezetésére, vagy a veszélyes technológia vagy tudományos ismeretek terjesztésének szigorú visszaszorítására.[29]
- *Nem turbulencia.* Az útvonal elkerüli a káoszból és a konfliktusokból eredő túlzott hatékonyságveszteségeket. A politikai rendszerek fenntartják a stabilitást és a rendet, sikeresen alkalmazkodnak a változásokhoz, és mérséklik a társadalmilag zavaró hatásokat.

[28] Az ideális összehangolási megoldás lehetővé tenné mind a külső, mind a belső viselkedés ellenőrzését (így

lehetővé téve a számítások nem kívánatos típusainak elkerülését anélkül, hogy a teljesítmény szempontjából nagy áldozatot kellene hozni; vö. az alább tárgyalt "elmebűnözés").

[29] A szingleton olyan világrend, amelynek legmagasabb szintjén egyetlen döntéshozó szerve van, amely képes "megakadályozni a saját létét és felsőbbrendűségét fenyegető (belső vagy külső) veszélyeket", és "hatékony ellenőrzést gyakorolhat saját területének főbb jellemzői felett (beleértve az adózást és a területi elosztást)" (Bostrom 2006).

Kiosztás

A vagyon, a státusz és a hatalom elosztása örökös politikai harc és vita tárgya. Talán nem sok remény van arra, hogy egy rövid tanulmányrészlet sok újszerű felismerést adjon ezekhez az évszázados vitákhoz. Vektorterületi megközelítésünk azonban lehetővé teszi számunkra, hogy megpróbáljunk némileg hozzájárulni ehhez a témához anélkül, hogy lényegesen foglalkoznunk kellene a fő vitás kérdésekkel. Így itt most néhány olyan különleges körülmény azonosítására összpontosítunk, amelyek a szuperintelligens mesterséges intelligencia fejlődését övezték, nevezetesen a kockázati externáliák, az átrendeződés, a tudatlanság fátyla és a sarjadzás. Ezeknek a körülményeknek (érvelésünk szerint) meg kellene változtatniuk az elosztással kapcsolatos bizonyos politikai megfontolások, normák és értékek relatív súlyát.[30]

Kockázati externáliák

Amint azt korábban említettük, azt állították, hogy a gépi intelligencia korszakába való átmenet bizonyos fokú egzisztenciális kockázattal jár majd. Ennek a kockázatnak minden ember ki lesz téve, függetlenül attól, hogy részt vesz-e a projektben, vagy beleegyezik-e vagy sem. Egy kislány egy azerbajdzsáni faluban, aki még soha nem hallott a mesterséges intelligenciáról, részesülne a gépi szuperintelligencia létrehozásának kockázatából. A méltányossági normák ezért megkövetelik, hogy ő is részesüljön a hasznokból, ha a dolgok jól alakulnak. Következésképpen, amennyiben a méltányossági normák részét képezik az egyes szereplők által alkalmazott értékelési standardoknak, az adott szereplőnek desiderátumként kell elismernie, hogy a mesterséges intelligencia fejlesztési útja ésszerű mértékű kompenzációt vagy haszonmegosztást biztosít mindazok számára, akiket kockázatnak tesz ki (ez a csoport legalábbis magában foglalja az összes olyan embert, aki életben van a veszélyes átmenet bekövetkezésekor).

A kockázati externáliákat a jelenlegi (fejlett mesterséges intelligencia) kontextuson kívül is gyakran figyelmen kívül hagyják, így ez a kívánalom általánosítható a *kockázatkompenzáció elvévé*, amely a közjót célzó politikai döntéshozatalt arra ösztönözné, hogy mérlegelje a más tevékenységéből eredő kockázatnak kitett személyek számára a valószínűsíthető károk kompenzálását, különösen olyan esetekben, amikor a tényleges kár bekövetkezése esetén a teljes kártérítés vagy lehetetlen (pl. mert az áldozat meghalt, vagy az elkövetőnek nincs elegendő pénze vagy biztosítási fedezete), vagy más okokból nem lenne elérhető.[31]

Átalakítás

Korábban már leírtuk a turbulencia korlátozását, mint *hatékonysággal kapcsolatos* kívánalom. A túlzott turbulencia gazdasági és társadalmi költségeket okozhat, és általánosabb értelemben csökkentheti a

[30] Hogy egyértelmű legyen, nem állítjuk, hogy az általunk meghatározott kívánatos szempontok az *egyetlen* olyan elosztási szempontok, amelyeket figyelembe kell venni. Lehetnek olyan kívánalmak is, amelyek más forrásból nyerik igazolásukat, mint a szuperintelligens mesterséges intelligencia forgatókönyvünkben fennálló különleges körülmények. (Lehetnek olyan további, elosztással kapcsolatos kívánalmak is, amelyeket ezekből a különleges körülményekből *lehetett* volna levezetni, de amelyeket nem vettünk figyelembe ebben a tanulmányban. Nem állítjuk a teljesség igényét.)

[31] Megjegyzendő, hogy az elv követése során ügyelni kell arra, hogy az elv végrehajtása ne akadályozza meg

indokolatlanul a társadalmilag kívánatos kockázatvállalást, például a kísérletezés és az innováció számos formáját.
Az ilyen tevékenységek negatív externáliáinak internalizálása a pozitív externáliák internalizálása nélkül rosszabb ösztönzőket eredményezhet, mintha egyik externália sem lenne internalizálva.

emberi értékek a jövőre nézve. A gépi intelligencia forradalmával járó turbulenciáknak azonban *allokációs* következményei is lehetnek, és ezek közül néhány további kívánalmakra utal.

Tekintsünk két lehetséges allokációs hatást: *koncentráció* és *permutáció*. A "koncentráció" alatt a jövedelem vagy a befolyás egyenlőtlenebbé válását értjük. A legvégső esetben egy nemzet, egy szervezet vagy egy egyén birtokolna és irányítana mindent. A "permutáció" alatt azt értjük, hogy a jövőbeli vagyon és befolyás kevésbé korrelál a jelenlegi vagyonnal és befolyással. A határesetben egy szereplő jelenlegi rangja (pl. jövedelem, vagyon, hatalom vagy társadalmi státusz tekintetében) és az adott szereplő jövőbeli rangja között nulla vagy akár antikorreláció is fennállna.

Nem állítjuk, hogy a koncentráció vagy a permutáció bekövetkezik, vagy hogy valószínű, hogy bekövetkezik. Csak azt állítjuk, hogy ezek kiemelkedő lehetőségek, és hogy a gépi intelligencia forradalma során kialakuló különleges körülmények között *nagyobb* valószínűséggel fordulnak elő szélsőséges mértékben, mint (hasonlóan szélsőséges mértékben) a fejlett mesterséges intelligencia kontextusán kívüli, ismertebb körülmények között. Bár ezt az állítást itt nem tudjuk teljes mértékben igazolni, illusztrációként megemlíthetünk néhány lehetséges dinamikát, amely ezt igazzá teheti. (1) A mai világban és a történelem során a bérjövedelmek egyenletesebben oszlanak el, mint a tőkejövedelmek.[32] A szuperintelligens mesterséges intelligencia az emberi munka erőteljes helyettesítésével nagymértékben megnövelheti a tényező
a tőke által kapott jövedelem részesedése.[33] Minden más esetben ez növelné a jövedelemegyenlőtlenséget, és így
növeli a koncentrációt.[34] (2) Bizonyos forgatókönyvekben olyan erős első lépő előnyök vannak a a szuperintelligencia létrehozása, hogy a kezdeti szuperintelligens mesterséges intelligencia vagy az azt irányító entitás döntő stratégiai előnyhöz jusson. Attól függően, hogy ez a mesterséges intelligencia vagy annak megbízója mit kezd ezzel az előnnyel, a jövőt végül teljes mértékben ez az első számú szereplő határozhatja meg, ami nagymértékben növelheti a koncentrációt. (3) Radikális és kiszámíthatatlan technológiai változások esetén nagyobb társadalmi-gazdasági hullámzás következhet be - egyes egyének vagy cégek jó helyzetben lévőnek bizonyulnak ahhoz, hogy az új körülmények között boldoguljanak, vagy szerencsésen fogadnak, és nagy hasznot húznak belőlük; mások humán tőkéje, befektetései és üzleti modelljei gyorsan erodálódnak. A gépi intelligencia forradalma felerősítheti ezt a változást, és ezáltal jelentős mértékű permutációt eredményezhet.[35]

---

[32] Piketty ch2014,. 7.

[33] Brynjolfsson & McAfee 2014.

[34] A gépi intelligencia korszakába való átmenet után is csökkentheti a permutációt, ha könnyebb lesz a tőkét a gyermekeinkre hagyni (vagy saját magunknak megőrizni, amíg élünk, ami a hatékony élethosszabbító technológia megjelenésével nagyon hosszú ideig lehet), mint a tehetségeket és készségeket a történelmileg megszokottabb körülmények között hagyni vagy megőrizni.

[35] A szereplők pozíciójukat úgy igyekezhetnek megőrizni, hogy folyamatosan diverzifikálják állományukat. Ennek elérését azonban jelentős korlátok és súrlódások nehezíthetik, amelyek a következőkhöz kapcsolódnak: (1) a diverzifikáció korlátai vagy költségei, (2) a diverzifikáció időbeli késleltetése, (3) egyes szereplők hajlandósága a nagy kockázatvállalásra. (1a) Egyes eszközosztályok (pl. lopakodó startupok, magáncégek, egyes nemzetgazdaságokban lévő részesedések) nem válnak tulajdonossá, vagy költséges keresési és befektetési folyamattal járnak. (1b) Számos szereplőnek jelentős diverzifikációs korlátokkal kell szembenéznie. Előfordulhat, hogy egy vállalat vagy egy ország erősen elkötelezett egy iparágban, és nem képes hatékonyan fedezni kitettségét, vagy gyorsan átirányítani tevékenységét, hogy alkalmazkodjon a gyorsan változó versenyhelyzethez. (2) A technológiai változás olyan gyorsan haladhat, hogy a befektetőknek nincs lehetőségük portfóliójuk "időben" történő újbóli kiegyensúlyozására. Mire egy

felértékelődő új eszközosztály felbukkan, lehet, hogy az ember már elvesztette növekedési értékének nagy részét. (3) Egyes szereplők úgy döntenek majd, hogy nagy tételeket tesznek a kockázatos eszközökre/technológiákra, amelyek ha nyernek, átrendezik a vagyoni rangsort; még a tökéletesen diverzifikált, felső kategóriás szereplőket is letaszíthatja a csúcspozíciójukból egy feltörekvő, aki a nagy átrendeződésben főnyereményt ér el.

(4) Az automatizált biztonsági és felügyeleti rendszerek megkönnyíthetik egy rezsim számára, hogy a szélesebb elit vagy a nyilvánosság támogatása nélkül fenntartsa magát. Ez lehetővé tenné a rezsim tagjai számára, hogy a nemzeti termelés nagyobb részét sajátítsák ki, és finomabb ellenőrzést gyakoroljanak a polgárok viselkedése felett, ami nagymértékben növelheti a vagyon és a hatalom koncentrációját.[36]

Amennyiben valaki (várakozásai szerint) elutasítja a vagyon és a hatalom elosztásának koncentrálódását vagy változásait - talán azért, mert súlyt helyez valamilyen társadalmi szerződéselméletre vagy más erkölcsi keretre, amely azt sugallja, hogy az ilyen változások rosszak, vagy egyszerűen azért, mert arra számít, hogy a vesztesek közé kerül -, akkor a folyamatosságot kívánatosnak kell tekintenie.[37]

A tudatlanság fátyla

A történelem jelenlegi szakaszában a jövő fontos aspektusai legalábbis részben rejtve maradnak. a tudatlanság fátyla mögött.[38] Senki sem tudja biztosan, hogy mikor, hol és hogyan jön létre a fejlett mesterséges intelligencia.
kit (bár bevallottan egyes helyszínek kevésbé tűnnek valószínűnek, mint mások). Mivel a legtöbb szereplőnek meglehetősen gyorsan csökkenő határhaszna van a vagyonban, és így a vagyonban is kockázatkerülő, ezért általában előnyös lenne, ha egy biztosítás-szerű rendszert fogadnának el, amely újraosztaná a gépi szuperintelligenciából származó nyereség egy részét.

Az is hihető, hogy a tipikus egyéneknek viszonylag gyorsan csökkenő határhaszna van a hatalomban. Például a legtöbb ember sokkal inkább szeretne biztos lenni abban, hogy hatalma van egy élet (a sajátja) felett, mint hogy 10% esélye legyen arra, hogy hatalma van tíz ember élete felett, és 90% esélye arra, hogy nincs hatalma. Emiatt az is kívánatos lenne, ha egy rendszerben a hatalom meglehetősen széles körű eloszlása megmaradna, legalábbis olyan mértékben, hogy az egyéneknek megmaradjon a saját életük és közvetlen körülményeik feletti megfelelő mértékű kontroll (pl. azáltal, hogy bizonyos mennyiségű

A katonai hatalom elosztása elvileg ki van téve a felgyorsult technológiai változás okozta átrendeződésnek is, többek között olyan módon, amely ellen nehéz védekezni katonai diverzifikációval vagy a meglévő erő felhasználásával egy olyan stabil megállapodásra való alkudozással, amely rögzíti a meglévő hatalmi hierarchiákat.

[36] Bueno de Mesquita & Smith 2011; Horowitz 2016.

[37] Kétféle permutációt különböztethetünk meg. (1) Olyan permutációk, ahol az egyén *várható* ex post vagyona (vagy hatalma, státusza stb.) megegyezik az ex ante vagyonával (hatalmával, státusával stb.). Az ilyen permutáció olyan, mint egy hagyományos lottó, ahol minél több szelvényünk van, annál több nyereményre számíthatunk. A kockázatkerülő egyének megpróbálhatnak védekezni az ilyen permutációk ellen a birtoklásuk diverzifikálásával; de ahogy az előző lábjegyzetben említettük, a kellően drasztikus átrendeződések ellen nehéz lehet védekezni, különösen olyan forgatókönyvek esetén, amelyekben a szerződések és a tulajdonjogok nagymértékű megsértése történik. (2) Olyan permutációk, ahol az egyén várható ex post vagyona nem függ az ex ante vagyonától. Gondoljunk erre úgy, mint a véletlenszerű szerepcserére: mindenki nevét egy nagy urnába helyezik, és minden egyén kihúz egy szelvényt - lemond arról, amije korábban volt, és helyette megkapja a másik személy vagyonát. Ha eltekintünk a társadalmi bomlás következményeitől, az ilyen típusú permutáció várhatóan nyereséget eredményezne azok számára, akik kezdetben rosszul jártak, a hivatalban lévő elit rovására. Azok azonban, akik nem önös érdekből támogatják a szegények javára történő újraelosztást, ezt jellemzően a gazdasági egyenlőtlenségek

csökkentésével akarják elérni, és nem azzal, hogy a szegények közül néhányan helyet cserélnek a gazdagokkal.

[38] Ez a John Rawls által javasolt "a tudatlanság fátyla" gondolatkísérlet kiterjesztése: "[A] felek... nem tudják, hogy a különböző alternatívák hogyan hatnak a saját konkrét esetükre, és kénytelenek az elveket kizárólag általános megfontolások alapján értékelni..... Először is, senki sem ismeri a társadalomban elfoglalt helyét, osztályhelyzetét vagy társadalmi státuszát; és nem ismeri a természeti javak és képességek elosztásában való sorsát sem....". (Rawls 137. o1971,.).

garantált hatalom vagy elidegeníthetetlen jogok). Nemzetközi megállapodás is van arról, hogy az egyéneknek jelentős jogokkal és hatalommal kell rendelkezniük.[39]

Cornucopia

A gépi szuperintelligenciára való áttérés hatalmas bónuszokat hozhat magával. Hanson becslése szerint például az olcsó, emberi szintű gépi intelligencia elegendő lenne ahhoz, hogy A gépi szuperintelligencia révén megvalósítható gazdasági potenciál [40]végső soron A csillagászati méretű lehet.[41]

Az ilyen növekedés lehetővé tenné, hogy a GDP kis hányadát felhasználva, számos olyan értéket, amelynek csökkenő hozama van az erőforrásokban (ésszerű kiadási sávokban), majdnem maximálisan ki lehessen használni.

---

[39] Ezt a megállapodást jól megalapozza többek között az Egyesült Nemzetek Alapokmánya (az Egyesült Nemzetek Alapokmánya, 1945) és az "Emberi Jogok Nemzetközi Törvénykönyve", amely az Emberi Jogok Egyetemes Nyilatkozatából (az Emberi Jogok Egyetemes Nyilatkozata, 1948), a Polgári és Politikai Jogok Nemzetközi Egyezségokmányából (a Polgári és Politikai Jogok Nemzetközi Egyezségokmánya, 1966) és a Gazdasági, Szociális és Kulturális Jogok Nemzetközi Egyezségokmányából (a Gazdasági, Szociális és Kulturális Jogok Nemzetközi Egyezségokmánya, 1966) áll, amelyeket szinte mindenki ratifikált (bár Kína és az USA nem ratifikálta). További támpontot nyújt a *jus cogens* ("kötelező jog") nemzetközi jogi elve, amely kötelező erejű nemzetközi jogi normákat alkot, amelyektől nem lehet eltérni. Bár a jus cogens pontos hatályát vitatják, általános konszenzus van abban, hogy többek között a rabszolgaság, a kínzás és a népirtás tilalmát is magában foglalja (Lagerwall 2015). A nemzetközi emberi jogi jog és a mesterséges intelligencia fejlesztése közötti potenciális kapcsolatról az egzisztenciális kockázatokkal kapcsolatban lásd Vöneky 2016.

[40] Hanson 189-194. o2016,.

[41] Az egyik technológiai terület, amelyről azt várhatjuk, hogy néhány éven belül az erősen szuperintelligens mesterséges intelligencia kifejlesztése után érettségre jut, az űrgyarmatosítás fejlett képességei, beleértve a von Neumann-szondák kibocsátásának képességét, amelyek képesek a fénysebesség bizonyos értelmes töredékével utazni intergalaktikus távolságokon, és egy olyan technológiai bázis létrehozását egy távoli erőforráson, amely képes további szondák előállítására és indítására (Freitas 1980; Tipler 1980). Feltételezve, hogy képesek ilyen von Neumann-szondák létrehozására, és hogy a megfigyelhető univerzumban nincsenek más intelligens civilizációk (Sandberg et al. 2018), akkor úgy tűnik, hogy az emberiség kozmikus adottságai 10-10 [1820]elérhető csillagot tartalmaznak (Armstrong & Sandberg 2013). Azzal a fajta asztrofizikai mérnöki technológiával, amelyről szintén azt várnánk, hogy a vonatkozó időskálán rendelkezésre áll (Sandberg hamarosan megjelenik), ez az erőforrásbázis elegendő lehet ahhoz, hogy (az univerzum fennmaradó élettartama alatt) néhány, valamilyen biológiai10[35] emberi élethez hasonló élőhelyet hozzon létre, vagy alternatívaként sokkal nagyobb számú (10[58] vagy annál is több) digitálisan megvalósított emberi elme számára (Bostrom 2014). Természetesen e lehetőségek nagy része csak nagyon hosszú időskálákon valósulhatna meg; de a türelmes szereplők számára a késedelmek talán nem sokat számítanak.

Megjegyzendő, hogy a szereplők nagyobb hányada lehet "beteg" a megfelelő értelemben, miután kifejlesztik az extrém élet meghosszabbítására vagy a felfüggesztett életműködésre alkalmas technológiai eszközöket (pl. az emberi elmék digitális tárolása által megkönnyítve). Azok a szereplők, akik arra számítanak, hogy az ilyen képességeket röviddel a szuperintelligens mesterséges intelligencia megjelenése után fogják kifejleszteni, türelmesek lehetnek - abban az értelemben, hogy nem számolnak le komolyan az időben rendkívül távoli gazdasági előnyökkel -, mivel nem triviális valószínűséget tulajdoníthatnak annak, hogy a hosszú késedelem után ők maguk is képesek lesznek fogyasztani e gazdasági előnyök egy részét. Egy másik fontos tényező, amely a szereplők szélesebb köre számára döntési szempontból relevánssá teheti a rendkívül távoli jövőbeli eredményeket, az, hogy megvalósíthatóvá válhat egy stabilabb társadalmi rend

vagy más megbízható kötelezettségvállalási technikák, ami növeli annak esélyét, hogy a közeli döntéseknek kiszámítható hatása legyen arra, ami nagyon hosszú távon történik.

Tegyük fel például, hogy a gazdaság olyan szintre bővülne, ahol a GDP 5%-ának elköltése elegendő lenne ahhoz, hogy az egész emberiség számára garantált éves alapjövedelmet biztosítson, amelynek összege
40 000 dollár plusz hozzáférés a futurisztikus minőségű egészségügyi ellátáshoz, szórakozáshoz és más csodálatos javakhoz.
és szolgáltatások.[42] Az ilyen politika elfogadása mellett szóló érvek így erősebbnek tűnnek, mint az alábbiak mellett szóló érvek.
a garantált alapjövedelem bevezetése ma, egy olyan időszakban, amikor a megfelelő politika sokkal kevésbé nagyvonalú juttatásokat eredményezne, a GDP nagyobb százalékának újraelosztását igényelné, és a munkaerő-kínálat drámai csökkenésével fenyegetne.

Hasonlóképpen, ha egy állam olyan gazdaggá válik, hogy GDP-jének mindössze 0,1%-át külföldi segélyekre költve a világon mindenkinek kiváló életminőséget tudna biztosítani (ahol egyébként széleskörű szegénység lenne), akkor különösen kívánatos lenne, hogy a gazdag állam legalább ilyen nagylelkű legyen. Míg egy szegény állam számára nem sokat számít, hogy a GDP 0,1%-át adja-e, vagy semmit - egyik esetben sem elég az összeg ahhoz, hogy nagy különbséget tegyen -, egy *rendkívül* gazdag állam számára döntő fontosságú lehet, hogy inkább 0,1%-ot adjon, mint 0%-ot. Egy igazán szélsőséges esetben talán nem is számít annyira, hogy egy szupergazdag állam 0,1%-ot, 1%-ot vagy 10%-ot ad: a lényeg az, hogy ne 0%-ot adjon.

Vagy vegyük például azt az esetet, amikor egy társadalmi tervezőnek kompromisszumot kell kötnie az állatjólét értéke és sok emberi fogyasztó azon vágya között, hogy a hús szerepeljen az étrendjében. Tegyük fel, hogy a tervezőt leginkább az emberi fogyasztói preferenciák érdeklik, de egy kicsit az állatok jóléte is. Alacsony GDP-szint mellett a tervező úgy dönthet, hogy engedélyezi a gyári állattartást, mert az csökkenti a hús költségét. A GDP növekedésével azonban eljön az a pont, amikor a tervező olyan jogszabályokat vezet be, amelyek elrettentik a gyári állattartástól. Ha a tervezőt *egyáltalán* nem érdekelné az állatjólét, ez a pont soha nem jönne el. A GDP szerény szintjén egy olyan tervező, aki sokat törődik az állatjóléttel, esetleg törvényt vezet be, míg egy olyan tervező, aki csak keveset törődik az állatjóléttel, talán engedélyezi a gyári állattartást. De ha a GDP kellően extravagáns szintre emelkedik, akkor lehet, hogy nem számít, hogy a tervező mennyire törődik az állatjóléttel, amíg *legalább egy icipicit* törődik vele.[43]

---

[42] A becsült világ 2017GDP-je névlegesen trillió81 USD volt (vagy vásárlóerő-paritáson számolva trillió128 USD dollár, Világbank, International Comparison Program adatbázis). Ez 11 000 USD (nominális) vagy 17 000 USD (vásárlóerő-paritáson) egy főre jutó GDP-nek felel meg. Ahhoz, hogy a 40 000 dolláros garantált éves alapjövedelem a világ GDP-jének 5%-ával elérhető legyen (7,6 milliárdos népesség2018 mellett), a világ GDP-jének 75,négymilliárd6 (10^15) 50dollárra kellene növekednie. Bár az 5% magas emberbaráti aránynak tűnhet, valójában ez a tíz leggazdagabb amerikai jelenlegi átlagának a fele. Bár a gazdasági termelékenység szükséges növekedése nagynak tűnhet, mindössze a világgazdaság hatszorosára van szükség. Az elmúlt évszázadban a világ egy főre jutó GDP-jének megduplázódása nagyjából 35 évente fordult elő. A fejlett gépi intelligencia valószínűleg az egy (emberi) személyre jutó vagyon növekedési ütemének jelentős növekedéséhez vezetne. Robin Hanson közgazdász szerint az emberi szintű gépi intelligencia megjelenése után, az emberi agy utánzása formájában, várhatóan évente vagy akár havonta is megduplázódhat a gazdasági teljesítmény (Hanson 189-191. o2016,.).

Vegyük észre azt is, hogy itt és másutt is azt feltételezzük, talán irreálisan, hogy vagy nem élünk számítógépes szimulációban, vagy igen, de az még jó ideig fog működni a gépi szuperintelligencia kialakulása után is (Bostrom 2003a). Ha olyan szimulációban élünk, amely röviddel a szuperintelligens mesterséges intelligencia létrejötte után véget ér, akkor a látszólagos kozmikus adottság illuzórikus lehet; és más megfontolások lépnek a képbe, amelyek meghaladják e tanulmány kereteit.

[43] Megfelelően fejlett technológiával a biotechnológiával előállított húspótlóknak teljesen fel kellene

számolniuk a húsevő fogyasztói preferenciák és az állatjólét közötti összeegyeztethetetlenséget. Még fejlettebb technológiával pedig a fogyasztók úgy alakíthatják át ízlelőbimbóikat, hogy az etikus, egészséges, fenntartható növényi ételeket részesítsék előnyben, vagy (feltöltés vagy más digitális elmék esetében) elektromos áramot és virtuális steakeket fogyasszanak.

Így úgy tűnik, hogy míg ma talán fontosabb lenne az altruizmus magasabb, mint alacsonyabb szintjének ösztönzése, egy cornucopia forgatókönyvben nem az altruizmus várható mennyiségének maximalizálása lenne a legfontosabb, hanem annak a valószínűségnek a minimalizálása, hogy az altruizmus szintje végül nulla legyen. A cornucopiás forgatókönyvekben, mondhatnánk, különösen kívánatos, hogy az epsilon-magnanimitás érvényesüljön. Ennél több is szép lenne, és ezt támasztja alá néhány más, ebben a tanulmányban említett kívánalom is; de különösen kívánatos, hogy a garantált alsó határ jelentősen a nulla szint felett legyen.

Általánosabban, úgy tűnik, hogy ha vannak olyan erőforrás-szegényebb értékek, amelyeknek kevés a támogatottsága (és nincs közvetlen ellenállásuk), és amelyek csak erősebben támogatott értékekkel versenyeznek az erőforrás-korlátozásokon keresztül, akkor kívánatos lenne, hogy ezek az erőforrás-szegényebb, gyengébb értékek legalább a rendelkezésre álló erőforrások egy kis hányadát megkapják egy cornucopiás forgatókönyvben, hogy valóban kielégíthetők legyenek.[44]

Egy olyan jövő, amelyben az epsilon-magnóniát biztosítják, intuitíve előnyösebbnek tűnik. Ezt az intuíciót többféleképpen is megalapozhatjuk. (1) Számos jelenlegi érdekelt fél preferenciasorrendjében magasabbra kerülne, különösen azoknál az érdekelteknél, akiknek erőforrás-szükségletű érdekei jelenleg az erőforrás-korlátok miatt dominálnak. (2) A normatív bizonytalanságot tekintve bölcs rendezés lenne: ha a domináns szereplők valamilyen pozitív valószínűséget tulajdonítanak különböző
erőforrás-szükségletű értékek vagy erkölcsi állítások igazak, és triviális lenne, hogy ezeket az értékeket egy kornukopikus forgatókönyvben is megillesse, akkor egy "erkölcsi parlament" [45]vagy más, a normatív bizonytalanság kezelésére szolgáló keretrendszer olyan politikákat támogathat, amelyek biztosítják az epsilon-magnóniás jövőt. (3) Azoknak a szereplőknek, akiknek vágyuk vagy erkölcsi kötelességük, hogy jótékonyak vagy nagylelkűek legyenek (vagy gyengébb esetben, hogy ne legyenek teljesen bunkók), lehet okuk arra, hogy különleges erőfeszítéseket tegyenek annak érdekében, hogy a jövő epsilon-magnanimitású legyen.[46]

Az elosztással kapcsolatos kívánalmak

Ezek a megfigyelések azt sugallják, hogy a mesterséges intelligenciával kapcsolatos hosszú távú eredmények allokációs tulajdonságai tekintetében az értékelési kritériumok a következők:

- **Egyetemes juttatás.** Mindenki, aki az átmenet idején életben van (vagy akit negatívan érinthet), részesül a juttatásból, kompenzációként a kockázat externáliáiért, amelyeknek ki volt téve.
- **Epsilon-magnanimitás.** Az erőforrásokkal jóllakható értékek széles skálája (olyanok, amelyek ellen a költségalapú megfontolásokon kívül nem sok kifogás van), akkor és úgy valósul meg, ha és amennyiben ez az összes erőforrás parányi töredékének felhasználásával lehetséges. Ez magában foglalhatja az alapvető jóléti ellátásokat és a jövedelemgaranciákat minden emberi egyén számára. Ez lehet

[44] Itt az epsilon-magnanizmus a gyakorlati értékpluralizmus egy gyenge formájának tekinthető.
[45] Bostrom 2009.
[46] Az epsilon-magánjövő úgy érhető el, hogy a jövőt sok különböző értéket képviselő szereplő alakítja, akik

mindegyike képes valamilyen nem elhanyagolható mértékű befolyást gyakorolni; vagy pedig úgy, hogy legalább egy rendkívül nagy felhatalmazással rendelkező szereplő egyénileg epsilon-magnanimous.

számos közösségi javakat, etikai eszményeket, esztétikai vagy szentimentális projekteket, valamint a nagylelkűség, kedvesség és együttérzés különböző természetes megnyilvánulásait foglalja magában.[47]

- **Folytonosság.** Az út ésszerű mértékű folytonosságot biztosít, hogy i. fenntartsa a rendet és biztosítsa a szükséges intézményi stabilitást ahhoz, hogy a szereplők a jelenlegi tudatlanság fátyla mögött kihasználhassák a kereskedelem lehetőségeit, beleértve a szociális biztonsági hálókat is; és ii. megakadályozza, hogy a koncentráció és a permutáció szükségtelenül nagy legyen.

Népesség

E címszó alatt gyűjtjük össze az új lények, különösen a digitális elmék létrehozásával kapcsolatos megfontolásokat, amelyek erkölcsi státusszal rendelkeznek, vagy amelyek egyébként nem instrumentális okokból fontosak a politikai döntéshozók számára.

A digitális elmék alapvető módon különbözhetnek a megszokott biológiai elméktől. A digitális elmék megkülönböztető tulajdonságai közé tartozhatnak: könnyen és gyorsan másolhatóak, különböző sebességgel futhatnak, látható fizikai alak nélkül létezhetnek, egzotikus kognitív architektúrával rendelkeznek, nem-animalista motivációs rendszerrel vagy esetleg pontosan módosítható céltartalommal rendelkeznek, pontosan megismételhetők, ha determinisztikus virtuális környezetben futnak, és potenciálisan meghatározhatatlan élettartammal rendelkeznek.

Az ilyen és más újszerű tulajdonságokkal rendelkező lények létrehozása összetett és széleskörű következményekkel járna a gyakorlati etikára és a közpolitikára nézve. Míg a legtöbb ilyen
következményeit félre kell tenni a jövőbeli vizsgálatokra, két nagy területet azonosíthatunk: a digitális elmék érdekeit és a populáció dinamikáját.[48]

A digitális elmék érdekei

A gépi intelligencia fejlődése lehetőséget teremthet a jogsértések és az elnyomás új kategóriáinak kialakítására. Az "elmebűnözés" kifejezést olyan számításokra használták, amelyek belső tulajdonságaik miatt erkölcsileg problematikusak, függetlenül a külvilágra gyakorolt hatásuktól: például azért, mert érző elméket instanciálnak, amelyekkel rosszul bánnak (Bostrom 2014). Az elmebűnözés kérdése már jóval az emberi szintű vagy szuperintelligens mesterséges intelligencia elérése előtt felmerülhet. Egyes nem emberi állatokról széles körben feltételezik, hogy érző és bizonyos fokú erkölcsi státusszal rendelkeznek. A jövőbeni mesterséges intelligenciák, amelyek hasonló képességekkel vagy kognitív architektúrával rendelkeznek, valószínűsíthetően hasonló mértékű erkölcsi státusszal rendelkezhetnek. Néhány olyan mesterséges intelligencia, amely funkcionálisan nagyon különbözik bármely állattól, szintén rendelkezhet erkölcsi státusszal.

A mentális élettel rendelkező digitális lények létrejöhetnek szándékosan, de véletlenül is létrejöhetnek. A gépi tanulásban például gyakran nagyszámú ágenst generálnak a képzési eljárások során - egy megerősítéses tanuló sok félig funkcionális változatát hozzák létre és vetik össze egymással az önjátékban, sok teljesen funkcionális ágenspéldányt hoznak létre...

[47] Ilyen körülmények között például megvalósíthatónak és kívánatosnak tűnik, hogy a segítséget

kiterjesszük a nem emberi állatokra, beleértve a vadon élő állatokat is, hogy enyhítsük nehézségeiket, csökkentsük szenvedésüket, és nagyobb örömet szerezzünk minden elérhető érző lénynek (Pearce 1995). [48] Elvileg ezek a megfigyelések a biológiai elmékre is vonatkoznak, amennyiben azok osztoznak a vonatkozó tulajdonságokban. Elképzelhető, hogy a rendkívül fejlett biotechnológia lehetővé teszi, hogy a biológiai struktúrák megközelítsenek néhány olyan tulajdonságot, amelyek a digitális megvalósítások számára könnyen elérhetőek lennének.

a hiperparaméterek söpörése során, és így tovább. Nem világos, hogy a mesterséges ágensek mennyire kifinomultak lehetnek, mielőtt elérnék az erkölcsileg releváns érzékenység bizonyos fokát - vagy mielőtt már nem lehetünk biztosak abban, hogy nem rendelkeznek ilyen érzékenységgel.

Több tényező együttesen jelzi az elmebűnözés lehetőségét, mint a mesterséges intelligencia fejlett fejlődésének kiemelkedő különleges körülményét. Az egyik az érző digitális entitások mint erkölcsi betegek újdonsága.
A politikai döntéshozók nem szokták figyelembe venni a digitális lények jólétét. Az a felvetés, hogy erre erkölcsi kötelezettséget kaphatnának, egyes kortársak számára ostobaságnak tűnhet, ahogyan a szabadidős állatkínzás kegyetlen formáit tiltó törvények is egykoron sokak számára butaságnak tűnt.[49] Az újszerűség kérdéséhez kapcsolódik az a tény, hogy a digitális elmék láthatatlanok lehetnek, mélyen egy mikroprocesszor belsejében futnak, és hogy talán nem képesek hangok, arckifejezések vagy más, emberi empátiát kiváltó viselkedésformák segítségével közölni a szorongást. E két tényező, az érző digitális lények újdonsága és potenciális láthatatlansága együttesen azt a veszélyt hordozza magában, hogy olyan eredményekbe is belenyugszunk, amelyeket saját, gondosabban megfogalmazott és alkalmazott erkölcsi normáink lelkiismeretlenségnek ítélnének.

Egy másik tényező az, hogy nem mindig egyértelmű, mi minősül a digitális elmével való visszaélésnek. Egyes kezelések, amelyek az érző biológiai organizmusok esetében helytelenek lennének, bizonyos digitális elmék esetében, amelyek úgy vannak kialakítva, hogy másképp értelmezik az ingereket, nem kifogásolhatóak. Ezek a bonyodalmak fokozódnak, ha a fejlettebb digitális elméket (pl. az emberhez hasonló digitális elméket) vesszük figyelembe, amelyeknek a szenvedéstől való mentességen kívül erkölcsileg jelentős érdekeik is lehetnek, például a túlélés, a méltóság, a tudás, az autonómia, a kreativitás, önkifejezés, társadalmi hovatartozás és politikai részvétel.[50] A kombinatorikus tér a különböző típusú elmék, különböző típusú, erkölcsileg jelentős érdekekkel, nehezen feltérképezhetőek és nehezen navigálhatóak.

A negyedik tényező, amely felerősíti a másik hármat, az, hogy olcsóbbá válhat a digitális elmék nagy számban történő előállítása. Ezáltal több ügynöknek lesz lehetősége az elmebűncselekmények elkövetésére, méghozzá nagy mennyiségben. Nagy számítási sebességgel vagy párhuzamosítással nagy mennyiségű szenvedést lehetne generálni kis falióraidő alatt. Valószínű, hogy a valaha létező összes elme túlnyomó többsége digitális lesz. Ezért a digitális elmék jóléte lehet a fő kívánalom az AI-fejlesztési útvonal kiválasztásakor azon szereplők számára, akik vagy jelentős súlyt fektetnek az etikai megfontolásokra, vagy valamilyen más okból erősen szeretnék elkerülni a nagy mennyiségű szenvedés okozását.

Népességdinamika

Számos aggály merül fel a nagyszámú új lény bevezetésének lehetőségével kapcsolatban, különösen akkor, ha ezek az új lények a személyiséggel kapcsolatos tulajdonságokkal rendelkeznek. Ezen aggályok némelyike az elmebűnözés lehetőségével kapcsolatos, amelyet az előző alfejezetben tárgyaltunk,

[49] A legkorábbi állatkínzásról szóló törvényeket övező gúnyolódásra lásd: Fisher Az2009. állatokkal való bánásmóddal kapcsolatos változó normákról lásd: Pinker ch2011,. és 36.
[50] De nem minden kifinomult elmének kell, hogy ilyen érdekei legyenek. Feltételezhetjük, hogy helytelen az embereket vagy más, az emberhez nagyon hasonló lényeket rabszolgasorba taszítani vagy kizsákmányolni. De lehet, hogy lehetséges egy olyan mesterséges intelligenciát tervezni, amely emberi szintű intelligenciával rendelkezik (de más tekintetben különbözik az emberektől, például motivációs rendszerében), és amelynek nem áll érdekében, hogy ne "rabszolgasorba taszítsák" vagy "kizsákmányolják". Lásd még Bostrom és Yudkowsky. 2014.

de más aggályok akkor is felmerülnek, ha feltételezzük, hogy nem történik elmebűncselekmény. Az egyik különleges körülmény, amely itt releváns, az, hogy a digitális replikációs arányok mellett a népesség száma rendkívül gyorsan változhat. Aktív népesedési politika, megfelelő, előre hozott intézkedésekkel, lehet, hogy szükséges a malthusiánus kimenetel (amikor az átlagjövedelem a létminimum közelébe esik) és más rossz eredmények megelőzéséhez.

Gondoljunk csak a fejlett országokban elterjedt gyermektartási rendszerre. Az egyének szabadon vállalhatnak annyi gyermeket, amennyit képesek létrehozni; az állam pedig közbelép, hogy támogassa azokat a gyermekeket, akiknek a szülei nem képesek gondoskodni róluk. A digitális lények esetében ez a rendszer nyilvánvalóan fenntarthatatlan. Ha a szülők tetszőleges számú gyermeket hozhatnának létre, és az erre való hajlandóság tartósan eltérő lenne, ez a rendszer gyorsan összeomlana. Igaz, hogy hosszabb távon a malthusi aggályok a biológiailag szaporodó személyek esetében is felmerülnek, mivel az evolúció az emberi hajlamokra hatva olyan típusokat választ ki, amelyek kihasználják a korszerű
jólét, hogy nagyobb családokat hozzon A digitális elmék számára azonban a malthusiánus kezdetű létre. [51]állapot lehet hirtelen.[52]

A társadalmak így dilemmával szembesülnének: *vagy* elfogadják a népességszabályozást, amely megköveteli, hogy a leendő nemzőképesek teljesítsenek bizonyos feltételeket, mielőtt új lényeket hozhatnának létre; *vagy* elfogadják annak kockázatát, hogy az új lények nagy száma csak a munkájukhoz szükséges minimális mennyiségű erőforrást kapja meg, miközben a lehető legkeményebben dolgoztatják őket, és amint már nem költséghatékonyak, elpusztítják őket. E lehetőségek közül az előbbi tűnik előnyösebbnek, különösen akkor, ha kiderülne, hogy a jövő gazdaságában a maximálisan produktív dolgozók tipikus mentális állapota nem rendelkezik pozitív affektusokkal vagy más kívánatos tulajdonságokkal.[53]

A malthusi eredmények az egyik példa arra, hogy a népességváltozás hogyan teremthet problémás körülményeket a helyszínen. A másik a demokrácia aláásása, amely akkor következhet be, ha a különböző demográfiai csoportok méreteit manipulálni lehet. Tegyük fel, hogy a digitális lények bizonyos típusai szavazati jogot kapnak, egyszemélyes szavazati alapon. Egy ilyen választójog megadása történhet azért, mert az emberek erkölcsi okokból szavazati jogot adnak a digitális elmék bizonyos osztályának, vagy azért, mert a nagy teljesítményű digitális elmék nagy populációja hatékonyan képes politikai befolyást gyakorolni. A választók ezen új szegmense aztán gyorsan bővülhet a másolás révén, egészen addig a pontig, ahol
ahol az eredeti emberi blokk szavazati ereje döntően elenyészik.[54] Minden másolat egy az adott sablon ugyanazokat a szavazási preferenciákat osztja meg, mint az eredeti, ami arra ösztönzi a digitális lényeket, hogy számos másolatot készítsenek magukról - vagy erőforrás-hatékonyabb helyettesítő anyagokról.

---

[51] A modern társadalomban a nagyobb családmérettel összefüggő tulajdonságok örökölhetőségének bizonyítékairól lásd Milot et al. 2001; Kong et al. 2017. Beauchamp 2016 szerint: "Az alacsony halandósággal rendelkező modern populációkban a fitnesz ésszerűen megközelíthető [az egyén által valaha szült vagy nemzett gyermekek számával]". [52] Az egyszerű érvelés a gazdaságilag nem produktív lények, például a gyermekek lehetőségére összpontosít, ami elegendő a következtetés megalapozásához. De akkor is lehetséges, hogy malthusi problémákba ütközünk, ha a létrehozott elmék gazdaságilag produktívak; lásd Hanson részletes vizsgálatát2016 egy ilyen
forgatókönyv. A Hanson-modellben a malthusi kimenetel elkerülése érdekében globális koordinációra lenne szükség. [53] A reproduktív paradigma egyik példája az lenne, ha egy új elme létrehozása előtt a leendő utódtól megkövetelnék, hogy elegendő gazdasági adottságot tegyen félre ahhoz, hogy az új elme számára további

transzferek nélkül is megfelelő életminőséget biztosítson. Amíg a világgazdaság növekszik, addig az alkalmi "ingyenes" utódok is megengedhetők lennének, olyan ütemben, hogy a népesség növekedési üteme ne legyen magasabb, mint a gazdaság növekedési üteme.

[54] Hasonló folyamat játszódhat le a biológiai állampolgárokkal is, bár hosszabb idő alatt, ha valamelyik csoport megtalálja a módját annak, hogy stabilan tartsa értékeit, miközben fenntartja a termékenység magas szintjét.

amelyek célja, hogy a kezdeményező szavazási preferenciáit megosszák, és megfeleljenek a választási követelményeknek - politikai befolyásuk növelése érdekében. Ez a demokratikus társadalmakat trilemma elé állítaná. *Vagy* (i) megtagadhatják az egyenlő szavazati jogot minden személytől (kizárva a választójogból az emberrel funkcionálisan és szubjektíven egyenértékű digitális elméket); *vagy* (ii) korlátozásokat szabhatnak az új személyek létrehozására (olyan típusúakra, amelyek választójogot élveznének, ha létrehoznák őket); *vagy* (iii) elfogadhatják, hogy a szavazati jog arányos lesz a szavazati jog helyettesítőinek létrehozására való képességgel és fizetési hajlandósággal, ami gazdaságilag nem hatékony kiadásokhoz vezet az ilyen helyettesítő eszközökre, és azoknak a politikai marginalizációjához, akik nem rendelkeznek erőforrásokkal vagy nem hajlandók szavazati jog megvásárlására költeni azokat.[55]

A népességgel kapcsolatos kívánalmak

Annak teljes körű számbavétele, hogy a fejlett mesterséges intelligencia különleges körülményeinek hogyan kellene befolyásolniuk a népességpolitikát, sokkal finomabb elemzést igényelne, de az előzőekben leírtak alapján két általános kívánatos adatot tudunk meghatározni:

- **Bűnmegelőzés. A** fejlett mesterséges intelligenciát úgy irányítják, hogy az érző digitális elmékkel való rossz bánásmód elkerülhető vagy minimálisra csökkenthető legyen.
- **Népesedéspolitika. A** nemzéssel kapcsolatos döntéseket, hogy milyen új lényeket hozzunk létre, összehangolt módon és kellő előrelátással hozzuk meg, hogy elkerüljük a nemkívánatos malthusi dinamikát és a politikai eróziót.

Folyamat

Az előző kívánalmakat az *eredmények* jellemzőivel fejezzük ki. Megfogalmazhatjuk a kívánatos adatokat olyan tulajdonságok formájában is, amelyeket a *folyamatra* szeretnénk vonatkoztatni, amelyen keresztül a jövő meghatározásra kerül. Itt három olyan különleges körülményre mutatunk rá, amelyek hatással lehetnek a kormányzásra, és amelyek a szuperintelligens mesterséges intelligencia megjelenése körül valószínűsíthetően fennállnak: a politikai kontextus újdonsága, mélysége és technikai kihívása; az események üteme; valamint az uralkodó elvek és normák aláásása.

Episztemikus kihívás (újdonság, mélység és technikai jelleg)

A gépi intelligencia forradalmának kontextusa szokatlan episztemikus követelményeket támasztana a politikai döntéshozatali folyamattal szemben.

[55] Az i. lehetőség többféle formában is megvalósulhat. Például át lehetne térni egy olyan rendszerre, amelyben a szavazati jogok öröklődnek. A kezdeti népesség egy része választójoggal rendelkezne (például a jelenlegi választójoggal rendelkező személyek és a nagykorúvá váláskor meglévő gyermekeik). Amikor e választópolgárok egyike új választójogosult lényt hoz létre - legyen az digitális másolat vagy helyettesítő személy, vagy biológiai gyermek -, akkor az eredeti szavazati jogát felosztják az ős és az utód között, így az egyes "klánok" szavazati ereje állandó marad. Ez megakadályozná, hogy a gyorsan növekvő klánok ténylegesen megfosszák a lassabban növekvő népességet a szavazati jogától, és megszüntetné a politikai befolyás megszerzése érdekében történő szaporodás perverz ösztönzését. Robin Hanson a sebességgel súlyozott szavazás alternatíváját javasolta, amely a gyorsabb számítógépeken futó digitális elméknek több szavazati jogot biztosítana (Hanson 265. o2016,.). Ez csökkentheti a szavazóinfláció problémáját (a képviselet

sokszorosításának egyik stratégiáját - a sok lassú, és ezért számítás szempontjából olcsó másolat futtatását - megakadályozva). Ez azonban extra befolyást adna azoknak az elméknek, amelyek elég gazdagok ahhoz, hogy megengedhessék maguknak a gyors végrehajtást, vagy amelyek történetesen gyors végrehajtást igénylő gazdasági szerepkörökben szolgálnak.

Először is, egy közelgő vagy bekövetkező gépi intelligencia forradalom kivételesen nagymértékű változást jelentene a politikai döntéshozatal kontextusában. Ez azt jelenti, hogy számos megszokott feltételezés - például az intézményi megállapodásokba, mentális szokásokba és kulturális normákba beágyazott - alkalmazhatatlanná válhat. Ez szükségessé tenné, hogy a helyzetet új szemszögből, az első elvekből kiindulva vagy rendkívül széles és sokrétű tapasztalati bázisra támaszkodva újragondolva lássuk a dolgokat.

Másodszor, és ehhez kapcsolódóan, a döntéshozók előtt álló kihívások ebben az összefüggésben olyan alapvető világnézeti kérdéseket foglalhatnak magukban, amelyek mély empirikus, filozófiai, stratégiai vagy vallási kérdéseket érintenek, és amelyeket gyakran bizonytalanság vagy ellentmondás homályosít. Ez rámutat a *bölcsesség* különleges szükségességére. Bár nehéz operacionalizálni, a bölcsesség alatt azt a képességet értjük, hogy a legfontosabb dolgokat megbízhatóan, legalább megközelítőleg helyesen fogalmazzuk meg. A bölcsesség magában foglal egyfajta szilárdan jó ítélőképességet, jól kalibrált hitet, és a képességet arra, hogy egy trükkös és zavaros helyzeten keresztül ésszerű utat találjunk, szem előtt tartva az átfogó képet. A bölcsességhez tartozik különösen, hogy kellő mértékű ismeretelméleti alázattal rendelkezünk ahhoz, hogy felismerjük tudásunk korlátait, és képesek legyünk meggondolni magunkat, még egészen alapvető dolgokról is, ahelyett, hogy végtelenségig kitartanánk egy katasztrofálisan téves terv mellett.

Harmadszor, mivel olyan döntéshozatali kontextust feltételezünk, amelyben az egyik abszolút kritikus tényező a technológiai találmány, a szokásosnál nagyobb jelentőséggel bír, hogy képesek legyünk megérteni a technológiát - különösen a mesterséges intelligencia technológiáját -, és megfelelő elvárásokat alakítsunk ki annak tulajdonságaival és lehetőségeivel kapcsolatban. Ez a kívánalom bizonyos mértékig kielégíthető a megfelelő műszaki szakértők bevonásával, akik tanácsot adnak a döntéshozóknak. Az irányítási mechanizmus egészének azonban olyannak kell lennie, hogy a megfelelő szakértőket válasszák ki, meghallgassák és megértsék. Egyébként pedig egy olyan döntéshozó, aki nem ismeri a tudományt és a technológiát, és képtelen követni egy matematikai vagy műszaki érvelést, és így arra szorítkozik, hogy a mesterséges intelligencia technológiát egy fekete dobozként fogja fel, amelyről a különböző akkreditált tudományos szakértők rejtélyes és néha egymásnak ellentmondó rendeleteket hoznak, valószínűleg hátrányban van egy olyan döntéshozóval szemben, aki képes a szóban forgó technológiát a mechanizmus szintjén ésszerű módon megérteni.

Pace

Számos forgatókönyv szerint a gépi szuperintelligenciára való áttérés során szokatlanul gyors ütemben bontakoznának ki világtörténelmi jelentőségű események. Ez azt sugallja, hogy a kormányzási folyamatok számára a szokásosnál is fontosabb lehet, hogy képesek legyenek gyorsan és határozottan cselekedni, hogy az események előtt járjanak. Különösen kívánatos lehet, hogy a szuperintelligens mesterséges intelligencia fejlődése olyan kormányzási környezetben történjen, amelyben az alkotmányos változtatások gyorsan végrehajthatók, és a globális kormányzási megállapodásokról való döntés és azok érvényre juttatása sokkal rövidebb idő alatt lehetséges, mint a multinacionális szerződések tárgyalása, ratifikálása és végrehajtása.

Aláássa a

A gépi intelligencia forradalmának kontextusa többféleképpen is különleges lehetőségeket kínálhat az elvek és normák aláásására vagy a meglévő hatalmi struktúrák bitorlására. Ezek közül néhányat már érintettünk az "átrendeződésről" szóló fenti vitánkban, a

hogy a társadalmi eredmények hogyan lehetnek kitéve a gazdagság és a befolyás szélsőséges mértékű permutációjának vagy koncentrációjának. De megközelíthetjük ezeket a kérdéseket egy folyamatorientált perspektívából is.

Vegyük figyelembe az olyan elveket, mint a legitimitás, a hozzájárulás, a politikai részvétel és az elszámoltathatóság. Ezekről széles körben úgy vélik, hogy a kormányzási rendszerek és a politikai döntéshozatali folyamatok kívánatos tulajdonságai. A gépi intelligencia forradalmának különleges körülményei azonban többféleképpen is alááshatják ezeket az elveket.

Vegyük például az önkéntes beleegyezés gondolatát, amely rendkívül fontos elv, és amely számos, az egyének és az államok közötti interakciót szabályoz. Sok olyan dolog, amit erkölcsileg helytelen vagy törvénytelen lenne egy egyénnel a beleegyezése nélkül megtenni, teljesen kifogástalan, ha a beleegyezésével történik. Ugyanez érvényes a vállalati entitások vagy államok közötti számos lehetséges interakcióra: nagyon gyakran nagy különbség, hogy valamit erőszakkal vesznek el vagy kényszerítenek ki, vagy önkéntesen beleegyeznek-e valamibe. Gondoljunk csak arra, hogy a fejlett mesterséges intelligencia kontextusában a beleegyezésnek tulajdonított központi szerep miként ásható alá, ha lehetővé válik egy
"szuper-győzködő", egy olyan rendszer, amely az érvelés és a retorika rendkívül ügyes alkalmazásával képes szinte bármilyen emberi személyt vagy csoportot (hasonlóan erős mesterséges intelligencia segítsége nélkül) szinte bármilyen álláspontról meggyőzni, vagy szinte bármilyen üzletet elfogadtatni velük. Ha lehetséges lenne egy ilyen szuper-győzködőt létrehozni, akkor nem lenne helyénvaló továbbra is a beleegyezésre mint eszközre hagyatkozni.
szinte elégséges feltétele annak, hogy számos ügylettípus erkölcsileg és jogilag kifogástalan legyen. Egy szuperügynökökkel rendelkező világban erősebb védelemre lenne szükség az emberi érdekek védelmére, hasonlóan a jelenleg a kiszolgáltatott személyek bizonyos csoportjainak, például a gyermekek és a kognitív zavarokkal küzdő felnőttek érdekeinek védelmére bevezetett extra biztosítékokhoz. Talán a beleegyezést csak akkor lehetne érvényesnek tekinteni, ha az emberi szerződő félnek hozzáférése van egy képzett mesterséges intelligencia tanácsadóhoz, vagy ha a tranzakciót egy, az emberi szereplő mellé rendelt "mesterséges intelligencia-felügyelő" hagyja jóvá, hogy megvédje őt a kizsákmányolástól.

Egy másik példa a politikai részvétel normája. Ezt a normát többféle okból is lehet indokolni. Egyfelől episztemikus előnyt jelenthet azáltal, hogy több információt és a nézőpontok szélesebb körét vonja be a döntéshozatali folyamatba. Másrészt biztosíthatja azt is, hogy a meghozott döntésekben sokféle érdek és preferencia tükröződjön. A másik oldalról pedig a politikai részvétel tekinthető olyan eredendő jószágnak, amelyet függetlenül attól, hogy milyen mértékben járul hozzá a demokráciához, értékelni kell.
olyan döntések meghozatala, amelyek minden érintett érdekét jobban szolgálják.[56] Ez a három indoklás
újra kell értékelni a szuperintelligens mesterséges intelligencia kontextusában. Lehetséges például, hogy a politikai döntések sok emberi vélemény által való befolyásolásának episztemikus értéke csökkenne vagy megszűnne, ha a szuperintelligens mesterséges intelligencia episztemikailag kellően magasabbrendű lenne az embereknél, és képes lenne önállóan felismerni és integrálni a bizonyítékok és meglátások minden olyan darabkáját, amelyet egy elosztott emberi episztemikus közösség képes lenne szolgáltatni. Az is elképzelhető, hogy a fejlett mesterséges intelligencia lehetővé tenné egy olyan mechanizmus létrehozását, amely nem igényli az emberi preferenciák folyamatos bevitelét ahhoz, hogy ezeket a preferenciákat a meghozandó döntésekbe beépítse - talán egy szuperintelligens mesterséges intelligencia képes lenne megtanulni egy olyan

preferenciafüggvényt, amely már előre jelzi az emberi preferenciák meglévő eloszlását és az idővel bekövetkező preferenciaváltozásokat, vagy az AI képes lenne erre következtetni másfajta emberi viselkedés megfigyeléséből. A politikai részvétel feltételezett belső értéke megmaradhatna

[56] Egy negyedik ok lehet annak biztosítása, hogy a döntéseket törvényesnek tekintsék.

még akkor sem lenne érintetlen, ha a két instrumentális indoklás eltűnne; vagy talán furcsa és perverz dolognak tartanánk, hogy részt akarunk venni a politikai ügyekben, miután világossá vált, hogy beavatkozásaink csak a politikai eredmények romlását szolgálják (mind a saját, mind a szélesebb társadalom érdekei szempontjából).

E két példa célja nem az, hogy a szuperintelligens mesterséges intelligencia korszakában a beleegyezésről vagy a politikai részvételről szóló konkrét állításokat terjesszünk elő, hanem az, hogy egy általánosabb pontot illusztráljunk: hogy vannak különböző elvek és normák, amelyek jelenleg mélyen beágyazódtak és gyakran támogatják anélkül, hogy minősítés, amelyet a radikális mesterséges intelligencia kontextusában újból meg kell vizsgálni.[57] Ezek közül néhány normákat és elveket ebben az összefüggésben esetleg el kell hagyni; másokat újra kell értelmezni és újra kell fogalmazni; és megint másokat a szokásosnál nagyobb éberséggel kell védeni. Ez rámutat a kormányzási folyamatok általános kívánalmára ebben az összefüggésben, nevezetesen arra, hogy azok képesek legyenek a vonatkozó normák és elvek megfelelő kiigazításához vezetni.[58]

A folyamathoz kapcsolódó kívánalmak

Az előző megfigyelésekből levezetünk egy sor kívánatos adatot, amelyek a szuperintelligens mesterséges intelligencia kontextusában a szakpolitikai döntések meghozatalának irányítási folyamataira vonatkoznak:

- **Elsődleges gondolkodás, bölcsesség, technikai megértés.** A szuperintelligens mesterséges intelligenciához való átmenetet egy olyan (egyéni vagy kollektív, központosított vagy elosztott) ügynökség irányítja, amely képes az első elveken alapuló gondolkodás, a bölcsesség és a technikai megértés szokatlan szintjeit hatékonyan integrálni a döntéshozatalba.
- **Gyorsaság és határozottság.** A szuperintelligens mesterséges intelligencia fejlesztése és alkalmazása olyan politikai környezetben történik, amelyben megvan a gyors döntéshozatal és a határozott globális végrehajtás képessége (vagy, alternatívaként, a fejlődés ütemének mérséklésére való képesség, hogy a lassabb döntéshozatali és koordinációs folyamatok hatékonyak lehessenek).
- **Alkalmazkodóképesség. A** szuperintelligens mesterséges intelligencia olyan társadalmi-politikai kontextusban kerül alkalmazásra, amelyben a szabályok, elvek, normák és törvények az új körülményekhez igazodóan alkalmazhatók.

[57] Ezek a normák és elvek azért nyerhettek teret, mert a korábbi évtizedek és évszázadok társadalmi-technológiai miliőjében segítettek a kormányzási kihívások kezelésében.

[58] Az e dokumentumban korábban tárgyaltak közül néhány további példát kínál arra, hogy a meglévő normákat milyen esetekben kellene visszavonni vagy újrafogalmazni. A korlátlan reprodukcióhoz való jog aligha védhető egy olyan kontextusban, ahol a malthusiánus aggodalmak nagymértékben megjelennek, mint például a digitális elmék esetében. A gondolatszabadságot hasonlóképpen korlátozni kell a mesterséges intelligencia elméi esetében, amelyek pusztán azáltal, hogy részletesen gondolkodnak egy szenvedő alanyról, képesek arra, hogy belsőleg szenvedő állapotba hozzák ezt az elmét, és így elmebűncselekményt kövessenek el. Bűncselekmények büntetése: a bebörtönzés néhány jelenlegi oka megszűnne, ha például a fejlett mesterséges intelligencia lehetővé tenné a bűnelkövetők hatékonyabb rehabilitációját vagy a társadalomba való visszaengedését anélkül, hogy más polgárokat veszélyeztetnének, vagy ha a hatékonyabb bűnmegelőzési módszerek bevezetése csökkentené a jövőbeni bűncselekményektől való elrettentés szükségességét. Az adott büntetés értelme: még ha az életfogytiglani börtönbüntetés néha igazságos

büntetés is, amikor a hátralévő élettartam jellemzően néhány évtized, nem biztos, hogy igazságos, ha a mesterséges intelligenciával támogatott orvostudomány lehetővé teszi az élettartam jelentős meghosszabbítását. Különböző méltóságon alapuló vagy vallási érzékenységek különleges védelmet és alkalmazkodást igényelhetnek a fejlett mesterséges intelligencia kontextusában. Magát a mesterséges intelligenciával kapcsolatos kutatást pedig másképp kell megközelíteni, mint a legtöbb alapkutatást, ahol a kíváncsiságtól vezérelt felfedezés, a nyitottság és a szellemi teljesítmény ünneplése gyakran a végső mérceként szerepel. A mesterséges intelligencia kutatása esetében a kutatási hozzájárulások értékelésének szempontjai közé a kutatási eredmények későbbi alkalmazásainak és stratégiai hatásainak megfontolását is be kell illeszteni.

Összefoglaló

Felhívtuk a figyelmet számos olyan különleges körülményre, amelyek a szuperintelligens mesterséges intelligencia fejlesztését és alkalmazását övezhetik, és amelyek sajátos kihívások elé állítják a kormányzást és a globális politikát. A normatív elemzés "vektormező" megközelítését alkalmazva arra törekedtünk, hogy ezekből a különleges körülményekből irányadó politikai implikációkat vonjunk le. Ezeket az implikációkat a kívánatos adatok - a jövőbeli politikák, kormányzási struktúrák vagy döntéshozatali kontextusok olyan jellemzői - összességeként jellemeztük, amelyek a legfontosabb szereplők, érdekeltek és etikai nézetek széles körének normái szerint javítanák a gépi intelligencia korszakára való áttérés kedvező eredményeinek kilátásait. Ezeket a kívánatos adatokat (amelyekről nem állítjuk, hogy kimerítőek) a táblázatban foglaltuk össze. 1.

<table>
<tr><td colspan="2">Hatékonyság</td></tr>
<tr><td>Technológiai lehetőség</td><td rowspan="4">Gyors előrehaladás. Ez két összetevőre oszlik: (a) olyan politikák, amelyek nagy valószínűséggel vezetnek a biztonságos szuperintelligencia végleges kifejlesztéséhez és annak alkalmazásához a jólét új forrásainak megcsapolásához; és (b) a gyors mesterséges intelligencia fejlődés, hogy a társadalmilag hasznos termékek és alkalmazások időben széles körben elérhetővé váljanak.
Mesterséges intelligencia biztonsága. Olyan technikákat dolgoznak ki, amelyek lehetővé teszik (túlzott költségek, késedelem vagy teljesítménycsökkenés nélkül) annak biztosítását, hogy a szuperintelligens mesterséges intelligencia a tervezett módon viselkedjen. A szuperintelligencia kialakulása és korai bevezetése során a körülmények is olyanok, amelyek a rendelkezésre álló legjobb biztonsági technikák és az általánosan óvatos megközelítés alkalmazását ösztönzik.
Feltételes stabilizáció. A fejlődési pálya és a tágabb politikai kontextus olyan, hogy ha drasztikus stabilizációs intézkedések hiányában katasztrofális globális koordinációs kudarc következne be, akkor a szükséges stabilizációt időben végrehajtják a katasztrófa elkerülése érdekében. Ez azt jelentheti, hogy (valamelyik szereplő vagy szereplők számára) megvalósítható lehetőségnek kell lennie egy egyeduralom létrehozására, vagy az intenzív globális felügyelet rendszerének bevezetésére, vagy a veszélyes technológia vagy tudományos ismeretek terjesztésének szigorú visszaszorítására.
Nem turbulencia. Az útvonal elkerüli a káoszból és a konfliktusokból eredő túlzott hatékonyságveszteségeket. A politikai rendszerek fenntartják a stabilitást és a rendet, sikeresen alkalmazkodnak a változásokhoz, és mérséklik a társadalmilag zavaró hatásokat.</td></tr>
<tr><td>Mesterséges intelligencia kockázata</td></tr>
<tr><td>A globális koordináció katasztrofális meghibásodásának lehetősége</td></tr>
<tr><td>A turbulencia csökkentése</td></tr>
<tr><td colspan="2">Kiosztás</td></tr>
<tr><td>Kockázati externáliák</td><td rowspan="3">Egyetemes juttatás. Mindenki, aki az átmenet idején életben van (vagy akit negatívan érinthet), részesül a juttatásból, kompenzációként a kockázat externáliáiért, amelyeknek ki volt téve.
Epsilon-magnanimitás. Az erőforrásokkal jóllakható értékek széles skálája (olyanok, amelyek ellen a költségalapú megfontolásokon kívül nem sok kifogás van), akkor és úgy</td></tr>
<tr><td>Átalakítás</td></tr>
<tr><td>A tudatlanság fátyla</td></tr>
</table>

Cornucopia	valósul meg, ha és amennyiben ez az összes erőforrás parányi töredékének felhasználásával lehetséges. Ez magában foglalhatja az alapvető jóléti ellátásokat és a jövedelemgaranciákat minden emberi egyén számára. Ide tartozhat számos közösségi jószág, etikai eszmény, esztétikai vagy szentimentális projekt, valamint a nagylelkűség, a kedvesség és az együttérzés különböző természetes megnyilvánulásai is. **Folytonosság**. Az út ésszerű mértékű folytonosságot biztosít, hogy i. fenntartsa a rendet és biztosítsa a szükséges intézményi stabilitást ahhoz, hogy a szereplők a jelenlegi tudatlanság fátyla mögött kihasználhassák a kereskedelem lehetőségeit, beleértve a szociális biztonsági hálókat is; és ii. megakadályozza, hogy a koncentráció és a permutáció szükségtelenül nagy legyen.
Népesség	
A digitális elmék érdekei Népességdinamika	**Bűnmegelőzés.** Elmés bűnmegelőzés. A fejlett mesterséges intelligenciát úgy irányítják, hogy az érző digitális elmékkel való bántalmazás elkerülhető vagy minimálisra csökkenthető legyen. **Népesedéspolitika. A** nemzéssel kapcsolatos döntéseket, hogy milyen új lényeket hozzunk létre, összehangolt módon és kellő előrelátással hozzuk meg, hogy elkerüljük a nemkívánatos malthusi dinamikát és a politikai eróziót.
Folyamat	
Episztemikus kihívás (újdonság, mélység és technikai jelleg) Pace Aláássa a	**Elsődleges gondolkodás, bölcsesség, technikai megértés.** A szuperintelligens mesterséges intelligenciához való átmenetet egy olyan (egyéni vagy kollektív, központosított vagy elosztott) ügynökség irányítja, amely képes az első elveken alapuló gondolkodás, a bölcsesség és a technikai megértés szokatlan szintjeit hatékonyan integrálni a döntéshozatalba. **Gyorsaság és határozottság.** A fejlett mesterséges intelligencia fejlesztése és alkalmazása olyan politikai környezetben történik, amelyben megvan a gyors döntéshozatal és a határozott globális végrehajtás képessége (vagy pedig a fejlődés ütemének mérséklésére való képesség, hogy a lassabb döntéshozatali és koordinációs folyamatok hatékonyak lehessenek). **Alkalmazkodóképesség. A** szuperintelligens mesterséges intelligencia olyan társadalmi-politikai kontextusban kerül alkalmazásra, amelyben a szabályok, elvek, normák és törvények az új körülményekhez igazodóan alkalmazhatók.

Táblázat A gépi intelligencia korszakára való áttéréssel várhatóan együtt járó különleges**1.** körülmények (bal oldali oszlop) és a megfelelő irányítási intézkedésekre vonatkozó kívánalmak (jobb oldali oszlop).

A táblázatban szereplő kívánalmak olyan kritériumokat 1határoznak meg, amelyek alapján a fejlett mesterséges intelligencia irányítására vonatkozó konkrét szakpolitikai javaslatok értékelhetők. A "szakpolitikai javaslatok" alatt nemcsak a hivatalos kormányzati dokumentumokat értjük, hanem a hosszú távú mesterséges intelligenciafejlesztésben érdekelt magánszereplők által kidolgozott terveket és lehetőségeket is. A desideráták tehát egyes vállalatok, kutatásfinanszírozók, akadémiai vagy nonprofit kutatóközpontok, valamint különféle más szervezetek és magánszemélyek számára is relevánsak.

Az e kívánalmaknak megfelelő konkrét javaslatok kidolgozása további kutatások feladata. Az ilyen konkrét javaslatokat valószínűleg konkrét szereplőkre kellene relativizálni, mivel az általunk meghatározott általános megfontolásoknak való megfelelés legjobb módja az adott szereplő kapacitásától, erőforrásaitól és politikai korlátaitól függ, akinek a javaslat szól. Ezen túlmenően az egyes szereplők további sajátos preferenciákkal is rendelkezhetnek, amelyeket a vektormező-elemzésünk nem tud teljes mértékben megragadni, de amelyeket figyelembe kell venni ahhoz, hogy egy politikai javaslatnak esélye legyen az elfogadásra.

Hivatkozások

Alexander, S., Elmélkedések2014. Molochról. *Slate Star Codex* (július 30.). Elérhető a következő címen: http://slatestarcodex.com/2014/07/30/meditations-on-moloch/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. és Mané, D., Concrete2016. problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Armstrong, M.S., Bostrom, N. és Shulman, C., Racing2016. to the precipice: a mesterséges intelligencia fejlődésének modellje. *AI & Society, 31*(2), pp. 201-206.

Armstrong, M.S. és Orseau, L., Biztonságosan2016. megszakítható szerek. *Konferencia a bizonytalanságról a mesterséges intelligenciában*.

Armstrong, M.S. és Sandberg, A., Örökkévalóság2013. hat órában: Az intelligens élet intergalaktikus terjedése és a Fermi-paradoxon élesítése. *Acta Astronautica*,*89* , pp. 1-13.

Barnett, M. és Duvall, R., 2005. Hatalom a nemzetközi politikában. *International organization*,*59* (1), pp. 39-75.

Beauchamp, J.P., Genetikai2016. bizonyítékok a természetes szelekcióra az embereknél a mai Egyesült Államokban. *Proceedings of the National Academy of Sciences,113* (28), pp. 7774-7779.

Beckstead, N., *On*2013. *the overwhelming importance of shaping the far future* (doktori disszertáció, Rutgers University-Graduate School-New Brunswick).

Bhuta, N., Beck, S., Geiβ, R., Liu, H., és Kreβ, C. (szerk.). 2016. *Autonóm fegyverrendszerek: Law, Ethics, Policy*. Cambridge: Cambridge University Press.

Bostrom, N., 2003a. Egy számítógépes szimulációban élünk?. *The Philosophical Quarterly,53* (211), pp. 243-255.

Bostrom, N., 2003b. Csillagászati hulladék: A késedelmes technológiai fejlődés alternatív költségei. *Utilitas*, *15*(03), pp. 308-314.

Bostrom, N., 2003c. A transzhumanista GYIK: v *World* 2.1.*Transhumanist Association*. Elérhető a következő címen: http://www.nickbostrom.com/views/transhumanist.pdf

Bostrom, N., Az emberi evolúció 2004.jövője. *Halál és halálellenesség: Kétszáz évvel Kant után, ötven évvel Turing után.* Ria University Press: Palo Alto, pp. 339-371.

Bostrom, N., Transzhumanista2005. értékek. *Journal of Philosophical Research*, *30*(Supplement), pp. 3-14.

Bostrom, N., Mi2006. az a singleton. *Nyelvészeti és filozófiai vizsgálatok*,*5* (2), pp. 48-54.

Bostrom, N., Levél2008. az utópiából. *Studies in Ethics, Law, and Technology*,*2* (1).

Bostrom, N., Morális2009. bizonytalanság - a megoldás felé? *Overcoming Bias* (január 1.). Elérhető a következő címen: http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html

Bostrom, N., 2013. Az egzisztenciális kockázatok megelőzése mint globális prioritás. *Global Policy*,*4*

(1), pp. 15-31. Bostrom, N. *Szuperintelligencia*2014.*: Paths, Dangers, Strategies*. Oxford: Oxford

University Press.

Bostrom, N. és Yudkowsky, E., A mesterséges intelligencia 2014.etikája. *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, pp. 316-334.

Bostrom, N. A 2018.sebezhető világ hipotézise. *Előkészületben*.

Brynjolfsson, E. és McAfee, A. *A* 2014.*második gépkorszak*: *Munka, haladás és jólét a briliáns technológiák korában*. Vancouver: WW Norton & Company.

Calo, R., Kukucskáló2010. HAL-ok: Making Sense of Artificial Intelligence and Privacy. *European Journal of Legal Studies*,*2* (3), p. 168.

Az Egyesült Nemzetek Alapokmánya. 1945.1 UNTS XVI, október 24Elérhető1945.: http://www.refworld.org/docid/3ae6b3930.html.

Christiano, P., Félig felügyelt2016. megerősített tanulás. *AI Alignment* (május 6.). Elérhető a következő címen: https://medium.com/ai-control/semi-supervised-reinforcement-learning-cf7d5375197f

Clark, J., Ki2016. irányítsa gondolkodó gépeinket? *Bloomberg* (augusztus 4.). Elérhető a következő címen: http://www.bloomberg.com/features/2016-demis-hassabis-interview-issue

Conitzer, V., Philosophy2016. in the Face of Artificial Intelligence. *arXiv preprint arXiv:1605.06048*.

de Mesquita, B.B. és Smith, A. *A* 2011.*diktátor kézikönyve: miért a rossz viselkedés szinte mindig jó politika*. New York: PublicAffairs.

Drexler, K.E. *A teremtés motorjai*:1986. *A nanotechnológia eljövendő korszaka*. New York: Anchor Books.

Evans, O., Stuhlmüller, A. és Goodman, N.D., Learning 2015.the preferences of ignorant, inconsistent agents. *Thirtieth AAAI Conference on Artificial Intelligence.*

FAT/ML., Fairness2018., Accountability, and Transparency in Machine Learning (Tisztesség, elszámoltathatóság és átláthatóság a gépi tanulásban). Elérhető a következő címen: https://www.fatml.org/

Fisher, D.R., Martin2009., Richard (1754-1834), Dangan és Ballynahinch, co. Galway és Manchester 16Buildings, Mdx. *A parlament története: az alsóház 1820-1832*. Cambridge: Cambridge University Press. Elérhető a következő címen: http://www.historyofparliamentonline.org/volume/1820-1832/member/martin-richard-1754-1834

Freitas, R.A., Egy 1980.önreprodukáló csillagközi szonda. *Journal of the British Interplanetary Society,33* (7), pp. 251-264.

Freitas, R.A. *Nanomedicina*1999*., I. kötet: alapvető képességek*. Georgetown, TX: Landes Bioscience, pp. 17-18.

Friend, T., Sam 2016.Altman's Manifest Destiny. *The New Yorker* (október 10.). Elérhető a következő címen: https://www.newyorker.com/magazine/2016/10/10/sam-altmans-manifest-destiny

Good, I.J., Spekulációk1965. az első ultraintelligens géppel kapcsolatban. *Advances in computers,6* (99), pp. 31-83.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. és Evans, O., When 2017.Will AI Exceed Human Performance? Evidence from AI Experts. *arXiv preprint arXiv:1705.08807*.

Hadfield-Menell, D., Dragan, A., Abbeel, P. és Russell, S., Cooperative2016. Inverse Reinforcement Learning. *arXiv preprint arXiv:1606.03137*.

Hale, T. és Held, D. *A transznacionális kormányzás kézikönyve*2011.. Cambridge: Polity.

Hanson, R. *A* 2016.*korszak Em: Munka, szerelem és élet, amikor robotok uralják a világot*. Oxford: Oxford University Press.

Horowitz, M., Kinek2016. kellenek majd a mesterségesen intelligens fegyverek? Az ISIS, a demokráciák vagy az autokráciák? *Bulletin of the Atomic Scientist Years70 Speaking Knowledge to Power.* Elérhető a következő címen: http://thebulletin.org/who%E2%80%99ll-want-artificially-intelligent-weapons-isis-democracies-or-a utocracies9692

Az alsóház tudományos és technológiai bizottsága. 2016. *Robotika és mesterséges intelligencia: Ötödik jelentés a 2016-17-es ülésszakról*. Elérhető a következő címen: http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

Lordok Háza mesterséges intelligenciával foglalkozó bizottsága. 2018. *AI az Egyesült Királyságban: Ready, Willing and Able*. Elérhető a következő címen: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

Polgári és politikai jogok nemzetközi egyezségokmánya. 1966. Treaty Series, vol. p999,. December 171,16Elérhető1966.: http://www.refworld.org/docid/3ae6b3aa0.html.

Gazdasági, szociális és kulturális jogok nemzetközi egyezségokmánya. 1966. Treaty Series, vol. p993,. December 3,16Elérhető1966.: http://www.refworld.org/docid/3ae6b36c0.html.

Nemzetközi Valutaalap, Világgazdasági2014. kilátások adatbázis. Elérhető a következő címen: http://www.imf.org/external/pubs/ft/weo/2014/02/weodata/index.aspx

Kong, A., Frigge, M.L., Thorleifsson, G., Stefansson, H., Young, A.I., Zink, F., Jonsdottir, G.A., Okbay, A., Sulem, P., Masson, G. és Gudbjartsson, D.F., Selection 2017.against variants in the genome associated with educational attainment. *Proceedings of the National Academy of Sciences,114* (5), pp. E727-E732.

Lagerwall, A. *Jus*2015. *Cogens*. Available at: http://www.oxfordbibliographies.com/view/document/obo-9780199796953/obo-9780199796953- 0124.xml

Lin, P., Abney, K. és Bekey, G. (szerk.). 2011. *Robot etika: The Ethical and Social Implications of Robotics*. Cambridge Massachusetts: The MIT Press.

Merkle, R.C., Az agy 1994.molekuláris javítása. *Cryonics magazin*,*15* .

Milot, E., Mayer, F.M., Nussey, D.H., Boisvert, M., Pelletier, F. és Réale, D., Evidence2011. for evolution in response to natural selection in a contemporary human population. *Proceedings of the National Academy of Sciences,108* (41), pp. 17040-17045.

Müller, V.C. és Bostrom, N., A mesterséges intelligencia jövőbeli 2016.fejlődése: A survey of expert opinion. *A mesterséges intelligencia alapvető kérdései*. Svájc: Springer International Publishing, pp. 553-570.

Nemzeti Tudományos és Technológiai Tanács. 2016. *Felkészülés a mesterséges intelligencia jövőjére*. Washington, D.C.: Tudományos és Technológiai Politikai Hivatal. Elérhető a következő címen: https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Nehal, B., Beck S., Geiss R., Liu, H. és Kress, K. (szerk.). 2016. *Autonóm fegyverrendszerek: Law, Ethics, Policy*. Cambridge: Cambridge University Press.

Nordhaus, W.D. *Közeledünk* 2015.*a gazdasági szingularitáshoz? Az információs technológia és a gazdasági növekedés jövője* (No. w21547). National Bureau of Economic Research.

OpenAI, biztonság2016.: Környezetek a különböző AI biztonsági tulajdonságok tesztelésére. Elérhető a következő címen: https://gym.openai.com/envs#safety

Pearce, D. *A* 1995.*hedonista imperatívusz*. Elérhető a következő címen: https://www.hedweb.com/hedab.htm

Piketty, T. *Capital*2014. *in the twenty-first century* (A. Goldhammer, Trans.) Cambridge Massachusetts: The Belknap Press.

Pinker, S. *Természetünk* 2011.*jobb angyalai*: *Az erőszak hanyatlása a történelemben és annak okai*. London: London: Penguin.

Rawls, J. *Az igazságosság* 1971.*elmélete*. Cambridge Massachusetts: The Belknap Press.

Roff, H.M., A 2014.stratégiai robotprobléma: Halálos autonóm fegyverek a háborúban. *Journal of Military Ethics*,*13* (3), pp. 211-227.

Russell, S. és Norvig, P. *Mesterséges*2010. *intelligencia: A Modern Approach*. 3. kiadás. Upper Saddle River, NJ: Prentice-Hall.

Russell, S., Dewey, D. és Tegmark, M., Research2016. priorities for robust and beneficial artificial intelligence. *arXiv preprint arXiv:1602.03506.*

Sandberg, A. *Grand Futures*. Hamarosan megjelenik.

Sandberg, A. és Bostrom, N., Egész2008. agyi emuláció: A Roadmap." *Technikai jelentés* 2008-3. Az Emberiség Jövője Intézet, Oxfordi Egyetem. Elérhető a következő címen:

http://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf

Sandberg, A., Drexler, E. és Ord, T., Dissolving2018. the Fermi Paradox. *arXiv preprint arXiv:1806.02404*.

Scherer, M.U., Mesterséges intelligencia rendszerek szabályozása2016.: Kockázatok, kihívások, kompetenciák és stratégiák. *Harvard Journal of Law and Technology*,*29* (2), pp. 353-400.

Soares, N. és Fallenstein, B., A szuperintelligencia és az emberi érdekek összehangolása2014.: A technical research agenda. *Gépi Intelligencia Kutatóintézet (MIRI) technikai jelentés,8* .

Taylor, J., Yudkowsky, E., LaVictoire, P. és Critch, A., Alignment2016. for Advanced Machine Learning Systems. Elérhető a következő címen: https://intelligence.org/files/AlignmentMachineLearning.pdf

Tipler, F.J., Földönkívüli1980. intelligens lények nem léteznek. *Quarterly Journal of the Royal Astronomical Society*,*21* , pp. 267-281.

Yudkowsky, E., A mesterséges2008.. intelligencia mint a globális kockázat pozitív és negatív tényezője. *Globális katasztrófakockázatok*. Oxford: Oxford University Press, pp. 308-345.

Az Emberi Jogok Egyetemes Nyilatkozata. 1948.217 A (III), december 10Elérhető1948. a következő címen: http://www.refworld.org/docid/3ae6b3712c.html.

Amerikai szenátus. Kereskedelmi albizottság, űrkutatási, tudományos és versenyképességi albizottság. 2016. *A mesterséges intelligencia hajnala*. Meghallgatás, november30. Washington. Elérhető a következő címen: http://www.commerce.senate.gov/public/index.cfm/2016/11/commerce-announces-first-artificial-int elligence-hearing

Vöneky, S., A tudományos kísérletek és a technológiai fejlődés egzisztenciális2016. kockázatai: Nehéz kérdések - nincs nemzetközi (emberi jogi) jog? Elérhető a következő címen: Vönöky, Vö: http://www.jura.uni-freiburg.de/institute/ioeffr2/forschung/silja-voeneky-hrp-precis.pdf

West, D.M. és Allen, J., How 2018.Artificial Intelligence Is Transforming the World (Hogyan alakítja át a mesterséges intelligencia a világot). *Brookings Institute* (április 24.). Elérhető a következő címen: https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/