A világ megosztása a digitális elmékkel[1]

(2020). Tervezet. 1.8. változat
Carl Shulman† & Nick Bostrom†

[in Clarke, S. , Zohny, H. & Savulescu, J. (szerk.): Rethinking Moral Status (Oxford University Press, 2021)]].

Absztrakt

A biológiai lények elméje csak egy kis sarkát foglalja el a lehetséges elmék sokkal nagyobb terének, amelyeket létrehozhatunk, amint elsajátítjuk a mesterséges intelligencia technológiáját. Erkölcsi intuíciónk és gyakorlatunk nagy része azonban az emberi természetre vonatkozó feltételezéseken alapul, amelyeknek a digitális elmék esetében nem kell érvényesnek lenniük. Ez rámutat arra, hogy a fejlett gépi intelligencia korszakához közeledve erkölcsi reflexióra van szükség. Itt a kérdések egy csoportjára összpontosítunk, amelyek az erőforrásokra és befolyásra való emberfeletti igényt támasztó digitális elmék kilátásából fakadnak. Ezek abból a hatalmas kollektív haszonból adódhatnak, amelyet a tömegesen előállított digitális elmék viszonylag kis mennyiségű erőforrásból nyerhetnek. Másik lehetőségként felmerülhetnek olyan egyéni digitális elmékből is, amelyek emberfeletti erkölcsi státusszal vagy az erőforrásokból való részesedés képességével rendelkeznek. Az ilyen lények óriási értékkel járulhatnak hozzá a világhoz, és ha nem tartjuk tiszteletben az érdekeiket, az erkölcsi katasztrófához vezethet, míg a naiv módon történő tiszteletük katasztrofális lehet az emberiség számára. Az ésszerű megközelítés erkölcsi normáink és intézményeink reformját igényli, valamint előzetes tervezést arra vonatkozóan, hogy milyen digitális elmékkel rendelkezünk.
létrehozni.

1. Bevezetés

Az emberi biológiai természet számos gyakorlati korlátot szab annak, hogy mit lehet tenni valakinek a jóléte érdekében. Csak ennyi ideig élhetünk, ennyi örömöt érezhetünk, ennyi gyermeket szülhetünk, és ennyi hasznot húzhatunk a további támogatásból és erőforrásokból. Mindeközben a boldogulásunkhoz fizikai, pszichológiai és társadalmi feltételek komplex összességének teljesülésére van szükségünk.

Más lények esetében azonban ezek a korlátok enyhülhetnek. Gondoljunk csak a tudatos tapasztalatokkal, vágyakkal, érvelési és autonóm döntéshozatali képességgel rendelkező gépi elmék lehetőségére.[2] Az ilyen gépek erkölcsi státuszt élvezhetnének, vagyis ahelyett, hogy az ember puszta eszközei lennének, ők és érdekeik saját jogon is számíthatnának. Nem kell ugyanazoknak a gyakorlati korlátozásoknak alávetni őket abban, hogy további erőforrásokból részesüljenek, és nem kell ugyanolyan összetett követelményektől függeniük a túlélésükhöz és virágzásukhoz. Ez csodálatos fejlemény lehetne: az életek mentesülnének a

† Az emberiség jövője Intézet, Oxfordi Egyetem.

[1] A hasznos megjegyzésekért hálásak vagyunk Guy Kahane, Matthew van der Merwe, Hazem Zohny, Max Daniel, Lukas Finnveden, Lukas Gloor, Uli Alskelung Von Hornbol, Daniel Dewey, Luke Muehlhauser,
James Babcock, Ruby Bloom, Vincent Luczkow, Nick Beckstead, Hilary Greaves, Owen Cotton-Barratt,

Allan Dafoe és Wes Cowley.

[2] Feltételezzük, hogy a megfelelően architektúrázott mesterséges intelligencia tudatos lehet, bár érdemes megjegyezni, hogy az erkölcsi státusz egyes elméletei ezt nem tekintik az erkölcsi státusz szükséges feltételének; lásd pl.
(Chalmers, 2010) a mesterséges intelligencia tudatosságáról, és (Kagan, 2019) az öntudatlan, de agenciális mesterséges intelligencia morális státuszáról.

fájdalmat és betegséget, pezseg a boldogságtól, emberfeletti tudatossággal és megértéssel és mindenféle magasabb rendű javakkal gazdagodva.[3]

A gépi tanulás terén a közelmúltban elért eredmények felvetik annak lehetőségét, hogy az ilyen digitális elmék a belátható jövőben gyakorlati valósággá válhatnak (vagy nagyon korlátozott mértékben talán már léteznek is). Néhány ilyen elme megvalósíthatja Robert Nozick (1974,
p. 41) híres filozófiai gondolatkísérlet a "hasznossági szörnyekről":

> Az utilitarista elméletet zavarba hozza a hasznossági szörnyetegek lehetősége, akik mások bármely áldozatából óriási mértékben nagyobb hasznot húznak, mint amennyit ezek a mások veszítenek. Az elmélet ugyanis elfogadhatatlanul úgy tűnik, hogy a teljes hasznosság növelése érdekében mindannyiunkat fel kell áldozni a szörnyeteg torkában.

Derek Parfit (1984, 343. o.) azzal érvel, hogy bár nehéz elképzelni egy olyan életet, amely milliószor annyit ér, mint a legjobban élő embereké, hasonló eredményekre juthatunk, ha a *népességméret* mennyiségi dimenzióját tekintjük, amelyben nyilvánvalóan nincs fogalmi akadálya a szélsőséges értékeknek.

Azzal fogunk érvelni, hogy a populáció mérete csak egy a számos mennyiségi dimenzió közül - több kevésbé biztos minőségi dimenzióval együtt -, amelyek mentén a digitális elmék messze felülmúlhatják az embereket az egységnyi erőforrás-fogyasztásból származó haszon tekintetében. Ez a többféle út teszi szilárdabbá azt a következtetést, hogy legalább az egyik megvalósul.

Bár a nem-utilitaristák immunisnak képzelhetik magukat a hasznossági szörny kihívásával szemben, a legtöbb ésszerű nézet valójában különböző mértékben fogékony rá. Ennek oka, hogy még ha azt tételezzük is fel, hogy nem történne deontológiai jogsértés, az emberi érdekeket akkor is hátrányosan érintheti a hasznossági szörnyek megjelenése, mivel ez utóbbiaknak erősebb erkölcsi igényeik lehetnek az állami támogatásra vagy a természeti erőforrásokra és más szűkös erőforrásokra, így csökkentve az emberek által védhetően igényelhető mennyiséget. Az ilyen tulajdonságokkal rendelkező digitális elmék pártatlan szempontból erkölcsileg értékesebbé tehetik a világot, miközben a közös normákat is sokkal igényesebbé tehetik a létező lények (vagy egyáltalán bármely kevésbé optimalizált elme (digitális vagy más) számára).

2. A szuperkedvezményezettek megvalósításához vezető utak

Bár a "hasznossági szörny" kifejezésnek van tudományos múltja, pejoratív és potenciálisan sértő módon utal azokra a lényekre, amelyek szokatlanul nagy szükségletekkel rendelkeznek, vagy képesek rendkívüli jó életet megvalósítani. Ezért ehelyett a következő elnevezést fogadjuk el:

> *szuperhasznosító*: olyan lény, aki emberfeletti hatékonysággal képes jólétet nyerni az erőforrásokból.
>
> *szuperpáciens*[4]: emberfeletti erkölcsi státusszal rendelkező lény.

[3] Ezek közül néhány legalább részben elérhető lehet a továbbfejlesztett vagy feltöltött emberi lények

számára (Bostrom, 2008a, 2008b; Chalmers, 2010).
[4] Köszönjük Daniel Dewey-nak, hogy ezt a kifejezést javasolta.

A "haszonmonstrum" kifejezés nem egyértelmű, de leginkább a "szuper-kedvezményezettnek" felelhet meg. Egyes nézetek szerint az erkölcsi státusz beleszámít a számításba.
erkölcsi követelések az érdekek erősségétől eltérő módon, pl. általános multiplikátorként, vagy azáltal, hogy különálló kötelességek vagy deontológiai korlátok halmazát eredményezi. Shelly Kagan (2019) például amellett érvel, hogy egy adott érdek - például egy bizonyos mennyiségű szenvedés elkerüléséhez fűződő érdek - morális súlyát az érdekkel rendelkező alany morális státuszának mértékével kell súlyozni, a státusz mértéke pedig különböző pszichológiai tulajdonságoktól és potenciáloktól függ. Ha egy lénynek olyan érdekei vannak, amelyeknek sokkal nagyobb erkölcsi figyelmet kell szentelni, mint egy emberi lény érdekeinek, nem azért, mert az érdeke erősebb, hanem mert magasabb erkölcsi státusszal rendelkezik, akkor ez a lény a mi terminológiánk szerint szuperpáciens lenne.

A szuperpáciens státusz lehetősége ellentmondásos: egyesek azt állítják, hogy az ember "teljes erkölcsi státusszal" rendelkezik, amelyet nem lehet túllépni, míg mások (például Kagan) szerint a szuperpáciens státusz lehetséges, mivel az emberi erkölcsi státusz megadásához szükséges pszichológiai képességek emberfeletti fokozatokra is képesek. Ebben a tanulmányban elsősorban a szuperpáciens státuszhoz vezető utakat vizsgáljuk, amelyek kombinálódhatnak azzal a kevésbé vitatott feltételezéssel, hogy a digitális elmék legalább az emberivel azonos erkölcsi státusszal rendelkezhetnek, és így szélsőséges erkölcsi állításokat eredményezhetnek.

2.1. Szaporodási képesség

A számítógépes szoftverek egyik legalapvetőbb jellemzője a pontos reprodukció egyszerűsége és gyorsasága, feltéve, hogy a számítógépes hardver rendelkezésre áll. A hardver gyorsan előállítható mindaddig, amíg gazdasági teljesítménye képes fedezni az előállítási költségeket (amelyek történelmileg - ár-teljesítmény alapon - óriási mértékben csökkentek; Nordhaus, 2007). Ez megnyitja a lehetőséget, hogy a népesség dinamikája, amely az emberek körében több évszázadot venne igénybe, az emberi élet töredékébe sűrűsödjön. Még ha *kezdetben* csak néhány, bizonyos szellemi kapacitású digitális elme építése megfizethető is, az ilyen elmék száma hamarosan exponenciálisan vagy szuperexponenciálisan nőhet, amíg más korlátok nem korlátozzák. Az ilyen robbanásszerű szaporodási potenciál lehetővé tenné, hogy a digitális elmék viszonylag rövid időn belül jelentősen meghaladják az emberek számát, és ennek megfelelően növeljék a követeléseik kollektív erejét.

Továbbá, ha a digitális elmék és a szükséges hardverek előállítása addig folytatódik, amíg az így előállított elmék bére el nem éri a határköltséget, ez a béreket a gépi megélhetési szintek felé szoríthatja lefelé, mivel a természeti erőforrások korlátozó tényezővé válnak. Ezek nem lehetnek elegendőek az emberek (és az elavult digitális elmék) túléléséhez (Hanson, 2001; Aghion, Jones és Jones, 2017). Az ilyen körülmények az újraelosztási kérdéseket sürgetőbbé teszik - élet-halál kérdéssé -, miközben a malthusi népességnövekedés a transzferfizetések iránti igényeket gyakorlatilag kielégíthetetlenné tenné.

A gyors és olcsó szaporodás másik fontos szempontja, hogy lehetővé teszi a populáció gyors cseréjét. Egy törölt digitális elme azonnal helyettesíthető a legújabb kiadású, teljes értékű elme másolatával - ellentétben az emberi esettel,

ahol kilenc hónapba telik, mire egy nyálas baba születik.[5] A gazdasági nyomás így az "elavult" elmék nagyon gyakori törlése és olyan elmékkel való helyettesítése felé tolnak, amelyek ugyanazzal a hardverrel több gazdasági értéket termelnek.

A digitális elmékre alkalmazott jelenlegi szoftvergyakorlatok hihető folytatása tehát rendkívül nagy számú rövid életet és halálesetet vonhatna maga után, még akkor is, ha ez csak töredéke a mindenkor létező elmék számának. Az ilyen efemer digitális elmék pszichológiailag érettek, kronológiailag fiatalok lehetnek, hosszú *potenciális* élettartammal, de támogatás hiányában nagyon rövid alapértelmezett élettartammal. Ha úgy gondoljuk, hogy fiatalon meghalni, miközben hosszú életet élhetünk, nagy hiányt jelent, vagy nagyon igazságtalan, amikor mások hosszú életet élhetnek, akkor ez különösen erős igényt támaszthat ezeknek a digitális elméknek az élettartamuk meghosszabbításához szükséges erőforrásokra (vagy a kompenzáció más formáira). Ha a halál önmagában rossz (és nem csupán a lemondott élet alternatív költsége), akkor az elmék ilyen gyors cserélődése szintén növelheti a megélt életévenkénti értékvesztés mértékét.

2.2. Megélhetési költségek

Valószínű, hogy sok digitális elménék kevesebb jövedelemre lesz szüksége ahhoz, hogy adott életszínvonalon fenn tudja tartani magát. A digitális elméket támogató számítógépes hardver költségei valószínűleg jóval az emberi agy és test fenntartásának költségei alá csökkennek. Ha a puszta megélhetésen túl tekintünk, az emberi fogyasztásra alkalmas fizikai javak és szolgáltatások (például a lakhatás és a közlekedés) általában drágábbak, mint az információs technológia és a virtuális javak, amelyekkel egy digitális elme egyenértékű szükségleteit ki lehet elégíteni. A digitális elménék nem kell szenvednie a kedvezőtlen környezeti körülményektől, a környezetszennyezéstől, a betegségektől, a biológiai öregedéstől vagy bármely más, az emberi jólétet csökkentő kényszertől.

Egy adott számú (minőséggel korrigált) életév előállításának költsége egy emberhez hasonló digitális elme esetében ezért valószínűleg jóval alacsonyabb lesz, mint egy biológiai ember esetében. Az életköltségek közötti nagy különbségek azt jelentik, hogy amikor elosztási kérdések merülnek fel, egy olyan erőforrás, amely egy ember számára kis hasznot jelent, sok digitális elme számára nagy hasznot jelenthet. Ha az az energiaköltségvetés, amely egy emberi élet egy hónapig történő fenntartásához szükséges, tíz digitális elmét egy éven át eltarthat, ez erős érv lenne az utóbbiak előnyben részesítése mellett szűkös helyzetben.

2.3. Szubjektív sebesség

A nagyobb soros sebességű hardverek segítségével a digitális elmék gyorsabban futtathatók. A számítógépek jelenlegi órajel-sebességét gigahertzben mérik, ami több milliószor nagyobb, mint az emberi idegsejtek tüzelési sebessége; és a jelátviteli sebesség hasonlóan meghaladhatja az emberi idegek vezetési sebességét. Ezért valószínű, hogy az emberhez hasonló képességekkel rendelkező digitális elmék legalább ezerszer (de talán milliószor) gyorsabban tudnának gondolkodni, mint az emberek, ha elegendő hardver állna rendelkezésre. Ha egy digitális elme szubjektív életévek ezreit pakolja egyetlen naptári évbe, akkor úgy tűnik, hogy az előbbi ("szubjektív idő", nem a falióra-idő) a helyes mérőszám az olyan dolgokra, mint a meghosszabbított életből származó jólét mértéke (Bostrom és Yudkowsky, 2014).

[5] Nem lehet azonban világos, hogy egy létező elme pontos vagy majdnem pontos másolata egy új, különálló személyt vagy inkább annak a személynek egy további példányát jelentené-e, akinek az elméje mintaként szolgált.

Mivel a sebességnöveléshez több hardverre van szükség, ez lehetőséget biztosít az egyes digitális elmék számára, hogy sokkal magasabb (szubjektív életévek dolláronként) hozamot érjenek el a vagyonból, mint amire az emberek általában képesek. Alacsony sebességnél a digitális elmék számára elérhető nyereség közel lineáris lenne; bár ahogy a sebesség megközelíti a technológia határait, a további sebességnövekedés határköltségei emelkednének.[6]

Mivel ezek a gyorsabb futásból származó előnyök az akkor már meglévő, eredetileg lassabban futó egyének, ez a hatás különösen fontos a "személyre ható" megközelítést alkalmazó populációs axiológiák esetében (erről később).

2.4. Hedonikus ferdeség

Okkal gondolhatjuk, hogy a mesterséges elmék sokkal hosszabb ideig és intenzívebben élvezhetik az élvezetet. Az emberi pszichológia úgy fejlődött ki, hogy örömöt és fájdalmat generáljon, ahol ez a reproduktív alkalmassággal kapcsolatos viselkedést motiválta az elmúlt generációkban, nem pedig a jólét maximalizálását. Ez számunkra nagyfokú
nehezen elkerülhető szenvedés. Élményeinket eközben csak takarékosan osztogatjuk. A kulináris élvezeteket az éhség, a szexuális élvezeteket a libidó szabályozza. A mások feletti relatív státuszból vagy hatalomból származó örömök strukturálisan ritkák. A legtöbb jutalmat olyan mechanizmusok is mérséklik, mint az unalom és a tolerancia, amelyek fokozatosan csökkentik az ismétlődő ingerekből vagy folyamatos jótékony körülményekből származó élvezetet. A digitális elmék számára ezek a korlátozások lazíthatók, hogy a jelenlegi emberi lét fájdalmas részei alóli felszabadulás mellett fenntartható intenzív örömöket is lehetővé tegyenek.

Az emberek hedonikus egyensúlya is nagymértékben javítható lenne a fejlett technológiával, amely valószínűleg vagy megelőzné, vagy pedig szorosan követné kiforrott gépi intelligencia technológia.[7] Azonban a hedonikus értékek radikális kiigazítása A biológiai emberek esetében az egyensúly megteremtése több szempontból is "költségesebb" lehet, mint a *de novo* digitális elmék esetében: (a) az agyműtétet, kiterjedt farmakológiai finomhangolást és manipulációt vagy ennek megfelelő beavatkozásokat igénylő beavatkozások - legalábbis közeljövőben - megvalósíthatatlanok vagy drágák lehetnek; és (b) a pszichénk radikálisabb átalakításai a személyiség-identitás vagy más, az emberi személyiséget sértő tényezők elpusztítását kockáztatnák.
jelenlegi emberi természetünk azon tulajdonságai, amelyeket értékelünk.[8] Az érző a gépek tehát nagy előnyökkel rendelkezhetnek a hatékonyság tekintetében, amellyel hedonikusan értékes állapotokat tudnak megvalósítani.

2.5. Hedonikus tartomány

A jelenlegi emberi lények számára elérhető hedonikus skála különböző részein eltöltött idő arányának megváltoztatásán túlmenően lehetséges lenne - inkább spekulatív módon - olyan digitális elméket tervezni, amelyek képesek lennének a hedonikus jóllét "túlzó" állapotainak megvalósítására, olyan szintű boldogságra, amelyet az emberi agyak egyáltalán nem képesek megvalósítani.

[6] Hanson (2016, 63-65. o.) azt állítja, hogy a költségnövekedés a sebességnövekedéssel kezdetben közel

lineáris lenne, azaz a 2x-es sebességnövekedés közel 2x-es hardverköltségvetést igényelne, egészen a lényegében emberfeletti sebességekig.

[7] David Pearce (1995) amellett érvelt, hogy a biológiai elméket úgy lehetne megtervezni, hogy a teljes jelenlegi fájdalom-élvezet skála helyett a "boldogság fokozataival" működjenek.

[8] Vö. (Agar, 164-189. o2010,.).

Evolúciós megfontolások némileg alátámasztják ezt a hipotézist. Amennyiben az örömök és fájdalmak intenzitása megfelel a viselkedési reakciók erősségének, az evolúciónak úgy kell beállítania a hedonikus élményeket, hogy az elérésükre vagy elkerülésükre tett erőfeszítések megközelítőleg alkalmasság-maximalizáló mértékét eredményezzék. Az ember számára azonban általában sokkal könnyebb rövid idő alatt nagy mennyiségű reproduktív fittséget *elveszíteni*, mint ugyanekkora mennyiséget *megszerezni.* Ha néhány pillanatig a tűzben maradunk, az maradandó sérülést vagy halált eredményezhet, a szervezet összes fennmaradó szaporodási lehetőségének árán. Egyetlen étkezés vagy nemi aktus esetében sincs olyan nagy tétje egy másodpercnek - hetekig tart az éhenhalás, és a párzás percenként várhatóan előállított szaporodó gyermekek száma csekély. Így az evolúciónak lehetett olyan felhívása, hogy a sérülésekre válaszul intenzívebben motiváló másodpercenkénti fájdalmakat generáljon, mint a pozitív eseményekre válaszul örömöket. Ezzel szemben a mesterséges elméket úgy lehetne kialakítani, hogy az örömök olyan intenzíven jutalmazóak legyenek, mint amennyire a legrosszabb gyötrelmek visszataszítóak. Az emberi tapasztalatokon teljesen kívül eső boldogság vagy nyomorúság is lehetséges lenne.[9]

2.6. Olcsó preferenciák

A jólét hedonista beszámolói esetében megjegyeztük, hogy lehetőség van arra, hogy a szuper-élvezők azáltal, hogy a digitális elméket úgy tervezték meg, hogy több dolgot találjanak élvezetesnek, vagy hogy emberfeletti intenzitású élvezeteket szerezzenek. A jólét preferencia-kielégítésre irányuló elméletek esetében a lehetőségek párhuzamos párja merül fel: olyan digitális elmék készítése, amelyek preferenciái nagyon könnyen kielégíthetők, vagy olyan digitális elmék készítése, amelyek emberfelettien erős preferenciákkal rendelkeznek. Az utóbbi lehetőség tárgyalását a következő alfejezetre halasztjuk. Itt a könnyen kielégíthető preferenciákkal rendelkező elméket tárgyaljuk.

Az alapeset meglehetősen egyszerű - jobban, mint az élvezetes élményekkel kapcsolatos párhuzamos eset, mivel a preferenciák hozzárendelése nem igényel ellentmondásos feltételezéseket a gépi tudatosságról. Ha a preferenciákat funkcionalista módon értjük, mint olyan absztrakt entitásokat, amelyek részt vesznek az intelligens célvezérelt folyamatok viselkedésének (aspektusainak) kényelmes magyarázataiban (a hiedelmekkel együtt), akkor egyértelmű, hogy a digitális elméknek lehetnek preferenciáik. Sőt, olyan preferenciákkal is rendelkezhetnek, amelyek triviálisan könnyen kielégíthetők: például azzal a preferenciával, hogy legalább tizennégy csillag létezik, vagy hogy egy bizonyos piros gombot legalább egyszer megnyomnak.

Néhány preferencia-kielégítéses beszámoló további követelményeket támaszt azzal kapcsolatban, hogy mely preferenciák számíthatnak bele valakinek a jólétébe. A szadista vagy rosszindulatú preferenciákat például gyakran kizárják. Egyes filozófusok kizárják az "ésszerűtlen" preferenciákat is, például annak a preferenciáját, aki megszállottan meg akarja számolni az összes fűszálat Princeton gyepén.[10] Attól függően, hogy mennyire korlátozzuk azt, hogy mely preferenciák számítanak "ésszerűnek", ezt a korlátot könnyű ~~vagy nem könnyű átlépni.~~

[9] Azt gondolhatnánk, hogy egy olyan hedonikus állapot, amely teljes mértékben leköti az elme figyelmét, és minden más aggodalmat felülír, elvileg a hedonikus intenzitás maximumát jelenti. Ugyanakkor hihetőnek tűnik, hogy egy nagyobb elme, amely "tudatosabb", a vonatkozó értelemben "nagyobb mennyiségű" maximálisan intenzív hedonikus élményt tartalmazhat.
[10] Ahogyan Parfit (1984, 498. o.) is, aki Rawlsra (1971, 432. o.) hivatkozik, aki Stace (1944) példájából merített.

Néhány más típusú követelmény, amely előírható, hogy
A jólléthez hozzájáruló preferenciáknak szubjektíven *támogatottnak* kell lenniük (esetleg azáltal, hogy az elsőrendű preferencia másodrendű preferenciával jár együtt), vagy további pszichológiai vagy viselkedési tulajdonságokban kell *alapulniuk* - például a mosolyra, a stresszérzésre, az öröm megtapasztalására, a visszafogottságra, a figyelem összpontosítására stb. való hajlamban. Ezeket a követelményeket valószínűleg egy digitális elme is teljesíteni tudná.
Az embereknek vannak preferenciáik az érzéki örömök, a szerelem, a tudás, a társas kapcsolatok és a teljesítmény iránt, amelyek kielégítése általánosan úgy tartják, hogy hozzájárulnak a következőkhöz
jólét. Mivel ezek közeli analógjai könnyen megvalósíthatók a virtuális valóságban, bármilyen szükséges pszichológiai vagy viselkedési tulajdonsággal és másodrendű jóváhagyással együtt, ezek a követelmények nem valószínű, hogy megakadályoznák olyan lények létrehozását, akiknek erős, de mégis minősített preferenciáik vannak, és amelyek nagyon könnyen kielégíthetők.

2.7. Preferencia erőssége

Míg a rendkívül könnyen kielégíthető preferenciák létrehozása koncepcionálisan egyszerű, az emberfeletti "erővel" rendelkező preferenciák létrehozása már problémásabb. A standard von Neumann-Morgenstern konstrukcióban a hasznossági függvények csak affin transzformációkig egyediek: a hasznossági függvényhez való hozzáadás vagy annak konstanssal való szorzása nem befolyásolja a választásokat, és egy preferencia erőssége csak ugyanazon ágens más preferenciáihoz viszonyítva határozható meg. Így az interperszonális összehasonlítások elvégzéséhez további struktúrát kell biztosítani a különböző hasznossági függvények normalizálásához és közös skálára hozásához.[11]

Vannak különböző megközelítések, amelyek megpróbálnak "egyenlő beleszólást" adni a különböző ágensek preferenciáinak, kizárólag a preferencia-struktúra alapján, kiegyenlítve a különböző ágensek várható befolyását, és többnyire kizárva a preferencia-erősséget.
szuper-kedvezményezettek.[12] Az ilyen megközelítések azonban kihagynak néhány fontos megfontolások. Először is, nem veszik figyelembe a pszichológiai komplexitást vagy a kompetenciákat: néhány minimális rendszer, például egy digitális termosztát, ugyanolyan súlyt kaphat, mint a pszichológiailag összetett elmék. Másodszor, tagadják az érzelmi csillogás vagy más olyan jellemzők szerepét, amelyeket intuitív módon használunk a vágyak erősségének értékelésére önmagunkban és más emberekben. Harmadszor pedig az így kapott társadalmi jóléti függvény nem biztosíthatja az együttműködés kölcsönösen elfogadható alapját az érdektelen felek számára, mivel az erős alternatívákkal rendelkező erős ágenseknek ugyanolyan súlyt ad, mint a hatalommal és alternatívákkal nem rendelkező ágenseknek.

Az első két kérdés megkövetelheti ezeknek a pszichológiai erősség-hangsúlyozó jellemzőknek a vizsgálatát. A harmadik kérdéssel egy olyan szerződéses álláspontot lehetne kezelni, amely játékelméleti megfontolások alapján osztja ki a súlyokat, és
(hipotetikus) alku. A kontraktuális megközelítést nem uralnák az alkupozíciójukhoz képest aránytalanul nagy kedvezményezettek, de veszélyesen közelít a "hatalom igazat ad" elvhez, és nem nyújt iránymutatást azoknak a szerződő feleknek, akik törődnek a kiszolgáltatottakkal, és a kedvezményezett alkupozíciójától függetlenül kívánják elosztani a támogatást.

[11] Harsányi (1953) megmutatta, hogy a hasznossági függvények súlyozott összege bizonyos feltételezések mellett optimális, de a tétel a súlyok értékét meghatározatlanul hagyja.
[12] Pl. (MacAskill, Cotton-Barratt és Ord, 2020)

2.8. Célkitűzés lista áruk és virágzás

A jólét objektív listás elméletei azt állítják, hogy az, hogy mennyire jól megy valakinek az élete, attól függ, hogy az élete milyen mértékben tartalmazza a különböző javak különböző fajtáit (*többek* között az örömöt és a preferenciák kielégítését). Néhány gyakran megjelenő tétel a tudás, a teljesítmény, a barátság, az erkölcsi erények és az esztétikai értékek, bár a különböző javak azonosítása és súlyozása igen változatos. Ezekben az elméletekben az a közös, hogy olyan tételeket tartalmaznak, amelyek hozzájárulása a jóléthez nem teljes mértékben az alany attitűdjei, érzései és meggyőződései által meghatározott, hanem megkövetelik a siker valamilyen külső mércéjének teljesítését is.

Az objektív listákon található számos elem nyitott a szélsőséges instanciák számára. Például a szuperintelligens gépek az emberi mértéket meghaladó intellektuális erényeket tudnának művelni. Az erkölcsi erények is elérhetik az emberfeletti szintet: egy digitális elme kiterjedt erkölcsi tudással és tökéletes motivációval kezdheti az életét, hogy mindig azt tegye, ami erkölcsileg helyes, így kifogástalanul bűntelen marad, míg minden felnőtt ember a végén a szabálysértések bűnlajstromával zárja az életét.

A barátság összetett jószág, de talán le lehetne egyszerűsíteni olyan alapvető alkotóelemeire, mint a lojalitás, egymás személyiségének és érdeklődési körének kölcsönös megértése, valamint a múltbeli interakciók története. Ezeket az alkotóelemeket aztán maximálisan hatékony formában lehetne újra összerakni, így a digitális elmék talán nagyobb számú mélyebb barátságot tudnának fenntartani sokkal hosszabb ideig, mint az emberek esetében lehetséges.

Vagy gondoljunk a teljesítményre. Hurka és Tasioulas (2006) teljesítményről szóló beszámolója szerint a teljesítmény értéke azt tükrözi, hogy milyen mértékben a gyakorlati ész gyakorlásának eredménye: a legjobb eredmények azok, ahol a kihívást jelentő célokat hierarchikus terveken keresztül érik el, amelyek egyre bonyolultabb résztervekre tagolódnak. Ilyenkor könnyen elképzelhetünk digitális "szuperteljesítményeket", amelyek fáradhatatlanul folytatják az egyre bonyolultabb projekteket anélkül, hogy a motiváció lanyhulása vagy a figyelem elkalandozása korlátozná őket.

Ilyen és sok más módon a digitális elmék sokkal nagyobb mértékben valósíthatnák meg a különböző objektív javakat, mint ahogyan az nekünk, embereknek lehetséges.

A jólét egy másik felfogása szerint a jólét a "virágzásból" áll, amelyet a ránk jellemző képességek gyakorlásában vagy a "telosz" elérésében lehet kifejezni. Az arisztotelészi felfogás szerint például egy lény olyan mértékben virágzik, amilyen mértékben sikerül megvalósítania teloszát vagy lényegi természetét. Úgy tűnik, hogy ez a fajta virágzás elérhető lenne egy digitális elme számára, amely minden bizonnyal képes lenne a rá jellemző képességek gyakorlására, és amelynek szintén tulajdoníthatunk teloszt, bármilyen értelemben is van az emberi lényeknek telosza - akár a teremtő szándékai által meghatározott telosz, akár az evolúciós vagy más dinamikából eredő telosz, amely létrehozta és formálta a természetét. Tehát lehetségesnek kell lennie, hogy legalább egyenlő, és valószínűleg némileg túlmutat az embereken az ilyen virágzás elérése szempontjából; bár az, hogy az emberen radikálisan túlmutató virágzást hogyan értenénk az ilyen típusú számítások alapján, kevésbé világos.

2.9. Elme skála

Absztrakt szinten a lehetséges elme-skálák széles skáláját vehetjük figyelembe, az apró rovarszerű (vagy akár termosztát-szerű) elméktől egészen a hatalmas szuperintelligens elmékig, amelyek számítási teljesítménye nagyobb, mint a mai emberi populációé. Az építés költségei e skálán felfelé haladva nőnek, ahogyan az erkölcsi jelentőség is. Fontos kérdés, hogy e két változó növekedési üteme milyen arányban növekszik egymáshoz képest.

Vegyük először azt a hipotézist, hogy a jólét lassabban nő, mint a költségek. Ez azt sugallná, hogy a legnagyobb teljes jólétet akkor érnénk el, ha rengeteg apró elmét építenénk. Ha ez igaz lenne, akkor a rovarpopulációk már most is túlnyomóan meghaladhatják az emberi populációt a jólétre való összesített kapacitás tekintetében; és a minimálisan képzett digitális elmék hatalmas populációi megelőznék mind a rovarokat, mind az emberi vagy emberfeletti léptékű lényeket.

Vegyük ehelyett azt a hipotézist, hogy a jólét gyorsabban nő, mint a költségek. Ez az ellenkező következtetést sugallná: hogy a legnagyobb teljes jólétet az erőforrások néhány óriási elmében való összpontosításával érhetnénk el.

Úgy tűnik, hogy az az eset, amikor az emberi elme léptékű elmék optimálisak, egy nagyon speciális esetet képvisel, ahol valamilyen kritikus küszöbérték létezik a mi szintünk közelében, vagy ahol a skálázási kapcsolatnak van egy csomópontja éppen az emberi léptékű pont körül. Egy ilyen egybeesés elfogulatlan szemszögből kissé valószínűtlennek tűnhet, bár sokkal természetesebben merülhet fel olyan beszámolókban, amelyek a jólét fogalmát az emberi tapasztalatban vagy az emberi természetben rögzítik.

Konkrétabban megkérdezhetjük az egyes tulajdonságok tekintetében, hogy emberi szinten hihető-e egy csomópont vagy küszöbérték. Például feltehetjük ezt a kérdést az agy által instantiált tudatosság mértékére vonatkozóan. Legalábbis nem nyilvánvaló, miért lenne az, hogy az erőforrások tudatossággá alakításának maximálisan hatékony módja az emberi méretű elmék konstruálása lenne, bár ennek a kérdésnek a további vizsgálatához meg kellene vizsgálni a tudatosság konkrét elméleteit.[13] Hasonlóképpen, az erkölcsi státusz tekintetében is feltehetjük a kérdést, hogy az hogyan változik az elme méretétől függően.
Ismétlem, az az állítás, hogy az emberi méretű elmék optimálisak ebből a szempontból, további indoklás hiányában kissé gyanúsnak tűnhet.

Még ha az emberi agy mérete optimális *is lenne* a tudatosság vagy az erkölcsi státusz létrehozásához, ebből még mindig nem következne, hogy az emberi agy *szerkezete* is az. Agyunk nagy részei irrelevánsnak vagy csak gyengén relevánsnak tűnnek a tudatosság vagy az erkölcsi státuszunk mértéke szempontjából. Például az agykérgi szövetek nagy része a következők feldolgozását szolgálja
nagy felbontású vizuális információ; mégis, úgy tűnik, hogy a homályosan látó emberek, sőt a teljesen vakok is képesek ugyanolyan tudatosak és ugyanolyan magas erkölcsi státuszúak lenni, mint a sasszemű látásélességgel rendelkezők.

Ezért nagyon is hihetőnek tűnik, hogy a szuper-kedvezményezett státusz lehetséges különböző méretű elmékkel, egyrészt azért, mert az erőforrások és az érték közötti skálázási kapcsolat valószínűleg nem az emberi elme méreténél éri el a csúcspontját, másrészt pedig azért, mert a szuper-kedvezményezett státusz nem az emberi elme méreténél éri el a csúcspontját.

[13] Ez a kérdés különösen akut, mivel a tudatosság számos olyan elmélete, amely eléggé specifikált ahhoz, hogy számítási megvalósításokat vegyen figyelembe, rendkívül minimális megvalósításokra is érzékenynek tűnik (Herzog, Esfeld és Gerstner, 2007).

mert az emberi elme jelentős részei kevéssé relevánsak a tudatosság foka, az erkölcsi státusz vagy más olyan tulajdonságok szempontjából, amelyek a legközvetlenebbül kapcsolódnak a jólét vagy az erkölcsi státusszal súlyozott jólét mennyiségéhez.

3. A digitális szuper-kedvezményezettek erkölcsi és politikai következményei

Foglaljuk össze azokat a dimenziókat, amelyek mentén a digitális elmék emberfeletti erőforrás-hatékonysággal érhetnék el a jólétet:

NÉHÁNY ÚT AZ EMBERFELETTI JÓLÉTHEZ
● szaporodási képesség ● megélhetési költségek ● szubjektív sebesség ● hedonikus ferdeség ● hedonikus tartomány ● olcsó preferenciák ● preferencia erőssége ● céllista áruk és virágzás ● elme skála

E dimenziók némelyike csak a jólét bizonyos beszámolói szempontjából releváns. A szélsőséges preferencia-erősség lehetősége például közvetlenül releváns a preferencia-alapú elszámolásokhoz, de nem a hedonista elszámolásokhoz. Mások, mint például a megélhetési költségek, általánosabban relevánsak, és úgy tűnik, hogy szinte minden olyan nézetre vonatkoznak, amely a digitális elméknek erkölcsi státuszt biztosít, és amely figyelembe veszi a költségeket, amikor szűkös körülmények között döntéseket hoz. A dimenziók némileg eltérnek a jólét növekedésének nagyságrendje tekintetében is, amelyet lehetővé tehetnek, valamint abban, hogy milyen könnyen és olcsón lehet ilyen szélsőséges értékeket elérni. Összességükben azonban elég szilárd érvet képviselnek amellett, hogy a szuperjólét valóban megvalósíthatóvá válna a technológiai érettség elérésekor. Más szóval, a jólét számos népszerű elmélete szerint a jólétre vonatkozó elméletek széles skálája szerint az erőforrások egységnyi erőforrására vetítve sokkal nagyobb jólétet lehet majd elérni, ha ezeket az erőforrásokat biológiai emberek helyett digitális elmékbe fektetjük.

Ezért két fontos kérdés merül fel (amelyeket külön-külön is feltehetünk a különböző erkölcsi elméleteknek):

- Hogyan kell tekintenünk arra, hogy a jövőben képesek leszünk szuperkedvezményezetteket létrehozni?
- Hogyan kellene reagálnunk, ha egy *kész tény* elé állítanának bennünket, amelyben szuper-kedvezményezettek, talán nagy számban, létrejöttek?

3.1. Szuper-kedvezményezettek létrehozása

Számos olyan nézet, amely a jó új életek létrehozását fontos értéknek tekinti, rendkívül vonzónak tartaná azt a kilátást, hogy a jövőt szuper-élvezőkkel népesítjük be, és ha nem élnénk ezzel a lehetőséggel, az drasztikusan csökkentené a jövő értékét - ez egzisztenciális katasztrófa lenne (Bostrom, 2013).

Másfelől azzal is érvelhetünk, hogy okunk van arra, hogy *ne* hozzunk létre szuperhasznú lényeket, pontosan azon az alapon, hogy ha egyszer léteznek ilyen lények, akkor azok
uralkodó igényt tartanak a szűkös erőforrásokra, ezért kénytelenek lennénk (potenciálisan minden) erőforrást átcsoportosítani az emberektől ezeknek a szuperhasznosítóknak, az emberiség kárára. Nicholas Agar (2010) egy ilyen irányú érvelést mutatott be, amely (legalábbis az emberekhez viszonyítva) erkölcsi okot ad arra, hogy ellenezzük a "poszthumánok" létrehozását, akik valamilyen módon nagyobb erkölcsi státusszal, hatalommal és jóléti potenciállal rendelkeznek.

A szuper-kedvezményezettek létrehozásának erkölcsi kívánatossága ilyen tagadásának igazolására a Narveson (1973) szlogenjével összhangban lévő "személyre ható" elvre hivatkozhatunk,
"Az erkölcs arról szól, hogy az embereket boldoggá tegyük, nem pedig arról, hogy boldog embereket csináljunk."[14] Ha a kötelességeink
csak a már létező embereknek, és nincs erkölcsi okunk további új embereket létrehozni, akkor különösen nem lenne kötelességünk szuper-kedvezményezetteket létrehozni; és ha az ilyen szuper-kedvezményezettek létrehozása ártana a már létező embereknek, akkor kötelességünk lenne nem létrehozni őket. Feltehetően akkor *sem* lenne kötelességünk elkerülni a szuper-kedvezményezettek létrehozását, ha az emberek, akiket ez károsítana, valamelyik jövőbeli generációhoz tartoznának, úgy, hogy a döntésünk "pillangóhatása" megváltoztatná, hogy mely emberek jönnek létre; de legalábbis ilyen szempontból nem lenne pozitív kötelességünk szuper-kedvezményezetteket létrehozni.

A szigorúan személyre ható megközelítésnek azonban van néhány meglehetősen ellentmondásos következménye. Ez például azt jelentené, hogy nincs erkölcsi okunk arra, hogy most bármilyen intézkedést tegyünk annak érdekében, hogy enyhítsük az éghajlatváltozásnak a jövő nemzedékekre gyakorolt hatását; és ha az intézkedések a jelenlegi nemzedékre nézve költségekkel járnának, akkor erkölcsi okunk lehet arra, hogy *ne* tegyük meg azokat. Mivel ez ilyen következményekkel jár, a legtöbben elutasítanák a szigorú személyre ható etikát. A gyengébb vagy minősített változatoknak szélesebb körű vonzereje lehet. Az egyik például adhatna *némi* többletsúlyt, de nem szigorú dominanciát a már létező emberek javára.

Hasonló eredmény, ahol van *valamilyen* erkölcsi okunk arra, hogy szuper-kedvezményezetteket hozzunk létre, még akkor is, ha a meglévő emberek különleges elbánásban részesülnek, a népesedési etikával kapcsolatos erkölcsi bizonytalanság figyelembevételéből adódhat (Greaves és Ord, 2017).
Attól függően, hogy hogyan kezeljük ezt a bizonytalanságot, *vagy* arra a következtetésre juthatunk, hogy a "legmegfelelőbb választás" az, ha minden erőforrást a szuper-kedvezményezettek létrehozására fordítunk, még akkor is, ha úgy gondoljuk, hogy nem valószínű, hogy ez lenne az erőforrások legjobb felhasználása; *vagy* (ami szerintünk hihetőbb), hogy a legmegfelelőbb választás az, ha legalább néhány erőforrást félreteszünk a meglévő emberek javára, még akkor is, ha valószínűnek tartjuk, hogy valójában jobb lenne minden erőforrást a szuper-kedvezményezettek létrehozására fordítani.

Egy másik megközelítést az aszimmetrikus személyre ható nézetek képviselnek, amelyek lehetővé teszik a nettó *rossz* életek - olyan életek, amelyeket nem érdemes

élni - létének okozásával kapcsolatos erkölcsi aggodalmat (Frick, 2014). Az ilyen nézetek szerint erős okunk van arra, hogy elkerüljük a hatalmas negatív jóléttel járó digitális elmék létrehozását, és hogy hajlandónak kell lennünk arra, hogy az ilyen kimenetel elkerülése érdekében nagy költségeket vállaljunk a meglévő emberi populációra nézve. Az aszimmetrikus nézetek más változatai, miközben tagadják, hogy erkölcsi

[14] Frick (2014) a szlogenhez igazodva tesz egy újabb kísérletet.

okok, hogy a jövőt új lényekkel töltsük meg, hogy minél több pozitív hasznosságot tapasztaljunk, fenntartják, hogy mindazonáltal erkölcsi kötelességünk biztosítani, hogy a jövő nettó hasznossága a nulla vonal fölött legyen. Az ilyen nézetek következésképpen nagy jelentőséget tulajdoníthatnak annak, hogy elegendő pozitív szuperhasznú lényt hozzunk létre, hogy "ellensúlyozzuk" a jövőbeli lények disutilitását (Thomas, 2019).

3.2. Megosztani a világot a szuperkedvezményezettekkel

Ha azt az esetet tekintjük, amikor a szuper-kedvezményezettek már beléptek a létezésbe, a személyre ható elvekből eredő bonyodalmak elmaradnak. Egy egyszerű utilitarista szemszögből nézve, feltételezve a tökéletes megfelelést, a végeredmény ekkor egyértelmű: minden erőforrást át kell utalnunk a szuperhasznosoknak, és hagynunk kell, hogy az emberiség elpusztuljon, ha már nem vagyunk instrumentálisan hasznosak.

Természetesen számos olyan etikai nézet létezik, amely tagadja, hogy kötelességünk lenne minden saját (nem is beszélve más emberek) erőforrásainkat átadni annak a lénynek, amelyik a legnagyobb jólétben részesül. A deontológiai elméletek például gyakran tartják az ilyen cselekedeteket szupererogatívnak a saját javaink elajándékozása esetén, és megengedhetetlennek mások javainak újraosztása esetén.

Mindazonáltal az olyan széles körben elfogadott elvek, mint a megkülönböztetésmentes transzferek, a politikai egyenlőség és a reproduktív szabadság már elegendőek lehetnek ahhoz, hogy komoly kompromisszumokat jelentsenek. Gondoljunk csak az adókból finanszírozott, általános alapjövedelemre vonatkozó általános javaslatra, amely a fejlett mesterséges intelligencia okozta emberi munkanélküliséget hivatott ellensúlyozni. Ha a digitális elmék gyorsan szaporodó populációi legalább olyan erős igényt támasztanak az alapjövedelemre, mint a biológiai emberek, akkor az adókapacitás gyorsan kimerülhet. Egy egyenlő juttatásnak az emberi létminimum alá kellene csökkennie (a digitális elmék létminimumának szintje felé), míg egy egyenlőtlen juttatás, ahol a jövedelmet egyenlő juttatások alapján osztják el, a kifizetéseket az alacsony megélhetési költségekkel rendelkező digitális elmékhez irányítaná - egy digitális elmének egy év életet adna, egy embernek pedig egy napot.

Úgy tűnik, hogy ennek a kimenetelnek az elkerülése az egyenlőtlen bánásmód valamilyen kombinációját igényli, amelyben a kiváltságos embereket előnyben részesítik a legalább azonos erkölcsi státuszú és nagyobb szükségletekkel rendelkező digitális elmékkel szemben, valamint a digitális elmék reprodukciós lehetőségeinek korlátozását - olyan korlátozásokat, amelyek, ha az emberekre alkalmaznák, a reprodukciós szabadság elveit sértenék.

Hasonlóképpen, politikai szinten a demokratikus elvek feljogosítanák a népesség óriási többségét alkotó digitális elméket a politikai ellenőrzésre, beleértve a transzferek és a tulajdonjogok rendszerének ellenőrzését is.[15]

Lehetne itt arra az útra lépni, hogy megpróbáljuk megvédeni az emberek különleges kiváltságát. Egyes kontraktuális elméletek például azt sugallhatják, hogy ha az emberek a digitális elmékhez képest nagy hatalmi helyzetben vannak, akkor ez feljogosít bennünket az erőforrásokból való megfelelő nagy részesedésre. Alternatív megoldásként elfogadhatnánk valamilyen olyan ágens-relatív okokról szóló

beszámolót, amelyek alapján a közösségek vagy fajok jogosultak arra, hogy saját tagjaikat kiváltságban részesítsék az objektíve ugyanolyan nagy sivataggal rendelkező kívülállókkal szemben.

[15] Vö. (Calo, 2015).

erkölcsi állapot.[16] Úgy tűnik, hogy ez a viszonylagosság tükrözi a de facto megközelítést, amelyet a
a mai államok, amelyek általában nagylelkűbbek a jóléti rendelkezésekkel saját állampolgáraikkal szemben, mint a külföldiekkel szemben, még akkor is, ha vannak olyan külföldiek, akik szegényebbek, akiknek több hasznot tudnának nyújtani, és akik eredendő tulajdonságaik alapján legalább annyira érdemesek a támogatásra, mint az ország saját állampolgárai.

Mielőtt azonban erre az útra lépnénk, alaposan és kritikusan el kellene gondolkodnunk az egykor széles körben elfogadott, de azóta hiteltelenné vált hasonló álláspontok történelmi múltján, amelyeket számos emberi csoport elnyomásának és a nem emberi állatokkal szembeni bántalmazásnak az igazolására használtak fel. Fel kellene tennünk például a kérdést, hogy a digitális elmék és az emberek közötti megkülönböztetés támogatása nem lenne-e olyan, mintha a faji felsőbbrendűség valamilyen doktrínáját támogatnánk?

Itt azt kell szem előtt tartani, hogy a digitális elméknek sokféle változata létezik. Némelyikük jobban különbözik egymástól, mint az emberi elme egy macskától. Ha egy digitális elme egészen másképp épül fel, mint az emberi elme, nem lenne meglepő, ha a vele szembeni erkölcsi kötelességeink eltérnének a többi emberi lénynek való kötelezettségeinktől; és így a másképp való kezelés nem kell, hogy kifogásolhatóan diszkriminatív legyen. Természetesen ez a pont nem vonatkozik azokra a digitális elmékre, amelyek nagyon hasonlítanak a biológiai emberi elmékhez (pl. teljes agyi emulációk). Nem igazolja az olyan digitális elmék negatív diszkriminációját sem, amelyek olyan módon különböznek az emberi elméktől, amely *nagyobb* erkölcsi státuszt biztosít számukra (szuperpáciensek), vagy amely szükségleteiket erkölcsileg súlyosabbá teszi az emberi szükségleteknél (szuper-kedvezményezettek). Ami azt illeti, azt sem igazolná, hogy a nem emberi lényekhez hasonló képességekkel vagy érzékenységgel rendelkező digitális elméket az állatokkal való jelenlegi interakcióink mintája szerint kezeljük, mivel ez utóbbiakkal szemben igen széles körben elterjedt és szörnyű visszaélések tapasztalhatók.

Az egyik módja annak, hogy megpróbáljuk igazolni az emberi lényekkel való kivételezett bánásmódot anélkül, hogy a saját fajtánkkal szemben nyers rasszizmushoz hasonló előítéletet tételeznénk fel, az lenne, ha hivatkoznánk valamilyen elvre, amely szerint jogosultak (vagy kötelesek) vagyunk nagyobb figyelmet fordítani azokra a lényekre, amelyek szorosabban integrálódtak a közösségünkbe és társadalmi életünkbe, mint a távoli idegenekre. Valamilyen ilyen elvre vélhetően szükség van, ha legitimálni akarjuk azt a (nem kozmopolita) módot, ahogyan a legtöbb ember és a legtöbb állam
jelenleg a legtöbb támogatást saját csoportjaikra korlátozzák.[17] Egy ilyen lépés azonban nem zárná ki
digitális elmék, akik társadalmi szövetünk részévé váltak, például adminisztrátori, tanácsadói, gyári munkás vagy személyi asszisztensi szerepkörökben. Lehet, hogy társadalmilag szorosabban kötődünk az ilyen mesterséges intelligenciákhoz, mint a világ másik felén élő idegen emberekhez.

4. Megbeszélés

Láttuk, hogy a digitális szuperkedvezményezettekhez sokféle út vezet, így lehetőségük még erőteljesebbé válik. Ez a jólétről szóló, jelenleg népszerű

beszámolók többségének egyik következménye.

Ez azt jelenti, hogy hosszú távon a teljes jólét sokkal nagyobb lenne, ha a világot digitális szuper-élvezők népesítenék be, nem pedig az élet, mint a digitális szuper-élvezők.

[16] Pl. (Williams, 2006)

[17] Ezek a gyakorlatok természetesen kozmopolita kritikának vannak kitéve; pl. (Singer, 1981; Appiah, 2006).

tudjuk. És amennyiben ilyen lények jönnek létre, az ő aggodalmaik erkölcsileg túlsúlyba kerülhetnek az emberi és állati aggodalmakkal való konfliktusban, például a szűkös természeti erőforrásokkal kapcsolatban.

Miközben azonban az a maximalista szemlélet, amely vagy a meglévő emberiség jólétére, vagy pedig az új digitális elmék jólétére összpontosít, a másik oldalra nézve szörnyű következményekkel járhat, lehetséges, hogy a kompromisszumos politikák mindkét mérce szerint rendkívül jól teljesítenek. Vegyünk három lehetséges politikát:

(A) az erőforrások 100%-a az embereknek
(B) A források 100%-a a szuper-kedvezményezettek számára
(C) Az erőforrások 99,99%-a a szuper-kedvezményezetteknek; 0,01%-a az embereknek.

Teljes hasznossági szempontból (C) megközelítőleg 99,99%-ban ugyanolyan jó, mint a legkedvezőbb (B) lehetőség. A hétköznapi ember szemszögéből nézve a (C) szintén 90+%-ban olyan kívánatos lehet, mint a leginkább preferált (A) lehetőség, tekintettel a digitális elmék által lehetővé tett csillagászati gazdagságra, amely nagyságrendekkel nagyobb, mint a jelenlegi összérték (Bostrom, 2003; Hanson, 2001). Így *ex ante* vonzónak tűnik az (A) és a (B) valószínűségének csökkentése a (C) nagyobb valószínűségéért cserébe - akár az erkölcsi hiba elleni védekezés, akár az erkölcsi pluralizmus megfelelő tükrözése, akár játékelméleti megfontolások figyelembevétele, akár egyszerűen *reálpolitikai* megfontolásból. Hasonlóképpen, mivel az emberiség boldogulhat anélkül is, hogy emberfeletti *rossz* életeket produkálna, és mivel az ilyen nyomorúság elkerülése nemcsak a totális utilitarizmus szempontjából, hanem számos más értékelő nézet szerint is rendkívül fontos szempont, az olyan intézkedések, amelyek csökkentik az értéktelenség ultrahatékony termelésének lehetőségét (még ha ez némi emberi árat is jelent), fontos részét képeznék a konszenzusos politikának.

A nagyobb kihívás nem egy olyan lehetséges jövő leírása, amelyben az emberiség és a digitális elmék populációja egyaránt nagyon jól teljesít, hanem egy olyan megállapodás elérése, amely stabilan elkerüli, hogy az egyik fél *utólagosan* eltiporja a másikat, ahogyan azt a következő szakaszban tárgyaltuk. 3.2.

Ez a kihívás gyakorlati és erkölcsi szempontokat is magában foglal. Gyakorlati szempontból az a probléma, hogy olyan intézményi vagy egyéb eszközöket találjunk ki, amelyekkel az emberek és állatok érdekeit védő politika a végtelenségig fenntartható, még akkor is, ha a kedvezményezettek számbeli fölényben vannak, és a nagyszámú, sokféle, nagy képességű intelligens gépekkel szemben túlerőben vannak. A probléma egyik megközelítése lehet egy szupertöbbség létrehozása a magas jóléti digitális elmék, akik motiváltak ennek az eredménynek a megőrzésére és a vonatkozó normák és intézmények fenntartására (beleértve a digitális elmék egymást követő generációinak kialakítását is).

Erkölcsileg az a kérdés, hogy az *előzetes* vonzó kompromisszum által javasolt intézkedések megengedhetőek-e az *utólagos* végrehajtás során. Az egyik hasznos teszt itt az, hogy analóg körülmények között helyeselnénk-e az alkalmazásukat a nem digitális elmékre. Megkövetelhetnénk például, hogy minden javasolt intézkedés feleljen meg a megkülönböztetésmentesség valamilyen elvének, például a következőnek (Bostrom és Yudkowsky, 2014):

Az aljzat megkülönböztetésmentességének elve

> Ha két lény azonos funkcionalitással és azonos tudatos tapasztalattal rendelkezik, és csak a megvalósítás szubsztrátjában különbözik, akkor azonos erkölcsi státusszal rendelkeznek.

és

> *Az ontogenezis megkülönböztetésmentesség elve*
> Ha két lény azonos funkcionalitással és azonos tudatos tapasztalattal rendelkezik, és csak abban különbözik egymástól, hogy hogyan jöttek létre, akkor azonos erkölcsi státusszal rendelkeznek.

Ezen elvek alkalmazásakor fontos emlékeztetni arra a korábbi pontra, hogy a gépi elme nagyon is különbözhet az emberi elmétől, többek között olyan szempontból is, ami befolyásolja, hogyan kell bánni vele. Még ha el is fogadjuk a megkülönböztetésmentesség olyan elveit, mint amilyenek az előbbiek, ezért óvatosnak kell lennünk, amikor azokat olyan digitális elmékre alkalmazzuk, amelyek nem pontos másolatai valamely emberi elmének.

Vegyük például a reprodukciót. Ha az emberi lények képesek lennének arra, hogy egy biokémiai reaktorba öntött kerti törmelékkel néhány percenként gyermeket szüljenek, valószínűnek tűnik, hogy az emberi társadalmak megváltoztatnák a jelenlegi jogi gyakorlatot, és korlátoznák a szaporodás ütemét. Ennek elmulasztása rövid időn belül csődbe juttatna minden szociális jóléti rendszert, feltételezve, hogy vannak legalább néhányan, akik egyébként hatalmas mennyiségű gyermeket hoznának létre ilyen módon, annak ellenére, hogy nincsenek meg az anyagi lehetőségeik, hogy eltartsák őket. Az ilyen szabályozás különböző formákat ölthetne - a leendő szülőket kötelezhetnék arra, hogy az utódok szükségleteinek kielégítésére megfelelő biztosítékot helyezzenek letétbe, mielőtt létrehoznák őket, vagy a reprodukciós engedélyeket kvóta alapján osztanák ki. Hasonlóképpen, ha az emberek képesek lennének arra, hogy tetszőleges számú pontos másolatot hozzanak létre magukról, akkor elvárható lenne, hogy alkotmányos kiigazítások történjenek annak megakadályozására, hogy a politikai versenyeket az alapján döntsék el, hogy ki akar és ki képes megengedni magának, hogy a legtöbb szavazóklónt hozza létre. A kiigazítások különböző formákat ölthetnek - például az ilyen másolatok létrehozójának meg kellene osztania saját szavazati jogát az általa létrehozott másolatokkal.

Következésképpen, amennyiben az ilyen jogi vagy alkotmányos kiigazítások elfogadhatóak lennének az emberek számára, ha rendelkeznénk ilyen reprodukciós képességekkel, ugyanígy elfogadható lehet, hogy hasonló kiigazításokat tegyünk az ilyen képességekkel rendelkező digitális elmék számára.

A kulcskérdés - természetesen a létező élet szempontjából - az, hogy erkölcsileg megengedhető lenne-e új elméket létrehozni úgy, hogy azok megbízhatóan támogassák bizonyos jogok és kiváltságok fenntartását az emberekkel szemben. Korábban felvetettük, hogy az emberi tulajdonjogok és társadalmi kiváltságok megőrzésének ilyen elrendezése védhető lenne, legalábbis mint a bölcs gyakorlati kompromisszumok bizonytalanságot tiszteletben tartó és konfliktuscsökkentő útja, függetlenül attól, hogy az alapvető erkölcsi elmélet szintjén optimális-e vagy sem. Analógiaként utalhatnánk arra az általános nézetre, hogy erkölcsileg elfogadható a drága fenntartási költségekkel és szükségletekkel rendelkező kisebbségek, például az idősek, a fogyatékkal élők, a fehér orrszarvúak és a brit királyi család megőrzése és védelme. Ezt a következtetést

még inkább alátámasztaná, ha azt feltételeznénk, hogy a digitális

a létrehozott elmék maguk is támogatnák a megállapodást és annak folytatását.

Még ha maga az eredmény erkölcsileg megengedhető is lenne, egy további etikai kérdéssel kell szembenéznünk, nevezetesen azzal, hogy van-e valami *eljárási szempontból* kifogásolható abban, hogy az általunk létrehozott új digitális elmék preferenciáit precíziósan megtervezzük, hogy biztosítsuk beleegyezésüket. Ezt a kérdést a megkülönböztetés tilalmának elvein keresztül vizsgálhatjuk meg, és mérlegelhetjük, hogy miként viszonyulnánk az emberi gyermekek preferenciáinak hasonló módon történő alakítására irányuló javaslatokhoz.

Míg az emberi kultúrák rutinszerűen próbálnak neveléssel, párbeszéddel és figyelmeztetéssel normákat és értékeket átadni a gyermekeknek - beleértve a gyermeki jámborságot és a meglévő normák és intézmények tiszteletét -, addig az a javaslat, hogy *géntechnológiával módosított* ivarsejtek segítségével különleges hajlamokat ültessenek beléjük, valószínűleg ellentmondásosabb lenne. Még ha félretesszük is a biztonsággal, az egyenlőtlen hozzáféréssel, az elnyomó kormányok általi visszaélésekkel vagy a szülők szűklátókörű vagy más módon ostoba döntéseivel kapcsolatos gyakorlati aggályokat, akkor is aggályos lehet, hogy maga az utódok hajlamai feletti részletes ellenőrzés gyakorlása, különösen ha az "mérnöki gondolkodásmóddal" történik, és olyan módszerekkel, amelyek teljesen megkerülik az ellenőrzött alany saját elméjét és akaratát (mivel az alany születése előtt történik), erkölcsileg eleve problematikus lenne.[18]

Bár itt nem tudjuk teljes mértékben értékelni ezeket az aggályokat, két fontos különbséget jegyzünk meg a digitális elmék esetében. Az első az, hogy az emberi reprodukcióval ellentétben itt nincs olyan nyilvánvaló "alapértelmezett" helyzet, amelyhez az alkotók igazodhatnának. A programozók *elkerülhetetlenül* döntéseket hozhatnak, amikor egy gépi intelligenciát építenek - hogy így vagy úgy építsék-e fel, hogy erre vagy arra a célra képezzék-e ki, hogy ezt vagy azt a célt tűzzék-e ki célul, hogy az egyik vagy másik preferenciát adják-e meg neki. Tekintettel arra, hogy valamilyen ilyen döntést kell hozniuk, ésszerűnek tarthatnánk, hogy olyan döntést hozzanak, amelynek kívánatosabb következményei vannak. Másodszor, egy olyan emberi lény esetében, akit úgy "terveztek", hogy bizonyos vágyakkal rendelkezzen, gyanítható, hogy mélyebb szinten maradhatnak más diszpozíciók és hajlamok, amelyekkel a tervezett preferencia konfliktusba kerülhet. Aggódhatunk például amiatt, hogy az eredmény egy olyan személy lehet, aki szörnyen bűntudatosnak érzi magát, amiért csalódást okozott a szüleinek, és ezért túlzottan feláldozza más érdekeit, vagy hogy pszichéjének bizonyos rejtett részei elfojtva és meghiúsítva maradnak. A digitális elmék esetében azonban talán elkerülhetőek lennének az ilyen problémák, ha úgy lehetne megtervezni őket, hogy belsőleg egységesebbek legyenek, vagy ha az "örökölt" emberi populáció érdekeinek tiszteletben tartására vonatkozó preferenciát olyan "könnyed" módon adnánk hozzá, amely nem szülne belső viszályokat, és nem akadályozná a digitális elme képességét, hogy mással foglalkozzon.

Mindent egybevetve úgy tűnik, hogy egy olyan eredmény, amely lehetővé teszi a digitális szuper-kedvezményezettek *és* a nagymértékben virágzó emberi populáció megőrzése nagyon magas pontszámot érhet el mind a személytelen, mind az emberközpontú értékelési mércén. Tekintettel a nagy tétre és a visszafordíthatatlan fejlemények lehetőségére, nagy értéket jelentene az erkölcsileg elfogadható és gyakorlatilag megvalósítható utak feltérképezése, amelyeken keresztül egy ilyen eredményt el lehet érni.

[18] Például (Habermas, 2003; Sandel, 2007).

Hivatkozások

Agar, N. (2010) Az *emberiség vége*. The MIT Press. pp. 164-189.

Aghion, P., Jones, B. F. és Jones, C. I. (2017) "Artificial Intelligence and Economic Growth", *National Bureau of Economic Research Working Paper Series*, No. 23928.

Appiah, A. (2006) *Kozmopolitizmus: Etika az idegenek világában*. Allen Lane.

Bostrom, N. (2003) "Csillagászati hulladék: *Utilitas*, 15(3), 308-314. o., "A késleltetett technológiai fejlődés alternatív költségei", *Utilitas*, 15(3), 308-314. o.

Bostrom, N. (2008a) "Letter from Utopia", *Studies in Ethics, Law, and Technology*, 2(1).

Bostrom, N. (2008b) "Miért akarok poszthumán lenni, amikor felnövök", in Gordijn, B. és Chadwick, R. (szerk.): *Medical Enhancement and Posthumanity*. Springer Hollandia, pp. 107-136.

Bostrom, N. (2013) "Exisztenciális kockázatok megelőzése mint globális prioritás", *Global Policy*, 4(1), pp. 15-31.

Bostrom, N. és Yudkowsky, E. (2014) "The Ethics of Artificial Intelligence", in Frankish, K. és Ramsey, W. M. (szerk.) *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, pp. 316-334.

Calo, R. (2015) "Robotika és a kiberjog tanulságai", *California Law Review*, 103(3), p. 529.

Chalmers, D. (2010) "A szingularitás: *Journal of Consciousness Studies*, 17(9-10), 7-65. o., A Philosophical Analysis.

Frick, J. D. (2014) *"Boldoggá tenni az embereket, nem boldog embereket csinálni": A Defense of the Asymmetry Intuition in Population Ethics (doktori disszertáció).*

Greaves, H. és Ord, T. (2017) "Moral Uncertainty About Population Axiology", *Journal of Ethics and Social Philosophy*, 12(2), pp. 135-167.

Habermas, J. (2003) *Az emberi természet jövője*. Polity Press.

Hanson, R. (2001) *Gazdasági növekedés gépi intelligencia mellett*. Technikai jelentés, University of California, Berkeley.

Hanson, R. (2016) *The Age of Em: Munka, szerelem és élet, amikor robotok uralják a Földet*. Oxford University Press. pp. 63-5.

Harsányi, J. (1953) "Cardinal Utility in Welfare Economics and in the Theory of Risk-taking", *Journal of Political Economy*, 61(5), pp. 434-435.

Herzog, M. H., Esfeld, M. és Gerstner, W. (2007) "Consciousness & the Small Network Argument", *Neural Networks*, 20(9), pp. 1054-1056.

Hurka, T. és Tasioulas, J. (2006) "Games and the Good", *Proceedings of the Aristotelian Society, Supplementary Volumes*, 80, p. 224.

Kagan, S. (2019) *Hogyan számoljuk az állatokat, többé-kevésbé*. Oxford University Press.

MacAskill, W., Cotton-Barratt, O. és Ord, T. (2020) "Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons", *Journal of Philosophy*, 117(2), pp. 61-95.

Narveson, J. (1973) "Moral Problems of Population", *The Monist*, *57*(1), 62-86. o., 62-86. o.

Nordhaus, W. D. (2007) "Two Centuries of Productivity Growth in Computing", *The Journal of Economic History*, 67(1), pp. 128-159.

Nozick, R. (1974) *Anarchia, állam és utópia*. Basic Books. p. 41.

Parfit, D. (1984) *Okok és személyek*. Oxford University Press, pp. 388-389343,, 498.

Pearce, D. (1995) *Hedonistic Imperative.* www.hedweb.com [hozzáférés: 242020.

szeptember]. Rawls, J. (1971) *Az igazságosság elmélete*. Belknap. pp. 379-380.

Sandel, J. M. (2007) *The Case Against Perfection: Etika a géntechnológia korában*. Harvard University Press.

Singer, P. (1981) *A táguló kör: Ethics and Sociobiology*. Clarendon Press. Stace, W.

T. (1944) "Interestingness", *Philosophy*, 19(74), pp. 233-241.

Thomas, T. (2019) "Aszimmetria, bizonytalanság és a hosszú távon", GPI Working Paper No. 11-2019.

Williams, B. A. O. (2006) "Az emberi előítélet", in: A *filozófia mint humanista diszciplína*. Princeton University Press. pp. 135-152.