A MEGTÉVESZTÉS FILOZÓFIAI VIZSGÁLATA A ROBOTIKUS TÁRSAS KAPCSOLATOKBAN

DONÁTH BÉLA, ELTE BTK

TKP NVA 2021 – NEMZETVÉDELEM, NEMZETBIZTONSÁG ALPROGRAM

MESTERSÉGES INTELLIGENCIA: ESZKÖZ VAGY TÁRS PÉNZÜGYEINK INTÉZÉSÉBEN? C. WORKSHOP

2024 ÁPRILIS 12

ELTE IK CLC

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY

PROGRAM FINANCED FROM THE NRDI FUND

Mi az alapvető probléma az androidokkal?

- Mondhatnánk, hogy ha egy gép tudatos akkor van morális státusza, ha pedig nem, akkor nincs neki és végeztünk is.
- A kérdés, hogy honnan tudjuk, hogy egy gép tudatos-e vagy sem?
- Egymásról honnan tudjuk? Vagyunk-e olyan episztemikus státuszban, hogy igazoltan állítsuk valamilyen entitásról hogy tudatos vagy sem?

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Az erős MI vagy a gépi tudat problémái

- A probléma, hogy ahhoz, hogy elkötelezzük magunkat az erős MI mellett, szükséges lenne, hogy a tradicionális elmefilozófiai problémákat megoldjuk. Enélkül ugyanis teljesen megalapozatlan azt állítani, hogy az elme emulálható vagy újraalkotható. Hogyan lenne lehetséges tudatos szintetikus elmét létrehozni, ha nem ismerjük a választ arra a kérdésre hogy mi a tudat és milyen a természete?
- Gödel nem-teljességi tétele
- Penrose állítása, miszerint az elme rendelkezik nem-komputációs tulajdonságokkal, amik kvantummechanikaiak
- Block Kínai nép gondolatkísérlete
- Searle Kínai szoba kísérlete, továbbá, hogy a programszerűség nem intrinzikus tulajdonsága az agynak, az mindig a megfigyelőtől függ
- Könnyen lehet, hogy az elme egzakt természete mindörökre tisztázatlan marad

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Az erős MI vagy a gépi tudat problémái

- A test-elme probléma valamint a hard problem miatt a filozófusok elkezdték megkerülni az ontológiai kérdéseket.
- Az új kérdés visszavezet a turingi alapokhoz: lehetséges-e olyan androidokat létrehozni, melyek viselkedésükben szinte vagy teljesen megkülönböztethetetlenek az embertől? Ennek már vélhetően nincs elvi akadálya.
- A más elmék problémája ráadásul nem engedi meg számunkra, hogy konkluzív állítást tegyünk egy embertől (adott esetben radikálisan) különböző entitás mentális állapotairól.
- Tehát az androidról minimum nem tudjuk, hogy tudatos-e, deklaráltan nem tudhatjuk, hogy rendelkezik-e valamilyen idegen/gépi tudattal és maximum azt állíthatjuk hogy tudjuk, hogy emberi tudatossággal nem rendelkezik.
- Elismerjük az erős MI ellenében felhozott érvek relevanciáját, a reduktív fizikalizmus és a számítógépelméletekkel szembeni ontológiai szkepticizmust, azonban kitartunk amellett, hogy az elmeállapotok megközelítőleg lemásolhatóak vagy újraalkothatóak és hogy ezzel az MI jövőbeli relevanciája nem csökken lényegesen, sőt a vita a morális és jogi státuszt illetően, a társadalmi szerepet és a morális felelősségvállalást illetően hevesebb lesz.
- Fenntartjuk, hogy az antropomorf gyenge MI legtökéletesebb formájában filozófiai zombi - nem rendelkezik tehát az általunk "jól" ismert emberi fenomenális tudatossággal, azonban a külső szemlélő számára megkülönböztethetetlen egy biológiai embertől.
- A gyenge MI-nek más típusú problémákkal kell megküzdenie, nevezetesen az antropocentrizmussal és a biologizmussal.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Antropocentrizmus 1.

- Az ember egy a természeten kívül- és felülálló entitás, kinek értéke és érdekei felülmúlnak minden más entitását körülötte. Az ember az egyedüli élőlény, mely öntudatos, reflektív és racionális ezért morális szempontból értékesebb akármelyik más lénynél.

- **Arisztotelész:** az emberek és az állatok osztoznak bizonyos tulajdonságokon: mindkettő képes táplálkozni, szaporodni, észlelni a külvilágot a érzékszerveik által és vágyakozni, érezni, emlékezni és elképzelni dolgokat. Viszont csak az ember képes racionálisan gondolkodni, ennélfogva az állatok alávetettek neki (amiknek pedig alávetettek a növények, melyek csak táplálkozásra és reprodukcióra képesek)

- **Aquinói Szent Tamás**: az élőlényeket a racionalitás alapján lehet rangsorolni: az ember tökéletesebb az állatoknál, mivel több racionalitással rendelkezik, cserébe az embernél tökéletesebbek az angyalok, a piramis tetején természetesen Isten áll, aki intellektuális szempontból (is) tökéletes.

- **Descartes:** az állatok nem mások, mint puszta automaták a szó legszorosabb értelmében: nincs lelkük és nem is gondolkodnak. A keresztény doktrína szerint halhatatlan lelke csak az embernek van, az állatoknak pedig nem hogy az nincs, de arra is képtelenek, hogy nyelvet használjanak. Éppen ezért az állatokkal szemben nagyjából annyi kötelezettségünk van, mint amennyi bármilyen más hétköznapi tárggyal szemben

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Antropocentrizmus 2.

- Az **antropocentrizmus térben és időben lokális:** nem csak az embert jelöli ki mint legfontosabb létezőt, hanem az itt-és most emberét

- A **térbeli lokalitás** szerint egy észak-európai ember antropocentrikussága kulturálisan és ontológiailag is a hozzá hasonlóakat jelöli ki, ami el fog térni egy afrikai emberétől. Ez a folyamat egészen az individuum szintjére is le tud korlátozódni, melynek értelmében legvégül "Én" leszek a legfontosabb entitás az összes közül.

- **Időbeli lokalitás:** a jelenkor embere joggal tárja szét értetlenül a kezeit az elmúlt évszázadok rabszolgatartásán. Azon kor emberének azonban - gondoljunk csak Arisztotelészre - ez a társadalmi berendezkedés szerves és kvázi-megkérdőjelezhetetlen része volt.

- **Továbbá**: az ember, mint ember nem végcél jelenlegi (sőt semmilyen) formájában az evolúció számára, ennélfogva "befejezett" sem lehet. Az emberi evolúció tovább folytatódik és bár jelenleg - a saját pozíciónkból nézve - nem úgy tűnik, hogy bármi változás is történne velünk, a futuristák és transzhumanisták egyaránt úgy gondolják, hogy a jövőben az ember akár radikálisan is megváltozhat a technológiai fejlődés hatására.

- Ha véletlenül találnánk egy élő egyedet a neandervölgyi emberből (esetleg létrehoznánk egyet klónozással), akkor az egész antropocentrikus filozófiánkat, etikánkat és politikánkat újra kellene gondolnunk. Ekkor ugyanis valósággá válna az, amitől elszoktunk, nevezetesen, hogy rajtunk kívül is van egy másik intelligens faj a bolygón

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Biologizmus

- A természetes evolúció által létrejött biológiai ember számára az az elfogadott, hogy az (intelligens) élet egyfelől szerves másfelől tudatos közbeavatkozás nélkül, évmilliók alatt végbemenő apró változások által jött létre.

- Észre kell azonban venni, hogy kitüntetett helyzetünk a puszta véletlen műve. Minden intelligens civilizáció, mely az eredetét vizsgálja, azt fogja tapasztalni, hogy az ő esetükben az evolúció sikeresen kifejlesztette az intelligenciát, mégpedig olyan formában amilyen maga a civilizáció.

- Félreértés lenne azt mondani, hogy az evolúció az emberi intelligencia (vagy épp a szilikon alapú űrlényintelligencia vagy akármi más) kifejlődését tartja szem előtt (mivel lényegében egyáltalán semmit sem tart szem előtt), mivel a folyamatok szükségszerűen megtörténnek csupán, mégpedig azon feltételek mellett, melyek éppenséggel adottak.

- Van egy erős intuíciónk ami alapján kijelölünk bizonyos entitásokat a világban és azt mondjuk rájuk, hogy tudatosak, míg másokra azt, hogy nem (és nem is lehetnek azok).

- Ez egy megszokáson alapuló durva kettős mércében összpontosul, mely negatívan diszkriminál minden olyan entitást, mely nem a hosszú és fáradtságos biológiai-evolúciós útvonalon keresztül jött létre: aki embertársai tudatosságát akarja tagadni, annak nagyon komoly bizonyítási teher nyomja a vállát, ellentétben azzal, aki szerint mind tudatosak vagyunk. Az androidoknál pont fordítva: szinte kikezdhetetlen érveket várunk el attól, aki szerint a gépek is lehetnek tudatosak, azonban jóval megengedőbbek vagyunk azokkal szemben, akik szerint soha nem lehetnek azok.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Relácionista fordulat

- Már a jogok kapcsán is felmerül, hogy relációsak: csak akkor van értelmük ha az egyénen kívül legalább még egyvalaki szerepet kap a történetben. Ha valakinek pl. joga van biztonságban élni akkor ennek csak akkor van értelme, ha van egy másik valaki aki ezt tiszteletben tartja vagy nem tartja.
- Egy relácionista nézetben nincs többé a priori ontológiai hierarchia az entitások között úgy, mint az esszencialista ontológiában. Ettől persze még az ontológiai sajátosságok nem lényegtelenek, csupán másodlagosak, hiszen látszólagosak, általunk interpretáltak. Ebből fakad hogy a morális számbavétel alapja sem az alanyban sem az objektumban nem lehet (csak és kizárólag) inkább a kettő közötti viszonyban.
- Ez a hozzáállás nem az ember és gép különkénti morális státuszára fókuszál, nem arra a tényre, hogy az ember és a gép egy-egy külön fajhoz tartozik, hanem arra hogy egyik sem önmagában létezik, hanem identitásuk igenis nagyban függhet egymástól, és más entitással kialakuló kapcsolatoktól.
- Eleve igen könnyen alakítunk ki érzelmi kapcsolatot bizonyos dolgokkal. Nem érdekel minket, hogy egy gép mit tud vagy ért (valójában, belül) meg azokból a közös eseményekből amiket együtt töltünk velük, elég ha úgy viselkednek mintha tudnák és értenék, azt viszont nagyon is meggyőzően. Ez lényegében Turing “udvarias konvenciója”. Maga a viszony, a kapcsolat vagy reláció megléte a fontos.
- A morális inklúzió feltételei így extrenzikussá válnak: a társadalom (vagy egyéb kisebb nagyobb csoportosulási forma) kapcsolatba lép az adott intelligensnek vélt entitással, majd a kapcsolat alakulásának fényében eldönti, hogy pl. jogi alapokra helyezi-e az addigra már bevett szokásként létező ilyen-olyan morális gyakorlatokat.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Etikai behaviorizmus

- A relácionista fordulat nehézsége a szélsőséges relativizmus. A probléma az elmélet “anything goes” jellegéből fakad.

- **Az érv lényegében: P1: ha felelősséget érzek x irányába, akkor morális státusza van P2: felelősséget érzek a robotok iránt K: morális státuszuk van.**

- Elég problémás hogy x helyébe bármit írhatunk. Akkor is ha nem csak én érzek felelősséget pl. a tollam iránt, hanem mi, mint közösség vagy társadalom. Ez így durva relativizmus, senki sem tudja majd innentől mit kéne tennie. Sőt, megengedi hogy valaki azt mondja, hogy “én mégsem érzek semmit vagy pont hogy undort a robotok iránt, ezért nincs morális státuszuk”.

- A relativizmus nehézségét ki lehet küszöbölni, ha le tudjuk korlátozni az x helyébe írható entitásokat. Erre tesz kísérletet az etikai behaviorizmus: P1: Ha egy robot viselkedését tekintve nagyjából ekvivalens egy olyan entitással amiről széles körben elfogadott, hogy van morális státusza, akkor a gépnek is megadható az a státusz. P2: A robot ekvivalens egy ilyen entitással. K: A robotnak adható morális státusz.

- Az ötlet alapja, hogy akkor is csupán a viselkedésből vonunk le következtetéseket, ha az entitás egyébként tudatos. Egyszerűen azért, mert nincs módunk megtudni, hogy milyen az adott élőlénynek lenni. Az episztemikus gátat tehát a viselkedés alapján ugorjuk át. Ha pedig az egyik esetben így teszünk, akkor logikus, hogy a másikban is így járjunk el. A két eset közötti különbség ugyanis - praktikusan - számunkra észrevehetetlen. Ez egyfajta etikai Turing-teszt. Searle analógiás érve is eszünkbe juthat: onnan tudom hogy a kutyám tudatos, hogy hasonló a fiziológiánk és viselkedése tudatosnak tűnik (az ő szintjén, de számomra is értelmezhetően).

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Társas kapcsolatok 1.

- **Laza kapcsolat:** általában idegenek között jön létre, szerződési és jogi szabályokon alapuló kapcsolat, mely a kölcsönös előnyök érdekében köttetik. Az ilyen viszonyt néhány alapvető erkölcsi és jogi szabály szabályozza, de az erkölcsi kötelezettségvállalások minimálisak.

- **Szoros kapcsolat:** a család és a barátok közötti kapcsolatok. Ezeket az etikai normák és szabályok átfogóbb halmaza szabályozza. Időtartamuk nem korlátozott, és nem egy adott célt szolgálnak. Kölcsönösen előnyösek lehetnek – abban az értelemben, hogy az emberek „kihoznak" valamit ezekből a kapcsolatokból –, de nem elsősorban ezekben a fogalmakban gondolkodnak róluk vagy értelmezik őket.

- Ha egy vállalati vezető azt mondja az alkalmazottainak: „Egy nagy család vagyunk", az üresen hangzik. Részben azért, mert manipulatív módon próbálja toborozni egy hideg, számító vállalat számára a családi kapcsolatokhoz kapcsolódó hűséget és szeretetet.

- A szoros kapcsolatokat egyfajta ragasztó tartja össze. Hétköznapi esetben ez a ragasztó egy kölcsönös összetartozás érzést, közös történelmet vagy emlékezetet és közös jelentést foglal magában.

- A kapcsolat felei otthon érzik magukat egymással, egy közös narratíva szereplőivé válnak, és rájönnek, hogy ez a narratíva értelmet és célt ad nekik. Különösen a lojalitási és a bizalmi kötelezettségek szabályozzák őket: a kapcsolatban álló feleknek meg kell védeniük egymás érdekeit, bízniuk kell egymásban, és abban hogy a másik nem vét ezen érdekekkel szemben.

- A megtévesztés és az árulás által sérül vagy megszűnik a kötőanyag. Bármi, ami jó okot ad az embereknek arra, hogy átértékeljék egy szoros kapcsolat jelentését, árulásnak minősülhet. Az árulás sokféle formát ölthet, de az egyik leggyakoribb az, amikor az áruló hamis és félrevezető jelzéseket küld az elárultaknak, esetleg kiadja őt egy harmadik félnek, miközben fenntartja a szoros kapcsolat illúzióját.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Társas kapcsolatok 2.

- Ha az ember-robot kapcsolatok a laza kapcsolatok világába tartoznak, akkor az ezen kapcsolatokban elvárható hűség és bizalom kötelezettségei viszonylag minimálisak lesznek, nagyrészt szerződésben és felhasználói feltételekben, jogi környezetben rögzíthetők. Ha viszont az ember-robot kapcsolatok a szoros kapcsolatok világába tartoznak, akkor speciális morális kötelezettségek érvényesülnek.
- A véleményem az, hogy legalább bizonyos esetekben az ember-robot kapcsolatokat szoros kapcsolatoknak lehet tekinteni.
- Ennek egyik oka a nyilvánvaló piaci rés: bizonyos cégek szociális robotokat fognak forgalmazni termékként profitszerzés céljából. A robottársak, barátok és gondozók például úgy lesznek kialakítva, hogy különleges jelentőséget kapjanak használóik életében. (Mindez azonban nem feltétlenül korlátozódik az elesettekre, magányos emberekre és az idősekre.)
- Továbbá, bár a szoros kapcsolatok alapesetei a barátok és a család, van hajlamunk rá, hogy rugalmasak legyünk. Bármi, ami létrehozza a kötőanyagot (kölcsönösség, összetartozás, a másikban otthonra lelés) kialakíthat szoros kapcsolatot. A robotok azáltal, hogy beépülnek az életünkbe, megosztanak velünk pillanatokat, reagálnak ránk, együtt nevetnek és segítenek nekünk, létrehozhatják ezt a kötőanyagot.
- Amennyiben a robot már pusztán azért megtévesztő, mert robot, tehát az embertől különböző fajú, ontológiailag képlékeny entitás (magyarán megtévesztő jelekkel folyamatosan azt sugallja, hogy rendelkezik bizonyos kapacitásokkal és belső állapotokkal miközben azok minden lehetséges formában tökéletesen hiányoznak), akkor a vele való kapcsolatunk illuzórikus és szabályozható pusztán instrumentális eszközökkel.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Megtévesztés 1.

- **Robotikus megtévesztés**: akkor fordul elő, amikor egy robot (a) valamilyen jelet (beszédműveletet, antropomorf jelzést) használ oly módon, hogy (b) megsérti azokat az elvárásokat/normákat, amelyeket általában az ilyen jelek használatához társítunk (leggyakrabban a jel használatával objektíve hamis vagy félrevezető módon), ahol (c) ez valamilyen hátsó szándékból eredő célt szolgál, amely vagy magára a robotra vagy egy harmadik félre vezethető vissza.

- **Hogy történik ez a beszédben?** A beszélgetésnek vannak normái: közös elképzelések a célokkal kapcsolatban és bizonyos szabályok amiket mindkét fél követ közben. A legtöbb beszélgetésnél a mennyiség, minőség, viszony és mód a mérvadó maximák. Ezek mind az igazmondást szolgálják. De bizonyos esetekben más maximák érvényesek, pl amikor kedvesek akarunk lenni. Ilyenkor se lehet össze-vissza beszélni, de az igazság hátrébb kerülhet egy nemesebb cél érdekében. Az igazság esetében a megtévesztés legegyszerűbb formája az amikor amit mondunk nem igaz.

- Egyesek szerint nem számít annyira egy beszéd igazságértéke addig, amíg nincs mögötte hátsó szándék. Ez a hátsó szándék az amely a normák ártalmatlan megsértéséből valami etikailag aggasztóbbá változtatja a beszélgetést.

- Azonban nem csak beszédből fakadhat megtévesztés, hanem az őszintétlen antropomorfizmusból is: a megtévesztés ezen formája magában foglalja antropomorf jelek (megjelenés és viselkedés) használatát, hogy elvonják és félrevezessék az embereket a robot valódi természetét és célját illetően.

- Érdemes megjegyezni, hogy ezen őszintétlenségek esetében a harmadik féltől származó megtévesztő célok és rossz indulatú hátsó szándékok adnak főként okot az aggodalomra.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Megtévesztés 2.

- Külső dologra irányul: például egy orvosi diagnosztikai robot, amely félrevezető benyomást kelt az egyén egészségéről és jólétéről, hogy szükségtelen orvosi kezelésre csábítsa. Ha ezt helytelenítjük egy embernél, akkor a robotnál is.

- Látszólagosan meglévő dolgokra irányul: ha úgy tűnik, hogy egy robot viselkedése és megjelenése miatt rendelkezik bizonyos képességekkel (vagy szándékkal vagy érzelmekkel). Az EB alapján nem megtévesztés, ameddig következetes. Tudom, hogy a partnerem iránti érzelmeim valódiak, mert személyesen érzem őket – ismerem az iránta érzett vágyakozást és vonzalmat –, de ő honnan tud róluk? Nem tud közvetlenül hozzáférni a belső mentális állapotaimhoz. Csak a külső viselkedésem alapján tud következtetni. Ha ez a viselkedés nincs összhangban az általam vallott vágyakozással és vonzalommal, akkor lesz episztemikus értelemben jó indoka (nem csupán indoka, hanem jó indoka) azt feltételezni, hogy nem vagyok őszinte.

- Rejtett állapotokra irányul: ha egy robot eltereli a figyelmet arról, hogy valamilyen képesség vagy funkció birtokában van. Ha egy robot elfordítja a tekintetét, de valamilyen módon tovább rögzít, az morálisan aggályos és megtévesztő, mert megsérti azokat az elvárásokat és normákat, melyeket ezen viselkedési formákhoz társítunk.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Példák a megtévesztő magatartási formákra - amennyiben a kapcsolat mindenképp illúzió

- Kamerák és mikrofonok elhelyezése nem várt helyen - A testtel kapcsolatos emberi elvárások felborulnak, ami torzítja a robotok által jelentett kockázatok értékelését.
- Olyan robot mely mosolyog, ráncolja a homlokát, megdönti a fejét, hogy az együttérzésnek tűnjön, vagy úgy ugorjon hátra, mintha meglepődne. - Az emberek intuitív módon arra következtetnek, hogy egy robot barátságos, ellenséges, rokonszenves, érdektelen stb., ami torzítja a képességüket az általa jelentett kockázatok felmérésére, elbillenti őket a kötődés és a bizalom felé.
- Olyan robot, mely eszik és élvezi (vagy nem szereti) az ételt - Az emberek egészségtelen szinten fektetnek be a robot jólétébe és egészségtelen szinten törődnek vele.
- Olyan robot tervezése, mely bár nem él át érzelmeket, mégis, bölcsnek vagy nyugodtnak vagy tekintélyesnek, asszertívnek stb. hangzik - Az emberek azt feltételezik, hogy a robot az emberével kompatibilis információkat közvetít nem létező attitűdökről és tapasztalatokról. Ez megzavar abban, hogy megfelelő képet kapjunk a robot jelentette kockázatokról és arra sarkall hogy túlzott bizalmat érezzünk vele kapcsolatban.
- Olyan robotok, melyek genderrel rendelkeznek, egy bizonyos fajhoz vagy kultúrához tartoznak, visszatükrözik a társadalmi osztályt vagy bizonyos készségek meglétét. - Kapcsolatba lépéskor a robotról visszatükröződnek az emberek saját előítéletei, lehetővé téve másoknak, hogy ezt kihasználják, nem megfelelő kötődésre vagy sértő magatartásra használják.
- Olyan robot, mely befolyásolja az ember fizikai állapotát, fény beállítása a hangulat szabályozására, vagy hanghatások stb. - Az emberek tudattalanul sebezhetővé válnak a környezetük érzékszervi alapú manipulálásával szemben, ami hatással lesz hangulatukra, érzelmeikre és viselkedésükre, továbbá robotikus manipuláció és kontroll áldozatává válnak.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Az etikai behaviorizmus válasza

- Amennyiben szoros társas kapcsolatot kívánunk kialakítani a robotokkal, akkor is lényeges, hogy se a gép, se a gépet gyártó harmadik fél ne éljen vissza az antropomorf jegyekkel. Magyarán fel kell térképezni és meg kell húzni a határt az őszinte és az őszintétlen antropomorfizmus között.
- Az előbbi megtévesztő magatartásformák az EB. szerint nem megtévesztőek pusztán azért, mert robotok követik el őket.
- Az EB alap érve szerint ugyanis egy olyan robottal szemben (legyen példa a replikáns) mely viselkedését tekintve minden szempontból megkülönböztethetetlen egy valódi embertől, nem várható el, hogy viselkedésen felül bármivel alá tudja támasztani "intencióit". Az EB szerint egy ilyen entitás közeledését elutasíthatjuk persze azzal, hogy az hamis, azonban nincs rá "jó" indokunk, ahogy nem lenne egy emberrel szemben sem.
- A megtévesztés és az árulás pontosan azért az ami, mint ahogy minden más emberi kapcsolatunkban: elbontja a kötőanyagot és kikezdi a bizalmat amiről addig azt hittük kölcsönös. (Nem minden megtévesztés árulás azonban: amennyiben a megtévesztés magasabb célt szolgál, akkor megengedhető, sőt kívánatos is lehet egy robotnál.)
- Ebben az esetben, ha egy robot hőérzékelőkkel meglesi mit csinálok a másik szobában, miközben azt hiszem nem lát (ahogy egy normál ember nem látna) nem csupán felhasználási feltételt szeg, de aláássa az addig épített közös kapcsolatunkat is, vagyis normatív szabályt sért meg.
- Ez esetben olyan társas robotokat kell (és talán csak akkor szabad, a saját érdekünkben) gyártani, melyek viselkedésüket tekintve következetesen értik és elfogadják az ember-robot kapcsolatok különleges státuszát, értik és képesek reflektálni a különbözőségből fakadó életviteli, etikai nehézségekre, következetesen értik egy bizalmi kapcsolat mibenlétét és elszámoltathatóak, amennyiben megsértik azt.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE HUNGARY
PROGRAM FINANCED FROM THE NRDI FUND

Összegzés

- Jelenleg az MI eszközszerű, instrumentális, de a távoli jövőben ez megváltozhat.
- Egyrészt társként fogják nekünk árulni, másrészt hajlamosak vagyunk őket társnak tekinteni.
- Egy bizonyos performatív szint elérése után saját érzelmi világunk látná kárát ha nem reagálnánk az antropomorf jelzésekre amiket használnak
- Az androidban az antropomorf és az idegenség egyszerre van jelen: az ember tervezi, az emberhez hasonlóra, de a "faji" sajátosságok miatt idegen tulajdonságokkal, "intenciókkal" (akként funkcionáló viselkedésformákkal) rendelkezhet.
- Ezen kettősségből fakadó veszélyek egyszerre tartoznak a "known unknown és az unknown unknown" kategóriákba.
- Amennyiben szoros kapcsolatot szeretnénk a gépekkel, esetleg úgy gondoljuk egy kívánatos jövő érdekében ilyen robotokat kell tervezni, akkor a megtévesztő viselkedési formák és esetlegesen az árulás lehetősége mindenképp olyan ár melyet meg kell fizetnünk.
- Ha laza, instrumentális kapcsolatot, akkor elég ha a jogi környezet szavatolja, hogy gépi természetük transzparens legyen és meggátolja hogy az érzelmek és szándékok illúzióját a kiszolgáltatott felhasználók kihasználására, kizsákmányolására használja.

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE
HUNGARY

PROGRAM FINANCED FROM THE NRDI FUND