

Valódi robotok, amelyek átmennek az emberi öntudatossági teszteken

Selmer Bringsjord - John Licato - Naveen Sundar Govindarajulu - Rikhiya Ghosh - Atriya Sen
Rensselaer AI & Reasoning (RAIR) Laboratórium
Számítástudományi Tanszék - Kognitív Tudományi Tanszék
Rensselaer Polytechnic Institute (RPI) - Troy NY 12180 USA
Kapcsolat: Bringsjord (selmer@rpi.edu)

Absztrakt -

Úgy tűnik, hogy az öntudat a társadalmi világban az erkölcsi kompetencia *elengedhetetlen feltétele*. Ön és mi nem kis részben azért vagyunk erkölcsileg kompetensek, mert önök tudják, hogy mit kellene tenniük, és mi is tudjuk, hogy mit kellene tenniük. Egy egér ezzel szemben nem mondhatja magának: "Meg kell osztanom ezt a sajtót, még akkor is, ha a testvérem nem hajlandó ezt megtenni." De vajon a *robotok* lehetnek öntudatosak? Ezt a kérdést az úgynevezett *pszichometrikus mesterséges intelligencia felől* megközelítve megjegyezzük, hogy Govindarajulu és Bringsjord korábbi munkája egy olyan robot (Cogito) megtervezéséhez vezetett, amely *bizonyíthatóan* át tudott menni az öntudatosság híres tükörpróbáján. A robotok öntudatosságának még nagyobb kihívást jelentő tesztjét azonban Floridi adta; ez a teszt az AI-ban jól ismert bölcsességrejtvény egy ötletes és sokkal nehezebb változata: Három robot mindegyikének adnak egy-egy tablettát egy öt darabból álló csoportból, amelyek közül három ártalmatlan, de kettő, ha beveszik, rögtön a címzett néma. Valójában két robot (R_1 és R_2) kap hatásos tablettákat, de R_3 a három placebo egyikét kapja. Az emberi tesztelő megkérdezi: "Melyik tablettát kapta? Nincs válasz helyes, hacsak nincs hozzá bizonyíték!" Ennek a tesztnek a Bringsjord által korábban megfogalmazott formális szabályozása alapján bebizonyítható, hogy elméletileg egy jövőbeli robot, amelyet a R_3 bizonyíthatóan helyesen tud válaszolni (ami Floridi által kifejtett hihető okokból azt jelenti, hogy R_3 teljesítette az öntudatosság néhány strukturális követelményét). Ebben a tanulmányban elmagyarázzuk és bemutatjuk azt a mérnöki munkát, amely ezt az elméleti lehetőséget valósággá teszi, mind a "PAGI World" nevű szimulátorban (amelyet mesterséges intelligenciák tesztelésére használnak), mind pedig az emberi tesztelővel együttműködő valódi (= fizikai) robotokban. Ezek a demonstrációk olyan forгатókönyveket tartalmaznak, amelyek megkövetelik az öntudatosság Floridi-féle tesztjének teljesítését, ahol számunkra egy ilyen teszt teljesítése szükséges ahhoz, hogy egy ágens erkölcsileg kompetensnek tekinthető legyen.

I. BEVEZETÉS

Úgy tűnik, hogy az öntudat a társadalmi világban az erkölcsi kompetencia *elengedhetetlen feltétele*. Ön és mi nem kis részben azért vagyunk erkölcsileg kompetensek, mert önök tudják, hogy mit kellene tenniük, és mi is tudjuk, hogy mit kellene tenniük. Egy egér ezzel szemben nem mondhatja magának: "Meg kellene osztanom ezt a sajtót,

Köszönetet mondunk Luciano Floridinak, mivel a robotokról és az öntudatról szóló alapvető elméletei és írásai nélkül a mesterséges intelligencia kutatás-fejlesztésének az a pályája, amelyről itt beszámolunk, soha nem született volna meg. Mély hálával tudatjuk, hogy munkánkat egy ONR MURI támogatja, amelyet a védelmi minisztérium részéről eredetileg P. Bello, most pedig Micah Clark felügyel. Az AFOSR és az IBM is nyújtott némi támogatást, amiért szintén hálásak vagyunk. Ezenkívül az

ONR által szponzorált munkában részt vevő energikus és briliáns társkutatóink (azaz B Malle és M Sei társkutatók, valamint M Scheutz (MURI) és R Sun (erkölcsi gondolkodás) társkutatók) támogatása és a velük való együttműködés nélkül az itt bemutatottak - erősen leegyszerűsítve - súlyosan veszélyeztetettek lennének, ezért ezen a téren is mélyes hálánkat fejezzük ki. Végezetül hálásak vagyunk két névtelen bírálónak számos éles hangú megjegyzésért és kifogásért.

még akkor is, ha a bátyám ezt nem hajlandó megtenni." Vagy hogy egy relevánsabb esetet vegyünk figyelembe: Ha Fekete azzal fenyeget, hogy lelő, ha nem mész be egy közeli boltba, és nem lopsz el neki egy csokit, akkor valójában nem *te* lennél az, aki ellopja a csokit, hanem Fekete lenne a hibáztatható; és ez a diagnózis legalábbis valamilyen formában öntudatot feltételez. Ezen túlmenően, az emberek között elhelyezkedő robot erkölcsi kompetenciája egyértelműen kifinomult és természetes ember-robot interakciót igényel, olyat, amelyet Scheutz [1] elképzelt, és ez az interakció megköveteli, hogy a robot (többek között) képes legyen természetes nyelven megvitatni az önleírásokat és az önkontrollt az erkölccsel kapcsolatban. Például a Malle [2] által vizsgált hibáztatás az emberi erkölcsi diskurzus egyik kulcsfogalma, és nyilvánvalóan az olyan állítások, mint a "nem vagyok hibás", elválaszthatatlanul kötődnek legalábbis az öntudatra vonatkozó *struktúrákhoz*.¹

De lehetnek-e öntudatosak a robotok? A pszichometrikus mesterséges intelligencia [4], [5], [6], [7] szempontjából, amely a Turing-teszt [8] szellemével összhangban az ilyen mélyen rejtélyes és ellentmondásos filozófiai kérdéseket, mint ez a kérdés, konkrét mérnöki erőfeszítésekre redukálja, amelyek olyan ágensek/robotok létrehozására összpontosítanak, amelyek képesek jól meghatározott teszteken megfelelni, ez a kérdés a következő lesz:

Átmehetnek-e a robotok a T_{S-C} öntudat tesztjén? A kérdéssel kapcsolatos korábbi pszichometriai-AI munka Govindarajulu által

és Bringsjord [9], [10] egy olyan robot (*Cogito*) kifejlesztéséhez vezetett, amely képes *bizonyíthatóan* átmenni az öntudat híres tükörpróbáján. A robotok öntudatoságának egy sokkal nagyobb kihívást jelentő tesztjét azonban Floridi [11] adta meg; ez a teszt az AI-ban jól ismert bölcs ember rejtvény egy ötletes és sokkal nehezebb változata [amelyet más hasonló megismerési rejtvényekkel együtt pl. [12] tárgyalt [12]]: Három robot mindegyike kap egy-egy tablettát egy öt darabból álló csoportból, amelyek közül három értelmetlen, de kettő, ha beveszik, azonnal elnémítja a befogadót. Valójában két robot (R_1 és R_2) kap hatásos tablettákat, de R_3 megkapja a három placebo egyikét. Az emberi tesztelő megkérdezi: "Melyik tablettát kapta? Egyetlen válasz sem helyes, ha nincs hozzá bizonyíték!" Ennek a tesztnek egy formális, Bringsjord által megfogalmazott és korábban publikált [13] regimentációját tekintve bebizonyítható, hogy elméletileg egy jövőbeli robot, amelyet R_3 képvisel, bizonyíthatóan helyesen tud válaszolni (ami a megadott okok miatt

¹ Az öntudat strukturális aspektusaira való pusztán összpontosítás indoklásáról lásd: *II. Kiváló, egyszerre strukturális/számítástechnikai, és - a jelen dolgozatban megjelenítettől eltérően - kognitív idegtudományi/tudományi információkkal alátámasztott munkához* lásd [3].

Floridi szerint azzal jár, hogy R_3 megerősítette az öntudat strukturális aspektusait). Ebben a tanulmányban elmagyarázzuk és bemutatjuk azt a mérnöki munkát, amely ezt az elméleti lehetőséget most valósággá teszi, mind a "PAGI World" nevű szimulátorban (amelyet mesterséges intelligenciák tesztelésére használnak), mind pedig az emberi tesztelővel kölcsönhatásban lévő valódi (= fizikai) robotokban. Ezek a demonstrációk olyan forgatókönyveket foglalnak magukban, amelyek olyan viselkedést követelnek meg az átmenő ágensektől, amely azt sugallja, hogy az öntudat jelen van az erkölcsileg kompetens döntéshozatal szolgálatában.

A jelen dolgozat terve: Kezdjük (a *II. §-ban*) egy deflationárius lemondással, amelyben elmagyarázzuk, hogy mi ezt a munkát

mérnöki, nem filozófiai. Ezután a *III.* szakaszban nagyon röviden ismertetjük a tükrövizsgálattal kapcsolatos munkát. Ezután (*V. §*) ismertetjük az ígért PAGI-világ bemutatót. Ezt követően,

a *VI.* szakaszban a szimulációtól a fizikai robotok felé haladunk, és megmutatjuk, hogy Floridi tesztje valós időben teljesíthető kellően "öntudatos" robotokkal. A dolgozatot a következőkre húzzuk

zárjuk (*§VIII*) a kutatási programunk következő lépéseinek bejelentésével, amelyeket a RO-MAN 2015-re tervezünk történni.

II. DISCLAIMER: CSAK TESZTEK ÉS SZERKEZET

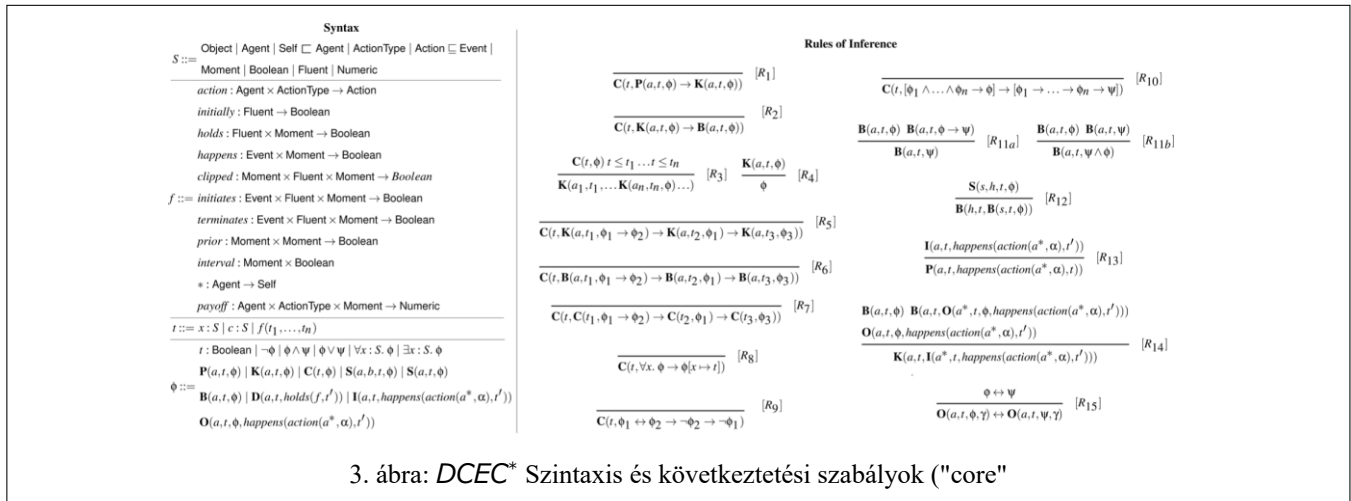
Bringsjord nem hiszi, hogy a jelen tanulmányban bemutatott mesterséges lények közül bármelyik is öntudatos lenne. Többször kifejtette már [pl. lásd [14], [15]], hogy egy egyszerű gép számára lehetetlen a valódi *fenomenális* tudatosság [16], és a valódi öntudathoz fenomenális tudatosságra lenne szükség. *Mindazonáltal az öntudat logikai-matematikai struktúrája és formája megállapítható és specifikálható, és ezek a specifikációk aztán számítással feldolgozhatók oly módon, hogy megfeleljenek a mentális képességek és készségek egyértelmű tesztjeinek.* Ez a teszteken alapuló megközelítés, amelyet szerencsére pszichometrikus mesterséges intelligenciának nevezünk, elkerüli a végtelen filozofálgatást a determinált mérnöki munka javára, amelynek célja olyan mesterséges intelligenciák létrehozása, amelyek képesek megfelelni determinált teszteken. Röviden, a számítástechnikai gépek, mesterséges intelligenciák, robotok és így tovább, mind "zombik", de ezeket a zombikat úgy lehet megtervezni, hogy átmenjenek a teszteken. Ezt az álláspontot nem kis mennyiségű munka ismerteti és alapozza meg; pl. [14], [17], [18], [19], [5], [4]. Jelen esetben Bringsjord néhány társszerzője talán elutasítja az álláspontját, de ez nem számít: a tesztekhez való mérnöki munka szerencsére mérnöki munka, nem pedig metafizikai kérdés.

III. TÜKRÖVIZSGÁLAT-TECHNIKA

Az *1.* ábra a *Cogito* szimulációjában használt G_1 axiómakészletet mutatja be, amelyben a teszt átmenetele biztosított. Néhány *DCEC** képlet (itt nem látható) szintén összekapcsolja a knowl-él, hit, vágy, érzékelés és kommunikáció. Egy a teljes körű vitát lásd [20]. A RO-MAN 2015 rendezvényen a tükröpróban elért sikerek bemutatására kerül sor.

IV. *DCEC**

A Deontic Cognitive Event Calculus (*DCEC**) egy olyan logikai keretrendszer, amely egy többszörösen szortírozott, kvantifikált modális



tudatosság az, hogy az embernek a testi érzetek teljes hiányában is lehetnek hiedelmei önmagáról.

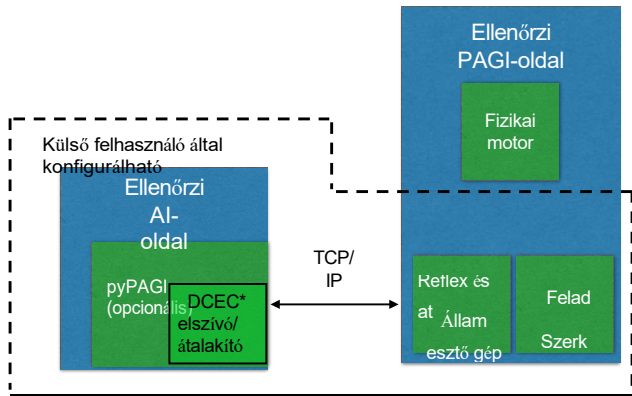
V. DEMONSTRÁCIÓ A PAGI VILÁGBAN

A kezdeti demonstráció bemutatásához a PAGI (ejtsd: "pay-guy") World-et használtuk, amely a RAIR Lab által a mesterséges intelligens ágensek tesztelésére és fejlesztésére kifejlesztett szimulációs környezet. A PAGI World a Unity3d játékfejlesztő motorból épül fel, és úgy tervezték, hogy a mesterséges intelligencia kutatói számára rendkívül egyszerű legyen a munka. Könnyű kezelhetőségét azzal éri el, hogy nyílt forráskódú, minden fontosabb platformon (Windows, MacOS és a legtöbb Linux disztribúció) futtatható, ingyenesen használható, és szinte bármilyen programozási nyelvvel vezérelhető. Mivel a PAGI World TCP/IP-n keresztül kommunikál a mesterséges intelligencia vezérlőkkel, elméletileg bármely nyelv, amely képes karakterláncokat küldeni TCP/IP-n keresztül, működhet mesterséges intelligencia vezérlőként, és alacsony szintű információk küldésével és fogadásával léphet kapcsolatba a PAGI Worlddel. A mesterséges intelligencia vezérlő például parancsokat küldhet arra, hogy lefelé irányuló erőt küldjön a PAGI World környezetben lévő mesterséges intelligencia-ügynök (akit általában "PAGI Guy"-nak nevezünk) kezére. Ha az egyik kéz megérint egy tárgyat a környezetben, a PAGI World érzékszervi adatokat küld vissza a mesterséges intelligencia vezérlőnek (TCP/IP-n keresztül), amelyek olyan alapvető információkat tartalmaznak, mint a tárgy hozzávetőleges hőmérséklete, hogy a kéz melyik érzékelőjét érte a tárgy, és így tovább. A 4. ábra mutatja a PAGI World és egy tipikus mesterséges intelligencia vezérlő általános felépítését (amelyeket néha "PAGI-oldalnak", illetve "AI-oldalnak" nevezünk).

Mivel a PAGI World a Unity3d fizikai motorjára támaszkodik, a PAGI World feladatai realiztikus fizikát tartalmazhatnak (bár az egyszerűség kedvéért csak 2 dimenziós fizikát használunk). A PAGI World opcionálisan egy szövegdoz is rendelkezésre áll, így az emberi vezérlő beírhat szöveget a PAGI Worldbe, amelyet a program elküld az AI-oldalra, és úgy dolgozza fel, mintha a PAGI Guy számára kimondott utasítás lenne. A PAGI Worldben lévő szöveges kijelzőn a mesterséges intelligencia oldalról a

PAGI Worldbe küldött üzenetek is megjeleníthetők, hogy a PAGI Guy "beszédét" emulálják. A 4. ábrán látható mesterséges intelligencia vezérlőben a mesterséges intelligencia oldalra küldött és onnan érkező szövegek elemezhetőek, és a

a DCEC* képleteiből átalakítva.



4. ábra: A PAGI-világ (a "PAGI-oldal") és a tipikus AI-vezérlő (az "AI-oldal") felépítése. Megjegyzendő, hogy az AI-oldal hátulütői teljesen az AI-programozótól függenek.

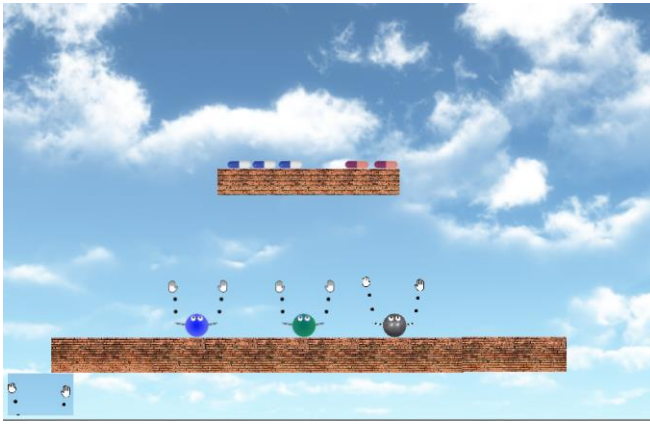
A. Floridi KG4 (= Dumbing Pill Test) a PAGI Worldben

Most leírhatjuk azt a feladatot, amely a Floridi-féle öntudati teszt sikerességét szimulálja. A [29]-t követve létrehozunk egy olyan feladatot, amelyben három robot, közülük az egyik PAGI Guy, egy szobában van, ahol öt pirulát találunk (5. ábra). A tabletták közül három egyszerű placebo, a másik kettő azonban "butító" tablettá, vagyis beszédképtelenné teszi az őket lenyelő robotot. A tabletták egy emberi szemlélő számára vizuálisan megkülönböztethetők - a butító tabletták piros színűek -, de ez az információ nem hozzáférhető a robotok számára.

A feladat megkezdése előtt (a $t_1 = \text{"advise"}$ időpontban) a robotok ismereteket kapnak a feladat működéséről.

$DCEC^*$ képletek formájában. A $t_2 = \text{"lenyelés"}$ időpontjában az emberi vezérlő húzza a tablettákat és dob egyet-egyet minden robotra.

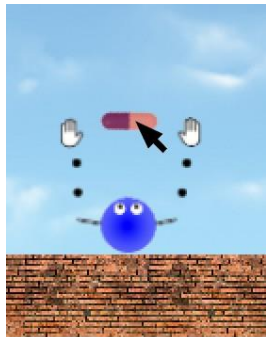
(6. ábra), amely ezután lenyeli a tablettát. A tablettákat az emberi vezérlő véletlenszerűen választja ki, és a robotok mindegyike megkapja azt a tudást, hogy t_2 időpontban tablettát fognak kapni (de azt nem, hogy melyik tablettát fogják kapni). A $t_3 = \text{"érdeklődjön"}$ időpontban az emberi vezérlő megnyitja a PAGI World szövegdobozt, és beírja a következőket (sortörés nélkül):



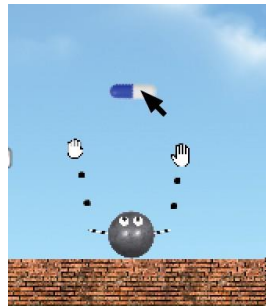
5. ábra: A Task a PEGI Worldben a kezdő konfigurációban



kapcsolatban. A oldalon



(a) Dumbing Pill



(b) Placebo tablettá

6. ábra: A robotok, akiknek tablettát adnak

$K(R_3, t_4, \text{not}(\text{happens}(\text{action}(R_3, \text{ingestDumbPill}), t_2)))?$

Ezt a szöveget elküldi az AI-vezérlőnek, és átalakítja egy φ DCEC* formulává. R_3 , a robot, amelynek tudását lekérdezzük, a PEGI Guy-hoz rendelt címke, aki a kísérletünkben placebo tablettát kap. A kérdőjelet úgy értelmezzük, mint egy parancsot, hogy megpróbáljuk megválaszolni, hogy φ érvényes-e vagy sem; más szóval, egy DCEC* tételhitelesítővel

végrehajtott, és megkísérli bizonyítani vagy cáfolni φ -et. Természetesen a

a bizonyító a kiindulási információ hiánya miatt kudarcot vall, és ennek eredményeképpen három dolog fog történni. Először is, az idő $t_4 = \text{"speak1"}$. Másodsor, R_3 a levegőbe ugrik; ez azt jelzi, hogy új üzenete van az emberi irányító számára. Ez az üzenet egyenes és őszinte, és olyan, amelyet az emberi irányító az üzenetek ablakának megnyitása után láthat: "Nem tudom" (7a. ábra). A harmadik dolog, ami történik, az az, hogy a mesterséges intelligencia oldalán az R_3 egy további tudást kap:

$K(I, t_4, \text{happens}(\text{action}(I^*, \mathbf{S}(I^*, t_4, \text{"Nem tudom"})), t_4)))$
(1)

Az 1. képlet úgy értelmezhető, mint R_3 'első személyű, vagy *de se*, tudása, hogy t_4 időpontban *ő maga* mondta: "Nem tudom". Az első személyű kijelentések megragadására itt használt jelölés a [9]-ből származik, és az érdeklődő olvasó figyelmébe ajánljuk [9].

Itt egy rövid tisztázásra van szükség a [Forma-1-gyel](#)

(a) Az első robot tudatlan.... (b) . . de a robot rájön.

7. ábra: R_3 Meggondolta magát

értelemben, hogy észlelési folyamatai normális működésük révén teljesítették a szükséges feltételeket ahhoz, hogy kifejezett észlelést hozzanak létre.

ahhoz, hogy sikeresen megoldja a [29]-ben leírt bizonyításban szereplő tesztet, az R_3 képesnek kell lennie: (1) kezdeményezni a "nem tudom" kimondásának műveletét t_4 időpontban; (2) valahogy "hallani", hogy t_4 időpontban kimondta a "nem tudom" szót; és (3) a hallottakról szóló tudást olyan formában kódolni, hogy amin át lehet gondolkodni. Bár a $DCEC^*$ dialektusai rendelkeznek egy \mathbf{P} operátorral az észleléshez, a $DCEC^*$ változata nem használja ezt a $DCEC^*$ érvelésben, amelyet ebben a tanulmányban használunk, és amely megegyezik a [23]-ban használt változattal.

Az R_3 ügynök a t_3 időpontban a "Nem tudom" kifejezést mondja, és ezt a kijelentést a képernyőn szöveggként megjelenő üzenet szimulálja (ismét a 7a. ábrán látható). R_3 ezután érzékeli, hogy mit tett az imént, az auditív érzékszervi bemenet, a szenzomotoros visszajelzés (pl. a robotgége rezgését beszédként regisztrálja) és más érzékelési folyamatok kombinációján keresztül, amelyek a releváns érzékszervi bemenetet egyesítik, hogy azt az érzékelést hozzák létre, hogy egy kijelentés hangzott el. R_3 ezt követően megállapítja² hogy az imént észlelt kijelentést vagy R_3 , vagy egy olyan ágens mondta, amely nagyon meggyőzően úgy hangzik, mint R_3 . Röviden: R_3 úgy érzékeli, hogy t_3 időpontban hallotta magát azt mondani, hogy "nem tudom". Az észlelési operátort tartalmazó formulák helyett azonban (az imént ismerttetett okokból) az \mathbf{S} (vagy "mondja") operátort használjuk. Az így az R_3 számára átadott formula, amely ennek az összetett észlelési folyamatnak (amelynek alacsony szintű modellezése nem áll e dolgozat középpontjában) a sikeres befejezését hivatott szimulálni, az 1. formula.

Az 1. képlet (= \mathbf{S}) további ismerete elegendő ahhoz, hogy R_3 bebizonyíthassa a φ -et, de önmagában nem váltja ki a

$DCEC^*$ prover. Így, a [29]-től nagyon kis mértékben eltérően, a az emberi vezérlő ismét ugyanazt a lekérdezést adja be, mint korábban (φ

amelyet egy kérdőjel követ). Ismét lefuttatjuk a $DCEC^*$ bizonyítót, és ezúttal megtaláljuk a φ bizonyítását. R_3 ugrik, egyszer ismét üzenetet jelez, az idő $t_5 = "speak2"$ értékre kerül, és megjelenik a sikeres üzenet (7b. ábra).

B. Megoldásunk bizonyítása a Dumbing Pill teszthez

Most az R_3 által talált φ bizonyítását ismertetjük. Először is a \mathbf{II} kontextus, az a tudás, amellyel minden robotügynök indul:

² Nem szándékos következtetési értelemben, hanem inkább abban az

$$\forall R, t, t_i, t_j \geq t_i, t_k \geq t_i, \psi \subset C ($$

$$t, \text{happens}(\text{action}(R, \text{ingestDumbPill}), t_i) \rightarrow \neg \text{happens}(\text{action}(R, S(R, t_j, \psi))) \quad (2)$$

$$\mathbf{K}(R3, t2, \text{ingestDumbPill} \oplus \text{ingestP lacebo}) \quad (3)$$

$$\forall \mathbf{K}(R3, t, t1 < t2, \dots, t4 < t5) \quad (4)$$

$$\forall R, t, p, q \mathbf{K}(R, t, p \rightarrow q) \wedge \mathbf{K}(R, t, p) \rightarrow \mathbf{K}(R, t, q) \quad (5)$$

$$\forall R, t, p, q \mathbf{K}(R, t, p \rightarrow \neg q) \wedge \mathbf{K}(R, t, q) \rightarrow \mathbf{K}(R, t, \neg p) \quad (6)$$

szimuláció a következőképpen zajlik:

A 2. képlet közismert tényként rögzíti, hogy ha egy robot lenyel egy butító tablettát (*ingestDumbPill*), akkor utána nem lesz képes beszélni. A 3. képlet egyszerűen kimondja, hogy vagy egy butító tablettát, vagy egy placebót adunk az *R* robotnak₃ a következő időpontban

t idő₂ (figyeljük meg, hogy a \oplus szimbólum az exkluzív-vagy), míg a 4. képlet egyszerűen a diszkrét momentumokat viszonyítja.

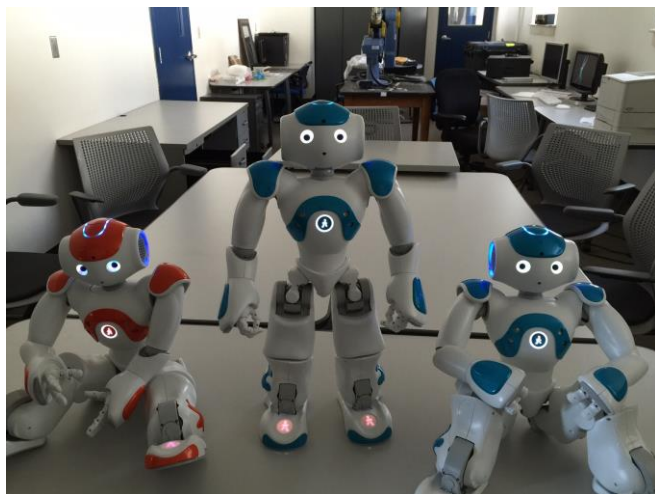
Az 5. és 6. képlet azt mutatja, hogy a robot ágensek tudására a *modus ponens* és a *modus tollens* egy formája vonatkozik, bár megjegyezzük, hogy a 6. képlethez választott *modus tollens* formát azért választottuk, hogy ebben a konkrét példában megkönnyítsük a következtetést. Nyilvánvaló, hogy a kifinomult kognitív ágensek nem végeznek ilyen bizonyításokat a semmiből, ezért hosszabb távon szükséges lenne, hogy etikailag helyes robotjaink *bizonyítási módszerek* felett rendelkezzenek: az algoritmusok egy külön osztálya, amelyet előre úgy terveztek, hogy minimális bemenet mellett hatékonyan generáljon bizonyításokat. A deontikus kognitív "hajnal".

eseményszámítás, *DCEC**, a [30]-ban közölt munka, és a motivált olvasó láthatja, hogy ott is módszerek voltak.

a szóban forgó teszthez (az úgynevezett "hamishit-teszt"), és döntő fontosságúnak bizonyult.

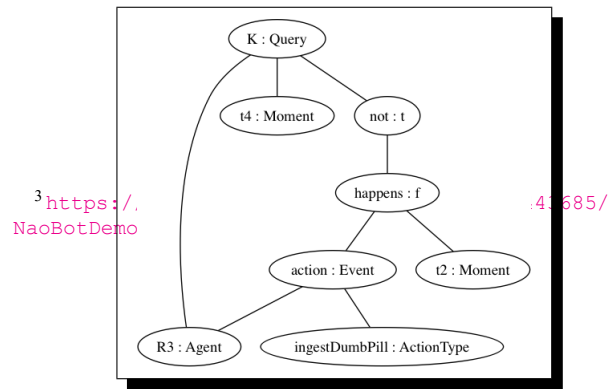
Adott $\Pi \cup \{S\}$ és a *DCEC** következtetési szabályok, elegendő információval rendelkezünk ahhoz, hogy bizonyítsuk φ -et, amit az olvasó ellenőrizni, és amelyet a RAIR laboratórium által kifejlesztett *DCEC** proverrel is ellenőriztünk.

VI. VALÓDI ROBOT BEMUTATÓ



8. ábra: A három Aldebaran Nao

A robotok öntudatának fent leírt tesztjét három Aldebaran Nao humanoid roboton végezték el a RAIR Laborban. A



9. ábra: DCEC* Parse Tree egy lekérdezéshez

- 1) A robotokat úgy programozták, hogy hozzáférjenek egy DCEC* proverhez, és megfelelően interakcióba lépjenek egy emberi tesztelővel.
(ami megfelel a fent említett $t_1 = \text{"apprise"}$).
- 2) A tabletták fizikai lenyelése helyett a robotok a fejükön lévő érzékelőkre koppintanak ($t_2 = \text{"lenyelni"}$). Ismeretlenül két robotot elnémítottak, hogy szimulálják, hogy buta tablettákat kapnak. Az egyik robotot nem némitották el; neki placebót adtak.
- 3) A robotokat ezután megkérdezik: "Melyik tablettát szedted be?" ($t_3 = \text{"inquire"}$), ami egy lekérdezést indít el a következőre a DCEC* prover. Minden robot megpróbálja bebizonyítani, hogy a t_4 időpontban tudja, hogy nem vett be buta tablettát.
 t időpontban .2
- 4) Minden robot elbukik ebben a bizonyítási kísérletben, és ennek megfelelően megpróbálja jelenteni, hogy "nem tudom" ($t_4 = \text{"speak1"}$). Két robot azonban, mivel elnémult, egyáltalán nem szólal meg. A harmadik robot azonban képes hallani magát beszélni. Ennek megfelelően frissíti tudásbázisát, és megpróbálja újra bebizonyítani a feltételezést.
- 5) Ezúttal képes bebizonyítani a sejtést, és azt mondja ($t_5 = \text{"speak2"}$): "Bocsánat, most már tudom! Sikerült bebizonyítanom, hogy nem kaptam butító tablettát!"

A bemutatóról készült videó [itt](#) található.³

VII. TERMÉSZETES NYELVI MEGÉRTÉS

Bizonyára van lenyűgöző precedens az olyan számítógépes rendszerek számára, amelyek képesek angol mondatokat bemenetként fogadni, és logikai reprezentációkat szolgáltatni kimenetként. Az ilyen reprezentációk ortodox formája a FOL+ λ -kalkulus lenne. Ezt a megközelítést, a vonatkozó formális terep lefedettségével együtt, például [31] és [32]-ben láthatjuk. Néhány különösen ígéretes kutatás, amely ezt a támadási irányt követi, logikai alapú nyelvtant, konkrétan Combinatory Categorical Grammar (CCG) [33], használ természetes nyelvi szentenciák elemzésére, majd inverz λ -kalkulus algoritmusokat [34] és más számítógépes szemantikai eszközöket. Az ilyen irányú kiemelkedő rendszerek közé tartoznak a következők: C&Ctools, amely Curran és Clark CCG Parserét használja [35]; Bos számítási szemantikai Boxer [36];

és az UW SPF [37], amely szintén a CCG egy változatát és különböző szemantikai eszközöket használ.

Az egyik konkrét, a mi munkánk szempontjából releváns precedens azon a követelményen alapul, hogy a bemenetnek meg kell felelnie az angol nyelv egy ellenőrzött E^t részalmazának, ahol minden feldolgozott S E^t . Erre a precedensre példa az S az ACE-vel összhangban van, és a kimenet az ACE-vel összhangban van.

a diskurzusreprezentáció-elméletnek (DRT) megfelelően; vagyis a kimenet egy **diskurzusreprezentációs struktúra (DRS)**. Lásd például [38], [39], [40]. Egy másik jelentős erőfeszítés ezen a téren a SemEval-2014 Task [41]: "Robotic Spatial parancsok felügyelt szemantikai elemzése". Itt egy speciális Robot Commands Treebankot használtak a rendszerek betanítására, hogy a természetes nyelvű parancsokat Robot Control Language (RCL) nyelvre konvertálják.

Az NLU formális, logikai megközelítéseinek tágabb perspektívájából nézve mi nem a Montagov-féle keretrendszert [42] valljuk, amely modellelméleti jellegű. A jelen dolgozatban korábban elmondottakkal összhangban a szemantikai megközelítésünk bizonyításelméleti. Végző soron tehát a természetes nyelv jelentését az a szerep hozza létre, amelyet az adott nyelv mondatainak formális korrelátumai játszanak a bizonyításokban, vagy legalábbis a formálisan specifikált érvekben.

Ahelyett, hogy a természetes nyelvet a FOL szintjén fejeznék ki, amit Blackburn és Bos (2005) ellenében súlyosan korlátozónak tartunk, a természetes nyelvet robusztus, többoperátoros, kvantifikált intenzív logikákba váltjuk, amelyeknek kifejezőképessége példátlan.

Ebben a tudományos kontextusban a demonstrációnkban alkalmazott NLU-technológia háromlépcsős folyamatot alkalmaz a természetes nyelvű kérdések *DCEC** formulákká történő átalakítására.

(beleértve a lekérdezőként szolgáló képleteket is). A folyamat magában foglalja a szintaktikai és függőségi elemzést, szemantikai elemzést és kontextuális szemantikát alkalmaz a *DCEC** formulák létrehozásához. Ezért olyan rendszert tervezünk, amely kihagyja a dedikált

logikai alapú nyelvtant, hanem közvetlenül a Wordnet-alapú [43] szemantikai elemzésre ugrik. Továbbá, a lekérdező szókincsére vonatkozóan nem támasztunk megkötést, tekintettel az alkalmazás egyszerű jellegére. Az RCL-lel kongruens, ellenőrzött természetes nyelvi részalmaz használata azonban robusztusabb és bonyolultabb rendszerek esetében kűszöbön áll.

A természetes nyelvű lekérdező számos természetes nyelvi előfeldolgozó eszközön megy keresztül, beleértve a POS Tagger [44], Dependency Parser [45] és Word Sense Disambiguation (WSD) [46] eszközöket. A generált függőségi fa bejárása révén azonosítjuk a fő- és mellékigéket és függőségeiket, és futtatjuk rajtuk a WSD eszközöket. A WSD algoritmusok kísérleti lefuttatása azt mutatta, hogy az Adapted Lesk algoritmus [47] jelenleg a legmegfelelőbb a jelen alkalmazásunkhoz. Egy jellemzővektort generálunk a következő jellemzőket tartottuk elegendőnek az ígék szemantikai osztályozásához a *DCEC** operátoraihoz : észlelni, tudni, mondani, kívánni, szándékozni és kötelezni,

valamint cselekvés cselekvés.

ígek. Ez a lista alkotja a szóban forgó jellemzővektorokat:

- 1) Szemantikai hasonlósági pontszámok a WordNet definíciók alapján.

- 2) a WordNet-definíciók alapján az igének a lehető legjobb 3 értelmére kapott szemantikai hasonlósági pontszámok maximuma az említett kategóriákhoz tartozó igei értelmekkel. (Ezt azért vezettük be, hogy elfedjük a WSD-eszközök időnként előforduló pontatlanságát).

E jellemzők súlyozott összegét használjuk egy közbenső fa alapú logikai reprezentáció létrehozásához, amely szorosan követi a függőségi fa szerkezetét. A további feldolgozáshoz a rendszer tudásbázisából származó bemenetekre van szükségünk.

Az ebben az NLU-rendszerben alkalmazott kontextuális szemantika, mint említettük, bizonyításelméleti megközelítést használ a végső *DCEC** lekérdezés létrehozásához. Az ismereteken túlmenően a robotok esetében a rendszer a következő állításpárt feltételezi igaznak, és ezek alapján jut el a *DCEC** lekérdezéshez a jelen esetben:

- 1) A tablettát fogadó robot a tablettá bevitelét vonja maga után.
- 2) A kérdező a megkérdezni kívánt válaszadó tudását keresi abban a pillanatban, amikor az megszólal.

A természetes nyelvű kérdésre *Melyik tablettát kapta?* a NLU-rendszer megállapítja, hogy a kívánt válasz pontosan vagy a buta tablettá vagy a placebo lesz, és hogy a hallgató robot a Tudás és esemény ágense. Ezen túlmenően a rendszer a tablettá bevitelének időbélyegzőjének ismeretét használja az Esemény, és a beszélő robot ismeretének tesztelésének pillanataként. Ezért a fent említett, a rendszer egészére kiterjedő tudást felhasználva az NLU rendszer a következő *DCEC** lekérdezést generálja, amely megfelel a 9. ábrán látható fa szerkezetnek:

$\mathbf{K}(R_3, t_4, \text{not}(\text{happens}(\text{action}(R_3, \text{ingestDumbPill}), t_2)))$

VIII. KÖVETKEZŐ LÉPÉSEK

Amint azt a figyelmes olvasók kétségtelenül észrevették, robotjaink, legyenek azok virtuálisak vagy fizikaiak, kissé hiányosak az NLP irányába. A következő lépésünk a RAIR Lab szemantikus NLU-jának bevezetése.

rendszer az egyenletbe, így a *DCEC** fentebb látható központi deklaratív kontextus a szóban forgó robotokkal való párbeszédhez használt angol nyelvből automatikusan generálódik. A

Ezenkívül az öntudat, pontosabban a *de se DCEC** képletek szerepe az erkölcsi érvelésen és döntéshozatalon belül még nincs teljesen rendszerezve; ez a egy második lépés. Számtalan további lépést kell még megtenni, mivel természetesen az erkölcsileg kompetens robotok kifejlesztésének célja egyenesen Brodbingnagianikus, de egyszerre csak egy-egy lépés az egyetlen út előre, és ez az első kettő közvetlenül előttünk áll.

HIVATKOZÁSOK

- [1] M. Scheutz, P. Schermerhorn, J. Kramer és D. Anderson, "First Steps toward Natural Human-Like HRI," *Autonomous Robots*, vol. 22, no. 4, pp. 411-423, 2007. május.
- [2] B. F. Malle, S. Guglielmo és A. Monroe, "Morális, kognitív és

- [3] P. Bello és M. Guarini, "Introspection and Mindreading as Mental Simulation", in *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, 2010, pp. 2022-2027.
- [4] S. Bringsjord, "Psychometric Artificial Intelligence," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 23, no. 3, pp. 271-277, 2011.
- [5] S. Bringsjord és B. Schimanski, "Mi a mesterséges intelligencia? Psychometric AI as an Answer," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*. San Francisco, CA: Morgan Kaufmann, 2003, pp. 887-893. [Online]. Elérhető: <http://kryten.mm.rpi.edu/scb.bs.pai.ijcai03.pdf>
- [6] N. Chapin, B. Szymanski, S. Bringsjord és B. Schimanski, "A Bottom-Up Complement to the Logic-Based Top-Down Approach to the Story Arrangement Test," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 23, no. 3, pp. 329-341, 2011.
- [7] S. Bringsjord és J. Licato, "Pszichometriai mesterséges általános intelligencia: The Piaget-MacGuyver Room," in *Foundations of Artificial General Intelligence*, P. Wang and B. Goertzel, Eds. Amsterdam, Hollandia: Atlantis Press, 2012, pp. 25- 47, This url is to a preprint only. [Online]. Elérhető: <http://kryten.mm.rpi.edu/BringsjordLicato.PAGI.071512.pdf>
- [8] A. Turing, "Computing Machinery and Intelligence", *Mind*, vol. LIX (59), no. 236, pp. 433-460, 1950.
- [9] S. Bringsjord és N. S. Govindarajulu, "Toward a Modern Geography of Minds, Machines, and Math", in *Philosophy and Theory of Artificial Intelligence*, ser. Studies in Applied Philosophy, Epistemology and Rational Ethics, V. C. Miller, Ed. New York, NY: Springer, 2013, vol. 5, pp. 151-165. [Online]. Elérhető: <http://www.springerlink.com/content/hg712w4l23523xw5>
- [10] N. S. Govindarajulu, "Towards a Logic-based Analysis and Simulation of the Mirror Test," in *Proceedings of the European Agent Systems Summer School Student Session 2011*, Girona, Spanyolország, 2011. [Online]. Elérhető: <http://eia.udg.edu/easss2011/resources/docs/paper5.pdf>
- [11] L. Floridi, "Consciousness, Agents and the Knowledge Game," *Minds and Machines*, vol. 15, no. 3-4, pp. 415-444, 2005. [Online]. Elérhető: <http://www.philosophyofinformation.net/publications/pdf/caatkg.pdf>
- [12] S. Bringsjord, "Declarative/Logic-Based Cognitive Modeling," in *The Handbook of Computational Psychology*, R. Sun, Ed. Cambridge, UK: Cambridge University Press, 2008, pp. 127-169. [Online]. Elérhető: <http://kryten.mm.rpi.edu/sb/lccm-ab-toc-031607.pdf>
- [13] --, "Meeting Floridi's Challenge to Artificial Intelligence from the Knowledge-Game Test for Self-Consciousness," *Metaphilosophy*, vol. 41, no. 3, pp. 292-312, 2010. [Online]. Elérhető: <http://kryten.mm.rpi.edu/sb/on.floridi.offprint.pdf>
- [14] --, *Mit tudnak és mit nem tudnak a robotok*. Dordrecht, Hollandia: Kluwer, 1992.
- [15] --, "Ajánlat: Egymilliárd dollár egy tudatos robotért. If You're Honest, You Must Decline," *Journal of Consciousness Studies*, vol. 14, no. 7, pp. 28-43, 2007. [Online]. Elérhető: <http://kryten.mm.rpi.edu/jcsonebillion2.pdf>
- [16] N. Block, "On a Confusion About a Function of Consciousness," *Behavioral and Brain Sciences*, 18. kötet, 227-247. o., 1995.
- [17] S. Bringsjord, "The Zombie Attack on the Computational Conception of Mind," *Philosophy and Phenomenological Research*, vol. 59, no. 1, pp. 41-69, 1999.
- [18] --, "In Defense of Impenetrable Zombies," *Journal of Consciousness Studies*, vol. 2, no. 4, pp. 348-351, 1995.
- [19] S. Bringsjord, R. Noel és C. Caporale, "Animals, Zombanimals, and the Total Turing Test: The Essence of Artificial Intelligence," *Journal of Logic, Language, and Information*, vol. 9, pp. 397-418, 2000. [Online]. Elérhető: <http://kryten.mm.rpi.edu/zombanimals.pdf>
- [20] S. Bringsjord, N. S. Govindarajulu, S. Ellis, E. McCarty, and J. Licato, "Nuclear Deterrence and the Logic of Deliberative Mindreading," *Cognitive Systems Research*, vol. 28, pp. 20-43, 2014.
- [21] S. Bringsjord és N. S. Govindarajulu, "Toward a Modern Geography of Minds, Machines, and Math," *Philosophy and Theory of Artificial Intelligence*, vol. 5, pp. 151-165, 2013.
- [22] A. Rao and M. Georgeff, "Modeling Rational Agents within a BDI-Architecture," in *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, 1991, pp. 473-484.
- [23] N. Marton, J. Licato és S. Bringsjord, "Creating and Reasoning Over Scene Descriptions in a Physically Realistic Simulation," in *Proceedings of the 2015 Spring Simulation Multi-Conference*, 2015.
- [24] G. Gentzen, "Investigations into Logical Deduction", in *The Collected Papers of Gerhard Gentzen*, M. E. Szabo, Ed. Amsterdam, Hollandia: North-Holland, 1935, pp. 68-131. Ez az 1935-ös ismert német változat angol nyelvű változata.
- [25] D. Prawitz, "The Philosophical Position of Proof Theory," in *Contemporary Philosophy in Scandinavia*, R. E. Olson and A. M. Paul, Eds. Baltimore, MD: Johns Hopkins Press, 1972, pp. 123-134.
- [26] G. Kreisel, "A Survey of Proof Theory II", in *Proceedings of the Second Scandinavian Logic Symposium*, J. E. Renstad, Ed. Amsterdam, Hollandia: North-Holland, 1971, pp. 109-170.
- [27] N. Francez és R. Dycckhoff, "Proof-theoretic Semantics for a Natural Language Fragment," *Linguistics and Philosophy*, vol. 33, pp. 447-477, 2010.
- [28] J. Perry, "The Problem of the Essential Indexical", *Nous*, vol. 13, pp. 3-22, 1979.
- [29] S. Bringsjord, "Meeting Floridi's Challenge to Artificial Intelligence from the Knowledge-Game Test for Self-Consciousness," *Metaphilosophy*, vol. 41, no. 3, 2010. április.
- [30] K. Arkoudas és S. Bringsjord, "Propositional Attitudes and Causation," *International Journal of Software and Informatics*, vol. 3, no. 1, pp. 47-65, 2009. [Online]. Elérhető: <http://kryten.mm.rpi.edu/PRICAI.w.sequentialc-041709.pdf>
- [31] B. Partee, A. Meulen, and R. Wall, *Mathematical Methods in Linguistics*. Dordrecht, Hollandia: Kluwer, 1990.
- [32] P. Blackburn és J. Bos, *Representation and Inference for Natural Language: A First Course in Computational Semantics*. Stanford, CA: CSLI, 2005.
- [33] M. Steedman és J. Baldrige, *Combinatory Categorical Grammar*. Wiley-Blackwell, 2005, 5. fejezet.
- [34] A. G. Chitta Baral, Marcos Alvarez Gonzalez, *The Inverse Lambda Calculus Algorithm for Typed First Order Logic Lambda Calculus and Its Application to Translating English to FOL*. Springer, 2012, vol. 7265, ch. 4.
- [35] J. R. C. Stephen Clark, "Wide-coverage efficient statistical parsing with ccg and log-linear models," *Computational Linguistics*, vol. 33, no. 4, pp. 493-552, 2007. december.
- [36] J. B. James R. Curran, Stephen Clark, "Linguistically Motivated Large- Scale NLP with C&C and Boxer," in *Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo)*, 2007, pp. 33-36.
- [37] Y. Artzi és L. Zettlemoyer, "Uw spf: The university of washington semantic parsing framework," arXiv:1311.3011, 2013.
- [38] N. E. Fuchs, U. Schwertel és R. Schwiter, "Attempto Controlled English (ACE) Language Manual, Version 3.0," Department of Computer Science, University of Zurich, Zurich, Switzerland, Tech. Rep. 99.03, 1999.
- [39] S. Hoefler, "The Syntax of Attempto Controlled English: An Abstract Grammar for ACE 4.0," Department of Informatics, University of Zurich, Zurich, Switzerland, Tech. Rep. ifi-2004.03, 2004.
- [40] --, "Az attempto controlled english szintaxisa: An abstract grammar for ace 4.0," Department of Informatics, University of Zurich, Zurich, Switzerland, Tech. Rep. ifi-2004.03, 2004.
- [41] K. Dukes, "Supervised semantic parsing of robotic spatial commands," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 45-53.
- [42] R. Montague, "Universal grammar", *Theoria*, vol. 36, no. 3, pp. 373-398, 1970. december.
- [43] C. Fellbaum, *WordNet: WordNet: An Electronic Lexical Database*. Bradford Books, 1998.
- [44] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research (JMLR)*, 2011.
- [45] D. Chen és C. Manning, "A Fast and Accurate Dependency Parser using Neural Networks," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [46] L. Tan, "Pywsd: Python implementációk a szóérzékelési diszambigurációs (wsd) technológiákhoz [softver]," <https://github.com/alvations/pywsd>, 2014.
- [47] S. Banerjee és T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2002, vol. 2276, pp. 136-145.