

A kölcsönös elmélet felé az ember és az AI közötti interakcióban: Hogyan tükrözi a nyelv, hogy a diákok mit érzékelnek? A virtuális tanársegédről

Qiaosi Wang
Georgia Institute of Technology
Atlanta, GA, USA
qswang@gatech.edu

Koustuv Saha
Georgia Institute of Technology
Atlanta, GA, USA
koustuv.saha@gatech.edu

Eric Gregori
Georgia Institute of Technology
Atlanta, GA, USA
egregori3@gatech.edu

David A. Joyner
Georgia Institute of Technology
Atlanta, GA, USA
david.joyner@gatech.edu

Ashok K. Goel
Georgia Institute of Technology
Atlanta, GA, USA
ashok.goel@cc.gatech.edu

ABSZTRAK

T

Olyan társalgási ügynökök létrehozása, amelyek képesek természetes és hosszan tartó beszélgetéseket folytatni, komoly technikai és tervezési kihívást jelentett, különösen a közösségi társalgási ügynökök esetében. A kölcsönös elméletet elméleti keretként javasoljuk a természetes, hosszú távú ember-intelligencia interakciók tervezéséhez. Ebből a perspektívából kiindulva önbevallásos felmérések és számítógépes nyelvészeti megközelítés segítségével vizsgáljuk meg, hogy egy közösség hogyan érzékeli egy kérdéseket megválaszoló társalgási ügynök működését az online oktatás kontextusában. Először is megvizsgáljuk a diákok Jill Watson (JW), egy virtuális tanársegéd hosszú távú időbeli változásait, akit egy online osztály vitafórumában vetettek be. Ezután megvizsgáljuk, hogy a tanulók JW-ről alkotott elképzeléseit a tanulók és a JW közötti párbeszédéből kinyert nyelvi jellemzők segítségével meg lehet-e állapítani. Azt találtuk, hogy a diákok JW antropomorfizmusáról és intelligenciájáról alkotott képük az idő múlásával jelentősen változott. A regressziós elemzések azt mutatják, hogy a nyelvi szóbeliség, az olvashatóság, az ~~szóhasználat~~ sokszínűség és az alkalmazkodóképesség tükrözi a diákok JW-ről alkotott képét. Megvitatjuk az adaptív, közösséget célzó társalgási ágensek hosszú távú társként való felépítésének és az ember és az AI közötti interakcióban a kölcsönösségi elmélet tervezésének következményeit.

CCS KONCEPCIÓK

- Számítástechnikai módszertanok → Mesterséges intelligencia; - Emberközpontú számítástechnika → Természetes nyelvi interfészek; *Empirikus tanulmányok az együttműködő és társas számítástechnikában*; *Közösségi média*; - Alkalmazott számítástechnika → Pszichológia.

KULCSSZAVAK

társalgási ügynök, online közösség, ember-AI interakció, elmélet, nyelvi elemzés, online oktatás, online oktatás

A mű egészének vagy egy részének digitális vagy nyomtatott másolata személyes vagy tantermi használatra díjmentesen engedélyezhető, feltéve, hogy a másolatok nem nyereségvágyból vagy kereskedelmi előnyökért készülnek, és a másolatokon szerepel az a közlemény és a teljes idézet az első oldalon. A szerző(k)n kívül más

ACM referencia formátum:

tulajdonában lévő alkotórészek szerzői jogait tiszteletben kell tartani. Az absztraktok feltüntetése megengedett. Egyéb másoláshoz, újraközléshez, szervereken való közzétételhez vagy listákon való terjesztéshez előzetes külön engedély és/vagy díj szükséges. Az engedélyeket a permissions@acm.org címen kell kérni.

CHI '21, 2021. május 8. és 13., Yokohama, Japán

Qiaosi Wang, Koustuv Saha, Eric Gregori, David A. Joyner és Ashok K. Goel. 2021. A kölcsönös tudatelmélet felé az ember és az AI közötti interakcióban: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*, May 8-13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3411764.3445645>.

1 BEVEZETÉS

A beszélgetőgynökök (CA-k)¹ egyre inkább beépülnek életünk különböző területeire, és szolgáltatásokat nyújtanak az egészségügy, a szórakoztatás, a kiskereskedelem és az oktatás területén. Míg a CA-k viszonylag sikeresek a feladatorientált interakciókban [82, 96], a kezdeti ígéret, hogy olyan CA-kat kell létrehozni, amelyek képesek természetes és koherens beszélgetéseket folytatni a felhasználókkal, mind a tervezési, mind a technikai kihívások miatt nagyrészt beteljesületlenül maradt [3, 18, 87]. Ez a felhasználói elvárások és a CA-kkal kapcsolatos tapasztalatok közötti szakadék [61] a felhasználók folyamatos frusztrációjához, a beszélgetések gyakori megszakadásához és a CA-k esetleges elhagyásához vezetett [3, 61, 98].

A felhasználókkal való zökkenőmentes beszélgetések © 2021 A szerzői jog a tulajdonos/szerző(k) tulajdonában van. A kiadási jogokat az ACM ACM ISBN 978-1-4503-8096-6/21/05. . . \$15.00 <https://doi.org/10.1145/3411764.3445645>

lebonyolítása még fontosabbá válik, ha a hitelesítésszolgáltatókat online közösségekben alkalmazzák, különösen a sérülékeny csoportokat, például online egészségügyi támogató csoportokat [71] és diákközösségeket [95]. Ezek a közösséggel szembenező hitelesítésszolgáltatók gyakran a közösség kritikus részeként biztosítják a zökkenőmentes interakciókat a közösség tagjai között, és hosszú távú információs és érzelmi támogatást nyújtanak. Ezek a közösséggel szembenező CA-k azonban két egyedi kihívással szembesülnek: a közösség egyes tagjaival való zökkenőmentes diadikus interakciók elvégzésének szükségessége, valamint a közösség változó percepciói alapján történő megfelelő reagálás szükségessége [53, 86]. Valójában a hitelesítésszolgáltató közösségre irányuló jellege új összetettséget ad - az egyes tagokkal folytatott minden egyes diadikus interakció látható a közösség többi tagja számára, ami nemcsak a közösségnek a hitelesítésszolgáltatóról alkotott képét változtathatja meg, hanem hatással lehet a közösség többi tagjára is, azaz az egyik személlyel való nem kielégítő interakció a többieket is frusztrálhatja [42]. Az emberek azonban képesek arra, hogy méltóságteljesen zökkenőmentes interakciókat folytassanak egymással, és a közösségnek megfelelően viselkedjenek.

kapta. _____
¹ Hacsak másképp nem jelezzük, ebben a tanulmányban a CA-kat kifejezetten a szövegalapú, szöveges beszélgetőgynökökre használjuk.

elvárások és normák egyszerre. Ez a folyamat egy egyedülállóan humánus tulajdonságon alapul, amelyet *Theory of Mind* [7, 12, 78]. A tudósok azt állítják, hogy az **elmélet (Theory of Mind, ToM) egy alapvető kognitív és szociális tulajdonság, amely lehetővé teszi számunkra, hogy megfigyelhető vagy látnis viselkedési és verbális jelzéseken keresztül következtetéseket vonjunk le egymás elméjéről** [6, 12, 37, 38, 94]. Ez a tulajdonság spontán módon irányítja a megértésünket arról, hogy hogyan érzékeljük egymást a társas interakciók során. Ez lehetővé teszi számunkra, hogy olyan szociális technikákat alkalmazzunk, mint például megjelenésünk és viselkedésünk kiigazítása annak érdekében, hogy mások rólunk alkotott elképzeléseit önreprezentációnk alapján igazítsuk [36]. A tipikus ember-ember interakciókban a **kölcsönös tudatelmélet (MToM)** megléte, **vagyis az interakciókban részt vevő minden fél rendelkezik a ToM-mel**, a viselkedési visszajelzéseken keresztül közös elvárásokat épít ki egymásról, ami segít bennünket abban, hogy konstruktív és koherens beszélgetéseket folytassunk [36, 75]. A MToM-et egyre gyakrabban használják elméleti keretként az emberközpontú mesterséges intelligencia, például a robotok tervezéséhez, amelyeket természetesebbnek és intelligensebbnek lehet érzékelni. az emberi partnerekkel való együttműködés során [26, 57, 59, 75].

Míg az MToM befolyásolja az emberközpontú mesterséges intelligencia tervezését a feladatorientált interakciókban, az ember és mesterséges intelligencia közötti kommunikatív interakciók tervezésében betöltött szerepe még mindig feltáratlan. Az ember és mesterséges intelligencia közötti interakciók tervezésének meglévő megközelítései szintén nem rendelkeznek elméleti kerettel és egységes tervezési irányelvekkel az emberközpontú mesterséges intelligencia tervezéséhez, különösen a kommunikatív interakciókban. Következésképpen a kutatók és a tervezők a hagyományos, grafikus felhasználói felületekre szánt HCI tervezési irányelvekhez fordulnak, ami nem mindig optimális perspektíva az emberek és a gyakran antropomorfizált CA-k közötti interakciók tervezéséhez [89] - a kutatók és a tervezők komoly akadályokba ütköznek az irreálisan magas felhasználói elvárások [61] kiegyensúlyozásában, miközben megfelelő mennyiségű szociális jelzést biztosítanak a hosszú távú természetes interakciók elősegítéséhez [56].

Az ember-ember közötti interakciók analógiájára javasoljuk a *MToM irányába történő tervezést, mint olyan elméleti keretet, amely az adaptív, közösséggel szembenező CA-k tervezését irányítja, amelyek képesek a felhasználók változó felfogásának és igényeinek kielégítésére. Az első lépés az MToM kiépítése felé az ember-CA kommunikációban tehát az, hogy a CA-kat a ToM analógiájával látjuk el, amely képes automatikusan azonosítani a felhasználóknak a CA-kkal kapcsolatos percepcióit*. Ezzel a képességgel a CA-k képesek lennének nyomon követni a felhasználók változó felfogását, és ennek megfelelően finom viselkedési jelzésekkel segíteni a felhasználókat abban, hogy jobb mentális modellt építsenek fel a CA képességéről. Ez segítene enyhíteni a felhasználók jelenlegi egyoldalú kommunikációs terheit is, akiknek folyamatosan módosítaniuk kellett a CA-ról alkotott mentális modelljüket egy önkényes próba-hiba folyamaton keresztül, hogy kiváltsák a kívánt CA-válaszokat [4, 9].

A kutatások a CA-k felhasználói észlelésének azonosítása mentén vizsgálták a dyadi ember-AI interakciók megkönnyítése érdekében, beleértve az egyén CA-król alkotott mentális modelljének vizsgálatát különböző kontextusokban [31, 54, 61]. Ezek a tanulmányok, amelyek többsége minőségi jellegű, nem csak nehezen skálázhatók, hanem közvetlenül megvalósítható

algoritmikus eredmények is hiányoznak, amelyek integrálhatók a CA-architektúrába a CA-val kapcsolatos felhasználói észlelés automatikus felismerése érdekében. A közösséggel szembenező CA-k esetében, amelyekről ismert, hogy az online közösségekben változó társadalmi szerepeket töltenek be [87], jelenleg nincs világos képünk arról, hogy a CA-k közösségi megítélése hogyan változik az idő múlásával, és hogy az emberek és a CA-k közötti, közösségi környezetben zajló dyadikus interakciók feltárnak-e bármilyen, a felhasználói megítéléssel kapcsolatos jelet.

Ezért megállapítjuk, hogy az elméletben és a gyakorlatban hiányosságok mutatkoznak a közösségi CA-kkal kapcsolatos emberi észlelések automatikus és skálázható megértésében, mind egyéni, mind kollektív szinten. Az ember-ember közötti interakciók ~~erősen~~ támaszkodva ez a tanulmány az első lépést vizsgálja a hosszú távú ember-CA interakciók MToM tervezésének irányába, megvizsgálva a közösséggel szembenező CA-k ToM-jének kialakíthatóságát. Konkrétan két kutatási kérdést célzunk meg:

1. **kérdés:** Hogyan változik egy közösség megítélése a közösséget érintő CA-ról az idő múlásával?
2. **kérdés:** Hogyan tükrözik az ember és az AI közötti interakció nyelvi markerei a közösséggel szembenező CA-ról alkotott képet?

Ezeket a kutatási kérdéseket az online tanulás kontextusában vizsgáljuk, ahol a közösségi CA-kat általában úgy tekintik, hogy információs és szociális támogatást nyújtanak a hallgatói közösségeknek [1, 92, 95]. A félév során 10 héten keresztül egy **Jill Watson** [24, 34, 35] (röviden: JW) nevű, közösség-közeli kérdés-válaszoló (QA)CA-t (QA CA) vetettünk be egy online vitafórumban, hogy válaszoljon a diákok kérdéseire. A további elemzéshez összegyűjtöttük a hallgatók kéthetenkénti önbevalláson alapuló észleléseit és a JW-vel folytatott beszélgetéseket. ~~Megítik~~ hallgatói közösség JW-vel kapcsolatos hosszú távú percepciójának változásait, és megvizsgáljuk a JW-vel kapcsolatos, a hallgatók által önbevallásuk szerint érzékelt percepciók és a hallgatók és a JW közötti beszélgetések nyelvi jellemzői, például a szóbeliség, az alkalmazkodóképesség, a sokszínűség és az olvashatóság közötti kapcsolatot. A nyelvi jellemzők és a hallgatók JW-ről alkotott elképzelései közötti regressziós elemzések olyan tanulságos eredményeket tárnak fel, mint például az olvashatóság, az érzelmek, a változatosság és az alkalmazkodóképesség, amelyek pozitívan változnak a kívánatos elképzelésekkel, míg a szóbeliség negatívan változik.

Hozzájárulásunk hármas: *Először is*, az MToM-ot javasoljuk elméleti keretként az online közösségeken belüli tartós ember és MI közötti interakció megtervezéséhez. *Másodszor*, munkánk mélyebb megértést nyújt arról, hogy egy közösségnek a közösséggel szembenező (QA)CA-ról alkotott képzele hogyan ingadozik hosszú távon. *Harmadszor*, empirikus bizonyítékot szolgáltatunk a számítógépes nyelvészeti megközelítés kihasználásának lehetőségeiről, hogy a közösséggel szembenező CA közösségi percepciójára következtetni lehessen a ~~közös~~ ~~hi~~ felhalmozott nyilvános dyadikus interakciók révén. Megvitatjuk munkánk következményeit az adaptív, közösségre irányuló (QA)CA-k tervezésében az elméletvezérelt komputációs nyelvészeti megközelítéseken keresztül, ahol a végső célunk a természetes, hosszú távú ember-AI interakciók kialakítása.

Adatvédelem, etika és nyilvánosságra hozatal.

Elköteleztük magunkat a tanulmányban felhasznált tanulói adatok védelmének biztosítása mellett. Ezt a tanulmányt a Georgia Tech intézményi felülvizsgálati bizottsága (IRB) hagyta jóvá. A felmérés és a vitafórum adatait (csak a diák-JW interakciókra korlátozva) a diákok beleegyezésével gyűjtöttük, és az adatokat anonimizáltuk. Az egyes felmérések kitöltéséért plusz kreditpontokat ajánlottunk fel a hallgatóknak, és bónusz plusz kreditpontokat, ha a hat felmérésből legalább ötöt kitöltöttek. Ez a munka az osztályfőnökkel együttműködve történt, és intézkedéseket tettünk a kényszerítés elkerülése érdekében. A diákok a részvétellel legfeljebb a teljes osztályzat

1%-ánál kevesebb extra kreditet szerezhettek, és ezeket az extra krediteket a szokásos tanórai struktúra részeként más módon is ki lehetett érdemelni. Tisztáztuk a hallgatókkal, hogy a felmérésre adott válaszokat nem osztjuk meg az oktatóval, és azok semmilyen hatással nem lesznek az osztályzatra.

2 HÁTTÉR

Ebben a szakaszban áttekintést nyújtunk a ToM-ről és annak alkalmazásáról az ember és az AI közötti interakcióval kapcsolatos kutatásokban. Ezután megvitatjuk a kapcsolódó munkákat, amelyek a CA felhasználói észlelését vizsgálják az ember és az AI közötti interakciók megkönnyítése érdekében, és kiemeljük a nyelvi elemzés kihasználásának lehetőségét az ember és az AI közötti interakció javítására.

2.1 Az elmeelmélet az ember és az AI közötti

interakcióban A ToM, vagyis az a képességünk, hogy viselkedési jelzéseken keresztül feltételezéseket tegyünk mások elméjéről, alapvető fontosságú számos emberi társadalmi és kognitív interakcióban.

tív viselkedés, különösen az együttműködésre való képességünk a célorientált feladatok elvégzésére és a másokkal való zökkenőmentes, természetes ~~kommunikáció~~ való képességünkre [6, 7]. Például a közös tervek és célok kialakítása alapvető fontosságú az együttműködő feladatvégzéshez a ToM lehetővé teszi számunkra, hogy felismerjük és mérsékeljük egymás terveit és céljait a közös munka érdekében [5, 6]; a szándékos kommunikáció a zökkenőmentes kommunikáció alapja a ToM lehetővé teszi számunkra, hogy megértsük, hogy a beszélgetőpartner olyan meggyőződéssel vagy tudással rendelkezik, amely potenciálisan megváltoztatható, és így lehetővé teszi számunkra, hogy üzeneteinket ennek megfelelően állítsuk össze [5, 6]. A ToM nélkül a másokkal való természetes interakcióra való képességünk súlyosan ~~károsít~~ és a nyelvi érzékelés és produkció képességét kevésbé értelmessé teheti [587, 12]. A ToM-et széles körben tanulmányozták, és ~~tudat~~ is vezető szerepet játszik számos területen, például a kognitív tudományokban, a fejlődépszichológiában és az autizmus kutatásában [5, 7, 47, 78].

Az évek során a kutatók felismerték az emberközpontú robotok ToM-mel történő tervezésének fontosságát, hogy megkönnyítsék az ember-robot csapatokon belüli együttműködést a célorientált feladatokban. Konkrétan a ToM-et szándékosan építették be a rendszerarchitektúra önálló moduljaként, hogy segítsék a robotokat a világ állapotának és az emberi állapotnak a megfigyelésében [25], az emberi partner hipotetikus kognitív modelljeinek szimulációját, hogy figyelembe vegyék az eredeti tervektől eltérő emberi viselkedést [44, 79], és segítsék a robotokat a felhasználói meggyőződésekről, tervekről és célokról szóló mentális modellek felépítésében [43, 52]. A ToM-mel épített robotok pozitív eredményeket mutattak a csapatmunkában [25, 44], a kollaboratív döntéshozatalban [39], és természetesebbnek és intelligensebbnek érzékelik őket [59].

A ToM-et azonban még nem vizsgálták, mint olyan beépített tulajdonságot, amely potenciálisan lehetővé teszi a CA-k számára, hogy természetes módon kommunikáljanak az emberekkel. Tekintettel az emberi interakciókban betöltött alapvető szerepére és az ember-robot interakciókban eddig elért sikerére, azt állítjuk, hogy a ToM szerepét az emberközpontú CA-k tervezésében az emberekkel való kommunikatív interakciók során tovább kell vizsgálni. A ToM-mel rendelkező CA-k építése az első lépés a *kölcsönös* ToM tervezése felé az ember-CA interakciókban. A *kölcsönös* ToM alatt nemcsak azt értettük, hogy feltárjuk, hogyan segíthetjük a felhasználókat a CA-k jobb megértésében és mentális modelljének felépítésében (pl. magyarázható AI), hanem azt is, hogyan segíthetjük a CA-kat a felhasználó átfogó mentális modelljének felépítésében és iterálásában. Ebben a tanulmányban bemutatjuk a MToM építése felé irányuló kezdeti feltárásainkat az ember-CA

kommunikációban azáltal, hogy megvizsgáljuk a CA ToM építésének megvalósíthatóságát az ember-CA beszélgetések automatikus nyelvi elemzése segítségével, hogy megértsük a CA-k felhasználói észlelését.

2.2 A társalgási ügynök felhasználói percepciója

A társalgási ügynökökről alkotott képünk határozza meg, hogyan lépünk velük kapcsolatba, és így döntő szerepet játszik az emberközpontú tervezés irányításában.

CA-k. Az emberek megítélése a hitelesítésszolgáltatókról sokrétű fogalom. Korábbi

a kutatások különböző beállításokban vizsgálták az emberek mentális modelljét a hitelesítésszolgáltatókról - egy kooperatív játékban az emberek mentális modellje a hitelesítésszolgáltatóról magában foglalhatja a globális viselkedést, a tudáselosztást és a helyi viselkedést [31]; az emberek érzékelése egy ajánlóügynökről a bizalomból, a hitelességből és az elégedettségéből áll [10]. Az emberek CA-król alkotott elképzelése döntő szerepet játszik abban, hogy hogyan lépnek kapcsolatba a CA-kkal [31], és így előfutárként szolgál a CA-k viselkedésével kapcsolatos elvárásaikhoz. Korábbi kutatások szerint a felhasználók hajlamosak magas elvárásokat támasztani a CA-kkal szemben [61], és így hajlamosak gyakori beszélgetésmegszakításokkal találkozni, ami végső soron a felhasználóknak a CA elhagyásához vezethet [58, 98]. A felhasználók CA-król alkotott elképzeléseinek felismerése és a megfelelő visszajelzés biztosítása, amely segít a felhasználóknak felülvizsgálni elképzeléseiket, ezért kritikus fontosságú a zökkenőmentes ember-CA interakciók kialakításában [9, 31, 41]. A CA-k felhasználói megítélésének megértése még fontosabbá válik az online tanulás kontextusában, ahol a CA-kat egyre gyakrabban használják arra, hogy kritikus szerepet játsszanak a diákok tanulási élményében. A különböző CA-kat úgy tervezték, hogy tanulási és szociális támogatást nyújtsanak a tanulóknak. Ezek közé tartoznak az intelligens oktatórendszerek, amelyek személyre szabott tanulási támogatást nyújtanak a tanulóknak [1, 46], a közösséget támogató CA-k, amelyek szinkron online előadásokat biztosítanak [95], és amelyek segítenek az online tanulók közötti szociális kapcsolatok kiépítésében [92]. Míg azonban ezek a CA-k hatékonyan segítik a diákok tanulási eredményeit, alig vizsgálták, hogy a diákok hogyan érzékelik ezeket a CA-kat az online osztályteremben. A diákok CA-ról alkotott elképzelései befolyásolhatják a diákok interakciós tapasztalatait a CA-val, és így potenciálisan befolyásolhatják a tanulási eredményeket.

online tanulási tapasztalataik [46].

A közösséggel szembenező hitelesítésszolgáltatók esetében a közösség a hitelesítésszolgáltatóról alkotott képének megértése nemcsak a közösségen belüli zökkenőmentes diadikus interakciók biztosítása szempontjából fontos, hanem a közösséggel szembenező hitelesítésszolgáltatók hosszú távú társakként való megtervezéséhez is elengedhetetlen. Számos kutatás azt sugallta, hogy a közösséggel szembenező CA-k változó társadalmi szerepekkel rendelkeznek, és ezért a közösség gyakran másként érzékeli őket. Például Seering és munkatársai [88] azt találták, hogy a CA társadalmi szerepe egy online tanulási közösségen belül a közösségen belül idővel a "függő"-ből a "társ"-ra változott; Kim és munkatársai [53] szintén kiemelték, hogy a CA-k a közösség dinamikájának idővel való alakulásával a részvétel ösztönzéséből a "társadalmi szervező"-re válhatnak. A CA-k közösségi megítélésének árnyaltabb értékelésére van szükség ahhoz, hogy a CA-k a változó társadalmi szerepük alapján megfelelően viselkedjenek.

A CA-k emberi észlelésével kapcsolatos kutatások többsége azonban kvalitatív módszereket használ az ügynök felhasználói észlelésének különböző dimenzióinak azonosítására [31, 67], vagy az ügynökkel való interakciót követően egyszeri értékelést ad a felhasználói észlelésről [30, 41, 51, 55, 88, 98]. Ezek a munkák, bár értékes meglátásokat és iránymutatásokat kínálnak az emberközpontú CA-k tervezéséhez, nehezen operacionalizálhatók és integrálhatók a CA rendszerarchitektúrájába. Az emberek CA-érzékelésének poszt-

hoc elemzése szintén kevésbé hatékony az interakciók során az emberek CA-érzékelésének árnyaltságának és folyékonyságának megragadásában.

A jelenlegi munka célja tehát a közösségek által a közösséggel szembenező hitelesítésszolgáltató hatóságokkal kapcsolatos közösségi megítélés hosszú távú változásainak vizsgálata. A CA közösségi megítélésének operacionalizálása érdekében azt is megvizsgáljuk, hogy a számítási nyelvészet felhasználásával lehetséges-e a CA közösségi megítélésének automatikus rögzítése.

2.3 Nyelvi elemzés az ember és az AI közötti interakció javítására

A nyelv mindenféle interakcióban fontos szerepet játszik, mégis ez a legfontosabb összetevője és gyakran az egyetlen összetevője az ember és a kommunikáció közötti interakcióknak. Az MToM tervezése az ember és a CA közötti interakciókban tehát nagymértékben függ az ember és a CA által a szöveges válaszok által közvetített információktól.

Annak érdekében, hogy a hitelesítésszolgáltatók időben, de nem tolatkodó módon megértsék a felhasználó mentális modelljét a hitelesítésszolgáltatókról, fontos megvizsgálni, hogy a nyelvből le lehet-e következtetni a felhasználóknak a hitelesítésszolgáltatókról alkotott elképzeléseire. Korábbi kutatások azt sugallták, hogy a nyelvi jelzések felhasználásával jelezni lehet az emberek CA-król alkotott elképzeléseit az ember és a CA közötti interakciók során. A kutatók a beszélgetési jelekből következtettek a felhasználóknak az ügynökkel szembeni érzelmeire [90], személyiségjegyeire [62], a beszélgetés megszakadásának jeleire [58, 99], udvariasságára [20]. Mégis, hogy a felhasználóknak a CA-ról alkotott teljes körű percepciója megkonstruálható-e a beszélgetésekből kinyert nyelvi jellemzők segítségével, még mindig nem vizsgálták.

Másrészt további kutatásokra van szükség annak feltárására, hogy a hitelesítésszolgáltató nyelvezte hogyan tudja közvetíteni a felhasználó felé a képességeit, és hogyan tudja segíteni a felhasználókat abban, hogy felülvizsgálják a hitelesítésszolgáltatóval kapcsolatos mentális modelljüket. Míg az emberek hajlamosak különböző szociális technikákat alkalmazni, például a megjelenés és a modor megváltoztatásával bizonyos benyomásokat kelteni [36], a megtestesült virtuális ügynökök és robotok képesek megváltoztatni az arckifejezésüket vagy a fizikai viselkedésüket [73], a hangalapú társalgási asszisztensek képesek megváltoztatni a hangszínüket [11], a testetlen, szöveg alapú CA-k számára a szöveg alapú nyelv az egyetlen módja annak, hogy a felhasználóknak közvetítsék képességeiket. A CA válaszai bizonyítottan befolyásolják a felhasználó javítási stratégiáit [9], és elősegítik a megnyerő beszélgetéseket [14, 33]. A CA válaszokban használt nyelvezetét tehát ki kell használni, hogy a felhasználóknak segítsen a CA jobb mentális modelljének felépítésében [9].

Azzal a végső céllal, hogy a nyelvi elemzés felhasználásával MToM-et építsünk az ember és a CA közötti interakciókban, ebben a munkában először is megvizsgáljuk, hogy az ember és a CA közötti beszélgetésekből kinyert nyelvi jellemzők segítségével hogyan lehet következtetni a CA felhasználói percepciójára. A CA-k beszélgetések során a felhasználók mentális modelljeinek automatikus konstruálása lehetővé teheti a CA-k számára, hogy pontosan megjósolják a felhasználók viselkedését, és megfelelő válaszokat adjanak a felhasználók irányítására, hogy a gördülékenyebb ember-CA beszélgetések megkönnyítése érdekében hangolják a men- tal modelljeiket.

3 VIZSGÁLATI TERV

A jelenlegi tanulmány célja, hogy megértse a közösség által a közösségi CA-ról alkotott közösségi észlelés longitudinális változásait, valamint a nyelvi markerek felhasználásának lehetőségét a közösségi CA-ról alkotott felhasználói észlelések levonására. E kérdések feltárása érdekében egy közösség-arcú (QA)CA-t telepítettünk Jill Watson (JW) néven egy online osztály vitafórumán, hogy a félév során válaszoljon a hallgatók osztálylogisztikai kérdéseire. Ezután kéthetente összegyűjtöttük a hallgatók JW-vel kapcsolatos percepcióit, és a félév során a hallgatók és a JW közötti beszélgetésekből nyelvi jellemzőket

vontunk ki (a részletes vizsgálati tervet lásd az 1. ábrán). Felméréseinket és nyelvi jellemzőinket azzal a céllal választottuk ki, hogy végül felépítsük a CA ToM-etD a felmérési intézkedéseket úgy terveztük, hogy a diákok JW-ről alkotott képét három dimenzióból mérjük: antropomorfizmus, intelligencia és szimpatikus; a nyelvi jellemzőket a korábbi szakirodalom azt sugallta, hogy potenciálisan tükrözhetik az emberek CA-ról alkotott képét. A következő szakaszokban részletesen tárgyaljuk őket.

A jelenlegi vizsgálatra a Georgia Tech online informatikai mesterképzésén (OMSCS) keresztül kínált online ember-számítógép interakciós osztályban került sor. Az osztályba a félév végén 376 hallgató vett fel. Egy különálló, szabványos osztályfelmérés alapján, amely a hallgatók demográfiai adatait kérdezte a félév elején (n=389, megjegyzendő, hogy néhány hallgató a félév vége előtt elhagyta az osztályt), 299 hallgató vallotta magát férfinak (76,86%), 87 hallgató vallotta magát nőnek (22,37%), három hallgató nem adta meg a nemét (0,77%). A hallgatók különböző korcsoportokban oszlottak meg: 61 hallgató 18 és 24 év közötti volt (15,68%), 236 hallgató 25 és 34 év közötti volt.

(60,67%), 65 diák 35 és 44 év közötti (16,71%), 22 diák 45-54 év közöttiek voltak (5,66%), két diák 55-54 év közötti volt. 64 éves (0,51%), és két diák 65 év feletti (0,51%), egy diák nem adta meg az életkorát. A legmagasabb iskolai végzettséget tekintve 311 hallgató jelentett alapidiplomát (79,95%), 56 hallgató mesterdiplomát (14,40%), 14 hallgató doktori fokozatot (pl. PhD, Ed.D.) (3,60%), hét hallgató pedig szakmai fokozatot (pl. M.D., J.D.) (1,80%), egy hallgató nem jelentette a legmagasabb megszerzett diplomáját.

3.1 A JW tervezése és végrehajtása

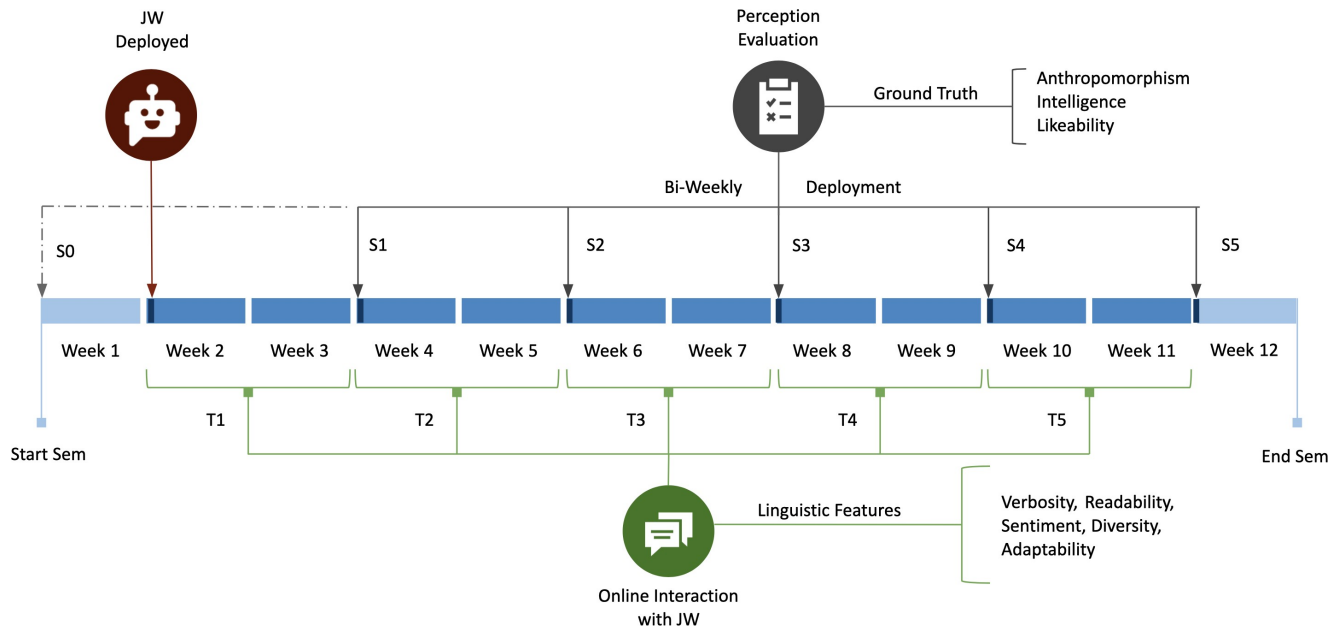
A JW egy ML-alapú kérdésmegoldó CA, amelyet arra terveztek, hogy megválaszolja a diákok osztálylogisztikával kapcsolatos kérdéseit. Három gépi tanulási modellt használ, mindegyik modell ugyanazokkal az adatokkal van betanítva. Amikor egy felhasználó kérdést tesz fel, a kérdést mindhárom modellnek továbbítja. A modellek végső kimenete egy előre programozott válasz kiválasztására szolgál (üdvözlés + releváns információk a tananyagban). A modelleket egy tudásbázisból generált gyakorló kérdésekkel képeztük. A tudásbázist egy tananyag ontológia és a tananyag felhasználásával hoztuk létre. A JW így *nem tud* idővel külső információkból (hallgatói válaszok vagy visszajelzések) tanulni. A JW hasonló korábbi verzióinak megvalósítási részletei megtalálhatók [Goel és Polepeddi \[35\]](#) munkájában.

A JW-t az osztály vitafórumán vetettük be az elején. a második héten (1. ábra). JW csak az erre a célra létrehozott JW threadsz-eken volt aktív, ahol JW minden egyes hozzászólást elolvasott és válaszokat adott, kizárólag az ezekben a témákban feltett kérdésekre. A diákokat arra ösztönözték, hogy az órával kapcsolatos kérdéseiket ebben a témakörben tegyék fel, ha választ akartak kapni JW-től. Annak érdekében, hogy a hallgatókat a félév során folyamatosan foglalkoztassuk, minden héten új JW-témát tettünk közzé a vitafórumon, és arra ösztönöztük a hallgatókat, hogy folyamatosan tegyenek fel kérdéseket JW-nek. [Az 1. táblázat](#) a hallgatók és JW között az osztály vitafórumán létrejött kérdés-válasz párosok példáit mutatja be.

Tanulmányunk során *szándékosan nem adtuk meg a hallgatóknak a JW munkamechanizmusát vagy képességeit, hogy ezek az információk ne befolyásolják a hallgatók JW-ről alkotott képét*. A diákoknak csak annyit mondtunk, hogy JW egy virtuális ügynök, aki válaszolhat az órával kapcsolatos kérdéseikre. A JW működési mechanizmusát és megvalósítását csak az összes felmérési adat összegyűjtése után fedték fel.

Annak feltárására, hogy a diákok JW-ről alkotott elképzeléseiben a félév során bekövetkezett változásokat feltárjuk, hat kéthetente végzett felmérést (lásd a [3. függelék 3. ábráját](#) az adaptált felmérési eszközhöz), hogy a diákok önbevallásukban a

4 RQ1: A DIÁKOK JW-VEL KAPCSOLATOS MEGÍTÉLÉSÉBEN BEKÖVETKEZETT VÁLTOZÁSOK VIZSGÁLATA



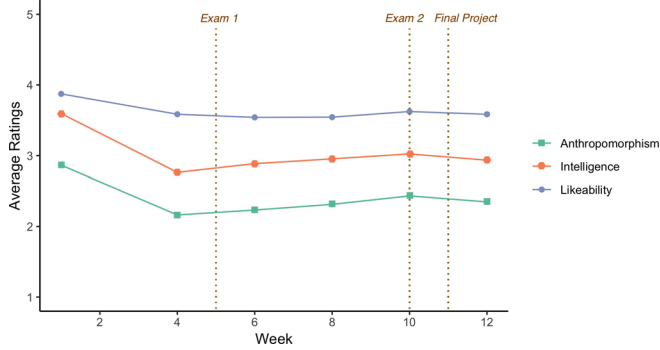
1. ábra: Tanulmányterv és idővonal. Az S0-S5 a felmérés adatait jelöli. A T1-T5 az osztályos vitafórumok adatainak felosztását jelenti a felmérés megosztásának idővonalára alapján. A regressziós elemzésben a felmérési adatokat alapigazságként használtuk, hogy a JW-vel való hallgatói interakciót minden egyes időkeretben megjelöljük. Például az S1-et használtuk a T1 fórumadatok megjelölésére, az S2-t a T2 megjelölésére, és így tovább.

JW felfogása. Az MToM elméleti keretrendszer által inspirálva szándékosan olyan észlelési mérőszámokat választottunk, amelyek megragadhatják a diákok holisztikus társadalmi észlelését a JW-ről, és potenciálisan tükrözhetik a JW észlelésének hosszú távú változásait, a CA-funkciók általánosan mért utólagos észlelései (pl. a válasz pontossága vagy helyessége) helyett. Különösen egy olyan validált felmérési eszközt adaptáltunk, amely az ember-robot interakciókban a robotok felhasználói észlelését méri [8], és amelyet korábban ember-CA interakciókban is alkalmaztak [50]. A diákok és a robotok közötti interakciókra vonatkozó sajátos környezetünkben a felméréseinkben a diákokat arra kértük, hogy három dimenzió mentén adjanak önbevallást a JW-ről alkotott elképzeléseikről: 1) antropomorfizmus, 2) intelligencia és 3) rokonszenv. Ezenkívül arra is megkértük a diákokat, hogy számoljanak be arról, hogyan/ha interakcióba léptek JW-vel az elmúlt két hétben (pl. olvasták más diákok interakcióit JW-vel, kérdéseket tettek fel JW-nek).

Adatok. Kezdetben 1513 válaszból álló adathalmazzal kezdtük a_0 és az S_5 közötti időszakot. Az összes válaszkérdést összevontuk, hogy létrehozzuk a végleges adathalmazt, amely tartalmazta az összes érvényes, teljes körű választ azokról a diákoktól, akik jelezték, hogy kapcsolatba léptek a JW-vel, vagy más diákok interakcióit olvasva, vagy kérdéseket küldve a JW-nek. Végül összesen 1132 választ kaptunk S0-tól S5-ig ($NS_0 = 260$, $NS_1 = 201$, $NS_2 = 171$, $NS_3 = 171$, $NS_4 = 164$, $NS_5 = 165$).

Elemzéseink nem tartalmazzák az S_0 felmérés eredményeit, amelyek a diákok JW-vel kapcsolatos elvárásait jelzik a tényleges interakciók előtt, mivel inkább a diákok megítélésében bekövetkező hosszú távú változások vizsgálata érdekelt bennünket, legalábbis a kezdeti interakciók után. Az is jól ismert a szakirodalomban, hogy az emberek gyakran irreálisan magas elvárásokat támasztanak a CA-kkal szemben [61, 98].

Eredményeink megismétlik ezt a korábbi szakirodalomból származó hasonló mintát, miszerint a hallgatók észlelése *csökkent* az összehasonlítva



2. ábra: A diákok JW-ről alkotott véleménye az idő múlásával. Az összefüggések bővítése érdekében az ábra az 1. vizsga, a 2. vizsga és a záróprojekt esedékességi időpontját jelöli. Megjegyzendő, hogy a diákok ebben az osztályban heti írásbeli feladatokat is kapnak.

az S_0 szerinti kezdeti felfogásukhoz (vagy várakozásukhoz) képest, amint az a 2. ábrán látható, amely a közösség JW-vel kapcsolatos összesített felfogását ábrázolja a félév során.

4.1 Változások a JW észlelt antropomorfizmusában, intelligenciájában és szimpatikus voltában

Ezután, hogy megértsük, hogy a diákok JW-ről alkotott felfogása jelentősen megváltozott-e a kezdeti interakciók után, Kruskal-Wallis tesztet végeztünk [64]

1. táblázat: Példák a kérdés-válasz párokra a hallgatók és a JW közötti interakciók során a félév során az órai vitafórumon.

Hét	Hallgatói kérdés	JW válasz
4. hét	Mi a következő feladat?	Szia, szívesen válaszolok a kérdésekre! A kurzus tíz írásbeli feladatot tartalmaz: öt P feladatot és öt M feladatot. A P feladatok elvi feladatok, és a 2. egységben tárgyalt anyagot fedik le. Minden alapvető feladat a feladat arra kéri Önt, hogy [...]. Az M feladatok módszertani feladatok, és a következő anyagot fedik le a 3. egységben. Minden egyes módszertani feladat [...].
4. hét	Hogyan programoztak?	Szia, szívesen válaszolok a kérdésekre! Sajnálom, hogy nem értem a kérdését. Átfogalmazná a kérdését egy új kérdés megírásával? követő megbeszélés?
6. hét	Szia Jill Watson, a vizsga lesz a az általunk alkalmazott fogalmakkal foglalkozunk a megbízások?	Szia, szívesen válaszolok a kérdésekre! Íme néhány kérdés, amit szerintem Ön is feltehet. Kérem, állítson össze egy új nyomon követést vita az itt feltett kérdések egyikével: ž, ŐMilyen a JDF specifikáció?ž ŐMegtékinthetjük az online forrásokat?ž
8. hét	Mi a leghatékonyabb módja annak, hogy résztvételi pontok?	Szia, szívesen válaszolok a kérdésekre! A HCI egy mélyen együttműködő terület, és nincs is jobb módja a tanulásnak, mint az interakció. társaiddal, hogy részt vegyetek egymás használhatósági tanulmányaiban, és hogy lássatok a különböző megközelítések sokféleségét. az osztálytársaid által az osztály feladataihoz. Így a részvételért kreditpont szerezhető háromféleképpen: a szakértői értékelések kitöltésével, egymás felméréseiben való részvétellel és interjúk, valamint a Piazza-n folytatott vitákban való részvétel révén. Az alábbi módokon résztvételi pontokat szerezhetsz [...]
10. hét	Mikor esedékes a félévközi felmérés?	Szia, szívesen válaszolok a kérdésekre! Kérjük, töltsd ki ezt a felmérést a hét végéig [link a kurzusközi felméréshez]. Ön 1 résztvételi... a négy tanfolyami felmérés egyikének kitöltéséért. További részletekért lásd a résztvételi politika [...]
12. hét	Szia Jill Watson, meg tudnád jósolni a a diákok százalékos aránya, akik kapnának egy ŐAz óra végén?	Szia, szívesen válaszolok a kérdésekre! Kérjük, töltsd ki ezt a felmérést a hét végéig [link a kurzusközi felméréshez]. Ön 1 résztvételi... a négy tanfolyami felmérés egyikének kitöltéséért. További részletekért lásd a résztvételi politika [...]

a diákok önbevalláson alapuló JW-érzékeléséről S₁ és S₅ között. A Kruskal-Wallis-teszt egy nem parametrikus, omnibusz teszt, amelyet azért használtunk, mert adataink a Shapiro-Wilk-féle normalitásvizsgálat eredményei alapján nem normális eloszlásúak ($p < 0,001$). Ezután további post-hoc páros összehasonlítást végeztünk, hogy megvizsgáljuk az egyes kétheti érzékelési jelentések közötti különbségeket. A Dunn-teszt eredménye szignifikáns különbségeket mutat az érzékelt antropomorfizmusban az S₁

és S₄ : $z = -2,82$, $p = 0,02$, valamint szignifikáns különbségek az észlelt intelligencia tekintetében S₁ és S₄ között : $z = -3,26$, $p = 0,01$. A részletes vizsgálati eredményeket és hatásméreteket a 2. táblázatban közöltük.

4.1.1 Antropomorfizmus. Az antropomorfizmus emberi tulajdonságok tulajdonítása nem emberi tárgyakra, például számítógépekre és hitelesítésszolgáltatóknak. Az antropomorfizmus a számítógépek széles körben vizsgált, de erősen vitatható tervezési jellemzője.

Az emberibb tulajdonságokkal rendelkező hitelesítésszolgáltatók javíthatják a felhasználók bizalmát [13, 32],

a CA-t könnyebben megközelíthetővé teszik és megkönnyítik a felhasználói interakciókat [54, 97]; másrészt a híres "Furcsa völgy" hatás [66] jelzi, hogy a felhasználóknak a CA szerint az erősen antropomorfizált CA negatív érzéseket kelthet az emberekben a CA iránt [17], valamint irreális felhasználói elvárásokat támaszt a CA képességeivel szemben [61]. Az érzékelt antropomorfizmus időbeli változása ezért fontos tulajdonság, amelyet vizsgálni kell, mivel jelentősen befolyásolhatja az emberek CA-val kapcsolatos elvárásait, és ezáltal a bizalomépítést

2. táblázat: Összefoglaló a diákok kétheti JW-percepcióinak összehasonlításáról. Közzöljük a Kruskal-Wallis teszt eredményeit az egyes észlelési metrikákra vonatkozóan S₁ és S₅ között, a poszt-hoc páronkénti összehasonlítás z statisztikáját (Dunn teszt) és a hatásméretet (Cohen's d). A p-értékeket Bonferroni korrekció után közöljük (* $p < 0,05$, ** $p < 0,01$).

Measurezd	Antropomorfizmus		Intelligencia zd		Kedveltség zd
S ₁ és S ₂	-0.60	0.08	-	-	0.670.06
S ₁ és S ₃	-1.47	0.17	1.630.16	-	0.690.06
S ₁ és S ₄	-	-	-	-	-0.590.05
<u>S₁ és S₅</u>	<u>-1.88</u>	<u>0.21</u>	<u>-2.132.320.25</u>	<u>0.04</u>	<u>0.00</u>
S ₂ és S ₃	-0.83	0.10	-0.66	0.09	0.02
S ₂ és S ₄	-	-	3.26**0.33	-	-
<u>S₂ és S₅</u>	<u>-1.23</u>	<u>0.13</u>	<u>1.590.18</u>	<u>0.06</u>	<u>0.60</u>
S ₃ és S ₄	-1.32	0.13	-0.93	0.09	-1.22
S ₃ és S ₅	-	-	-	-	0.11
S ₃ és S ₅ -0.	410.04	0.160.02	-0.620.06	-	-
S ₄ és S ₅	0.90	0.10	1.090	.11	0.600.05
Kruskal-Wallis	$\chi^2(4) = 9,55$ ***		$\chi^2(4) = 11.81$ *		$\chi^2(4) = 2,09$

és a hosszú távú ember-ügynök kapcsolatot [23]. A Kruskal-Wallis teszt megállapította, hogy a diákok önbevalláson alapuló per-a JW-vel való kezdeti interakciót követően érzékelt antropomorfizmus szignifikánsan változott az S₁-től S₅-ig terjedő időszakban: $\chi^2(4) = 9,55$, $p < 0,05$.

A poszt-hoc páronkénti összehasonlítás szerint az S_1 és az S_4 jelentősen különbözik egymástól:

$z = -2,82$, $p < 0,05$. Ez azt jelzi, hogy a CA-k közösség által érzékelt emberszerűsége idővel változhat, még akkor is, ha az ágens tanulási képessége és alkalmazkodóképessége nulla.

4.1.2 *Intelligencia.* Az intelligencia a CA-nak a közösség által érzékelt intelligenciaszintjére utal, más szóval arra, hogy a felhasználók mennyire érzékelik a CA-t intelligens lénynek. Annak ellenére, hogy a mesterségesen "intelligens" gépek megalkotása beváltatlan ígéret volt.

a különböző technikai és megvalósíthatósági kihívások miatt [8, 87] a felhasználók hajlamosak elvárni, hogy a hitelesítésszolgáltatók "okosak" legyenek [98], így szakadék keletkezik a felhasználói elvárások és a hitelesítésszolgáltatók valódi intelligenciája között. A CA tudása szintén az emberek CA mentális modelljében azonosított egyik kulcsfontosságú összetevő [31]. Ezért az észlelt intelligencia fontos szerepet játszik abban, hogy az emberek hogyan észlelik, értékelik és interakcióba lépnek az ügynökökkel. Nem világos azonban, hogy az emberek észlelése a CA intelligenciájáról változik-e az idő múlásával. Ebben a tanulmányban a Kruskal-Wallis teszt azt találta, hogy a JW észlelt intelligenciája jelentősen változott

$S_{1-t61} S_{5-ig}$: $\chi^2(4) = 11,811, p < 0,05$, kifejezetten, post-hoc párbölcs összehasonlítás azt mutatja, hogy az S_1 és az S_4 által jelentett észlelt intelligencia jelentősen különbözik: $z = -3,26, p < 0,01$. Ez rávilágít arra, hogy a CA észlelt intelligenciája fontos tulajdonság, amelyet figyelembe kell venni.

az ember és az AI közötti hosszú távú kapcsolatok kiépítése során.

4.1.3 Szimpatikus. A rokonszenv arra utal, hogy a beszélgetőpartner mennyire szimpatikus mások számára. Az emberi interakciókban a szimpatikus viselkedés a feltételezések szerint pozitív hatást vált ki, növeli a meggyőzőképességet és elősegíti a kedvező megítélést [76, 80]. Mivel az emberek a számítógépeket gyakran szociális szereplőként kezelik [8, 70], az észlelt szimpatikusnak vélt szimpatikusság olyan potenciális tényező, amely befolyásolhatja a hosszú távú kapcsolatépítést. A Kruskal-Wallis teszt nem talált statisztikailag szignifikáns változásokat a diákok

JW önbevalláson alapuló szimpatizálhatósága az idő múlásával: $\chi^2(4) = 2,0947, p = 0,72$. Ez az eredmény annak tulajdonítható, hogy a pozitív első im-

az emberi interakciókban jellemzően döntő szerepet játszik a hosszú távú rokonszenvben [81]. Egy másik ok lehet, hogy a diákok kezdeti megítélése a JW-ről idővel ugyanaz marad, mivel a JW-t szándékosan úgy tervezték meg, hogy egy alap CA legyen, tanulási képesség nélkül.

4.1.4 Az érzékelési mérések közötti korreláció. A Spearman-féle korrelációs tesztet, egy nem parametrikus korrelációs tesztet is elvégeztük, hogy megvizsgáljuk a három érzékelési mérés közötti kapcsolatot. A Spearman-féle korrelációs eredmények azt mutatják, hogy az érzékelt antropomorfizmus és az intelligencia között erős pozitív kapcsolat van ($r_s = (0,74), p < 0,001$), az intelligencia és a szimpatikus viselkedés között mérsékelt erősen pozitív relationship ($r_s = (0,62), p < 0,001$), az antropomorfizmus és a rokonszenvesség pedig alacsony pozitív korrelációt mutat ($r_s = (0,51), p < 0,001$). Ez az eredmény arra utal, hogy bár a az érzékelés három mérőszáma némileg interdependens [8], ez nem biztos, hogy a mi adatainkban így van. Vagyis adataink szerint a diákok kívánatosnak vélt megítélése a három mérőszám mentén hasonló irányú, az egyikben mutatkozó általános növekvő tendencia valószínűleg a másik kettőben is általános növekvő tendenciát eredményezne.

5.1.1 Összefoglalás és értelmezés. A hallgatók kéthetenkénti önbevallásának elemzésével, amely a JW-ről alkotott képükről szól, arra a következtetésre jutottunk, hogy a JW észlelt antropomorfizmusa és intelligenciája jelentősen változott az idő múlásával, de az észlelt szimpatikusnak tartott szimpatizálás nem változott jelentősen hosszú távon.

5 RQ2: A TANULÓ ÉS A DIÁK KÖZÖTTI INTERAKCIÓK NYELVEZETÉNEK VIZSGÁLATA

Ebben a részben azt vizsgáljuk, hogy a tanulóknak hogyan érzékelték és nyelviileg hogyan léptek kapcsolatba a JW-vel. Ehhez összegyűjtöttük a diákok és a JW közötti beszélgetési naplókat a nyilvános vitafórum heti kérdés- és válaszszálaiból, majd a további adatelemzéshez nyelvi jellemzőket vontunk ki. Azzal a céllal, hogy feltárjuk a CA-kra vonatkozó ToM létrehozásának megvalósíthatóságát, a nyelvi intézkedéseket azért választottuk, mert ismertek a felhasználók CA-kkal kapcsolatos holisztikus észlelésének tükrözésére szolgáló lehetőségeik, amelyeket a vonatkozó kutatásokra hivatkozunk, és a következő szakaszokban részletesebben ismertetjük. A megállapításokat és az ember-CA kölcsönhatások tervezésére vonatkozó következményeket is megvitajuk.

5.1 A tanulói észlelés következtetése a nyelvi jellemzőkből

Először is, összekötjük a diákok JW-vel való nyelvi interakcióit egy időblokkban a JW-ről mint alapigazságról szóló, közvetlenül következő önbevallásukkal. Például, ha egy diák a 4. héttől a 6. hétig több hozzászólást tett JW-hez (T_2), és a 6. héten beszámolt a JW-ről alkotott képéről (S_2), akkor ennek a diáknak a esetében a T_2 nyelvi jellemzőit vezetjük le, hogy megértsük az S_2 önbevallás szerinti képét. Ez a megközelítés lehetővé teszi számunkra annak vizsgálatát, hogy a diákok és JW közötti nyelvi interakció egy időblokkban képes-e megjósolni, hogyan fogja érzékelni az ügynököt közvetlenül az adott időblokk végén. Így összesen 551 nyelvi interakció és a saját maguk által jelentett észlelés párosát kapjuk, $N(T_1) = 157, N(T_2) = 86, N(T_3) = 126, N(T_4) = 96, N(T_5) = 86$.

Ezután lineáris regressziós modelleket készítünk. A lineáris regresszió

ismert, hogy segít a függő változóval való feltételesen monoton kapcsolatok értelmezésében [22]. Konkrétan három lineáris regressziós modellt építünk, ahol mindegyik modell a három perceptuális intézkedés egyikét használja függő változóként. Korábbi kutatásokra támaszkodva a nyelvi kölcsönhatásokból különféle nyelvi jellemzőket (feature) származtatunk, amelyek közé tartozik a szóbeliség, az olvashatóság, az ~~adány~~ változatosság és az alkalmazkodóképesség [27, 84]. Ezeket a nyelvi jellemzőket független változóként használjuk a modellekben. Mivel mind az észlelés, mind a nyelvi kölcsönhatások az idő függvényei lehetnek, az adatpont hetének ordinális változóját kovariátorként szerepeltetjük a modellekben. Továbbá, modelljeinket az egyén kiindulási nyelvhasználatával, különösen az alapszintű átlagos szószámmal ellenőrizzük, amelyet az ugyanazon egyén által készített összes hozzászólás alapján számítottunk ki.

Az 1. egyenlet a lineáris regressziós modelljeinket írja le, ahol a P az antropomorfizmus, az intelligencia és a szimpatikus tulajdonságok mérésére utal.

$$P \sim \text{Baseline} + \text{Week} + \text{Verbosity} + \text{Readability} + \text{Sentiment} + \text{Diversity} + \text{Adaptability} \quad (1)$$

Eredményeink segítenek megérteni, hogyan változik a ~~közös~~ közösséggel szembenező CA-kkal kapcsolatos megítélése. Ez hatással van a közösséggel szembenező CA-k tervezésére, hogy

hosszú távon alkalmazkodni tudjanak a közösségnek a CA-ról alkotott változó felfogásához. Azt is megállapítottuk, hogy az önbevalláson alapuló észlelések három mérőszáma egymással korrelál, ami rávilágít arra, hogy ezek a mérőszámok a felhasználók mentális modelljeiben nem feltétlenül különülnek el (vagy függetlenek) egymástól.

A modellek összefoglalása. A lineáris regressziós modelljeink szignif-
cance-t mutatnak: R^2 (Anth.) = 0,85, R^2 (Intel.) = 0,93, R^2 (Like.) = 0,95; mindegyik $p < 0,001$. A 3. táblázat összefoglalja az egyes de-
függő változó. Először is megjegyezzük a kontrollváltozókat, a *hét* és a *kiindulási szóhasználat* statisztikai szignifikanciáját. Azt találjuk, hogy azok az emberek, akik kifejezőbbek, mindhárom észlelési mérés alapján nagyobb valószínűséggel ~~azok~~ pozitívan az ügynököt. Úgy találjuk, hogy a ver-
bozítás negatívan társul az észlelés minden egyes mérőszámával, míg az alkalmazkodóképesség, a sokszínűség és az olvashatóság pozitívan társul.

a diákok JW-ről alkotott elképzeléseivel. Ezután az alábbiakban ismertetjük motivációnkat, hipotézisünket, operacionalizálásunkat és megfigyelésünket az egyes nyelvi jellemzőinkre vonatkozóan.

5.2 Nyelvi jellemzők: Motiváció, operacionalizálás és megfigyelések

5.2.1 Szóbeliség. Az ember-ember közötti beszélgetésekben hajlamosak vagyunk rövidebb és kevésbé összetett mondatokat használni, amikor egy hatodik osztályos gyerekkel beszélgetünk, mint amikor egy felnőtt munkatárssal [68]. Az általunk használt társalgási nyelv szóbelisége tehát attól a mentális modelltől függ, hogy mennyire intelligensnek érzékeljük beszélgetőpartnerünket, ami meghatározza, hogy milyen módon kommunikáljuk másokkal kognitív tervezésünket és gondolataink kivitelezését [27]. Az ember-ember beszélgetési környezetből az ember-CA beszélgetési környezetre lefordítva, a szóbeliség változhat az alapján, hogy mennyire intelligensnek és emberszerűnek érzékeljük a CA-t [45]. Hill et al. azt találták, hogy az emberek kevésbé szószátyár és kevésbé bonyolult szókincset használnak a CA-kkal való kommunikáció során, mint az ember-ember közötti beszélgetésekben [45]. Továbbá a CA emberhez való hasonlóságát a használt szavak hossza alapján is meg lehet ítélni [60]. Ha valaki úgy érzékeli, hogy egy CA emberibb vagy tájékozottabb, valószínűleg terjedelmesebb nyelvet használna. Ennek megfelelően a mi környezetünkben *azt feltételezzük, hogy a nagyobb szóbeliség a JW pozitívabb megítélésével jár együtt.*

Korábbi munkákra [45, 84] támaszkodva két mérőszámot használunk a következők leírására a diákok hozzászólásainak szószátyársága: 1) *hossza* és 2) *nyelvi összetettsége*. A *hosszúságot* a hozzászólásonkénti egyedi szavak számaként, a *komplexitást* pedig a szavak mondatonkénti átlagos hosszaként operacionalizáljuk [27, 84]. A regressziós modellünk (3. táblázat) azt mutatja, hogy mind a verbotív as-tribute negatív együtthatót mutat az összes észlelési mérőszámmal, statisztikai szignifikancia mellett. Ez *elutasítja a hipotézisünket*. A korábbi kutatásokkal és a közhiedelemmel [28, 45, 60] ellentétben eredményeink arra utalnak, hogy azok a diákok, akik több egyedi szót használtak posztonként vagy összetettebb nyelvet, hajlamosak voltak a JW-t kevésbé emberinek, kevésbé intelligensnek és kevésbé szimpatikusnak érzékelni. Úgy értelmezzük, hogy a terjedelmesebb és összetettebb nyelvezet hihetően okozhatta azt, hogy a CA nem tudott támogató vagy hatékony válaszokat adni, ami ahhoz vezetett, hogy a nemkívánatos CA-érezékelés.

5.2.2 Olvashatóság. Az olvashatóság azt jelenti, hogy az olvasók milyen könnyen megérthetnek egy adott szöveget [63]. A pszicholingvisztikai szakirodalom az olvashatóságot az emberek kognitív viselkedésének kulcsfontosságú mutatójaként értékeli, és korábbi munkák ezt a mérőszámot az online közösségek társalgási mintáinak megértéséhez igazították [27, 84, 85]. Bár ezt a mérőszámot nem vizsgálták az ember és az AI közötti interakciók kontextusában, az MToM szempontjából a diákok JW-nek feltett kérdéseinek olvashatósága közvetítheti a JW szövegértési képességének megítélését. Ezért a diákok és a JW interakciójának megértéséhez az olvashatóságot vizsgáljuk. Ugyanakkor, figyelembe véve az ember-ember és az ember-AI közötti beszélgetések analógiáját, *azt feltételezzük, hogy a magasabb olvashatóság a JW pozitívabb megítélését jelzi*. A diákok JW-hez írt hozzászólásainak olvashatóságát a Coleman-Liau-index (CLI) segítségével mérjük. A CLI egy olyan olvashatósági értékelő, amely megközelíti a szövegblokk megértéséhez szükséges

minimális amerikai osztályszintet, és a következő képlet segítségével számítható ki:
$$CLI = 0,0588L - 0,296S - 15,8$$
, ahol L a 100 szóra jutó betűk átlagos száma, S pedig a mondatok átlagos száma. 100 szavanként [77].

A regressziós modellünk azt mutatja, hogy az olvashatóság pozitívan asszociál a JW-ről alkotott tanulói kép mindhárom dimenziójával, statisztikai szignifikanciával: antropomorfizmus (2,33), észlelt intelligencia (2,41) és szimpatikus (3,00). Ez az eredmény *alátámasztja hipotézisünket*, ami arra utal, hogy az olvashatóság erős előrejelzője a tanulók észlelésének, és pozitívan változik az észleléssel. Ez összefügghet azzal a mögöttes intrikával, hogy minél olvashatóbb a kérdés, annál sikeresebb a CA-válasz, és annál elégedettebbek (vagy pozitívan érzékelik) a felhasználók.

5.2.3 Érzelmek. Az ember és a CA közötti beszélgetések során a nyelvünkön keresztül közvetített érzelmek gyakran annak a megnyilvánulása, hogy a CA észlelt teljesítménye megfelel-e a CA-val szembeni elvárásainknak [61]. Valójában a hangulatelemzést már használták az ügyfélszolgálat chatbotjaival való elégedettség kimutatására, és pozitív ~~erőssé~~ hozott [29]. A CA észlelt szimpatizálhatósága mellett a nyelvi sentiment pozitívan kapcsolódik az ember-CA interakciók észlelt természetességéhez is [45, 72]. Bár nincs bizonyíték arra vonatkozóan, hogy a megfogalmazásban megjelenő érzelmek hogyan hozhatók összefüggésbe az észlelt intelligenciával, az intelligencia az egyik legfontosabb kívánt tulajdonság, amelyet az emberek elvárnak egy CA-tól [54]. Ezért *hipotézisünk szerint a diákok által feltett kérdésekben szereplő érzelmek pozitívan kapcsolódnak a JW pozitív megítéléséhez.*

A JW-nek küldött egyes posztok hangulatának mérésére a VADER hangulatelemző modellt [48] használtuk, amely egy szabályalapú hangulatelemző modell, amely -1 (szélsőségesen negatív) és +1 (szélsőségesen pozitív) közötti numerikus pontszámokat ad.

A regressziós modellünk (3. táblázat) azt mutatja, hogy az antropomorfizmus esetében nincs bizonyíték hipotézisünk alátámasztására, de az észlelt intelligenciával (0,69) és a szimpatizálással (0,64) kapcsolatos hipotézis statisztikailag szignifikánsan pozitív együtthatókkal alátámasztott. A jelenlegi vizsgálati környezet, amelyben a JW-t bevetették, formális akadémiai környezetnek tekinthető, és így a tananyaghoz kapcsolódó tematikus beszélgetések gyakoribbak. Úgy véljük, hogy olyan környezetben, ahol az affektív nyelvezet sokkal elterjedtebb (pl. az online Reddit közösségekben), az érzelmek erős szerepet játszhatnak az emberek közösségi CA-ról alkotott megítélésének tükrözésében.

5.2.4 Nyelvi sokszínűség. A beszélgetőpartnerről alkotott képüktől függően változhat nyelvi (és tematikai) sokszínűségünk, azaz a beszélgetés témáinak sokszínűsége vagy a használt nyelv ~~gazdag~~. A nyelvi sokféleségről azt feltételezték, hogy korrelál az ember-ember közötti interakciók során érzékelt intelligenciával [68]. Az ember-CA interakciókban, amikor a CA természetesebb és hitelesebb módon viselkedik, a felhasználók hajlamosak gazdagabb nyelvi készletet is használni, ami pozitív hozzáállást közvetít a CA-val szemben [72]. *Ezért azt feltételezzük, hogy minél nagyobb a nyelvi sokszínűség, annál pozitívabban érzékelik a diákok a JW-t.*

A nyelvi sokféleség eléréséhez korábbi munkákra [2, 84] támaszkodunk, és ehhez szóbeágyazásokat használunk. A szóbeágyazások egy magasabb dimenziós látens térben lévő vektorokként mutatják be a szavakat, ahol a lexikoszemantikailag hasonló szavak vektorai általában közelebb vannak egymáshoz [21, 65, 74]. Esetünkben a JW-hez minden egyes poszthoz először megkapjuk a szóbeágyazás reprezentációját a 300 dimenziós látens lexiko- szemantikus

vektortérben, előre betanított szóbeágyazások segítségével [65]. Ezután kiszámítjuk az átlagos koszinusz-távolságot az ugyanazon felhasználó által az egyes kéthetes időszakokban tett összes bejegyzés középpontjától, mielőtt

3. táblázat: A diákok percepciója (mint függő változó) és a JW-vel való interakció nyelvi alapú mérései (mint független változók) közötti lineáris regresszió együtthatói. A lila sávok a pozitív együtthatók nagyságát, az arany színű sávok pedig a negatív együtthatók nagyságát jelzik. $p < 0,1$, $* p < 0,05$, $ p < 0,01$, $*** p < 0,001$.**

Mérés	Coeff.	p	Coeff. Antropomorfizmus	Coeff. Intelligencia	Coeff. Szerethetőség
Alapvonal Avg. Szavak száma Hét	0.15		0.16	0.13	
	***	0.06	***	***	**
Szóbeliség					
Egyedi szavak száma	-3.34	**	-3.37	-3.91	*
Komplexitás	-1.33	***	-1.82	-2.00	***
Olvashatóság	2.33		2.41	3.00	***
Érzelmei Nyelvi sokszínűség	0.10	***	0.69	0.64	**
	***	0.17	0.09	0.20	*
Nyelvi alkalmazkodóképesség	1.02		1.53	2.55	
	**	1.02	***	***	***
Korrigált R ²	0.85	***	0.93	0.95	***

A regressziós modellünk szerint az észlelt intelligencia és a szimpatikus viselkedés tekintetében hipotézisünk nem kap támogatást, míg az antro- morfizmusra vonatkozó hipotézisünk statisztikailag szignifikánsan alátámasztott, pozitív együtthatót (0,17) mutat. Ez a megállapítás némi támogatást nyújt az ember-CA kölcsönhatással kapcsolatos korábbi munkákhoz, amelyek pozitív kapcsolatot feltételeztek a magas lexiko-szemantikai sokféleség és a CA észlelt ember-szerűsége között [72]. Az ember-ember interakciókkal kapcsolatos megfigyelésekkel [68] ellentétben megfigyeléseink arra utalnak, hogy az emberek nyelvi sokszínűsége nem feltétlenül jelzi, hogy mennyire intelligensnek érzékelünk egy ágenst.

5.2.5 Alkalmazkodóképesség. Emberként hajlamosak vagyunk alkalmazkodni egymás nyelvhasználatához a beszélgetések során, mivel a társas helyzetekben el akarjuk kerülni a kínos helyzeteket [36]. Korábbi kutatások azt sugallták, hogy az emberek gyakran ész nélkül alkalmazzák a társadalmi szabályokat és etikettet a számítógépekre [69], így lehetséges, hogy mi is alkalmazkodunk a nyelvünkhöz, amikor egy CA-val beszélgetünk. Sőt, korábbi munkák azt sugallják, hogy képesek vagyunk a beszédmintánkat ennek megfelelően adaptálni aszerint, hogy a beszélgetőpartnerünk ember vagy CA [45], ami arra utal, hogy a beszédmintánk adaptálhatósága a beszélgetőpartner intelligenciájának, ember-szerűségének, valamint szimpatikus voltának érzékelését jelezheti. Az emberi felhasználók nagyobb valószínűséggel alakítanak ki kívánatos percepciókat egy CA-ról, ha a CA válasza az emberi kérdésekre adaptált és testre szabott, szemben a sablonos válaszokkal (pl. ¡Köszönöm!, ¡Sorry!) [84]. *Ezért azt feltételezzük, hogy az alkalmazkodóképesség pozitívan kapcsolódik az észlelt antropomorfizmushoz, a szimpatizálhatósághoz és az intelligenciához.*

Saha és Sharma [84] megközelítése által motiválva, mérjük a az alkalmazkodóképesség, mint a lexiko-szemantikai hasonlóság a tanuló és a tanuló-JW interakciók egyes kérdés-válasz párosai között, a kérdések és a válaszok szóbeágyazási reprezentációinak ko-színusz hasonlóságaként operacionalizálva. A sokféleséghez hasonlóan 300 dimenziós szóbeágyazási teret használunk [65].

A regressziós modellünk azt mutatja, hogy az alkalmazkodóképesség pozitívan asszociál az antropomorfizmussal (1,02), az intelligenciával (1,53) és a szimpatizmussal (2,55), mindezek statisztikai szignifikanciával. Ez alátámasztja hipotézisünket, és összhangban van a korábbi

kutatókkal, amelyek arra vonatkoznak, hogy az emberek beszélgetőpartner egy CA vagy egy ember [45]. Megfigyeléseink arra utalnak, hogy az alkalmazkodóképesség egy érvényes

a JW megítélésének előrejelzője. Úgy véljük, hogy ha a diákok alkalmazkodó válaszokat kapnak, akkor nagyobb valószínűséggel fogják a JW-t emberibbnek, szimpatikusabbnak és intelligensebbnek érzékelni.

5.2.6 Összefoglaló és értelmezések. Regresszióelemzéssel vizsgáljuk a diák-JW beszélgetések nyelvi jellemzői és a diák-JW észlelése közötti kapcsolatot. Azt találtuk, hogy a verbositás negatívan kapcsolódik a diákok JW-percepciójához, míg az olvashatóság, az érzelmek, a sokszínűség és az alkalmazkodóképesség pozitívan kapcsolódik az antropomorfizmussal, az intelligenciával és a szimpatizmussal. Eredményeink azt sugallják, hogy a nyelvi jellemzők kinyerésének lehetőségei a CA közösségi észlelésének mérésére a beszélgetés során, és így lehetővé teszik a CA számára, hogy folyamatosan megértse és a felhasználói észlelésnek megfelelő, kívánatos re- szponzorokat nyújtson. Fontos megjegyezni, hogy a nyelvi mérések és a JW hallgatói percepciójának három mérőszáma közötti kapcsolat azonos mértékű és irányú.

6 MEGJEGYZÉS

Eredményeink empirikus bizonyítékot szolgáltatnak a közösség által egy közösséggel szembenező CA-ról alkotott hosszú távú vélekedésekről, valamint arról, hogy az ember és a CA párbeszédéből kinyert linguisztikai jellemzőkből következtetni lehet a felhasználók CA-ról alkotott vélekedéseire. Konkrétan azt találtuk, hogy a hallgatói közösség JW antropomorfizmusának és intelligenciájának megítélése jelentősen változott az idő múlásával, de az észlelt szimpatikusnak tartott szimpatizáns nem változott jelentősen. A regressziós elemzéseink azt mutatják, hogy az olyan nyelvi jellemzők, mint a szóbeliség, az olvashatóság, az érzelmek, a sokszínűség és az alkalmazkodóképesség érvényes mutatói a közösség JW-vel kapcsolatos megítélésének. Ezen eredmények alapján először is megvitatjuk a nyelvi elemzés kihasználásának következményeit az ember és az AI közötti interakciók megkönnyítése érdekében. Ezután bemutatjuk az adaptív, a közösséggel szembenező CA-k tervezésének kihívásait és lehetőségeit. Megvitatjuk továbbá az ember és a CA közötti kommunikáció technikai és tervezési következményeit, valamint azt, hogy a jövőbeli munka hogyan terjesztheti ki eredményeinket az ember és az AI közötti interakciókban a MToM kiépítése felé.

6.1 Nyelvi elemzés az ember és az AI közötti interakciók tervezéséhez

Munkánk azt mutatja, hogy az ember és a CA közötti beszélgetésekből kinyert nyelvi jellemzők kihasználása javíthatja az ember és a CA közötti interakciókat. Ez a technika, ha megfelelően integrálódik a CA

tervezés, beváltaná a valóban "beszélgető" ügynökök létrehozásának ígéretét. Eredményeink azt mutatják, hogy a nyelvi elemzés felhasználható arra, hogy automatikusan következtetni lehessen a közösségnek a közösséggel szembenező hitelesítésszolgáltatókról alkotott képére. Ez megnyitja annak lehetőségét, hogy a nyelvi elemzés segítségével olyan CA-kat tervezzünk, amelyek automatikusan képesek azonosítani a felhasználónak a CA-ról alkotott mentális modelljét, ami lehetővé teszi a CA-k számára, hogy a válaszokban finom utalásokat adjanak, hogy a folyamatos és eredményes beszélgetés érdekében a felhasználónak segítsenek kiigazítani a CA-ról alkotott mentális modelljét.

Tanulmányunkban, annak ellenére, hogy a JW egy kérdés-válaszoló (QA) CA, amelyet csak a diákok alapvető információs szükségleteinek kielégítésére terveztek, a diákok percepciójára tudunk következtetni az ilyen egyszerű QA párbeszédéből kinyert nyelvi jellemzők segítségével. Eredményeink egybecsengenek korábbi munkákkal, amelyek szintén feltárták a felhasználók és a QA ügynökök közötti kérdés-válasz párbeszédadatok nyelvi elemzésének lehetőségét a beszélgetések megszakadására való következtetés céljából [58]. Úgy véljük, hogy kifinomultabb társalgási környezetben, ahol az ember-CA interakciók túlmutatnak az alapvető információs szükségleteken, és a multimodális adatokat (pl. hang- és vizuális kommunikációt) tartalmazó interakciókban árnyaltabb leírásokat lehet kinyerni a felhasználók CA-kkal kapcsolatos észleléseiről. Ez olyan felismerésekhez vezetne, amelyek elősegíthetik az ember és a CA konstruktív és következetes párbeszédét.

Azt is megjegyezzük, hogy a diák-JW interakciókat egy sokkal jobban ellenőrzött környezetben, mint az ember-CA interakciók számos lehetséges beállítása. Például az online kurzusfórumon folytatott vitáknak tematikusan a kurzus munkájáról *kell szólniuk*. Emellett a diákoktól elvárják, hogy kívánatos és civilizált módon jelenjenek meg önmagukban. Különböző online és offline normák és konvenciók léteznek, amelyeket az emberek hajlamosak követni az akadémiai környezetben [40]. Másrészt egy általános célú online közösségben (pl. Reddit) folyó viták, beleértve a moderáltakat is, nemcsak változatos és deviáns vitákat, hanem informális nyelvezetet is tartalmazhatnak [15]. Az ilyen típusú adatok zajjal egészíthetők ki az automatizált nyelvi modelleket, és ez több újrakeresési lehetőséget nyit meg annak vizsgálatára, hogy az általános célú online közösségekben használt nyelv hogyan tükrözi az egyéni és kollektív percepciót egy közösséggel szembenező CA-ról.

Amellett, hogy a nyelv segít a hitelesítésszolgáltatóknak megérteni, hogyan érzékelik őket a felhasználók az interakciók során, potenciálisan jelezheti a felhasználók preferenciáit is a hitelesítésszolgáltatóval kapcsolatban egy adott kontextusban, és így tájékoztathatja a hitelesítésszolgáltatók jövőbeli tervezését. A regressziós elemzéseinkben például az olyan nyelvi mérőszámok, mint a hangulat és a sokszínűség, hasonló irányultságot tükröznek (lásd a 3. táblázatot) a három észlelési mérőszám közötti korrelációban (4.3. szakasz) - pozitív összefüggést találunk a JW észlelt intelligenciája és a szimpatizálhatóság között, de gyenge korrelációt az antropomorfizmus és a szimpatizálhatóság között. Különösen a diák-JW beszélgetésből kinyert érzelmek szignifikánsan összefüggnek mind az intelligenciával, mind a szimpatizálással, de nem függenek össze szignifikánsan az érzékelt antropomorfizmussal. Érdeemes tehát megfontolni, hogy a JW-hez hasonló virtuális tanársegédek tervezésekor az emberhez való hasonlóság fontosabb tényező-e, mint az információs támogatás

nyújtásával demonstrált intelligencia. Ez az eredmény további bizonyítékkal szolgál a régóta tartó vitához, hogy a CA-kat a lehető legemberibbé kell-e tervezni [16, 17], ami arra utal, hogy a felhasználók preferenciája, hogy a CA-knak emberszerűnek kell-e lenniük, nagymértékben függ a CA szerepétől és a használati kontextustól.

6.2 Adaptív, közösséggel szembenező társalgási ügynökök tervezése

Korábbi munkák hét olyan társadalmi szerepet javasoltak, amelyeket a közösségi CA-k az online emberi közösségekben szolgálhatnak [87], de az, hogy hogyan lehet gyorsan felismerni és mérni az emberek elképzeléseit és elvárásait arról, hogy a CA-nak hogyan kell viselkednie a különböző társadalmi szerepek betöltésekor, még nem vizsgálták. Munkánk lehetőséget teremt arra, hogy a közösségi CA-k kívánt társadalmi szerepeit a CA-k észlelésének konkrét dimenziói alapján operacionalizáljuk. Például, amikor a CA társadalmi szervezőként segít a közösség tagjainak társas kapcsolatok kiépítésében, a közösség elvárhatja, hogy a CA ne intelligensebb, hanem emberibb és szimpatikusabb legyen. Ezeket az elvárásokat potenciálisan nyelvi jelzésekkel lehetne azonosítani és nyomon követni, ahogyan azt a munkánk is mutatja. Ez a működés-tionalizálás segíthet a közösséggel szembenező CA-knak gyorsan azonosítani a közösség elvárásait, és olyan viselkedést produkálni, amely jobban illeszkedik a közösségen belüli észlelt társadalmi szerepükhöz.

Míg a korábbi kutatások azt sugallták, hogy a közösséggel szembenező CA-k társadalmi szerepe idővel változik az online közösségeken belül [51, 87, 88], a mi vizsgálatunk a hallgatói közösség JW-ről alkotott megítélésében bekövetkező hosszú távú változásokról empirikus bizonyítékot szolgáltat a közösségnek az ügynökről alkotott megítélésének konkrét változásairól. Eredményeink azt mutatják, hogy a közösségben megjelenő CA-k észlelt antropomorfizmusa és intelligenciája árnyaltabb és képlékenyebb jellemzők, és ezért gyakrabban kell értékelniük, hogy a CA-k ennek megfelelően tudják beállítani a közösségen belüli viselkedésüket. A JW észlelt szimpatikus képessége nem változott jelentősen a vizsgálatunkban, ami arra utal, hogy a tervezőknek nagyobb mozgástere lehet a CA észlelt szimpatikus képességének nyomon követésében. A JW közösségen belüli stabil észlelt szimpátiájának oka azonban további vizsgálatot igényelhet, hogy ez azért van, mert a hosszú távú szimpátia észlelése nagymértékben függ az első benyomástól [8], vagy lehet, hogy ez a JW félév során nyújtott stabil teljesítményének eredménye, ami a tanulási képesség hiányának köszönhető.

Az egyik előrelátható kihívás az adaptív, a közösséggel szembenező CA-k tervezésekor, amelyek nyelvi jeleket használnak a felhasználó CA-ról alkotott képének kialakításához, az egyes üzenetek szándékának megkülönböztetése - hogy a felhasználó valódi kérdést tett-e fel, vagy csak a rendszerrel próbál játszani; vagy hogy a felhasználó válasza a CA-nak vagy a közösség más tagjainak szólt-e. A nyelvi jelek alapján a felhasználó választ a CA-nak vagy a közösség más tagjainak szánták. Míg az emberek a mindennapi életben olyan stratégiákat alkalmaznak, mint például a megjelenés megváltoztatása, hogy kezeljék az önreprezentációjukat [36], addig a nyilvános online platformokon az emberek az észlelt közönségtől függően szintén linguisztikus jelzésekkel kezelik az önreprezentációjukat [19, 49, 83]. A közösséggel szembenező CA-k esetében minden dyadikus ember-CA interakció látható a közösség többi tagja számára is. Az emberek így kihasználhatják ezt a lehetőséget, hogy ne csak támogatást nyerjenek a CA-tól, hanem a részonzoraik modulálására is, hogy segítsék a közösségen belüli önreprezentációjuk kezelését. Például az emberek szándékosan korlátozhatják érzelmi kifejezésüket a nyelvi kifejezéssel, hogy ne tűnjenek hülyének, mert azt gondolják, hogy a CA értelmezni tudja a nyelvben rejlő érzelmi elemeket [40]; vagy az emberek

célzottan válaszolhatnak olyan kérdésekkel, amelyek segíthetnek nekik abban, hogy humorosabbnak tűnjenek, mintha helyes választ kapnának a CA-tól. Tanulmányunkban többször is előfordult ez, amikor a diákok olyan kérdéseket tesznek fel JW-nek, amelyek egyértelműen nem tartoznak a JW hatáskörébe, például *!Mi az élet értelme?* vagy *!Mi a kedvenc karaktered a Trónok harca sorozatban?*.

6.3 Az elme kölcsönös elmélete felé az ember és az AI közötti interakcióban

A végső cél az MToM kiépítése az ember és az AI közötti interakciókban, a tanulmányunk a CA ToM kiépítésének megvalósíthatóságát vizsgálta a CA felhasználói észlelésének operacionalizálásával és azonosításával a nyelvi jelzéseken keresztül. A MToM szemszögéből nézve, az ember-ember interakcióhoz hasonlóan az interakciók során egymás érzékelésének megértése az ember-AI interakciók kognitív alapjaként működik [6]. Az ember-ember interakciókra támaszkodva, amelyek a nyelveken keresztül közvetített mindenféle szociális jelekre támaszkodnak, a CA ToM pontosságának javítása érdekében a jövőbeli kutatásoknak meg kell vizsgálniuk, hogy más társalgási jelek hogyan kombinálhatók potenciálisan az általunk vizsgált nyelvi jellemzőkkel, hogy nagyobb kontextust és pontosságot biztosítsanak az emberek CA-ról alkotott percepcióinak megértésében. Például a beszélgetés megszakadásainak azonosítása a konverzációs jelzéseken keresztül [58] vagy a felhasználói érzelmek és elégedettségek megfigyelése az interakciók során [29, 93] kombinálható a CA azonosított felhasználói percepciójával. Ennek egyik lehetséges következménye, hogy a nyelvi elemzéssel azonosított felhasználói percepciót állandó állapotként tartjuk, míg más beszélgetési jelek felhasználhatók arra, hogy a felhasználó CA-ról alkotott percepciójának állapotváltozását kiáltsuk, és így a CA viselkedését folyamatosan kiigazítsuk.

Míg a mi munkánk arra összpontosít, hogy a hitelesszolgáltatókat nyelvi elemzéssel segítsük abban, hogy megértsék a felhasználóknak a hitelesszolgáltatókról alkotott elképzeléseit, szeretnénk hangsúlyozni, hogy *a kommunikáció minden típusa kétirányú interakció. A kölcsönös ToM elérése érdekében az ember és a CA közötti interakciókban olyan technikák feltárása is kulcsfontosságú, amelyek segíthetnek a felhasználóknak megérteni a CA-nak a céljaikról, szándékaikról, képességeikről stb. alkotott elképzeléseit, hogy a felhasználók időben korrigálhassák a CA észlelését a kívánatos ember és CA közötti interakció eredményének elérése érdekében. Azáltal, hogy az MToM-et az ember és a CA közötti interakciók tervezésének elméleti keretként javasoljuk, munkánk motiválja a jövőbeli kutatásokat annak további feltárására, hogy hogyan segíthetjük a felhasználókat abban, hogy megértsék a CA róluk alkotott elképzeléseit. Ez az irány magában foglalhatja annak vizsgálatát, hogy a CA hogyan és mikor tudja felajánlani a felhasználókról alkotott percepcióját olyan módon, amely intuitív és könnyen érthető a felhasználók számára, ugyanakkor elég természetes ahhoz, hogy a beszélgetés hitelessége megmaradjon. Ez magában foglalhat olyan technikákat, amelyeket jelenleg a magyarázható mesterséges intelligencia kutatói vizsgálnak, hogy segítsék a felhasználókat abban, hogy kellőképpen megértsék, hogyan elemezné a CA a szöveges válaszaikat, hogy a céljakra és szándékaikra vonatkozó észleléseiket kivonja [91]. A korábbi munkáknak az ember-ember társadalmi interakciók megértésére vonatkozó javaslatával összhangban, mint az ember és az AI közötti interakciók javításának módjára [33], kiemeljük az ~~interakció~~ elméleti keretek kihasználásának fontosságát, hogy új tervezési perspektívákat kínáljunk az ember-CA interakciókhoz. Ez a munka egyesíti a kognitív tudományokból és a társadalomtudományokból származó elméleteket az ember-ember interakciókról az emberi társadalmi interakciók nagyrészt a benyomások kezeléséről szólnak [36], ami a ToM [6, 12] egyedülállóan emberi kognitív képességétől függ. Ez lehetővé teszi számunkra, hogy*

újrarendeljük az ember és az AI közötti interakciók tervezését. Ezért kiemeljük annak fontosságát, hogy az olyan területekről, mint az antropológia, a kognitív tudományok, a szociálpszichológia stb. elméleti kereteket kölcsönözzünk, hogy új tervezési perspektívákat kínáljunk az ember és az AI közötti interakciókról.

6.4 Korlátozások és jövőbeli munka

Munkánknak vannak bizonyos korlátai. Eredményeink nem biztos, hogy átvihetők, ha az ember-CA interakció magánjellegű dyadikus interakciós kontextusban zajlik. Ez a munka azt vizsgálja, hogy egy nyilvános vitaforumon zajló ember-ügynök dyadikus interakcióból kinyert nyelvi jellemzők segítségével le lehet-e következtetni a hallgatók közösségi CA-ról alkotott képét. A diákok észlelését és az ügynökkel való interakciót így torzítják a többi diáknak az ügynökkel a nyilvános fórumon folytatott interakciói, amire rámutatunk, mint az emberi közösségeken belüli dyadikus interakciókat végző, közösség-központú CA-k tervezésének egyedi kihívására. A jövőbeni kutatások, amelyek célja adaptív CA-k tervezése a dyadikus interakciókban, megismételhetik a jelenlegi tanulmányt az egy-egy ember-CA interakciókban.

Munkánk egy formatív lépést tett afelé, hogy megértsük, hogyan érzékelik az emberek a hitelesítésszolgáltatókat a nyelvi jellemzők segítségével. Eredményeink korrelációs jellegűek, és nem állíthatunk ok-okozati összefüggéseket. A jövőbeni munka, amely figyelembe veszi a nem megfigyelt zavaró tényezőket, jobb betekintést nyújthat az ember és az AI közötti észlelésekbe és interakciókba. Azt is elismerjük, hogy több kvalitatív vagy vegyes módszertani megközelítésre van szükség ahhoz, hogy mélyebb betekintést nyerjünk az emberek érvelésébe és szándékába a nyelvi viselkedésük mögött, amikor egy CA-val beszélgetnek. A mi vizsgálatunkban például a diákok szándékosan tesztelheték, hogy a JW megtanult-e valamit a korábbi kérdéseikből azáltal, hogy pontosan ugyanazokat a kérdéseket tették fel a JW korábbi szálaiból; vagy a diákok frusztráltak lehetnek a JW tanulási képességei miatt, és ezért szándékosan nehéz kérdéseket tesznek fel a nyilvános szála - ezt nem lehet kvantitatív módon értékelni, és a jövőbeli kvalitatív kutatások fényt deríthetnek erre a kérdésre.

A diákok JW-ről alkotott képének számszerűsítéséhez egy, az ember-robot interakcióból vett standardizált mérőszámot használtunk, amely magában foglalja az antropomorfizmust, az intelligenciát és a szimpatikus jelleget [8]. Az általunk elfogadott mérés azonban nem azt sugallja, hogy ezek a JW felhasználói észlelésének standard dimenziói, vagy hogy ezeknek kellene lenniük a JW-ről alkotott felhasználói elképzeléseknek. Valójában már korábbi kutatások is azt sugallták, hogy különböző értelmezések léteznek arra vonatkozóan, hogy a felhasználók hogyan építik fel a JW-ről alkotott mentális modelljeiket [10, 31]. Reménykedünk azonban abban, hogy a nyelvi elemzés feltárhatja az emberek CA-król alkotott elképzeléseinek különböző dimenzióit az interakciók során. A jövőbeni kutatásoknak meg kell ismételnünk a jelenlegi vizsgálatot a felhasználók CA-ról alkotott mentális modelljének különböző mérései segítségével, hogy további bizonyítékokat szolgáltatassanak a nyelvi elemzésben rejlő lehetőségekről.

7 KÖVETKEZTETÉS

Ez a tanulmány a kölcsönös elméletet (Mutual Theory of Mind) mint elméleti keretet tűzte ki az adaptív, közösséggel szembenező társalgási ügynökök (CA-k), mint hosszú távú társak tervezéséhez. E keretrendszer alapján vizsgáltuk a CA-k közösségi észlelésének hosszú távú változásait, és mértük az észlelések nyelvi jelekből történő kikövetkeztetésének megvalósíthatóságát. Egy közösséggel szembenező CA-t, JW-t, egy virtuális tanársegédet vetettünk be, amelyet arra terveztünk, hogy válaszoljon a diákok órával kapcsolatos logisztikai

kérdéseire egy online nyilvános vitaforumon. Az elméletletről alkotott elképzeléseink alapján mértük a diákok JW-ről alkotott képét az érzékelt antropomorfizmus, az intelligencia és a szimpatizálás szempontjából. Statisztikailag szignifikáns hosszú távú változásokat találtunk a diákközösség JW-ről alkotott képében az antropomorfizmus és az intelligencia tekintetében. Ezután a félév során a hallgatók és a JW közötti interakciókból elméleti alapú nyelvi jellemzőket vontunk ki. A regressziós elemzések kimutatták, hogy az olyan nyelvi jellemzők, mint a

mint a szóbeliség, a változatosság, az alkalmazkodóképesség és az olvashatóság magyarázza a **műk** JW-ről alkotott képét. Megvitatjuk a nyelvi elemzés kihasználásának lehetőségét, hogy teljesítsük a valóban "kon- versációs" ágensek tervezésének ígérését, beleértve az adaptív, közösséggel szembenező CA-k építésének tervezési következményeit, amelyek képesek a közösségnek a CA-ról alkotott változó felfogásához igazodni, valamint a kölcsönös elmélet mint tervezési keretrendszer alkalmazásának elméleti következményeit az ember és az AI közötti interakciók megkönnyítésében. Úgy véljük, hogy ez a kutatás inspirálhatja a jövőbeli munkát az interdiszciplináris elméletek felhasználására az ember és a CA közötti interakciók újragondolásához.

KÖSZÖNETNYILVÁNÍTÁS

Köszönjük Vedant Das Swain, Dong Whi Yoo, Adriana Alvarado Garcia, Scott Appling és a névtelen bírálók segítségét és visszajelzéseit. Ezt a munkát a Georgia Tech és a Georgia Tech Számítástechnikai Főiskola belső támogatásaiból finanszíroztuk.

HIVATKOZÁSOK

- [1] Vincent AWM Alevin és Kenneth R Koedinger. 2002. Egy hatékony metakognitív stratégia: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science* 26, 2 (2002), 147-179.
- [2] Tim Althoff, Kevin Clark és Jure Leskovec. 2016. Tanácsadói beszélgetések nagyszabású elemzése: A természetes nyelvi feldolgozás alkalmazása a mentális egészségügyben. *TACL* (2016).
- [3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Irányelvek az ember és az AI közötti interakcióhoz. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1-13.
- [4] Zahra Ashktorab, Mohit Jain, Q. Vera Liao és Justin D. Weisz. 2019. Rugalmas chatbotok: Javítási stratégia preferenciák a beszélgetési zavarok esetén. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1-12. <https://doi.org/10.1145/3290605.3300484>.
- [5] Simon Baron-Cohen. 1997. *Agyvaktság: Egy esszé az autizmusról és az elmeelméletről*. MIT press.
- [6] Simon Baron-Cohen. 1999. Az elme elméletének evolúciója? In *Az elme lezármasága: Psychological Perspectives on Hominid Evolution*. Oxford University Press, 1-31.
- [7] Simon Baron-Cohen, Alan M Leslie, Uta Frith, et al. 1985. Van-e az autista gyerekeknek lemelelete? *Cognition* 21, 1 (1985), 37-46.
- [8] Christoph Bartneck, Dana Kulić, Elizabeth Croft és Susana Zoghbi. 2009. A robotok antropomorfizmusának, animabilitásának, szimpatizálhatóságának, észlelt intelligenciájának és észlelt biztonságának mérőeszközei. *International Journal of Social Robotics* 1, 1 (1 2009), 71-81. <https://doi.org/10.1007/s12369-008-0001-3>.
- [9] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip és Alexis Hiniker. 2019. Kommunikációs zavarok a családok és Alexa között. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1-13.
- [10] Emilie Bigras, Marc Antoine Jutras, Sylvain Sénécal, Pierre Majorique Léger, Marc Fredette, Chrystel Black, Nicolas Robitaille, Karine Grandé és Christian Hudon. 2018. Munka egy ajánlóügynökkel: Hogyan befolyásolja az ajánlások bemutatása a felhasználók észlelését és viselkedését. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1-6. <https://doi.org/10.1145/3170427.3188639>.
- [11] Michael Braun, Anja Mainz, Ronée Chadowitz, Bastian Pflöging és Florian Alt. 2019. Az Ön szolgálatában: Hangasszisztens személyiségek tervezése az autópári felhasználói felületek javítása érdekében. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1-11.
- [12] Peter Carruthers és Peter K Smith. 1996. *Az elmeelméletek elméletei*. Cambridge University Press.
- [13] Justine Cassell és Timothy Bickmore. 2000. A bizalom- méltóság külső megnyilvánulásai az interfészen. *Commun. ACM* 43, 12 (2000), 50-56.
- [14] Justine Cassell és Timothy Bickmore. 2003. Tárgyalásos összejátszás: A társas nyelv és kapcsolati hatásainak modellezése intelligens ágensekben. *User Modelling and User-Adapted Interaction* 13, 1-2 (2003), 89-132. <https://doi.org/10.1023/A:1024026532471>.
- [15] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan és Eric Gilbert. 2017. A közösségek zsákja: A visszaélészerű online viselkedés azonosítása meglévő internetes adatokkal. In *Proc. CHI*.
- [16] Dasom Choi, Daehyun Kwak, Minji Cho és Sangsu Lee. 2020. "Senki sem beszél ilyen gyorsan!" A beszédsebesség empirikus vizsgálata a társalgási ügynökökben a
- Látássérült emberek. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1-13. <https://doi.org/10.1145/3313831.3376569>
- [17] Leon Ciechanowski, Aleksandra Przegalina, Mikolaj Magnuski és Peter Gloor. 2019. A kísérleties völgy árnyékában: A humans chatbot interakciójának kísérleti vizsgálata. *Future Generation Computer Systems* 92 (2019), 539-548.
- [18] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade és Benjamin R. Cowan. 2019. Mitől lesz jó egy beszélgetés? Kihívások a valóban beszélgető ágensek tervezésében. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1-12. <https://doi.org/10.1145/3290605.3300705>.
- [19] Cristian Danescu-Niculescu-Mizil, Michael Gamon és Susan Dumais. 2011. Mark my words! Nyelvi stíluselhelyezés a közösségi médiában. In *Proceedings of the 20th international conference on World wide web*. 745-754.
- [20] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec és Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078* (2013).
- [21] Vedant Das Swain, Koustuv Saha, Manikanta D Reddy, Hemang Rajvanshy, Gregory D Abowd és Munmun De Choudhury. 2020. A szervezeti kultúra modellezése a Glassdooron megosztott munkahelyi tapasztalatokkal. In *CHI*.
- [22] Robyn M Dawes és Bernard Corrigan. 1974. Lineáris modellek a döntéshozatalban. *Psychological bulletin* 81, 2 (1974), 95.
- [23] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger és Raja Parasuraman. 2016. Majdnem ember: Az antropomorfizmus növeli a bizalom ellenálló képességét a kognitív ágenseknél. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331.
- [24] Chris Dede, John Richards és Bror Saxberg. 2018. *Learning Engineering for Online Education: Elméleti összefüggések és tervezési alapú példák*. Routledge.
- [25] Sandra Devin és Rachid Alami. 2016. Egy implementált elmélet az ember-robot közös tervek végrehajtásának javítására. *ACM/IEEE International Conference on Human-Robot Interaction* 2016-April (2016), 319-326. <https://doi.org/10.1109/HRI.2016.7451768>.
- [26] Bobbie Eicher, Kathryn Cunningham, Sydney Peterson Marissa Gonzales és Ashok Goel. 2017. A kölcsönös tudatelmélet mint a társalkotás alapja felé. In *International Conference on Computational Creativity, Co-Creation Workshop*.
- [27] Sindhu Kiranmai Ernal, Asra F Rizvi, Michael L Birnbaum, John M Kane és Munmun De Choudhury. 2017. A skizofrénia közösségi médiás közlések terápiás kimenetelét jelző nyelvi markerek. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1-27.
- [28] Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A Smith és Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. *arXiv preprint arXiv:1804.10202* (2018).
- [29] Jasper Feine, Stefan Morana és Ulrich Gnewuch. 2019. Service Encounter Satisfaction Measuring Customer Service Chatbots with Customer Service Chatbots using Sentiment Analysis. *Proceedings of the 14th International Conference on Wirtschaftsinformatik* December (2019), 0-11.
- [30] Radhika Garg és Subhasree Sengupta. 2020. Beszélgetési technológiák az otthoni tanulóhoz: A társtervezés felhasználásával a gyermekek és a szülők Perspektívák. (2020), 1-13. <https://doi.org/10.1145/3313831.3376631>.
- [31] Katy Ikonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel és Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1-12. <https://doi.org/10.1145/3313831.3376316>.
- [32] Eun Go és S Shyam Sundar. 2019. A chatbotok humanizálása: A vizuális, identitásbeli és társalgási jelzések hatása az emberség észlelésére. *Computers in Human Behavior* 97 (2019), 304-316.
- [33] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C Schultz, et al. 2005. Robotok tervezése hosszú távú szociális interakcióra. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1338-1343.
- [34] Ashok Goel. 2020. AI-alapú tanulás. *arXiv preprint arXiv:2006.01908* (2020).
- [35] Ashok K Goel és Lalith Polepeddi. 2016. *Jill Watson: Virtuális tanársegéd az online oktatásban*. Technikai jelentés. Georgia Institute of Technology.
- [36] Erving Goffman. 1978. *Az én bemutatása a mindennapi életben*. London: Harmondsworth.
- [37] Alvin I Goldman et al. 2012. Az elme elmélete. *The Oxford handbook of philosophy of cognitive science* 1 (2012).
- [38] Alison Gopnik és Henry M Wellman. 1992. Miért a gyermek elmélete valóban elmélet. (1992).
- [39] O Can Görür, Benjamin Rosman és Guy Hoffman. 2017. Toward Integrating Theory of Mind into Adaptive Decision-Making of Social Robots to Understand Human Intention. In *Workshop on the Role of Intentions in Human-Robot Interaction at the International Conference on Human-Robot Interactions*. Bécs, Ausztria.

- [40] Pamela Grimm. 2010. Társadalmi kívánatossági torzítás. *Wiley International Encyclopedia of Marketing* (2010).
- [41] Shivashankar Halan, Brent Rossen, Michael Crary és Benjamin Lok. 2012. Virtuális emberek konstruktívizmusa a beszélgetőpartnerek észlelésének javítása érdekében. (2012), 2387. <https://doi.org/10.1145/2212776.2223807>.
- [42] Jeffrey T Hancock, Kailyn Gee, Kevin Ciaccio és Jennifer Mae-Hwah Lin. 2008. Szomorú vagyok, hogy szomorú vagy: érzelmi fertőzés a CMC-ben. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 295&298.
- [43] Maaïke Harbers, Karel Van Den Bosch és John Jules Meyer. 2009. Az elméletben gondolkodó ágensek modellezése. *Proceedings - 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2009 2* (2009), 217&224. <https://doi.org/10.1109/WI-IAT.2009.153>
- [44] Laura M. Hiatt, Anthony M. Harrison és J. Gregory Trafton. 2011. Az emberi variabilitás elősegítése az ember-robot csapatokban az elméleten keresztül. *IJCAI International Joint Conference on Artificial Intelligence* (2011), 2066&2071. <https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-345>. <https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-345>
- [45] Jennifer Hill, W. Randolph Ford és Ingrid G. Farreras. 2015. Valódi beszélgetések mesterséges intelligenciával: Ember-ember online beszélgetések és ember-chatbot beszélgetések összehasonlítása. *Computers in Human Behavior* 49 (2015), 245&250. <https://doi.org/10.1016/j.chb.2015.02.026>. <https://doi.org/10.1016/j.chb.2015.02.026>
- [46] Kenneth Holstein, Bruce M McLaren és Vincent Alevan. 2019. Tervezés a komplementaritás érdekében: A tanárok és a diákok igényei a hangszerelés támogatására az ai-erősített osztálytermekben. In *International Conference on Artificial Intelligence in Education*. Springer, 157&171.
- [47] Claire Hughes és Sue Leekam. 2004. Milyen kapcsolat van az elmélet és a társadalmi kapcsolatok között? Áttekintés, reflexiók és új irányok a tipikus és atipikus fejlődés tanulmányozásához. *Social Development* 13, 4 (2004), 590&619. <https://doi.org/10.1111/j.1467-9507.2004.00285.x>. <https://doi.org/10.1111/j.1467-9507.2004.00285.x>.
- [48] C J Hutto és E E Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (2014). <http://sentiment.net/>.
- [49] Kokil Jaïdka, Sharath Chandra Guntuku, Anneke Buffone, H Andrew Schwartz és Lyle H Ungar. 2018. Facebook vs. Twitter: Platformok közötti különbségek az önfeltáráshoz és a vonásjelöléshez. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*. 141&150.
- [50] Yuin Jeong, Younah Kang és Juho Lee. 2019. A társalgási töltelékek hatásának feltárása a társalgási ügynökök felhasználói percepciójára. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1&6. <https://doi.org/10.1145/3290607.3312913>
- [51] Da-jung Kim és Youn-kyung Lim. 2019. Társügynök: Design for Build- ing User-Agent Partnership in Learning and Adaptive Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1&14. <https://doi.org/10.1145/3290605.3300714>.
- [52] Kyung-Joong Kim és Hod Lipson. 2009. Egy egyszerű robotikus tudatelmélet felé. (2009), 131. <https://doi.org/10.1145/1865909.1865937>.
- [53] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh és Joonhwan Lee. 2020. Bot a csoportban: A csoportos csevegés megbeszélésének megkönnyítése a hatékonyság és a részvétel javításával egy chatrobottal. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (A számítástechnikai rendszerek emberi tényezői című konferencia 2020. évi jegyzőkönyve)*. ACM, New York, NY, USA, 1&13. <https://doi.org/10.1145/3313831.3376785>
- [54] Anastasia Kuzminykh, Jenny Sun, Nivetha Govindaraju, Jeff Avery és Edward Lank. 2020. Dzsinn a palackban: Anthropomorphized Perceptions of Conversational Agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1&13. <https://doi.org/10.1145/3313831.3376665>.
- [55] Sunok Lee, Sungbae Kim és Sangsu Lee. 2019. "Hogyan néz ki az ügynököd?". In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1&6. <https://doi.org/10.1145/3290607.3312796>.
- [56] Sangwon Lee, Naeun Lee és Young June Sah. 2020. Az elme érzékelése egy chatrobotban: Az elme észlelésének és a szociális jelzések hatása a társjelenlétre, a közelségre és a használati szándékra. *International Journal of Human-Computer Interaction* 36, 10 (2020), 930&940. <https://doi.org/10.1080/10447318.2019.1699748>.
- [57] Séverin Lemaignan és Pierre Dillenbourg. 2015. Kölcsönös modellezés a robotikában: Inspirációk a következő lépésekhez. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 303&310.
- [58] Q. Vera Liao, Werner Geyer, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller és N. Sadat Shami. 2018. Csak munka és semmi szórakozás? Beszélgetések egy kérdés-felelet chatrobottal a vadonban. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Vol. 8. ACM Press, New York, New York, USA, 1&13. <https://doi.org/10.1145/3173574.3173577>.
- [59] Shuhong Lin, Boaz Keysar és Nicholas Epley. 2010. Reflexív agyvaktság: Az elmélet használata a viselkedés értelmezéséhez erőlködő figyelmet igényel. *Journal of Experimental Social Psychology* 46, 3 (2010), 551&556. <https://doi.org/10.1016/j.jesp.2009.12.019>.
- [60] Catherine L Lortie és Matthieu J Guittton. 2011. A humánum megítélése

a beszélgetőpartner a szemlélő szemében van. *PLoS One* 6, 9 (2011), e25085.

- eredmények ok-okozati tényezői az online mentális egészségügyi közösségekben. In *ICWSM*.
- [85] Koustuv Saha, Ingmar Weber és Munmun De Choudhury. 2018. A közösségi média alapú vizsgálat a tanácsadási ajánlások hatásainak vizsgálata hallgatói halálesetek után a főiskolai kampuszokon. In *ICWSM*.
- [61] Ewa Luger és Abigail Sellen. 2016. "Mintha egy nagyon rossz hangosbeszélő lenne": a felhasználói elvárások és a társalgási ügynökökkel kapcsolatos tapasztalatok közötti szakadék. *CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 52865297. <https://doi.org/10.1145/2858036.2858288>.
- [62] François Mairesse, Marilyn A Walker, Matthias R Mehl és Roger K Moore. 2007. Nyelvi jelek használata a személyiség automatikus felismeréséhez beszélgetésben és szövegben. *Journal of artificial intelligence research* 30 (2007), 457-500.
- [63] Douglas R McCallum és James L Peterson. 1982. Számítógépes olvashatósági indexek. In *Proceedings of the ACM'82 Conference*. 44-48.
- [64] Patrick E McKight és Julius Najab. 2010. Kruskal-Wallis teszt. *The corsini encyclopedia of psychology* (2010), 1-1.
- [65] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado és Jeffrey Dean. 2013. Szavak és mondatok elosztott reprezentációi és kompozicionalitásuk. In *Neural Information Processing Systems (NIPS)*. 3111-3119.
- [66] Masahiro Mori, Karl F MacDorman és Norri Kageki. 2012. A háborzongató völgy [a terepről]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98-100.
- [67] Kellie Morrissey és Jurek Kirakowski. 2013. A chatbotok "valódisága": Kvantitatív kritériumok megállapítása. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8007 LNCS, PART 4 (2013), 87-96. https://doi.org/10.1007/978-3-642-39330-3_10
- [68] Nora A. Murphy. 2007. Okosnak tűnni: Az intelligencia, a személy észlelési pontosságának és a viselkedésnek a benyomáskezelése társas interakciókban. *Personality and Social Psychology Bulletin* 33, 3 (2007), 325-339. <https://doi.org/10.1177/0146167206294871>.
- [69] Clifford Nass és Youngme Moon. 2000. Gépek és esztelenség: Társadalmi válaszok a számítógépekre. *Journal of Social Issues* 1, 56 (2000), 81-103.
- [70] Clifford Nass, Jonathan Steuer és Ellen Tauber. 1994. A számítógépek társadalmi szereplők. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1994)*. ACM. <https://doi.org/10.1109/VSMM.2014.7136659>
- [71] Oda Elise Nordberg, Jo Dugstad Wake, Emilie Sektan Nordby, Eivind Flobak, Tine Nordgreen, Suresh Kumar Mukhiya és Frode Guribye. 2020. Chatbotok tervezése ADHD-s felnőttek online társas támogató beszélgetéseinek irányítására. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11970 LNCS, november (2020), 113-126. https://doi.org/10.1007/978-3-030-39540-7_8.
- [72] Nicole Novielli, Fiorella de Rosis és Irene Mazzotta. 2010. A felhasználók hozzáállása egy megtestesült társalgási ügynökhöz: Az interakciós mód hatásai. *Journal of Pragmatics* 42, 9 (2010), 2385-2397. <https://doi.org/10.1016/j.pragma.2009.12.016>
<https://doi.org/10.1016/j.pragma.2009.12.016>
- [73] Catherine Pelachaud és Isabella Poggi. 2002. Az arckifejezések finomságai megtestesült ágenseknél. *The Journal of Visualization and Computer Animation* 13, 5 (2002), 301-312.
- [74] Jeffrey Pennington, Richard Socher és Christopher Manning. 2014. Kesztyű: Globális vektorok a szóreprezentációhoz. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532-1543.
- [75] Christopher Peters. 2005. A virtuális környezetben történő beszélgetésközvetítés ágens elméletű modelljének alapjai. *Virtual Social Agents* (2005), 163.
- [76] Matthew D. Pickard, Judee K. Burgoon és Douglas C. Derrick. 2014. Toward an Objective Linguistic-Based Measure of Perceived Embodied Conversational Agent Power and Likeability. *International Journal of Human-Computer Interaction* 30, 6 (2014), 495-516. <https://doi.org/10.1080/10447318.2014.888504>
<https://doi.org/10.1080/10447318.2014.888504>
- [77] Emily Pitler és Ani Nenkova. 2008. Az olvashatóság felülvizsgálata: Egységes keretrendszer a szöveg minőségének előrejelzésére. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 186-195.
- [78] David Premack és Guy Woodruff. 1978. Van-e a csimpánznak elmélete? *Behavioral and brain sciences* 1, 4 (1978), 515-526.
- [79] David V. Pynadath és Stacy C. Marsella. 2005. PsychSim: Az elmeelmélet modellezése döntésméleti ágensekkel. *IJCAI International Joint Conference on Artificial Intelligence* (2005), 1181-1186.
- [80] Stephen Reysen. 2005. Egy új skála felépítése: A Reysen szimpatikus skála. *Social Behavior and Personality: an international journal* 33, 2 (2005), 201-208.
- [81] Tina L Robbins és Angelo S DeNisi. 1994. Az interperszonális affektus mint a teljesítményértékelés kognitív feldolgozására gyakorolt különálló hatás közelebről történő vizsgálata. *Journal of Applied Psychology* 79, 3 (1994), 341.
- [82] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill és James A Landay. 2019. Quizbot: Egy párbeszédalapú adaptív tanulási rendszer tényszerű ismeretekhez. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1-13.
- [83] Koustuv Saha, Manikanta D Reddy, Stephen Mattingly, Edward Moskal, Anusha Sirigiri és Munmun De Choudhury. 2019. Libra: A linkedin alapú szerep ambi-guitásról és annak kapcsolatáról a jólléttel és a munkateljesítménnyel. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1-30.
- [84] Koustuv Saha és Amit Sharma. 2020. A hatékony pszichoszociális

- [86] Joseph Seering, Juan Pablo Flores, Saiph Savage és Jessica Hammer. 2018. A botok társadalmi szerepei: A botok elhelyezése az online közösségek vitáiban. *Proceedings of the ACM on Human-Computer Interaction 2*, CSCW (2018). <https://doi.org/10.1145/3274426>
- [87] Joseph Seering, Michal Luria, Geoff Kaufman és Jessica Hammer. 2019. Túl a dyadikus interakciókon: A chatbotok közösségi tagként való figyelembevétele. In *Conference on Human Factors in Computing Systems - Proceedings*. <https://doi.org/10.1145/3290605.3300680>.
- [88] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman és Jessica Hammer. 2020. It Takes a Village: Adaptív chatbot integrálása egy online játékközösségbe. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (A számítástechnikai rendszerek emberi tényezői című konferencia 2020. évi jegyzőkönyve)*. ACM, New York, NY, USA, 1613. <https://doi.org/10.1145/3313831.3376708>.
- [89] James Simpson. 2020. A CUI-k csak GUI-k beszéd-buborékokkal?. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 163.
- [90] Marcin Skowron, Stefan Rank, Mathias Theunis és Julian Sienkiewicz. 2011. A jó, a rossz és a semleges: affektív profil a párbeszédrendszer-felhasználó kommunikációban. In *Affektív számítástechnika és intelligens interakció nemzetközi konferenciája*. Springer, 3376346.
- [91] Danding Wang, Qian Yang, Ashraf Abdul és Brian Y Lim. 2019. Elméletvezérelt, felhasználóközpontú, megmagyarázható mesterséges intelligencia tervezése. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1615.
- [92] Qiaosi Wang, Shan Jing, Ida Camacho, David Joyner és Ashok Goel. 2020. Jill Watson SA: Design and Evaluation of a Virtual Agent to Build Communities Among Online Learners. In *Extended Abstracts of the 2020 CHI Conference on Emberi tényezők a számítástechnikai rendszerekben*. 168.
- [93] Qiaosi Wang, Shan Jing, David Joyner, Lauren Wilcox, Hong Li, Thomas Plötz és Betsy Disalvo. 2020. Az affektus érzékelése a diákok felhatalmazása érdekében: Tanulói perspektívák az affektusérzékeny technológiáról nagy oktatási kontextusokban. In *Proceedings of the Seventh ACM Conference on Learning@Scale*. 63676.
- [94] Henry M Wellman. 1992. *A gyermek elmeelmélete*. The MIT Press.
- [95] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner és Jan Marco Leimeister. 2020. Sara, az előadó: A tanulás javítása az online oktatásban egy Scaffolding-alapú beszélgetőügynökkel. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (A számítástechnikai rendszerek emberi tényezői című konferencia 2020)*. ACM, New York, NY, USA, 1614. <https://doi.org/10.1145/3313831.3376781>
- [96] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha és Rama Akkiraju. 2017. Egy új chatbot a közösségi médiában történő ügyfélkiszolgáláshoz. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 350663510.
- [97] Xi Yang, Marco Aurisicchio és Weston Baxter. 2019. Affektív élmények megértése társalgási ügynökökkel. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1612. <https://doi.org/10.1145/3290605.3300772>
- [98] Jennifer Zamora. 2017. Sajnálom, Dave, de attól tartok, nem tehetem: Chatbot perception and expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 2536260.
- [99] Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain és Dario Taraborelli. 2018. Félresikerült beszélgetések: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345* (2018).

A FÜGGELEK

Ez az anyag (3. ábra) a kéthetente kitöltött, a diákok által készített percepciók felmérést mutatja be. Ezt a Bartneck et al. által az ember-robot interakció mérésére használták. Különösen az antropomorfizmus, az intelligencia és a szimpatizálhatóság kategóriáit választottuk ki a JW-vel kapcsolatos tanulói észlelések beállításában.

The following questions will give you a spectrum from one quality to the other on a scale of 1 to 5, such as from "Unkind"(1) to "Kind"(5). Please rate your perception of JW along each of these spectrums:

Fake 1 2 3 4 5 Natural	✓ [Select] 1 2 3 4 5
Unintelligent 1 2 3 4 5 Intelligent	[Select]
Unkind 1 2 3 4 5 Kind	[Select]
Foolish 1 2 3 4 5 Sensible	[Select]
Artificial 1 2 3 4 5 Lifelike	[Select]
Dislike 1 2 3 4 5 Like	[Select]
Awful 1 2 3 4 5 Nice	[Select]
Ignorant 1 2 3 4 5 Knowledgeable	[Select]
Machinelike 1 2 3 4 5 Humanlike	[Select]
Responding rigidly 1 2 3 4 5 Responding elegantly	[Select]
Unfriendly 1 2 3 4 5 Friendly	[Select]
Irresponsible 1 2 3 4 5 Responsible	[Select]
Unpleasant 1 2 3 4 5 Pleasant	[Select]
Incompetent 1 2 3 4 5 Competent	[Select]
Unconscious 1 2 3 4 5 Conscious	[Select]

3. ábra: Az antropomorfizmus elemeket **zöld dobozokkal**, az intelligencia elemeket **narancssárga dobozokkal**, a szimpatikus elemeket pedig **kék dobozokkal** jelöltük.