Tartalomjegyzék

AZ AI FILOZÓFIÁJA ÉS A FILOZÓFIA AI-JA

John McCarthy
Informatikai Tanszék Stanford
Egyetem
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

2006. június 25.

Absztrakt

Az X filozófiája, ahol az X egy tudomány, azt jelenti, hogy a filozófusok elemzik az X fogalmait, és néha kommentálják, hogy milyen koncepciók valószínűsíthetően koherensek vagy nem koherensek. A mesterséges intelligencia (AI) szorosabb tudományos kapcsolatban áll a filozófiával, mint más tudományok, mivel az AI számos fogalmat osztozik a filozófiával, például az akciót, a tudatot, az ismeretelméletet (mit lehet értelmesen mondani a világról), sőt még a szabad akaratot is. Ez a cikk a mesterséges intelligencia filozófiájával foglalkozik, de a mesterséges intelligencia szempontjából is elemez néhány, a filozófiában és a mesterséges intelligenciában közös fogalmat. Az X filozófiája gyakran magában foglalja az X gyakorlóinak adott tanácsokat arra vonatkozóan, hogy mit tehetnek és mit nem.

Részben megfordítjuk a szokásos irányt, és tanácsot adunk a filozófusoknak, különösen az elmefilozófusoknak. Az AI álláspontja szerint a filo- szofikus elméletek csak akkor hasznosak az AI számára, ha nem zárják ki az emberi szintű mesterséges rendszereket, és alapot nyújtanak a hiedelmekkel rendelkező, érvelő és tervező rendszerek *tervezéséhez*. A mesterséges intelligencia kutatások különösen nagy hangsúlyt fektetnek a szituációban rendelkezésre álló cselekvések és a több cselekvés közül mindegyik megtételének következményeinek formalizálására. Ennek érdekében a mesterséges intelligencia

elsősorban a jelenségek egyszerű közelítésével foglalkozott.

A mesterséges intelligencia és a filozófia számára egyaránt kulcsfontosságú probléma a társas érzékszervi tudás és képességek megértése. A *kom- monális értelemben vett informatikai helyzet* fogalmát kezeljük, azt a helyzetet, amelyben egy ember vagy egy számítógépes program van, amikor a rendelkezésre álló tudás részleges mind a tárgyak, mind az elmélet tekintetében, és rosszul definiált fogalmakat kell használni. Az általánosságban rosszul definiált fogalmak speciális kontextusokban pontosak lehetnek.

1 Bevezetés

Richmond Thomason (Thomason 2003) írta

> A mesterséges intelligencia és a filozófiai logika közötti kapcsolatok egy nagyobb történet részei. Nehéz olyan nagy filozófiai témát találni, amely ne keveredne az észérvekkel kapcsolatos kérdésekkel. Az implikatúráknak például olyan következtetéseknek kell megfelelniük, amelyeket a diskurzus racionális értelmezője el tud végezni. Bármi legyen is a kauzalitás, az oksági kapcsolatoknak a mindennapi józan ész keretei között is kikövetkeztethetőnek kell lenniük. Bármi legyen is a hiedelem, a racionális ágensek számára lehetővé kell tenni, hogy más ágensek hiedelmeire vonatkozóan hihető következtetéseket vonjanak le. A racionális ágensek viselkedését meghatározó céloknak és állandó korlátozásoknak lehetővé kell tenniük ésszerű tervek kialakítását.

A mesterséges intelligencia és a filozófia kapcsolata számos olyan fogalmat foglal magában, amelyek mindkét alanyi körbe tartoznak - például cselekvés, célok, tudás, hit és tudatosság. A mesterséges intelligencia azonban a *tervezői álláspontot* képviseli e fogalmakkal kapcsolatban; azt kérdezi, hogy milyen tudásra, hitre, tudatra stb. van szüksége egy számítógépes rendszernek ahhoz, hogy intelligensen viselkedjen, és hogyan építse be ezeket egy számítógépes programba. A filozófusok általában absztraktabb nézőpontot képviselnek, és azt kérdezik, hogy mi a tudás stb. A *tervezői álláspont* rokon Daniel Dennett *tervezői álláspontjával*(Dennett 1978), de nem ugyanaz. A de- sign álláspont egy létező műtárgyat vagy szervezetet abból a szempontból vizsgál, hogy mire tervezték, vagy mire fejlődött ki. A tervezői álláspont azt vizsgálja, hogyan kell megtervezni egy műtárgyat. Ez szükségessé teheti, hogy tudást, meggyőződéseket stb. adjunk neki, valamint a tervek megtervezésének és végrehajtásának képességét.

A filozófiai kérdések különösen fontosak a mesterséges intelligencia szempontjából, amikor emberi szintű mesterséges intelligenciára törekszünk.

Az 1970-es évek óta a legtöbb mesterséges intelligencia-kutatás azonban nem az emberi szintű mesterséges intelligenciára irányul, hanem a mesterséges intelligencia elméleteinek és technikáinak konkrét problémákra való alkalmazására.

El kell ismernem, hogy elégedetlen vagyok a legtöbb mesterséges intelligencia kutatótársam ambíciójának hiányával. Sok hasznos és érdekes programot írnak a mesterséges intelligencia és a filozófia közös fogalmainak használata nélkül. Például a Garri Kaszparov sakkvilágbajnokot legyőző Deep Blue program által használt nyelvezetet nem lehet arra használni, hogy kifejezzük: "Sakkprogram vagyok, de sokkal több lényegtelen lépést veszek figyelembe, mint az ember.", és ebből következtetéseket vonjunk le. A program tervezői nem látták szükségét ennek a képességnek. Hasonlóképpen a DARPA Grand Challenge járművezetési versenyén induló programok közül egyik sem tudta, hogy az egyik a 20 versenyző program közül. A DARPA versenybírói megakadályozták, hogy a járművek lássák egymást azzal, hogy szükség esetén szünetet kellett tartaniuk. Egy fejlettebb versenyen, ahol az egyik jármű megelőzheti a másikat, szükség lehet némi tudatosságra a "más elmékről".

Az 1950-es évek AI-kutatói valóban az emberi szintű intelligenciára gondoltak. Alan Turing, a mesterséges intelligencia úttörője volt az első, aki azt is hangsúlyozta, hogy a mesterséges intelligenciát számítógépes programok fogják megvalósítani. Most nagyobb az érdeklődés az emberi szintű mesterséges intelligencia és az annak elérésére szolgáló módszerek iránt, mint az elmúlt 40 évben.

(Nilsson 2005) egy kritériumot kínál arra, hogy mikor érte el az emberi szintű mesterséges intelligencia szintjét. Eszerint a rendszernek meg kell tudnia tanítani az emberek által végzett munkák széles körét - különösen, hogy képesnek kell lennie arra, hogy átmenjen a vizsgákon, amelyeket az emberek kiválasztására használnak ezekhez a munkákhoz, elismerve, hogy a vizsgák letétele lehetséges anélkül is, hogy megfelelő józan ésszel rendelkezne a munka elvégzéséhez. Nilsson nem részletezi, hogy milyen tanításról van szó, és az ő kritériuma gyengébb, mint Lenat azon követelménye, hogy a rendszer képes legyen az emberek számára írt tankönyvekből tanulni. Egyetértek azzal, hogy ez az emberi szintű mesterséges intelligencia egyik követelménye.

(McCarthy 1996a) szintén tárgyalja az emberi szintű mesterséges intelligencia kritériumait, hangsúlyozva a józan informatikai helyzetet.

Még ha az emberi szintű mesterséges intelligenciára irányuló munka növekszik is, a mesterséges intelligencia kutatása és a filozófiai kutatás közötti fontos módszertani különbségek valószínűleg megmaradnak. Vegyük például a hit fogalmát. A filozófusok a hitet általánosságban vizsgálják. A mesterséges intelligencia kutatása valószínűleg a nagyon korlátozott hiedelmekkel rendelkező rendszerekkel folytatódik, és onnan építkezik tovább. Talán ezek a fentről lefelé és alulról felfelé irányuló megközelítések.

A következő példával kapcsolatban a mesterséges intelligencia és a filozófia néhány közös fogalmát tárgyaljuk.

Egy rendőr megállít egy autót, és azt mondja,

"Megbüntetem gondatlan vezetésért. Ha egy másik autó jött volna a dombon, amikor maga megelőzte azt a BMW-t, akkor frontális ütközés történt volna."

Vegyük észre, hogy a példa egy kontrafaktuális feltételes módot tartalmaz: "ha átmentél volna . . . . " és egy nem kontrafaktuális következmény ". . . gondatlan vezetés". Talán kevésbé nyilvánvaló, hogy a mondatot megértő rendszernek egy megfelelő kontextusba kell ugrania, és abban a kontextusban kell érvelnie, a kontextusban értelmes fogalmakat használva. Így egy konkrét hipotetikus frontális ütközésről van szó, nem pedig például statisztikákról arról, hogy milyen gyakran végződik halálos kimenetelű frontális ütközés.

Az X filozófiája, ahol az X egy tudomány, gyakran magában foglalja, hogy a filozófusok elemzik az X fogalmait, és véleményezik, hogy mely fogalmak valószínűleg koherensek vagy nem koherensek. A mesterséges intelligencia szükségszerűen sok fogalmat oszt meg a filozófiával, pl. cselekvés, tudatosság, ismeretelmélet (mit van értelme mondani a világról), sőt még a szabad akarat is.

Ez a cikk a mesterséges intelligencia filozófiáját tárgyalja, de a 6. szakasz megfordítja a szokásos irányt, és a mesterséges intelligencia szempontjából elemzi a filozófia néhány alapfogalmát. Az X filozófiája gyakran magában foglalja az X gyakorlóinak adott tanácsokat arra vonatkozóan, hogy mit tehetnek és mit nem. A 6. szakasz megfordítja a szokásos irányt, és tanácsokat ad a filozófusoknak, különösen az elmefilozófusoknak. Az egyik pont az, hogy a filozófiai elméleteknek csak akkor lehet értelmük számunkra, ha nem zárják ki az emberi szintű mesterséges rendszereket. A filozófiai elméletek akkor a leghasznosabbak, ha a *tervezői álláspontot képviselik*, és javaslatokat tesznek arra vonatkozóan, hogy milyen tulajdonságokat tegyünk az intelligens rendszerekbe.

Az elmefilozófia az elmét mint jelenséget tanulmányozza, és azt vizsgálja, hogy a gondolkodás, a tudás és a tudat hogyan hozható kapcsolatba az anyagi világgal. A mesterséges intelligencia a gondolkodó és cselekvő számítógépes programok tervezésével foglalkozik. Ez a filozófiában vizsgált problémák néhány eltérő megközelítéséhez vezet, és amellett fogunk érvelni, hogy ez új megfontolásokat vagy legalábbis eltérő hangsúlyokat ad, amelyeket a filozófusoknak figyelembe kell venniük. Megragadom az alkalmat, hogy ebben a kézikönyvben meglehetősen szemtelenül bemutassak néhány gondolatot és formalizmust.

Néhány formalizmust, például a nem monotonikus érvelést és a szituációs kalkulációt, nagymértékben használják a mesterséges intelligencia rendszerekben. Másokat még nem használtak számítógépes programokban, de úgy gondolom, hogy az általuk megoldott problémák fontosak lesznek az

emberi szintű mesterséges intelligencia számára.

2 Néhány történelmi megjegyzés

Bár voltak előzmények, a komoly AI-munka az 1950-es évek elején kezdődött, amikor nyilvánvalóvá vált, hogy az elektronika elég fejlett ahhoz, hogy univerzális számításokat végezzen. Alan Turing (Turing 1947) felismerte, hogy az általános célú számítógépek programozása jobb, mint a speciális célú gépek építése. Ez a megközelítés attól függött, hogy a mesterséges intelligencia kutatói hozzáférjenek-e a számítógépekhez, ami az 50-es évek elején még marginális volt, de az 1950-es évek végére már szinte általános volt.[1]

Az 1956-os dartmouthi műhely, amelynek 1955-ös javaslata bevezette a fogalmat. *a mesterséges intelligencia* kiváltotta a mesterséges intelligenciát mint megnevezett területet.[2]

Az én (McCarthy 1959) munkám indította el a logikai mesterséges intelligencia, azaz a matematikai logikai nyelvek és érvelés használata a józan ész reprezentálására. A logikai mesterséges intelligencia fejlődése folyamatos volt, de még mindig messze van az emberi szinthez képest.

Az Ernst-Newell-Simon *Általános Problémamegoldó* (GPS) (Ernst és Newell 1969) azon az elképzelésen alapult, hogy a problémamegoldást a start- a kezdeti kifejezéssel, és azt adott szabályok alkalmazásával célkifejezéssé alakítja át. Sajnos ez az elképzelés nem volt megfelelő a problémamegoldás általános céljára.

Az első sakkprogramokat az 1950-es években írták, és a heurisztikák és a gyorsabb számítógépek kombinációja révén a 90-es évek végén érték el a világbajnoki szintet. Sajnos a bajnoki szintű sakkozáshoz megfelelő ötletek nem megfelelőek az olyan játékokhoz, mint *a go,* amelyek jobban elágaznak, mint a sakk, és amelyek egy szituáció részeinek felismerését igénylik.

Marvin Minsky (Minsky 1963) összegezte az 1963-ban rendelkezésre álló elképzeléseket. (McCarthy és Hayes 1969) a szituációs kalkulus formalizmusát egy

nagy AI közönség.

Pat Hayes (Hayes 1979) és (Hayes 1985) egy sor olyan gondolatot terjesztett elő, amelyek a későbbi mesterséges intelligencia kutatásokra is hatással voltak.

David Marr (Marr 1982) a 2 és fél dimenziós ábrázolás ötletével nagy hatással volt a számítógépes látás területén végzett munkára.

A Stanfordi Mesterséges Intelligencia Laboratórium bemutatta az első robotkarokat, amelyeket a TV-kamerákból származó adatok alapján programokkal irányítottak. (Moravec 1977)

[1] 1948-ban kezdtem el gondolkodni a mesterséges intelligenciáról, de a számítógépekhez való hozzáférésem 1955-ben kezdődött. Ez térített át Turing véleményére.

[2]Newell és Simon, akik elsőként kezdték el, és akiknek határozott eredményeik voltak, amelyeket Dartmouthban bemutattak, néhány évig a ***komplex információfeldolgozás*** kifejezést használták, ami nem tett igazságot saját munkájuknak.

leírt egy kocsit, amely egy TV-kamerát tartalmazott, amelyet egy időben megosztott számítógépről rádión keresztül irányítottak.

Az 1960-as éveken túl nem nagyon fogok elmenni a mesterséges intelligencia kutatásának általános leírásában, mert a saját érdeklődési köröm túlságosan speciális lett ahhoz, hogy igazságot tegyek a munkának.

3 A AI filozófiai előfeltételei

Az, hogy lehetséges lenne olyan intelligens gépeket létrehozni, mint az ember, magában foglal néhány filozófiai előfeltevést, bár a filozófusok többsége valószínűleg elfogadja ezt a lehetőséget. Az általunk javasolt mód, ahogyan intelligens gépeket akarunk építeni, több előfeltevést tartalmaz, amelyek közül néhány valószínűleg ellentmondásos lesz.

Ez a rész kissé dogmatikus, mivel nem nyújt részletes érveket állításaihoz, és nem tárgyal más filozófiai álláspontokat, csak az ellentétek bemutatásával.

A mi módszerünket *logikai mesterséges intelligenciának* nevezzük, és magában foglalja a számítógépben lévő tudás logikai nyelveken történő kifejezését és logikai következtetésekkel történő következtetést, beleértve a nem monoton következtetést is. A mesterséges intelligencia másik fő megközelítése az emberi neurofiziológia tanulmányozását és utánzását foglalja magában. Ez is működhet.

Íme, a logikai mesterséges intelligencia filozófiai előfeltételei. Ezek a legfontosabbak az emberi szintű mesterséges intelligenciára irányuló kutatások szempontjából. Rengeteg van belőlük. A jelenlegi mesterséges intelligencia azonban túlságosan korlátozott célokat tűzött ki maga elé ahhoz, hogy fontos legyen a filozófia helyes megvilágítása.

objektív világ A világ az embertől függetlenül létezik. A matematika és a természettudományok tényei függetlenek attól, hogy vannak-e emberek, akik megismerhetik őket. Az intelligens marslakóknak és robotoknak ugyanazokat a tényeket kell ismerniük, mint az embereknek.

A robotnak azt is el kell hinnie, hogy a világ tőle függetlenül létezik, és hogy nem tud mindent megtudni a világról. A tudomány azt mondja, hogy az emberek egy olyan világban fejlődtek ki, amelyben korábban nem voltak emberek. Ezt figyelembe véve furcsa a világot az érzékszervi adatokból származó emberi konstrukciónak tekinteni. Még furcsább egy robotot úgy programozni, hogy a világot saját konstrukciójának tekintse. Az, hogy a robot mit hisz a világról általában, nem merül fel a mai korlátozott robotok számára, mert az általuk programozott nyelvek nem

képesek a világról általánosságban tett állításokat kifejezni. Ez korlátozza azt, amit megtanulhatnak, vagy amit elmondhatnak nekik...

és ezért mit tudunk rávenni őket, hogy tegyenek meg értünk.[3]

A példában sem a sofőrnek, sem a rendőrnek nem okoz gondot az objektív világ létezése. Egy robot sofőrnek vagy rendőrnek sem kellene.

Az igazság korrespondenciaelmélete Egy logikai robot logikai mondatokkal reprezentálja azt, amit *a* világról *gondol*. E hiedelmek egy részét mi építjük be; mások a megfigyeléseiből származnak, megint mások pedig a tapasztalataiból adódnak. A mondatokon belül kifejezésekkel utal a világ tárgyaira.

Minden esetben megpróbáljuk úgy megtervezni, hogy amit a világról hinni fog, az a lehető legpontosabb legyen, bár általában nem a lehető legrészletesebb. A robot hibakeresése és fejlesztése magában foglalja a világról alkotott téves hiedelmek felderítését és az információszerzés módjának megváltoztatását, hogy maximalizáljuk a megfelelést a világ tényei és az általa hitt dolgok között.

A referencia korrespondenciaelmélete A mesterséges intelligenciának szüksége van a *referencia korrespondenciaelméletére* is, vagyis arra, hogy egy mentális struktúra egy külső tárgyra hivatkozhat, és a referencia pontossága alapján megítélhető. A robot által az entitásokra való hivatkozáshoz használt kifejezéseknek meg kell felelniük az entitásoknak, hogy a mondatok tényeket fejezzenek ki ezekről az entitásokról. Gondolunk itt anyagi tárgyakra és más entitásokra is, például egy tervre vagy a héliumatom elektronikus szerkezetére. A referencia megfelelés ellenőrzésének egyszerű esete az, amikor egy robotot arra kérünk, hogy vegye fel a B3-as blokkot, és akkor ezt a blokkot veszi fel, és nem egy másik blokkot.

A tudományhoz hasonlóan a robotok elméleteit is kísérletileg tesztelik, de a robotok által használt fogalmakat aligha határozzák meg kísérletekkel. Tulajdonságaik részben axiomatizálva vannak, és néhány axióma a megfigyeléseken keresztül a fogalmakat reprezentáló kifejezéseket a világ tárgyaihoz kapcsolja.

Egy robotrendőrnek hibakeresésre lenne szüksége, ha azt hinné, hogy egy autó 20 mérföld/órás sebességgel halad, miközben valójában 75 mérföld/órás sebességgel megy. Szintén hibakeresésre szorulna

[3]A fizika, a kémia és a biológia már régóta azon a szinten van, ahol az érzékelést a tudomány szempontjából jobban meg lehet érteni, mint a (Russell 1914) projektet végrehajtani, hogy a tudományt az érzékelés szempontjából építsük fel. A józan ész és a tudományos megismerés igazolása az emberi érzékelésről és annak a világhoz való viszonyáról alkotott teljes tudományos kép, nem pedig az érzékelésből való konstrukció

szempontjából.

ha a belső vizuális memóriája egy tehenet emelt ki, miközben egy bizonyos autót kellett volna kiemelnie.

A referenciák korreszpondenciaelmélete szükségszerűen bonyolultabb lesz, mint az igazság elmélete, mivel a kifejezések a világ tárgyaira vagy a szemantikai értelmezések tárgyaira vonatkoznak, míg a mondatok igazságértékekre. Sajnos, a valóságos világra vonatkozó referenciaelméleteket nem nagyon tanulmányozták. A kognitív tudósok és a velük szövetséges filozófusok ***a szimbólum-alapozás problémájára*** hivatkoznak, de nem vagyok benne biztos, hogy mire gondolnak.

valóság és megjelenés A megfelelési elmélet fontos következménye, hogy szem előtt kell tartani a ***megjelenés***, a robot érzékelőin keresztül érkező információ és a ***valóság*** közötti kapcsolatot. Csak bizonyos egyszerű esetekben, például amikor egy program beírt lépésekkel sakkozik, a robotnak elegendő hozzáférése van a valósághoz ahhoz, hogy ezt a különbséget figyelmen kívül lehessen hagyni. Egy fizikai robot, amely a táblát nézegetve és a figurákat mozgatva sakkozik, két szinten működne: egy absztrakt szinten, amely (mondjuk) algebrai jelölést használ a pozíciók és lépések számára, és egy konkrét szinten, ahol egy figurának egy mezőn meghatározott alakja, helye és orientációja van, ez utóbbi szükséges az ellenfél lépésének felismeréséhez és a saját lépésének megtételéhez a táblán. A látórendszerének a TV-képekből kellene kiszámítania a pozíciók algebrai reprezentációit.

Az evolúció véletlenje, hogy a denevérekkel ellentétben nekünk nincs olyan ultrahangos érzékünk, amely a tárgyak belső szerkezetéről adna információt.

A józan ész és a tudomány szerint a világ háromdimenziós, és a tárgyak általában összetett belső struktúrával rendelkeznek. Az emberek és az állatok érzékszervei az evolúció véletlenjei. Nincs közvetlen hozzáférésünk a tárgyak belső szerkezetéhez, vagy ahhoz, hogy azok hogyan épülnek fel atomokból és molekulákból. Az érzékszerveink és az érvelésünk összetett módon tájékoztatnak bennünket a világ tárgyairól.

Egyes robotok közvetlenül, memória vagy következtetések nélkül reagálnak a bemenetekre. A mi tudományos (azaz nem filozófiai) állításunk szerint ezek nem megfelelőek az emberi szintű intelligencia számára, mert egy robotnak túl sok olyan fontos entitásról kell következtetnie, amelyet nem lehet közvetlenül teljes mértékben megfigyelni.

Az információszerzésről érvelő robotnak magának is tisztában kell

lennie ezekkel az összefüggésekkel. Annak érdekében, hogy egy robot ne higgye mindig, hogy

amit a saját szemével lát, különbséget kell tennie a látszat és a valóság között.

Egy robotrendőrnek is szkeptikusnak kellene lennie azzal kapcsolatban, hogy az, amire emlékszik, hogy látta (megjelenés), megfelel-e a valóságnak.

harmadik személyű nézőpont Azt kérdezzük: "Honnan tudja (vagy ő)?", "Mit érzékel?", ahelyett, hogy azt kérdeznénk: "Honnan tudom és mit érzékelek?". Ez összeegyeztethető az igazság és a referencia korrespondenciaelméleteivel. Ez vonatkozik arra, ahogyan a robotokra tekintünk, de arra is, ahogyan azt szeretnénk, hogy a robotok az emberek és más robotok tudásáról érveljenek.

A járművezető és a rendőr közötti interakció során mindketten a másik tudására következtetnek.

tudomány A tudomány alapvetően helyes abban, amit a világról mond, és a tudományos tevékenység a legjobb módja annak, hogy több ismerethez jussunk. A korábbi tudományos ismeretek 20. századi korrekciói a régi elméleteket többnyire jó közelítésként hagyták meg a valósághoz. Amióta a tudomány elvált a filozófiától (mondjuk Galilei idején), a tudományos elméletek megbízhatóbbak, mint a filozófia, mint tudásforrás.

A rendőr jellemzően a radarjára hagyatkozik, bár nem valószínű, hogy sokat tud a mögötte álló tudományról.

elme és agy Az emberi elme az emberi agy tevékenysége. Ez egy tudományos állítás, amelyet a tudomány által eddig felfedezett összes bizonyíték alátámaszt. Az elme és a test szétválasztásának dualista intuíciója azonban azzal a ténnyel függ össze, hogy gyakran fontos, hogy cselekvés nélkül gondolkodjunk a cselekvésről. A dualista elméleteknek lehet némi haszna pszichológiai absztrakcióként. Egy programozott robot esetében az elme és az agy (a program és a számítógép) közötti elválasztás eléggé élesre tehető.

a józan ész A világ és a közvélekedés józan ész alapján történő észlelése és a közvélemény is többnyire helyes. Ha az általános józan ész téved, azt a tudomány gyakran korrigálhatja, és a korrekció eredményei a józan ész részévé válhatnak, ha nem túl matematikaiak. Így a józan ész magába olvasztotta a tehetetlenség fogalmát. Matematikai általánosítása, az impulzusmegmaradás törvénye azonban

csak az emberek kis hányadának a józan eszébe jutott el - még azok között is, akik fizikából tanultak. Azoknak, akik aszteroidákra költöznek, be kell építeniük az intuíciójukba az impulzus és még a szögimpulzus megőrzését is.

Szókratésztől kezdve a filozófusok számos hiányosságot találtak a köznapi értelemben vett használatban, például a szavak jelentéséről alkotott köznapi felfogásban. A korrekciók gyakran kidolgozások, a köznapi értelemben vett használatban elmosódott megkülönböztetéseket tesznek. Sajnos sok fogalom lehetséges elaborációjának nincs vége, és az elméletek nagyon bonyolulttá válnak. A kifejtések némelyike azonban bizonyos körülmények között elengedhetetlenül fontosnak tűnik a zavar elkerülése érdekében.

A robotoknak a legegyszerűbb, józan ész szerinti használatra lesz szükségük, és szükség esetén a kidolgozottságot is tolerálniuk kell. Ehhez három fogalmat vetettünk fel - a kontextus mint formális objektum (McCarthy 1993) és (McCarthy és Buvač 1997), az *elaboráció-tűrés* (McCarthy 1999b) és a *közelítő objektumok*. (McCarthy 2000)[4]

a tudomány a józan észbe ágyazva A tudomány a józan észbe ágyazva van. Galilei megtanította nekünk, hogy az a távolság s, amit egy leejtett test leesik...

---

[4]Hilary Putnam (Putnam 1975) két, korábbi filozófusok által javasolt, a jelentéssel kapcsolatos felfogást tárgyal, amelyeket nem tart megfelelőnek. Ezek a következők

> (I) Hogy egy kifejezés jelentésének ismerete csupán egy bizonyos "pszichológiai állapot" (a "pszichológiai állapot" azon értelmében, amelyben az emlékezeti állapotok és a pszichológiai diszpozíciók "pszichológiai állapotok"; természetesen senki sem gondolta, hogy egy szó jelentésének ismerete folyamatos tudatállapot lenne).
>
> (II) Hogy egy kifejezés jelentése (az "intenció" értelmében) meghatározza a kiterjesztését (abban az értelemben, hogy az intenció azonossága az ex-tenzió azonosságát vonja maga után).

Tegyük fel, hogy Putnamnak igaza van az (I) és (II) általános helyességére vonatkozó kritikájában. Saját elképzelései sokkal kidolgozottabbak.

Kényelmes lehet egy robot számára, hogy többnyire egy nagyobb, ***Cphil1*** kontextuson belüli kontextusban dolgozzon, amelyben (I) és (II) (vagy valami még egyszerűbb) érvényes. Azonban ugyanennek a robotnak, ha emberi szintű intelligenciával akar rendelkezni, képesnek kell lennie *túllépni a* ***Cphil1-en***, amikor olyan kontextusokban kell dolgoznia, amelyekre Putnam kritikája a ***Cphil1*** feltételezéseivel kapcsolatban érvényes.

Érdekes, de talán nem szükséges az AI számára, hogy jellemezzük ezeket a körülményeket. olyan álláspontok, amelyekben az (I) és a (II) helyes.

a t időpontban a következő képlettel adható meg

$$s = \frac{1}{2} gt^2$$

Ezen információk használatához az angol vagy olasz (vagy ezek logikai megfelelője) ugyanolyan lényeges, mint a képlet, és a képlet használatához vagy ellenőrzéséhez szükséges mérések elvégzéséhez a józan ész világismeretére van szükség.

a józan ész kifejezhető matematikai logikában A józan ész ismerete és érvelése kifejezhető logikai formulák és logikai rea- soning formájában. A jelenlegi matematikai logika néhány kiterjesztésére van szükség.

a mesterséges intelligencia lehetősége Egyes filozófusok szerint a mesterséges intelligencia vagy ellentmondásos (Searle 1984), vagy eredendően lehetetlen (Dreyfus 1992), vagy (Penrose 1994). Ezeknek az érveknek a módszertani ba- sisának kell tévesnek lennie, és nem csak maguknak az érveknek.

A mesterséges intelligenciának **az** elmét komponensek szerint kell kezelnie, ahelyett, hogy az elmét olyan egységnek tekintené, amely szükségszerűen rendelkezik az emberekben előforduló összes mentális tulajdonsággal. Így néhány nagyon egyszerű rendszert tervezünk meg az általunk kívánt hiedelmek szempontjából, és a hibás hiedelmek azonosításával hibaelhárítjuk őket. A szisztematikus elmélete lehetővé teszi, hogy olyan egyszerű entitásoknak, mint a termosztátok, minimális hiedelmeket tulajdonítsunk, analóg módon ahhoz, ahogyan a számrendszerben a 0 és az 1 szerepel. Így egy egyszerű termosztátnak csak az lehet a lehetséges hiedelmek halmaza, hogy a szobában túl meleg van, vagy hogy túl hideg van. Nem kell tudnia, hogy ő egy termosztát. Ez vitához vezetett a filozófusokkal, például John Searle-lel, akik szerint a hiedelmeket csak olyan rendszereknek lehet tulajdonítani, amelyeknek nagy a mentális tulajdonságkészlete. (McCarthy 1979a) részletesen foglalkozik a termosztát példával.

gazdag ontológia Elméleteink sokféle entitást tartalmaznak - anyagi tárgyakat, helyzeteket, tulajdonságokat mint tárgyakat, kontextusokat, tételeket, egyéni elképzeléseket, kívánságokat, szándékokat. Még akkor is, ha egyfajta A entitást a többivel együtt definiálhatunk, gyakran inkább külön kezeljük A*-t, mert* később esetleg meg akarjuk változtatni a többi entitáshoz való viszonyáról alkotott elképzeléseinket.

A mesterséges intelligenciának számos kapcsolódó fogalmat kell figyelembe vennie, ahol sok filozófus a minimális ontológiákat támogatja. Tegyük fel, hogy egy ember lát egy kutyát. A látás az ember és a kutya közötti kapcsolat, vagy az ember és a kutya megjelenése közötti kapcsolat? Egyesek azzal cáfolják, hogy a látást az ember és a kutya közötti kapcsolatnak neveznék, hogy rámutatnak, hogy az ember valójában egy hologramot vagy a kutya képét látja. Az AI-nak szüksége van az ember és a kutya megjelenése közötti kapcsolatra, az ember és a kutya közötti kapcsolatra, valamint a kutyák és a kutyák megjelenése közötti kapcsolatra. Egyiket sem kell a legalapvetőbbnek tekinteni.

Mind a járművezető, mind a rendőr olyan feldúsított ontológiákat használ, amelyek olyan fogalmakat tartalmaznak, amelyeknek a definíciója az alapfogalmak szempontjából ismeretlen vagy akár meghatározatlan. Így mindkettőjüknek van olyan fogalma az autóról, amely nem az autó részeinek előzetes ismeretén alapul. A rendőrnek vannak fogalmai és nevei az olyan szabálysértésekről, amelyekért bírságot kell kiszabni, és azokról, amelyekért letartóztatás szükséges.

természetes fajták A robot által hivatkozandó entitások gyakran olyan tulajdonságokkal *rendelkeznek, amelyekről* a robot nem tudhat mindent. A legjobb példa erre egy olyan *természetes fajta*, mint a citrom. Egy gyerek, aki citromot vásárol egy boltban, elég tulajdonságot ismer az általa látogatott boltokban előforduló citromokról ahhoz, hogy megkülönböztesse a citromot a többi gyümölcstől az adott boltban. A gyermek számára kon- venció, hogy a citrom és a narancs között nem létezik a gyümölcsök kontinuuma. A hegyek és hegyek megkülönböztetése több problémát és nézeteltérést okoz. A szakértők több tulajdonságát ismerik a citromnak, mint mi laikusok, de senki sem ismeri mindet. A mesterséges intelligencia rendszereknek is különbséget kell tenniük azon tulajdonságok halmazai között, amelyek elegendőek egy tárgy felismeréséhez bizonyos típusú helyzetekben, és egy általános fajta között.

Érdekes módon a filozófiában vizsgált fogalmak közül sok nem természetes fajta, például a tétel, a jelentés, a szükségszerűség. Amikor természetes fajtáknak tekintjük őket, gyakran eredménytelen viták folynak arról, hogy valójában mik is ezek. A mesterséges intelligenciának szüksége van ezekre a fogalmakra, de képesnek kell lennie arra, hogy korlátozott fogalmakkal dolgozzon.

közelítő entitások Számos, a társalgásban és írásban sikeresen használt, köznapi értelemben vett kifejezést és tételt nem lehet a párbeszéd

résztvevői által elfogadott, ha- és csak-ha-definícióval ellátni. Ilyen például az "x hisz y", amely sok filozófiai figyelmet kapott, de olyan kifejezések is, mint a "hely(x)", amelyek nem.

Néhányan azt mondták, hogy a számítógépek használata megköveteli a fogalmak pontos meghatározását, de én nem értek ezzel egyet. A számítógépes programoknak sok közelítő entitást kell figyelembe venniük, belsőleg és a kommunikációban. A pontosság azonban gyakran akkor érhető el, ha a kifejezéseket és kijelentéseket az adott helyzetnek megfelelő kontextusban értelmezzük. Az emberi használatban maga a kontextus általában nincs explicite meghatározva, és az emberek megértik egymást, mert a közös kontextus implicit.

A közelítő entitások első osztályú jellegének hangsúlyozása újdonság lehet. Ez azt jelenti, hogy a közelítő entitások felett számszerűsíthetünk, és azt is kifejezhetjük, hogy egy entitás hogyan közelítő. (McCarthy 2000) közelítő entitásokat és közelítő elméleteket kezel.

A "Ha egy másik autó jött volna a dombon, amikor ön elhaladt. . . . " nagyon közelítő. A járművezető és a rendőr közötti kommunikációhoz megfelelő, de a pontosabb meghatározásra tett kísérletek valószínűleg nem érnének egyet.

Van némi átfedés a közelítő entitások és a homályosságról szóló filozófiai viták között. A mi szempontunk azonban a közelítő entitások szükségessége a mesterséges intelligenciában.

A determinizmus és a szabad akarat összeegyeztethetősége Egy logikus robotnak mérlegelnie kell a döntéseit és azok következményeit. Ezért úgy kell tekintenie magát, mintha egyfajta *szabad akarattal* rendelkezne (és valóban rendelkezik is), annak ellenére, hogy determinisztikus eszköz. A példában a bírónak fel lehet ajánlani azt a kifogást, hogy a sofőr nem tudott visszalépni, miután elkezdett előzni, mert valaki közvetlenül mögötte volt.

(McCarthy 2005) a determinisztikus szabad akarat egy egyszerű formáját formalizálja. Egy robot vagy ember cselekvése néha két szakaszból áll. Az elsőben egy nemdeterminisztikus elméletet, pl. *helyzetkalkulust* használunk a választások és következményeik halmazának kiszámítására, valamint a cselekvések végrehajtásából adódó helyzetek értékelésére. A második szakaszban kiválasztja azt a cselekvést, amelynek következményeit a legjobbnak ítéli. A szabad akarat érzékelése az első szakasz végén kialakult helyzet. A választási lehetőségek kiszámításra kerültek, de a cselekvés még nincs eldöntve vagy végrehajtva. Ez az egyszerű elmélet önmagában is hasznos lehet, de ki kell dolgozni, hogy az emberi szabad akarat további aspektusait is figyelembe lehessen venni. Az igény mind filozófiai, mind filozófiai

és praktikus a robottervezéshez. Az emberi szabad akarat egyik olyan aspektusa, amely valószínűleg szükségtelen a robotok esetében, az akarat gyengesége.

elme-agy megkülönböztetés Nem vagyok benne biztos, hogy ez a pont filozófiai vagy tudományos. Az elme némileg megfelel a szoftvereknek, talán a program és a tudás közötti belső megkülönböztetéssel. A szoftver nem csinál semmit hardver nélkül, de a hardver lehet egészen egyszerű, pl. egy univerzális Turing-gép vagy egy egyszerű tárolt programú számítógép. Egyes hardverkonfigurációk sok különböző programot tudnak egyidejűleg futtatni, azaz sok elme lehet ugyanabban a számítógéptestben. A szoftverek más szoftvereket is képesek értelmezni.

Az ezzel kapcsolatos zűrzavar a Searle-féle kínai szoba tévedés alapja (Searle 1984). A hipotetikus kínai szobában az ember egy kínai személyiség szoftverét inter- pretálja. Egy program értelmezéséhez nem szükséges, hogy rendelkezzünk a program által birtokolt tudással. Ez nyilvánvaló lenne, ha az emberek képesek lennének más személyiségeket gyakorlati sebességgel értelmezni, de a kínai szoba szoftvere, amelyet egy segítség nélküli ember értelmez, 10^{-9} sebességgel futhat, mint egy valódi kínai.[5]

A legtöbb mesterséges intelligenciával kapcsolatos munka nem feltételez ennyi filozófiát. Például a jelenetek és más bemenetek osztályozásának nem kell feltételeznie, hogy az osztályozandó megjelenések mögött valóság van. A látszat mögötti valóság figyelmen kívül hagyása azonban nem vezet emberi szintű mesterséges intelligenciához, és néhány rövid távú mesterséges intelligencia cél is szenvedett a helytelen, filozófiai feltételezések miatt, amelyek szinte mindig implicitek.

Az emberi szintű mesterséges intelligenciának is vannak tudományos előfeltevései.

4 A AI tudományos előfeltételei

A logikai mesterséges intelligencia néhány premisszája tudományos abban az értelemben, hogy tudományos ellenőrzés vagy cáfolat tárgyát képezi. Ez igaz lehet a fentebb filozófiai értelemben felsorolt premisszák némelyikére is.

veleszületett tudás Az emberi agynak fontos veleszületett tudása van, pl. hogy a világ háromdimenziós tárgyakat tartalmaz, amelyek általában megmaradnak.

[5]Ha Searle megelégedne egy Joseph Weizenbaum (Weizenbaum 1965) szintű interakcióval, akkor egy ember számítógépes segítség nélkül is értelmezhetné a szabályokat - ahogy Weizen- baum nemrég tájékoztatott.

még akkor is, ha nem figyelik meg. Ezt a tudást az evolúció tanulta meg. A veleszületett tudás létezését nem a fogalom filozófiai elemzésével állapították meg, hanem pszichológiai kísérletekkel és elméletalkotással tanulják meg. Az ilyen tudás megszerzése az érzékszervi adatokból való tanulással meglehetősen nehéz lesz, de lehetséges.

Valóban érdemes minél több tudást beépíteni robotjainkba. Douglas Lenat CYC projektje egy kísérlet arra, hogy nagy mennyiségű józan tudást helyezzenek egy adatbázisba.

Az emberi veleszületett tudás azonosítása a közelmúlt pszichológiai kutatásainak tárgya. Lásd (Spelke 1994) és a (Pinker 1997) értekezését, valamint a Pinker által megadott hivatkozásokat. Különösen a csecsemők és a kutyák tudják veleszületetten, hogy vannak állandó tárgyak, és keresik őket, amikor eltűnnek a látóterükből. Jobb, ha ezt beépítjük a robotjainkba, csakúgy, mint a pszichológusok által azonosított egyéb veleszületett tudást. Az evolúció a
sok fáradságot okoz olyan ismeretek elsajátítása, amelyek megszerzéséhez nincs szükségünk robotokra, hogy tapasztalatból tanuljunk. Talán a természetes jellegű fogalmak gyermekkori preferenciája olyasmi, amit a robotokba be kellene építeni.

középen ki Az emberek középméretű tárgyakkal foglalkoznak, és a közepétől felfelé és lefelé fejlesztik tudásunkat. A világról szóló formális elméleteknek is abból a közepéből kell kiindulniuk, ahonnan a tapasztalataink tájékoztatnak bennünket. A legalapvetőbb fogalmakból kiinduló erőfeszítések, pl. egy alapontológia elkészítése, valószínűleg nem lesznek olyan sikeresek, mintha a közepéről indulnánk. Az ontológiának összeegyeztethetőnek kell lennie azzal a ténnyel, hogy a kiinduló ontológiánkban szereplő alapegységek nem a világ alapegységei. Több alapvető entitást,

pl. az elektronok és a kvarkok, kevésbé ismertek, mint a középső entitások.

logikai szint Allen Newell, aki nem használta a logikai mesterséges intelligenciát, mégis azt javasolta (Newell 1993), hogy az emberi racionalitásnak van egy olyan elemzési szintje, amelyet ő *logikai szintnek* nevezett, és amelyen az embereket úgy lehet tekinteni, hogy *azt teszik, amiről úgy gondolják, hogy céljaikat elérik.* A Carnegie-Mellon csoport által épített rendszerek közül sok, pl. a SOAR, először a logikai szinten készült.

Az intelligencia egyetemessége A célok elérése a világban megköveteli, hogy egy korlátozott tudással, számítási képességgel és

megfigyelési képességgel rendelkező ágens bizonyos módszereket alkalmazzon. Ez független attól, hogy az ágens

ember, marslakó vagy gép. Például a sakkszerű játékok hatékony lejátszásához valami olyasmi szükséges, mint az alfa-béta metszés.

A logika univerzális kifejezőképessége Ez a tétel analóg a Turing-tézishez, miszerint a Turing-gépek számítási szempontból univerzálisak - bármi, amit bármilyen gép kiszámíthat, azt egy Turing-gép is kiszámíthatja. A *kifejezhetőségi tézis* az, hogy bármi, ami kifejezhető, kifejezhető elsőrendű logikában, függvények és predikátumok megfelelő gyűjteményével.

A gondolat némi kidolgozására van szükség ahhoz, hogy olyan egyértelmű legyen, mint a Turing-tézis. Az elsőrendű logika nem a legjobb módja mindannak, amit ki lehet fejezni, mint ahogy a Turing-gépek sem a legjobb módja a számítások kifejezésére. Az elsőrendű logikában axiomatizált halmazelmélettel azonban, ami erősebb rendszerekben kifejezhető, az nyilvánvalóan az elsőrendű logikában is kifejezhető.

A G¨odel teljességi tétele azt mondja, hogy minden p mondat, amely igaz egy a mondathalmaz összes modelljében, levezethető. A nemmonoton következtetésre azonban szükség van, és az emberek ezt használják is, hogy egyszerű modellekben igazak legyenek a következmények. Nagyon valószínű, hogy reflexiós elvekre is szükség van.

Arra számítunk, hogy ezek a filozófiai és tudományos előfeltevések egyre fontosabbá válnak, ahogy a mesterséges intelligencia elkezdi kezelni az emberi szintű intelligenciát.

5 A józan ész és a józan ész in- formális szituáció

A fő akadálya annak, hogy a számítógépes programok emberi szintű intelligenciával rendelkezzenek, az, hogy még nem értjük, hogyan adjunk nekik emberi szintű kom- monális érzéket. A józan ész nélkül semmilyen számítógépes teljesítmény nem ad emberi szintű intelligenciát. Amint a programok rendelkeznek józan ésszel, a számítógépes teljesítmény és az algoritmusok tervezésének fejlesztése közvetlenül alkalmazható lesz arra, hogy intelligensebbé tegyük őket. A józan ész megértése számos filozófiai probléma megoldásának kulcsa is.

A logikai mesterséges intelligencia és a tudás reprezentációjával foglalkozó közösségek arra vállalkoznak, hogy logikai formulákkal

tanulmányozzák a világot és reprezentálják a köznapi tudást. A konkurens megközelítés az agy tanulmányozásán és azon alapul, hogy a köznapi értelemben vett tudás hogyan reprezentálódik a szinapszisokban és más neurológiai struktúrákban.

A CYC (Lenat 1995) egy olyan tudásbázis, amely több millió, józan ésszel felfogható tényt tartalmaz. Douglas Lenat (Matuszek et al. 2005) többször hangsúlyozta, hogy a józan ész kulcsfontosságú szintjét akkor érjük el, ha a programok a világhálóról tudományra, történelemre, aktuális ügyekre stb. vonatkozó tényeket tanulhatnak. A fent idézett 2005-ös tanulmány szerint

> A CYC-projekt eredeti ígérete - hogy a valós világból származó tudás olyan alapot biztosít, amely elegendő ahhoz, hogy támogassa azt a fajta nyelvtanulást, amelyre az emberek képesek - még nem teljesült.

Vegye észre, hogy a hiányosság inkább a józan ész, mint az angol nyelvtudás. Egyetértek.

Ez a szakasz a józan ész különböző aspektusainak nem hivatalos összefoglalása. A mesterséges intelligencia és a filozófia szempontjából egyaránt kulcsfontosságú jelenség az, amit mi a *józan ész informatikai helyzetének* nevezünk.

Mi a józan ész?

A *józan ész* a tudás, az érvelési képességek és talán más képességek bizonyos gyűjteménye.

A (McCarthy 1959) című könyvemben azt írtam, hogy az 1958-ig írt számítógépes programokból hiányzott a józan ész. A józan ész nehezen értelmezhető jelenségnek bizonyult, és a 2005-ös programok is nélkülözik a józan ész, vagy csak *korlátozott informatikai helyzetekben* rendelkeznek józan ésszel. Az 1959-es dolgozatban azt írtam: "Azt fogjuk tehát mondani, hogy **egy programnak van józan esze, ha automatikusan levezet magának egy kellően széles osztálynyi azonnali következményei mindannak, amit mondanak neki, és amit már tud.** "

A józan ésszel rendelkező programok ára (McCarthy 1959) még mindig hiányoznak, és ráadásul az abban a dokumentumban megfogalmazott elképzelések sem elegendőek. A logikai dedukció elégtelen, és nem monoton gondolkodásra van szükség. A józan ész ismerete is szükséges.

Itt van egy szerintem naprakészebb megfogalmazás.

Egy programnak van józan esze, ha elegendő józan tudással rendelkezik a világról és megfelelő következtetési módszerekkel ahhoz, hogy az ésszerű következmények egy elég széles osztályára következtessen bármiből, amit és amit már tud. Ráadásul sok olyan következtetés, amelyet az emberek nyilvánvalónak tartanak, nem levezethető. Némelyik mentális szimulációval

történik, és némelyik nem monoton érvelést foglal magában.

Ha a józan ész eszméjének részeként némi intelligenciát követelünk meg, az egy másik megfogalmazást ad.

Egy program akkor rendelkezik józan ésszel, ha képes hatékonyan cselekedni a *kom- mon értelemben vett informatikai helyzetben*, a rendelkezésre álló információkat felhasználva céljai eléréséhez.

Egy program, amely eldönti, hogy mit tegyen, rendelkezik bizonyos beépített információkkal, más információkat a bemeneteiből vagy megfigyeléseiből nyer; megint más információkat pedig az érvelés generál. Így egy bizonyos *informatikai helyzetben van*. Ha a felhasználandó információ józan ész jelleggel bír, akkor az úgynevezett *józan ész informatikai szituációban lesz*.

Szembe kell állítanunk az általános, *józan informatikai helyzetet* a kevésbé általános, *korlátozott informatikai helyzetekkel*. Ez utóbbiak ismertebbek a tudományban és valószínűleg a filozófiában is.

5.1 Korlátozott informatikai helyzetek

A jelenlegi (2006-os) tudomány és technológia megköveteli, hogy ahhoz, hogy egy bizonyos területen számítógépes programot írjunk, adatbázist építsünk, vagy akár egy formális elméletet írjunk, le kell határolni a figyelembe vett fogalmak halmazát.

A matematika és a természettudományok jelenlegi formális elméletei *korlátozott informatikai helyzetekkel* foglalkoznak. A tudós informálisan, ad- vance dönti el, hogy milyen jelenségeket vesz figyelembe. Például az égi mechanika nagy része a newtoni gravitációs elmélet keretében történik, és nem veszi figyelembe az olyan lehetséges további hatásokat, mint az üstökösből származó gázok vagy a napszél által kifejtett elektromágneses erők. Ha több jelenséget is figyelembe kell venni, a tudósoknak új elméleteket kell alkotniuk - és természetesen ezt meg is teszik.

Hasonlóképpen a jelenlegi mesterséges intelligencia formalizmusok csak korlátozott informatikai helyzetekben működnek. Azt, hogy milyen jelenségeket kell figyelembe venni, még a formális elmélet megalkotása előtt eldönti az ember. Ilyen korlátozások mellett az érvelés nagy része monoton lehet, de az ilyen rendszerek nem érhetik el az emberi szintű képességeket.

Ehhez a gépnek magának kell eldöntenie, hogy milyen információk relevánsak, és ez az érvelés elkerülhetetlenül részben nem monoton lesz.

Az egyik példa a mesterséges intelligenciában sokat vizsgált egyszerű "blokkok világa", ahol egy x blokk pozícióját teljes egészében egy $At(x, l)$ vagy $On(x, y)$ mondat jellemzi, ahol l egy hely, vagy y egy másik blokk. A

nyelv nem engedi meg, hogy azt mondjuk, hogy egy blokk részben egy másikon van. Ráadásul az $On(x, y)$ használata nem igényli az "on" szó jelentésének vagy az általa képviselt fogalomnak az előzetes elemzését. Csak bizonyos egyszerű axiómákat használunk.

Ez azért működik, mert a most épülő egyszerű blokkhalmozó program kontextusában az egyik blokk biztosan egy másikra kerül, vagy nem egy másikra, feltéve, hogy a program soha nem kényszeríti a robotot arra, hogy egy blokkot kétértelmű pozícióba helyezzen. Patrick Winston kiterjesztette a blokkok világát, hogy egy blokkot két másik is megtámaszthasson, és olyan szerkezeteket tárgyalt, mint az ívek. Lásd (Winston 1977).

Egy másik példa a MYCIN (Davis et al. 1977) szakértői rendszer, amelyben az ontológia (a figyelembe vett objektumok) betegségeket, tüneteket és gyógyszereket tartalmaz, de nem tartalmaz betegeket (csak egy van), orvosokat vagy időben bekövetkező eseményeket. Így a MYCIN-nek nem lehet megmondani, hogy az előző, azonos tünetekkel rendelkező beteg meghalt. Lásd (McCarthy 1983) a MYCIN-ről szóló további megjegyzéseket.

A korlátozott informatikai helyzetben lévő rendszereket kívülről tervezik újra, ha az általuk figyelembe vett jelenségek halmaza nem megfelelő. Az embert azonban senki sem tervezheti újra kívülről, ezért az embernek képesnek kell lennie arra, hogy új jelenségeket vegyen figyelembe. Egy emberi szintű mesterséges intelligenciarendszernek ugyanerre a képességre van szüksége, hogy új jelenségeket vegyen figyelembe.

Általában a gondolkodó ember olyan helyzetben van, amit mi a *józan ész információ- matikus helyzetének* nevezünk. Az ismert tények szükségszerűen hiányosak. [6]

5.2 Az általános józan informatikai helyzet

Egy állat, ember vagy számítógépes program *informatikai helyzete* alatt a rendelkezésére álló információfajtákat és a rendelkezésére álló következtetési módszereket értem. A *hétköznapi értelemben vett informatikai helyzet* egy átlagos megfigyelési képességgel, átlagos veleszületett tudással és átlagos következtetési képességgel rendelkező ember hely*zete,* különösen a bekövetkező események következményeiről, beleértve az általa esetlegesen végrehajtott cselekvések következményeit is. A speciális információk, például a tudományról és az emberi intézményekről, például a jogról, megtanulhatók és beágyazhatók az ember józan észbeli információiba. Közel 50 évnyi erőfeszítés ellenére csak szerény előrelépés történt az emberi szintű józan ész képességekkel rendelkező számítógépes rendszerek létrehozása felé. Sokkal több előrelépés

[6]Amint azt a 4. szakaszban tárgyaltuk, a közepes méretű, csak részben megfigyelhető objektumok világában élünk. A tudományos fantasztikus irodalom, valamint a tudományos és filozófiai spekulációk gyakran hódoltak a *Laplace-féle fantáziálásnak* a szuperlényekről, akik az összes részecske helyzetének és sebességének ismeretében képesek megjósolni a jövőt. A spekulációnak nem ez az iránya. A hihetőbb szuperlények jobban

használnák az érzékszervek által elérhető információkat - talán több és érzékenyebb érzékszervvel rendelkeznének, például ultrahanggal, ami lehetővé tenné a tárgyak belső felszínének meglátását. Mindazonáltal a jövő előrejelzésére és az általuk választott cselekvések következményeinek előre látására való képességüket továbbra is korlátoznák a kaotikus folyamatok.

speciális rendszerekkel végeztek korlátozott informatikai helyzetekben.

Senki sincs tisztában azzal, hogy mi a józan informatikai helyzet. Szerintem ennek megértése az AI, és talán a filozófia és a kognitív tudományok legnagyobb problémája. Azonban legalább a következő jellemzőkkel rendelkezik.

cselekedetekkel és egyéb eseményekkel kapcsolatos hiedelmek A rendőr úgy véli, hogy az egyik autó megelőzte a másikat. Az események hatásairól alkotott hiedelmei miatt azt hiszi, hogy ha egy másik autó jött volna a dombon, akkor frontális ütközés történt volna.

elaborációtűrő elméletek Az ágens által használt elmélet nyitott az új tényekre és új jelenségekre. A sofőr és a rendőr például figyelembe vehetné az esetleges ködöt, vagy a sofőr azt állíthatná, hogy ha egy másik autó jött volna, akkor látta volna a fényszórók visszatükröződését a domb tetején lévő pajtán. A rendőr elmélete azt javasolta, hogy azt válaszolja: "Ezt mondd a bírónak".

Egy másik példa: Egy háziasszony, aki vacsorát vásárol, a hentespultnál áll, és azt hiszi, hogy az aznap délután repülővel érkező fia szereti a steaket. Elhatározza, hogy megnézi, hogy a repülőgép időben érkezik-e. Hirtelen a józan ész ismereteinek egy egészen más területe válik relevánssá, amely nem része a vacsoravásárlás forgatókönyvének, azaz a légitársaság járatinformációs száma, és hogy hogyan tudja megszerezni, ha az nincs a mobiltelefonja telefonlistáján. A 6. szakaszban többet olvashatunk a kidolgozási toleranciáról.

hiányosan ismert és hiányosan meghatározott entitások A vizsgált tárgyak és egyéb entitások hiányosan ismertek, és nem jellemzi őket teljes mértékben az, amit róluk tudunk. A sofőr és a rendőr valós autója hiányosan ismert, és a hipotetikus autó, amely a dombon át jöhetett volna, meglehetősen homályos. Nem lenne helyénvaló, ha a sofőr megkérdezné a rendőrtől, hogy "Milyen autóra gondolt?". A legtöbb figyelembe vett entitás önmagában még csak nem is teljesen meghatározott. A hipotetikus autó, amely talán átjött a dombon, rosszul definiált, de a tényleges autók is azok.

nemmonoton gondolkodás Az Elaborációs tolerancia egy követelményt támaszt a logikával szemben, mégpedig a *nemmonoton gondolkodás* képességét. A rendszernek olyan következtetésekre kell jutnia, amelyeket az eredeti tényeknek nem ellentmondó további tények megváltoztathatnak. Például, amikor egy madarat említenek,

az ember általában arra következtet, hogy tud repülni. Ha megtudjuk, hogy ez egy pingvin, ez megváltozik. Két fő formalizmus létezik a nem monoton gondolkodásra, a *körülírás* és az *alapértelmezett logika*. A Prolog programok is végeznek nemmonoton következtetést, amikor a *negációt mint hibát* használják.

A *körülírás* (McCarthy 1980), (McCarthy 1986) és (Lifschitz 1993) minimalizálja egy predikátum kiterjesztését, néhány más predikátum kiterjesztését fixen tartva, és megengedve, hogy még több más predikátum kiterjesztése is változhasson a mini- mum elérése érdekében. A körülírás a variációszámítás logikai analógja.
a matematikai analízisben, de eddig nem rendelkezik ilyen elegáns elmélettel. Íme a körülírás egy alapvető formája.

Legyen a egy axióma a p (minimalizálandó), z (ami változtatható) és c (ami állandó) argumentumokkal. Ekkor a p körülírása, $Circum(a, p, z, c)$ a következőképpen definiálható

$$Circum[a, p, z, c] := a(p, z, c) \wedge (\forall p^l z^l)(a(p^l, z^l, c) \rightarrow \neg p^l < p), \quad (1)$$

ahol a következő definíciókkal rendelkezünk

$$p^l < p \equiv p^l \leq p \wedge p^l /= p,$$
és
$$p^l \leq p \equiv (\forall x)(p^l(x) \rightarrow p(x)). \quad (2)$$

A jelenségek csak egy részének figyelembe vétele nem monoton érvelési lépés. Nem számít, hogy a figyelembe nem vett jelenségeket szándékosan hagyjuk ki, vagy azok ismeretlenek az érvelő számára.

Bár a nem monoton gondolkodás mind az ember, mind a gép számára alapvető fontosságú, hibához vezet, ha egy fontos tényt nem veszünk figyelembe. Ezek a leggyakrabban észlelt hibák. [7]

[7] Íme egy bővebb példa a tudománytörténetből.
A 19. század közepétől kezdve Lord Kelvin (William Thomson) vállalkozott arra, hogy meghatározza a Föld korát. Mérésekkel rendelkezett a hőmérséklet mélységgel való növekedésének mértékéről és a kőzetek hővezető képességéről. Abból a feltételezésből indult ki, hogy a Föld eredetileg olvadt volt, és kiszámította, mennyi időbe telt volna, amíg a Föld lehűlt a jelenlegi hőmérsékletére. Először 98 millió évre becsülte, majd ezt a becslést később 20-40 millió évre csökkentette. Ezzel konfliktusba került a geológusokkal, akiknek már nagyobb becsléseik voltak, amelyek az üledékes kőzet éves rétegeinek számolásán alapultak.

(Koons 2005 tavasza) jó tárgyalást tartalmaz a nem monoton érvelés különböző fajtáiról.

összefüggésekben és összefüggésekről való gondolkodás A Sher- lock Holmes-történetek összefüggésében Holmes detektív, és anyja leánykori neve meghatározatlan. Az amerikai jogtörténet kontextusában Holmes bíró, és az anyja leánykori neve Jackson. A kötött elméletek, általában rögzített kontextussal rendelkeznek.

A köznapi informatikai helyzetben lévő ügynök gyakran szembesül új kontextusokkal. A 7. szakasz a kontextusokban és a kontextusokról szóló információkkal, valamint a különböző kontextusokban lévő információk közötti kapcsolatokkal foglalkozik.

A fizikai tárgyak ismerete A pszicho- logikai kísérletek egyre több bizonyítékot szolgáltatnak arra (Spelke 1994), hogy a csecsemők veleszületett tudással rendelkeznek a fizikai tárgyakról és azok állandóságáról, amikor azok eltűnnek a látóterükből. Minden józan ész rendszerébe ezt be kell építeni. (McCarthy 1996c), "A jól megtervezett gyermek" című könyvében tárgyalja, hogy milyen információkat kell beépíteni egy robotba a világról.

tárgyak összetétele Tekintsünk egy részekből álló tárgyat. Logikailag akkor következetes, ha az, amit a részekről és azok összerakásáról tudtunk, lehetővé teszi számunkra, hogy meghatározzuk az összetett objektum viselkedését. Valóban ez gyakran igaz a természettudományokban és a mérnöki tudományokban, és gyakran ez a célja a tudományos elmélet keresésének. . Így igen hasznos, hogy a molekulák tulajdonságai az atomok tulajdonságaiból és kölcsönhatásaikból következnek.

A józan informatikai helyzet logikailag nem olyan kényelmes. Egy tárgy tulajdonságai gyakran könnyebben hozzáférhetőek, mint a részek és azok kapcsolatainak tulajdonságai.

Kelvin számításai helyesek voltak, de rossz választ adtak, mert Becquerel 1896-os felfedezéséig senki sem tudott a radioaktív bomlásról, a Földet melegen tartó fő energiaforrásról.

Kelvin érvelése nem volt monoton. Feltételezte ugyanis, hogy minden olyan energiaforrás létezik, amelynek létezésére tudományos ismeretei alapján következtetni lehetett.

A nem monoton gondolkodásra a tudományban éppúgy szükség van, mint a mindennapi életben. Mindig lehetnek olyan jelenségek, amelyekről nem tudunk. Valóban lehet, hogy a radioaktivitáson kívül más energiaforrás is van a Földben.

A tapasztalat azt mutatja, hogy a gondos nem monoton érvelés, amely figyelembe veszi az összes általunk fellelhető és megérthető információforrást, általában jó eredményeket ad, de soha nem lehetünk olyan biztosak, mint a tisztán matematikai eredményekben.

Például egy baseball-labdának látható és tapintható felülete van, és láthatjuk és érezhetjük a varrásokat, valamint érezhetjük a megfelelőségét és a legegyszerűbb hőátadási tulajdonságait. Olvasásból vagy egy szétszedett baseball-labda láttán a belsejéről is tudunk valamit. Ez a szerkezeti ismeret azonban kevésbé használható, mint a baseball-labda egészének ismerete.

Az a jelenség, hogy gyakran többet tudunk az egészről, mint a részekről, nem csak a fizikai tárgyakra vonatkozik. A folyamatokra is vonatkozhat. A jelenség még a matematikában is létezett. Euklidész ge- ometriája erős logikai struktúra volt, de az alapfogalmak homályosak voltak.

A térbeli régiók ismerete Nem tudom, hogyan fogalmazhatnám meg ezt pontosan, és a pszichológiai szakirodalomban sem ismerek átfogó vitákat, de bizonyos ilyen ismeretek veleszületettnek tekinthetők. Az evolúciónak majdnem 4 milliárd éve volt arra, hogy ezt veleszületetté tegye. Az autópályán lévő tér ismerete közös a példában szereplő sofőr és a rendőr számára.

lokalizáció Nem várjuk, hogy a Holdon történő események befolyásolják az asztalon lévő tárgyak fizikai helyét. Számolhatunk azonban azzal a lehetőséggel, hogy egy távcsövön keresztül nézelődő csillagász annyira megijed, ha egy meteoritot lát a Holdba ütközni, hogy leesik a székéről, és leüt egy tárgyat az asztalról. A távoli kauzalitás különleges jelenség. Csak akkor vesszük figyelembe, ha konkrét okunk van rá.

más szereplők ismerete A csecsemők nagyon korán megkülönböztetik az arcokat más tárgyaktól. Feltehetően a csecsemőknek vannak bizonyos veleszületett elvárásaik arról, hogy más szereplők hogyan reagálnak a csecsemő cselekedeteire.

önreferencia Általában maga az informatikai helyzet olyan tárgy, amelyről tények ismertek. Ezt az emberi képességet nem sok emberi gondolkodás használja, és nagy valószínűséggel az állatok sem rendelkeznek vele.

introspektív tudás Ez talán kifejezetten emberi tulajdonság, de bizonyos mértékű introspektív tudás már korán, legalább ötéves korban a józan ész részévé válik. Ebben az életkorban egy tipikus

a gyermek emlékezhet arra, hogy korábban azt hitte, hogy egy dobozban édesség van, még akkor is, ha megtudta, hogy valójában zsírkrétákat tartalmaz.

ellentételezések A józan ész gyakran magában foglalja az ellentételezések ismeretét, valamint azt a képességet, hogy megfigyelésből következtetni tudjunk rájuk, és ezekből nem ellentételező következtetéseket tudjunk levonni. A példában a rendőr abból az ellentényből következtet arra, hogy a sofőrnek bírságot kell adnia, hogy ütközés történt volna, ha egy másik autó jön át a dombon. Az emberek olyan kontrafaktuális tapasztalatokból tanulnak, amelyeket a valóságban inkább nem tennének meg.

Korlátozott informatikai szituációk kontextusokban A korlátozott informatikai szituációk fontos kapcsolatban állnak a köznapi értelemben vett informatikai szituációkkal. Tegyük fel például, hogy egy asztalon van néhány kocka. Ezek nem tökéletes kockák, és nincsenek pontosan egymás mellé igazítva. Ettől függetlenül egy egyszerű blokkvilág-elmélet hasznos lehet egy torony építésének megtervezéséhez a blokkok mozgatásával és festésével. Az egyszerű blokkok világának korlátos elmélete, amelyben a blokkok csak az $on(x, y, s)$ relációval kapcsolódnak egymáshoz, a toronyépítő előtt álló, józan ész szerinti informatikai helyzethez kapcsolódik. Ez a kapcsolat kényelmesen kifejezhető a 7. fejezetben és (McCarthy és Buva ˇ c 1997) tárgyként tárgyalt kontextuselmélet segítségével. A tömbök világelmélete a c köznapi értelem elméletének egy al-kontextusában, $cblocks\text{-}ban$ érvényesül, és a mondatok bármelyik irányban *átemelhetők* c és $cblocks$ között.

tanulás A gyermek tapasztalatokból és elmondásból is tanulhat tényeket. Egészen kicsi gyerekeknek lehet mesélni a Mikulásról. Sajnos az eddig (2006. január) kifejlesztett mesterséges intelligencia-rendszerek nem képesek megtanulni a weboldalakon természetes nyelven kifejezett tényeket.

Közelebbről nem várjuk el, hogy az egymással nem érintkező vagy köztes objektumokon keresztül összekapcsolt objektumok hatással legyenek egymásra. Talán sok olyan józan tudást kell logikai elméletként kifejezni az asztali léptékű tárgyak fizikai mozgásáról és egymásra hatásáról, amit a józan észnek megfelelően kell megfogalmazni.

Az e követelmények által támasztott nehézségek az oka annak, hogy Leibniz, Boole és Frege célja, hogy a logikai számítást az emberi ügyek

eldöntésének fő módjaként használják, még nem valósult meg. Céljuk megvalósításához a logika azon kiterjesztéseire van szükség, amelyek túlmutatnak azokon, amelyek a logikai gondolkodáshoz szükségesek.

korlátozott informatikai helyzetek. A köznapi informatikai helyzetben működő számítógépes programoknak az eddig használtakon túlmutató eszközökre is szükségük van.

A fenti nézettel ellentétben Nagel (Nagel 1961) a józan ész ismeretét ugyanolyan ismeretként kezeli, mint a tudományos ismereteket, csak nem szisztematikusan tesztelt és igazolt. Ez igaz a józan ész bizonyos ismereteire, de a józan ész sok ismerete olyan entitásokra vonatkozik, amelyek szükségszerűen rosszul definiáltak, és olyan ismeretekre ezek viszonyairól, amelyek szükségszerűen pontatlanok.

Shannon kvantitatív információelmélete kevéssé alkalmazhatónak tűnik a köznapi informatikai helyzetre. A Chaitin- Kolmogorov-Solomonoff-féle számításelmélet sem. Egyik elmélet sem foglalkozik azzal, hogy mi a common sense információ.

6 A filozófia mesterséges intelligenciája - néhány tanács

Van Benthem (van Benthem 1990) szerint a mesterséges intelligencia más eszközökkel folytatott filozófia. Ez is része annak, amit az AI-nak tennie kell.

A mesterséges intelligencia kutatása a mesterséges intelligencia és a filozófia közös problémáit más módon támadja. Egyes filozófiai kérdések esetében a mesterséges intelligencia megközelítése előnyös. A mesterséges intelligencia viszont már hasznot húzott az analitikus filozófiában és a filozófiai logikában végzett munkából, és a további kölcsönhatások mindkét törekvést segítik. Ez a fejezet azt indokolja, hogy a filozófusok miért lehetnek érdekeltek a mesterséges intelligencia megközelítéseiben néhány konkrét közös problémára, és hogyan profitálhat a mesterséges intelligencia a kölcsönhatásból.

Az emberi szintű józan ész eléréséhez számos filozófiai probléma legalább részleges megoldása szükséges, amelyek közül néhány már régóta ismert a filozófiai, mesterséges intelligencia és/vagy kognitív tudományok szakirodalmában, míg mások még nem kerültek meghatározásra. E problémák azonosítása fontos a filozófia, a mesterséges intelligencia és a kognitív tudomány számára.

Bizonyos *meggyőződések*, *tudás*, *szabad akarat*, *szándék*, *tudatosság*, *képességek* vagy *akaratok* tulajdonítása egy gépnek vagy számítógépes programnak akkor *jogszerű*, ha ez a tulajdonítás ugyanazt az információt fejezi ki a gépről, mint amit az emberről. *Hasznos*, ha a leírás segít megérteni a gép szerkezetét, múltbeli vagy jövőbeli viselkedését, vagy azt, hogy hogyan javítsuk vagy javítsuk meg. *Logikailag* talán soha nem

szükséges, még az emberek esetében sem, de ha ésszerűen röviden kifejezzük azt, amit valójában tudunk a gép állapotáról.

egy adott helyzetben egy gépnek mentális tulajdonságokat vagy ezekkel izomorf tulajdonságokat kell tulajdonítani. A hit, a tudás és a akarás elméletei a gépek számára egyszerűbb környezetben konstruálhatók, mint az emberek számára, és később alkalmazhatók az emberekre. A mentális tulajdonságok hozzárendelése a legegyszerűbb az ismert szerkezetű gépek, például a termosztátok és a számítógépes operációs rendszerek esetében, de akkor a *leghasznosabb*, ha olyan entitásokra alkalmazzuk, amelyek szerkezete nagyon hiányosan ismert.

Bár meglehetősen liberálisak vagyunk abban, hogy *bizonyos* mentális tulajdonságokat még a meglehetősen primitív gépeknek is tulajdonítsunk, konzervatívnak kell lennünk a kritériumok tekintetében, amelyek alapján egy *adott* tulajdonságot tulajdonítunk. A hozzárendeléseket (Dennett 1978) *szándékos álláspont* elfoglalásának nevezi.

Még fontosabb, mint a meglévő gépeknek mentális tulajdonságokat tulajdonítani, a gépeket úgy tervezni, hogy azok a kívánt mentális tulajdonságokkal rendelkezzenek.

Íme néhány jellemzője a mesterséges intelligencia és a filozófia gyakori problémáinak néhány mesterséges intelligencia-megközelítésének.

A mesterséges intelligencia kicsiben kezdődik. Szerencsére a mesterséges intelligencia kutatása gyakran a fogalmak kis változataival is beéri. A fogalmaknak és kapcsolataiknak ezek a kis változatai korlátozott kontextusban érvényesek. Három példát tárgyalunk itt és a 7. szakaszban, amely a kontextusról szól. Ezek a hit, a blokkok világában való cselekvés és a vásárolt tárgyak tulajdonjoga.

Az épület intelligens hőmérséklet-szabályozó rendszerét úgy kell de-jelezni, hogy tudjon az egyes helyiségek hőmérsékletéről, a különböző szelepek állapotáról, a helyiségek lakóiról stb. Mivel a rendszer ezeket a tényeket nem mindig ismeri helyesen, nekünk és neki is hiedelmeknek kell tekintenünk őket. Az időjárás-előrejelzéseket mindig bizonytalannak, azaz hiedelemnek kell tekinteni.

Érdemes először a legegyszerűbb hiedelmeket, például egy termosztát hiedelmeit vizsgálni.

Egy egyszerű termosztátnak mindössze három lehetséges hiedelme lehet: a hőmérséklet túl hideg, rendben van, vagy túl meleg. Az aktuális meggyőződése szerint viselkedik: bekapcsolja a fűtést, változatlanul hagyja, vagy kikapcsolja. Nem hiszi el, hogy termosztát, és nem hiszi el, hogy a szoba túl hideg.

Természetesen ennek az egyszerű termosztátnak a viselkedése mindenféle hiedelem tulajdonítása nélkül is megérthető. A hitelméletet

ilyen egyszerű esetekkel kezdeni ugyanazzal az előnnyel jár, mintha az 1-es számot is belevennénk a számba.

rendszer. (Ha egy sziklának nem tulajdonítunk hiedelmeket, az olyan, mintha 0-t is tartalmazna.) Egy egész épület hőmérséklet-szabályozó rendszerének megfelelően bonyolultabb hiedelmeket tulajdonítunk. A hiedelmek és más mentális tulajdonságok hozzárendelését alaposabban tárgyalja (McCarthy 1979a).

A gyermeknek jót tesz, ha tudja, hogy ő is egy gyermek a többi között. Hasonlóképpen, egy hőmérséklet-szabályozónak is hasznára válhat, ha tudja, hogy ő egy hőmérséklet-szabályozó a többi ilyen rendszer között. Ha az interneten keresztül megtudja, hogy egy másik rendszer a tetőn lévő hőhöz igazodik, akkor ennek megfelelően módosíthatja a programját.

A naiv józan észnek gyakran igaza van a kontextusban. Erre példa az "x okozta y-t" közgondolkodás.

Van olyan szövegkörnyezet, amelyben a "Az ablakot Susan baseball-labdája törte be" igaz, és a "Az ablak azért tört be, mert az építésvezető elmulasztotta, hogy rácsot tegyen elé" még csak nem is szerepel az igazgató által használt nyelvezetben, amikor a labdát eldobó lány büntetéséről beszél. Az ilyen korlátozott összefüggések gyakran használatosak és hasznosak. A kauzalitás általánosabb összefüggéseivel való kapcsolatuk tanulmányozást és logikai formalizálást igényel.

A cselekvés elmélete és a keretprobléma A világban a célok elérésének feltételei általában nagyon bonyolultak, de a mesterséges intelligencia kutatása egyre kifinomultabb elméleteket és számítógépes programokat fejlesztett ki.

A mesterséges intelligencia már régóta (legalábbis az 1950-es évek óta) foglalkozik azzal, hogy olyan cselekvéssorozatokat találjon, amelyekkel elérhetők a célok. Ehhez a mesterséges intelligenciának szüksége van az egyes cselekvések hatásainak, a kiindulási helyzetből eredő helyzetek fájának és a cselekvéssorozatok hatásainak elméleteire. Az erre a célra leggyakrabban használt mesterséges intelligencia formalizmus a (McCarthy és Hayes 1969) által bevezetett *szituációkalkulus*[8] . A filozófiával való kapcsolatát (Thomason 2003) tárgyalja. Alapos tárgyalásokat tartalmaz (Shanahan 1997) és (Reiter 2001), és egy új, *a* szokásos *hatásaxiómák* mellett *előfordulási* axiómákat is tartalmazó változatot mutat be (McCarthy 2002). Három probléma, *a keretprobléma*, a *minősítési probléma* és *a elágazási probléma* merült fel, amelyeket a mesterséges intelligencia irodalomban és (Thomason 2003-ban) is részletesen tárgyalnak. A keretproblémát, amelyet a

[8]Az eseményszámítás (Mueller 2006) alternatívát jelent.

filozófusok, arra vonatkozik, hogy hogyan kerüljük el annak kijelentését, hogy mely *folyományok* (egy helyzet aspektusai) maradnak változatlanok, amikor egy cselekvés megtörténik, pl. elkerülve annak explicit kijelentését, hogy egy tárgy színe nem változik, amikor a tárgyat elmozdítják.

Az alapvető szituációs kalkulus egy nemdeterminisztikus (elágazó) cselekvéselmélet. A mesterséges intelligencia a cselekvés determinisztikus (lineáris) elméleteivel is foglalkozott. A (McCarthy 2002) új formalizmusa lehetővé teszi (McCarthy 2005) egyfajta determinisztikus szabad akarat kezelését, amelyben egy nem determinisztikus elmélet a determinisztikus számítási mechanizmus részeként szolgál.

A mesterséges intelligencia egyszerű példákat vett figyelembe, amelyek utólagosan elabilizálhatók. A jól ismert *blokkok* világát olyan logikai mondatokkal kezelik, mint az $On(Block1, Block2)$ vagy $On(Block1, Block2, S0)$, amelyekben a helyzet explicit. Egy másik formalizmus a $Value(Location(Block1), S0) = Top(Block2)$ mondatot használja. Az is lehet, hogy

$$(\forall s)(\ldots \rightarrow Location(block, Result(Move(block, l), s)) = l)$$
és (3)
$$(\forall s)(\ldots \rightarrow Color(block, Result(Paint(block, c), s)) = c$$

ahol . . . a cselekvés sikerének bizonyos előfeltételeit jelenti. Egyrészt ilyen egyszerű akciómodelleket építettek be olyan robotkarokat vezérlő programokba, amelyek sikeresen mozgatják a blokkokat. Másrészt a *keretprobléma* felmerült annak megadásakor, hogy egy blokk mozgatása nem változtatta meg más blokkok helyét vagy a blokkok színét. Ez a probléma, valamint társai, a minősítési probléma és a elágazási probléma a mesterséges intelligencia kutatásában merült fel, de felmerül a filozófiában a cselekvés hatásainak vizsgálatakor is.

Megjegyezzük, hogy a blokkok világának itt részben leírt korlátozott elméletében csak egy szereplő van, és egy blokk soha nem lehet részben az egyik blokkban, részben a másikban. E bonyodalmak tanulmányozására további kidolgozások történtek, de az a módszertan, hogy először az egyszerű eseteket végezzük el, jó eredményekhez vezetett. A teljes cselekvéselmélet nulláról való elkészítése még mindig csak egy homályosan meghatározott projekt.

nem monoton érvelés A nem monoton érvelés lényegében ugyanaz a téma, mint a filozófiában régóta vizsgált megdönthető érvelés. Mi az újdonság

az 1970-es évek óta a nem monoton gondolkodás formális rendszereinek fejlesztése, például az elmaradások logikája (Reiter 1980) és a körülírás logikája (McCarthy 1980) és (McCarthy 1986). Az 1970-es évekből származó számítógépes rendszerek is vannak, amelyek nem monoton következtetést végeznek, pl. a Mi- croplanner és a Prolog. A nem monoton gondolkodás kiemelkedő szerepet játszik olyan programokban, amelyek terveket készítenek a célok elérésére.

A Stanford Encyclopedia of Philosophy legújabb cikkei kapcsolatot teremtettek a nem monoton gondolkodással kapcsolatos mesterséges intelligencia és a megdönthetőséggel kapcsolatos filozófiai munkák között. Kényelmes hivatkozások: (Thomason 2003), (Koons Spring 2005) és (Antonelli 2003).

kidolgozási tolerancia A józan ész jelenségeinek explicit formalizációja szinte soha nem teljes. Mindig van több információ, amit figyelembe lehet venni. Ez független attól, hogy a jelenségeket hétköznapi nyelven vagy logikai mondatokkal írjuk le. Az elméleteket mindig ki kell dolgozni. Attól függően, hogy az elméletet eredetileg hogyan írták meg, az elmélet ***elviselhet*** egy adott kidolgozottságot pusztán mondatok hozzáadásával, ami általában nem monotonitást igényel az elméletből való következtetések levonásához, vagy az elméletet el kell vetni, és egy új elméletet kell építeni a semmiből. (McCarthy 1999b) bevezeti az ***elaboráció-tolerancia*** fogalmát, és a jól ismert misszionáriusok és kannibálok rejtvény 19 elaborációjával illusztrálja. Az elaborációk angolul egyszerűnek tűnnek, de az olvasó józan eszére támaszkodnak. A logikai megfogalmazások némelyike néhány elaborációt csupán mondatok hozzáadásával tolerál, mások viszont nem. Az egyik cél egy olyan logikai nyelv megtalálása, amelyben az összes kifejtés additív.

(Lifschitz 2000) a fent említett 19 kifejtésből 9-et teljesít.
McCain és Turner ok-okozati kalkulátorában (McCain és Turner 1998).
(Shanahan 1997) részletesen tárgyalja az elaborációs toleranciát.

Nem tudok a filozófiai szakirodalomban javasolt elméletek kidolgozottsági toleranciájáról szóló vitákról.

A világgal kellően összetett módon kölcsönhatásba lépő robot a világ azon részének lényegében egyedi értelmezését eredményezi, amellyel kölcsönhatásba lép. Ez egy empirikus, tudományos feltevés, de sokan, különösen filozófusok (lásd (Quine 1960),

(Quine 1969), (Putnam 1975), (Dennett 1971), (Dennett 1998)), úgy tűnik, természetesnek veszik a tagadást. A rövid leírások világában gyakran sok értelmezés létezik, de a hosszú leírások szinte mindig legfeljebb egyet engednek meg. Amennyire én látom, (Quine 1960) nem tárgyalta a nagy kontextus hatását a fordítás meghatározhatatlanságára - mondjuk *gavagai*.

A legegyszerűbb példa egy angol kifejezés egyszerű helyettesítő rejtjelezése. Így az XYZ kódot "macska" vagy "kutya" néven lehetne megfejteni. Egy angol mondat egyszerű helyettesítési kriptogramjának általában többféle értelmezése van, ha a szöveg 21 betűnél rövidebb, és általában egyetlen értelmezése van, ha a szöveg 21 betűnél hosszabb. Miért 21? Ez az angol nyelv redundanciájának egy mértékegysége (Shannon 1949). Egy ember vagy egy robot világgal való interakcióinak redundanciája ugyanilyen valós - bár nyilvánvalóan sokkal nehezebb számszerűsíteni.

közelítő tárgyak és elméletek Az az elképzelés, hogy a filozófiai érdeklődésre számot tartó entitások nem mindig jól definiáltak, ha tetszik, Arisztotelésznek tulajdonítható.

> A tárgyalásunk akkor lesz megfelelő, ha olyan világos, amennyire a téma engedi, mert a pontosság nem minden tárgyalásnál egyformán keresendő, mint ahogyan a kézművesség minden termékénél sem.
>
> *-Nikomakhész etikája*.

Nem tudom, hogy Arisztotelész továbbgondolta-e ezt a gondolatot.

Azt javasoltam (McCarthy 2000), hogy a mesterséges intelligencia megköveteli az ap- proximális entitások formalizálását, ami néha szilárd logikai elméleteket eredményez a szemantikai futóhomokra való rátalálásokon. Így tehát bizonyos, hogy a Mount Everestet 1953-ban megmászták, még akkor is, ha nem biztos, hogy milyen szikla és jég alkotja a Mount Everestet. Sokkal közelítőbb, bár még mindig hasznos fogalom az *"Az Egyesült Államok 1990-ben azt akarta"*, hogy "Irak kivonuljon Kuvaitból". Az egyik javaslat az, hogy használjunk szükséges feltételeket egy tételhez és elégséges feltételeket, de ne törekedjünk olyan feltételekre, amelyek egyszerre szükségesek és elégségesek. Ezek az elképzelések a homályosság fogalmához kapcsolódnak, amelyet filozófusok már megvitattak, de a Stan-

A Ford Encyclopedia of Philosophy nem tárgyalja, hogyan lehet formalizálni a lényegében homályos fogalmakat.

kontextusok mint objektumok Ez egy olyan terület, ahol a Stanford En- cyclopedia of Philosophy alapján úgy tűnik, hogy még nincs kapcsolat a (McCarthy 1993) által megkezdett, meglehetősen kiterjedt mesterséges intelligencia-kutatás és a filozófiai kutatások között. Mivel a *mesterséges intelligenciában* (és a hétköznapi nyelvben) az *információ* mindig kontextusban jelenik meg, a 7. fejezetet a kontextusok mint objektumok elméletének vázlatának szenteljük.

fogalmak mint objektumok A természetes nyelvben a fogalmakról állandóan beszélünk. Ennek ellenére Carnap azt írta

> . . úgy tűnik, aligha javasolja bárki is, hogy különböző változókat használjunk a tételek és az igazságértékek, vagy különböző változókat az egyének és az egyes fogalmak számára.
>
> ((Carnap 1956) , 113. o.

Talán Carnap (Church 1951) a kivételre gondolt. Ehelyett a modális logikát a propozíciókkal kapcsolatos bizonyos állítások kifejezésére használják, és az egyes fogalmak alig vannak formalizálva.

az emberi szintű mesterséges intelligencia megköveteli, hogy képes legyen kifejezni bármit, amit az emberek természetes nyelven kifejeznek, valamint képes legyen kifejezni a kifejezésekre és azok szemantikájára vonatkozó kijelentéseket.

(McCarthy 1979b) azt javasolja, hogy különböztessük meg a tételeket az igazságértékektől és az egyes fogalmakat az alaptartomány objektumaitól - és használjunk különböző változókat rájuk. Íme néhány példa a jegyzetelésre. A ***Mike*** értéke egy személy, míg az ***MMike*** értéke egy fogalom - ebben az esetben az a szándék, hogy ez a Mike fogalma legyen, de hogy az legyen, az nem tipográfiai konvenció. Íme néhány mondat egy első rendű nyelvből, fogalmakkal és tárgyakkal.

$$
\begin{aligned}
&\mathit{Denot}(\mathit{MMike}) = \mathit{Mike},\\
&\mathit{Male}(\mathit{Mike}),\\
&\mathit{Denot}(\mathit{MMale}(\mathit{MMike})),\\
&\mathit{Denot}(\mathit{HHusband}(\mathit{MMary})) = \mathit{Mike},\\
&\mathit{Husband}(\mathit{Mary}) = \mathit{Mike},\\
&\mathit{HHusband}(\mathit{MMary} /= \mathit{MMike},
\end{aligned}
\qquad (\forall x)
$$

$$(x \not= \mathit{férj}(\mathit{Mike}) \rightarrow \neg \mathit{Létezik}(\mathit{HHusband}(\mathit{MMike})). \tag{4}$$

A *Denot*(*MMike*) */= Mike mondat* bizonyos körülmények között igaz lehet.

A fogalmak és az objektumok megkülönböztetése lehetővé teszi, hogy kényelmesen kifejezzünk néhány olyan állítást, amelyek az egyszerűbb jelölések számára rejtélyesek. Így Russell "Azt hittem, hogy a jachtod hosszabb, mint amilyen" mondatát (McCarthy 1979b) kezeli.

Ez a példa és más példák is az objektumokból származó függvényeket használják a róluk alkotott fogalmakra. Így írhatjuk azt, hogy *CConcept1*(*Cicero*) = *CCicero*. Ha *Cicero* = *Tully*, akkor *CConcept1*(*Tully*) = *CCicero* lesz. Bár rendes körülmények között nem akarnánk, hogy *TTully* = *CCicero legyen,* de mivel a fogalmakat nem jellemzi az írásukhoz használt tipográfia, ez nem lenne ellentmondás.

Egyes objektumok standard fogalmakkal rendelkeznek, például a számok. Szeretnénk azt írni, hogy *Concept1*(3) = 33, de ez ütközik a tizedesjegyekkel, ezért jobb, ha azt írjuk, hogy *Concept1*(3) = 3^{l} 3. Tekintsük az igaz mondatokat

¬*Knew*(*Kepler, CComposite*(*NNumber*(*PPlanets*))) és
Tudta(*Kepler, CKkompozit*(*CConcept1*(*Szám*(*bolygók*)))). (5)

Az első szerint Kepler nem tudta, hogy a bolygók száma összetett. A második azt mondja, hogy Kepler tudta, hogy a szám, ami történetesen a bolygók száma, összetett. Lásd még (Maida és Shapiro 1982) és (Shapiro 1993) a fogalmak reprezentálásának egy másik mesterséges intelligencia megközelítését.

Ezek a megfontolások csak egy kis lépést jelentenek abba az irányba, amely mind a mesterséges intelligencia, mind a filozófia számára szükséges, hogy a fogalmakat első osztályú objektumként kezeljük. (McCarthy 1997) a modális logika alkalmatlansága mellett érvel a modalitás teljes körű kezelésére. A cikk heves válaszokat váltott ki.

A hivatkozás korrespondenciaelmélete Ez bonyolultabb, mint az igazság korrespondenciaelmélete, mert az entitások, amelyekre egy terminus hivatkozhat, nem csak igazságértékek. Javasoljuk, hogy a filozófusok tanulmányozzák a referencia formalizálásának problémáját. A referenciára még a modális logikának sincs analógja.

megjelenés és valóság A tudomány azt mondja nekünk, hogy korlátozott érzékszerveink, és valójában minden olyan érzékszervünk,

amelyet robotokba építhetünk, túlságosan korlátozottak ahhoz, hogy megfigyeljük a következő dolgokat

a világot teljes részletességgel, azaz atomi szinten. A mesterséges intelligenciának általában, és különösen a robotikának együtt kell élnie ezzel a ténnyel, és ezért szükség van a látszat és a valóság közötti kapcsolatok elméletére. Ennek a teóriának mindkettő különböző részletességi szintjeit figyelembe kell vennie. Ezzel még nem jutottam messzire, de (McCarthy 1999a) ad egy kis példát a kétdimenziós megjelenés és a háromdimenziós valóság közötti viszonyra. A realista, különösen a materialista filozófusoknak ezt a kapcsolatot is formalizálniuk kell.

tudatosság, különösen az én-tudat Az emberek bizonyos mértékben képesek megfigyelni és következtetni saját belső állapotukról. Például arra a következtetésre juthatok, hogy nincs módomban megtudni - hacsak nem hívom fel -, hogy a feleségem ebben a pillanatban az irodájában van-e. A saját belső állapotról való ilyen tudatosság fontos olyan célok eléréséhez, amelyek önmagukban nem járnak tudatossággal. (McCarthy 1996b) tárgyalja, hogy milyen tudatosságra lesz szüksége egy robotnak az általunk rábízott feladatok elvégzéséhez.

7 Információk kontextusokban és a kontextusokkal kapcsolatban.

Az információ mindig kontextusban kerül továbbításra. Az ember valóban kontextusban gondolkodik. Az információ filozófiája számára a kontextusokban lévő információ és a kontextusok közötti kapcsolatok fontosabbak, mint egy szöveg Shannon-entrópiája.

Ez a szakasz a kontextusok első osztályú objektumként való formalizálását tárgyalja. A ba- sic reláció az $Ist(c, p)$. Ez azt állítja, hogy a p *tétel* igaz a *kontextusban* c. A legfontosabb formulák a különböző kon- textusokban igaz tételeket kapcsolják össze. A kontextusok mint formális objektumok bevezetése lehetővé teszi, hogy a korlátozott kontextusú axiomatizációkat úgy bővítsük ki, hogy *túllépjenek az* eredeti korlátokon. Ez szükségesnek tűnik ahhoz, hogy a logikát használó AI-programok bizonyos olyan képességekkel rendelkezzenek, amelyekkel az emberi tényreprezentáció és az emberi gondolkodás rendelkezik. A *transzcendencia* teljes megvalósítása a matematikai logika további kiterjesztéseit igényli, azaz a nem monoton következtetési módszereken túl, amelyeket először a mesterséges intelligenciában találtak ki, és amelyeket most a logika új területeként tanulmányoznak.

A $Value(c, term)$ kifejezés, amely a kifejezés $term$ értékét adja meg a c

kontextusban, ugyanolyan fontos, mint az ***Ist***(***c, p***), sőt az alkalmazások szempontjából talán még fontosabb is.

Íme, a kontextus formalizált elméletének néhány jellemzője.

1. Sokféle kontextus létezik, pl. a newtoni gravitáció kontextusa és ezen belül egy adott űrhajó pályájának kontextusa, a bináris $On(x, y)$ és $Above(x, y)$ relációkat formalizáló elmélet kontextusa, egy szituációszámítás kontextusa a terner relációkkal $On(x, y, s)$ és $Above(x, y, s)$, egy adott beszélgetés vagy előadás kontextusa, egy francia nyelvű csoportelméleti vita kontextusa, a Sherlock Holmes-történetek kontextusa.

2. Kell lennie egy nyelvnek, amely kifejezi egy kifejezés értékét egy adott kontextusban. Például, van egy

$$Co : Value(Context(ThisArticle), Author) = JohnMcCarthy.$$

3. Az elméletnek nyelvet kell biztosítania a kontextusok viszonyainak kifejezésére, pl. hogy az egyik kontextus időben vagy térben specializál egy másikat, hogy az egyik kontextus több csoportelméletet feltételez, mint a másik, hogy az egyik ugyanazt a témát tárgyalja, de más nyelven.

4. Kell lennie olyan nyelvnek, amely kifejezi az összefüggő kontextusokban igaz mondatok közötti kapcsolatokat, valamint az összefüggő kontextusokban lévő kifejezések közötti kapcsolatokat. Ha $c1$ a co specializációja, az ilyen szabályokat *emelési szabályoknak* nevezzük.

5. Íme egy példa az adatbázisokhoz kapcsolódó emelési szabályra. Tegyük fel, hogy a GE (General Electric) sugárhajtóműveket ad el az AF-nek (U.S. Air Force), és mindkét szervezetnek van egy adatbázisa a sugárhajtóművekről, beleértve az árat is. Tegyük fel, hogy az AF kontextus (adatbázis) feltételezi, hogy egy hajtómű ára tartalmazza a pótalkatrész-készletet, míg a GE kontextus külön árazza azokat. Megkaphatjuk az *emelési képletet*

$$Ist(Külső, Value(AF, Ár(motor)) = Value(GE, Ár(motor)) + Value(GE, Price(Spare\text{-}Parts\text{-}Kit(engine)))),$$

egy külső kontextusban történő kifejezések $Külső$ kapcsolat az AF-kontextusban lévő kifejezés és a GE-kontextusban lévő kifejezések között. Mások az ilyen formulákat *hídképleteknek* nevezik.

(McCarthy 1993) példát mutat egy általános szabály felemelésére, amely az $On(x, y)$ és $Above(x, y)$ predikátumokra vonatkozik egy

három érvvel rendelkező helyzetre.

On(*x, y, s*) és *Above*(*x, y, s*) relációk, amelyekben a harmadik argumentum *s* egy helyzet.

6. Olyan érvelőt képzelünk el, amely mindig egy kontextusban van. Az aktuális kontextus specifikációit és egyéb módosításait tudja *megadni*, majd abban érvelni. Ezután *kiléphet* a belső kontextusból, és visszatérhet a külső kontextusba. Az emberi szintű mesterséges intelligencia rendszerekben nem lesz külső kontextus. Mindig lehetőség lesz arra, hogy az eddig megnevezett legkülső kontextuson *túllépjen*, és egy új kontextusban érveljen, amelyben az előző kontextus egy objektum.

(McCarthy 1993) és (McCarthy és Buvač 1998) a formalizált kontextusokra vonatkozó, jobban körülhatárolt elméletet mutatnak be. Lásd még (Guha 1991).

Nem szerepel ezekben a tanulmányokban az az újabb elképzelés, hogy az, amit egyes AI-kutatók "játékelméleteknek" neveznek, bizonyos kontextusokban érvényes lehet, és hogy egy érvelő gondolkodó gondolkodásának fontos részét ilyen korlátozott kontextusban végezheti. Vegyünk például egy egyszerű elméletet a vásárlásról és a birtoklásról. Egy kisgyerek szemszögéből a boltban, miután megtanulta, hogy nem vehet le csak úgy valamit a polcról, tudja, hogy a szülőnek meg kell vennie valamit, hogy odaadhassa a gyereknek. Nevezzük ezt az összefüggést *Owno-nak*. A vásárlás részletei meghatározatlanok, és ez az egyszerű fogalom akár több évig is tarthat. A következő kifinomultsági szint a tárgy árának kifizetését foglalja magában. Ez a fogalom nem csak a gyermek számára tart tovább, de egy felnőtt egy élelmiszerboltban általában ebben az *Own1* kontextusban operál, amely egy egyszerű szituációkalkulus axiomatizációt enged meg. A szupermarketeken kívül a tulajdonlás bonyolultabbá válik, pl. jelzáloggal terhelt házvásárlás. Bizonyos ilyen tulajdonjogi kontextusokat a közvélemény, másokat pedig az ügyvédek és az ingatlanbefektetők értenek, de senkinek sincs teljes elmélete. a tulajdonjog.

8 Következtetések és megjegyzések

A mesterséges intelligencia néhány filozófiai és tudományos előfeltevésen alapul. A mesterséges intelligencia legegyszerűbb formái kevesebb előfeltételt támasztanak, mint az emberi szintű mesterséges intelligenciára irányuló kutatások. Az emberi szintű mesterséges intelligencia általunk hangsúlyozott jellemzője az, hogy képes tanulni a tapasztalataiból anélkül,

hogy további programozásra szorulna.

A mesterséges intelligencia kutatásának konkrétsága számos olyan felfedezéshez vezetett, amelyek a filozófia számára is relevánsak, és ezekre csak most kezdenek felfigyelni a kutatók.

filozófusok. Az ebben a fejezetben tárgyalt témák közül három a formalizált nem monoton érvelés, a formalizált kontextusok és a csak közelítő jelentéssel rendelkező fogalmak kezelésének szükségessége. A fejezetben foglaltakon kívül különösen ajánljuk Richmond Thoma- son (Thomason 2003) című művét.

Hivatkozások

Antonelli, A. 2003. Nem monoton logika. In E. N. Zalta (szerk.), *The Stan- ford Encyclopedia of Philosophy*.

Carnap, R. 1956. *Jelentés és szükségszerűség*. University of Chicago

Press. Church, A. 1951. Az absztrakt entitások szükségessége a

szemantikai elemzésben.
Proceedings of the American Academy of Arts and Sciences 80(1):100-112. Újranyomtatva a *The Structure of Language*. szerkesztette Jerry A. Fodor és Jerrold Katz, Prentice-Hall 1964.

Davis, R., B. Buchanan és E. Shortliffe. 1977. Termelési szabályok, mint egy tudásalapú konzultációs program reprezentációja. *Artificial Intelligence* 8(1):15-45.

Dennett, D. 1978. *Agyviharok: Philosophical Essays on Mind and Psychology*. Cambridge: Bradford Books/MIT Press.

Dennett, D. 1998. *Brainchildren: Essays on Designing Minds*. MIT Press.

Dennett, D. C. 1971. Intentional systems. *The Journal of Philosophy* 68(4):87-106.

Dreyfus, H. 1992. *Amit a számítógépek még mindig nem tudnak*. M.I.T. Press.

Ernst, G. W. és A. Newell. 1969. *Gps: A CASE Study in Generality and Problem Solving*. New York: Academic Press.

Guha, R. V. 1991. *Kontextusok: A Formalization and Some Applications*. Doktori értekezés, Stanford Egyetem. Megjelent a STAN-CS- 91-1399-Thesis technikai jelentésként is, az MCC Technical Report Number ACT-CYC-423-91, és elérhető a http://www-formal.stanford.edu/guha/guha.ps címen.

Hayes, P. J. 1985. A második naiv fizikai kiáltvány. In H. J.R. és M. R. C. (szerk.), *Formal Theories of the Commonsense World*, 1-36. Ablex.

Hayes, P. J. 1979. A naiv fizika manifesztuma. In D. Michie (Ed.), *Expert systems in the microelectronic age*. Edinburgh, Skócia: Edinburgh University Press.

Koons, R. 2005 tavasza. Megdönthető érvelés. In E. N. Zalta (szerk.), *The Stanford Encyclopedia of Philosophy*.

Lenat, D. B. 1995. Cyc: Nagyszabású beruházás a tudásinfrastruktúrába. *Communications of the ACM* 38(11).

Lifschitz, V. 1993. Körírás[9] . In *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 3: Nonmonotonic Reasoning and Uncertain Reasoning*. Oxford University Press.

Lifschitz, V. 2000. Misszionáriusok és kannibálok az ok-okozati számításban. In A. G. Cohn, F. Giunchiglia és B. Selman (szerk.), *KR2000: Principles of Knowledge Representation and Reasoning, Proceedings of the Seventh International conference*, 85-96. Morgan-Kaufman.

Maida, A. S. és S. C. Shapiro. 1982. Intenzionális fogalmak a propozicionális szemantikai hálózatokban. *Cognitive Science* 6(4):291-330. Újranyomtatva in
R. J. Brachman és H. J. Levesque, szerk. Readings in Knowledge Representation, Morgan Kaufmann, Los Altos, CA, 1985, 170-189.

Marr, D. 1982. *Látás*. New York: Freeman.

Matuszek, C., M. Witbrock, R. Kahlert, J. Cabral, D. Schneider, P. Shah és D. Lenat. 2005. A józan ész keresése: Populating cyc from the web. In *Proceedings of the Twentieth National Conference on Artificial Intelligence,* Pittsburgh, Pennsylvania, 2005. július*.

McCain, N. és H. Turner. 1998. Kielégíthetőségi tervezés ok-okozati teóriákkal. In *KR*, 212-223.

McCarthy, J. 1959. Programok józan ésszel[10] . In *Mechanisa- tion of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory*, 77-84, London, U.K. Her Majesty's Stationery Office. Újranyomtatva in (McCarthy 1990).

McCarthy, J. 1979a. Mentális tulajdonságok tulajdonítása a gépeknek[11] . In M. Ringle (szerk.), *Filozófiai perspektívák a mesterséges intelligenciában*. Harvester Press. Újranyomtatva in (McCarthy 1990).

[9]http://www.cs.utexas.edu/users/vl/mypapers/circumscription.ps[10] http://www-formal.stanford.edu/jmc/mcc59.html

[11]http://www-formal.stanford.edu/jmc/ascribing.html

McCarthy, J. 1979b. Az egyes fogalmak és tételek elsőrendű elméletei[12] . In D. Michie (Ed.), *Machine Intelligence*, Vol. 9. Edin- burgh: Edinburgh University Press. Újranyomtatva in (McCarthy 1990).

McCarthy, J. 1980. Circumscription-A Form of Non-Monotonic Reasoning[13] . *Artificial Intelligence* 13:27-39. Újranyomtatva in (McCarthy 1990).

McCarthy, J. 1983. Some Expert Systems Need Common Sense[14] . In H. Pagels (szerk.), *Computer Culture: The Scientific, Intellectual and Social Impact of the Computer*, 426. kötet. A New York-i Tudományos Akadémia Évkönyvei.

McCarthy, J. 1986. Applications of Circumscription to Formalizing Common Sense Knowledge[15] . *Artificial Intelligence* 28:89-116. Újranyomtatva in (McCarthy 1990).

McCarthy, J. 1990. *A józan ész formalizálása: Papers by John Mc- Carthy*. Ablex Publishing Corporation.

McCarthy, J. 1993. Notes on Formalizing Context[16] . In *IJCAI93*.

McCarthy, J. 1996a. From Here to Human-Level AI[17] . In *KR-96*, 640-646.

McCarthy, J. 1996b. A robotok tudatosítása mentális állapotukról[18] . In S. Muggleton (szerk.), *Gépi intelligencia 15*. Oxford University Press. Megjelent 2000-ben. A webes változat az 1995-ben a Machine Intelligence 15 konferencián bemutatott változat továbbfejlesztett változata.

McCarthy, J. 1996c. The well-designed child. http://www-formal.stanford.edu/jmc/child.html.

McCarthy, J. 1997. Modalitás si! modális logika, no! *Studia Logica* 59(1):29-32.

[1213] http://www-formal.stanford.edu/jmc/concepts.htmlhttp://www-formal.stanford.edu/jmc/circumscription.html[14] http://www-formal.stanford.edu/jmc/someneed.html[15] http://www-formal.stanford.edu/jmc/applications.html[16] http://www-formal.stanford.edu/jmc/context.html[17] http://www.formal.stanford.edu/jmc/human.html[18] http://www-formal.stanford.edu/jmc/consciousness.html http://www-formal.stanford.edu/jmc/consciousness.html

McCarthy, J. 1999a. Appearance and reality[19] . *web csak most, és talán a jövőben*. nem teljesen publikálható papíron, mert tartalmaz egy lényeges beágyazott appletet.

McCarthy, J. 1999b. Elaborációs tolerancia[20] . *egyelőre csak a weben*.

McCarthy, J. 2000. Közelítő tárgyak és közelítő elméletek[21] . In A. G. Cohn, F. Giunchiglia és B. Selman (szerk.), *KR2000: Proceedings of the Seventh International conference*, 519-526. Morgan-Kaufman.

McCarthy, J. 2002. Cselekvések és egyéb események a helyzetszámításban[22] . In B. S. A.G. Cohn, F. Giunchiglia (szerk.), *Principles of knowledge representation and reasoning: Proceedings of the eighth international conference (KR2002)*. Morgan-Kaufmann.

McCarthy, J. 2005. Egyszerű determinisztikus szabad akarat. Lásd http://www- formal.stanford.edu/jmc/freewill2.html.

McCarthy, J. és S. Buva ˇ c. 1997. A kontextus formalizálása (bővített jegyzetek). In A. Aliseda, R. v. Glabbeek, and D. Westerst˚ahl (Eds.), *Computing Natural Language*. Center for the Study of Language and Information, Stanford University.

McCarthy, J. és S. Buva ˇ c. 1998. A kontextus formalizálása (bővített jegyzetek). In A. Aliseda, R. v. Glabbeek, and D. Westerst˚ahl (Eds.), *Computing Natural Language*, Vol. 81 of *CSLI Lecture Notes*, 13-50. Center for the Study of Language and Information, Stanford University.

McCarthy, J. és P. J. Hayes. 1969. Néhány filozófiai probléma a mesterséges intelligencia szempontjából[23] . In B. Meltzer and D. Michie (Eds.), *Machine Intelligence 4*, 463-502. Edinburgh University Press. Újranyomtatva in (McCarthy 1990).

[19][20] http://www-formal.stanford.edu/jmc/appearance.htmlhttp://www-formal.stanford.edu/jmc/elaboration.html[21] http://www.formal.stanford.edu/jmc/approximate.html[22] http://www-formal.stanford.edu/jmc/sitcalc.html[23] http://www-formal.stanford.edu/jmc/mcchay69.html

Minsky, M. L. 1963. Lépések a mesterséges intelligencia felé. In E. A. Feigen- baum és J. Feldman (szerk.), *Computers and Thought*, 406-450. McGraw- Hill. Eredetileg megjelent: *Proceedings of the Institute of Radio Engi- neers*, January, 1961 **49:**8-30.

Moravec, H. P. 1977. Az automatikus vizuális akadályelkerülés felé. In *IJCAI*, 584.

Müller, E. T. 2006. *Common Sense Reasoning*. Morgan Kaufmann. Nagel,

E. 1961. *A tudomány szerkezete*. Harcourt, Brace és a világ. Newell, A.

1993. Elmélkedések a tudásszintről. *Mesterséges intelligencia* 59(1-2):31-38.

Nilsson, N. J. 2005. Emberi szintű mesterséges intelligencia? *The AI Magazine* 26(4):68-75.

Penrose, R. 1994. *Az elme árnyai: A tudat hiányzó tudományának keresése*. Oxford: Oxford University Press.

Pinker, S. 1997. *Hogyan működik az elme*. Norton.

Putnam, H. 1975. A "jelentés" jelentése. In K. Gunderson (szerk.), *Nyelv, elme és tudás*, a *Minnesota Studies in the Philosophy of Science* VII. kötete, 131-193. University of Minnesota Press.

Quine, W. V. O. 1969. Propositional objects. In *Ontológiai relativitás és más esszék*. Columbia University Press, New York.

Quine, W. v. 1960. *Szó és tárgy*. MIT Press.

Reiter, R. 1980. A Logic for Default Reasoning[24] *Artificial Intelligence* 13 (1-2):81-132.

Reiter, R. 2001. *Tudás a gyakorlatban*. M.I.T. Press.

Russell, B. 1914. *Ismereteink a külső világról*. Open Court.

Searle, J. R. 1984. *Elmék, agyak és tudomány*. Cambridge, Mass.: Harvard University Press.

[24].

Shanahan, M. 1997. *A keretprobléma megoldása, a köznapi értelemben vett tehetetlenségi törvény matematikai vizsgálata*. M.I.T. Press.

Shannon, C. 1949. Kommunikációs elmélet a tengeri rendszerekről. *Bell System Technical Journal* 28:656-715. http://www.cs.ucla.edu/ jkong/research/security/shannon.html.

Shapiro, S. C. 1993. Hiedelemterek mint tételek halmazai. *Jour- nal of Experimental and Theoretical Artificial Intelligence* 5:225-235. http://www.cse.buffalo.edu/tech-reports/SNeRG-175.ps.

Sorensen, R. 2003 ősze. Homályosság. In E. N. Zalta (szerk.), *The Stanford Encyclopedia of Philosophy*.

Spelke, E. 1994. Kezdeti ismeretek: hat javaslat. *Cognition* 50:431-445. http://www.wjh.harvard.edu/ lds/pdfs/Spelke1994.pdf.

Thomason, R . 2003. Logika és mesterséges intelligencia. In . E. N. Zalta (szerk.), *A Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/fall2003/entries/logic-ai/.

Turing, A. M. 1947. Előadás a londoni matematikai társaság előtt. In *The Collected Works of A. M. Turing*, Vol. Mechanical Intelligence. North-Holland. Nyilvánvalóan ez volt a mesterséges intelligencia első nyilvános bemutatása, gépirat a King's College archívumában, a könyv 1992-ben.

van Benthem, J. 1990. Kunstmatige intelligentie: Een voortzetting van de filosofie met andere middelen. *Algemeen Nederlands Tijdschrift voor Wijsbegeerte* 82:83-100.

Weizenbaum, J. 1965. ELIZA-a számítógépes program az ember és gép közötti természetes nyelvi kommunikáció tanulmányozására. *Communications of the Association for Computing Machinery* 9(1):36-45.

Winston, P. H. 1977. *Mesterséges intelligencia*. Reading, Mass.: Addison Wesley Publishing Co.