A mesterséges intelligencia filozófiai problémái

Szigorlati dolgozat
ELTE-BTK filozófia szak

Hallgató:
Csáji Balázs Csanád

Témavezető:
Farkas Katalin

Budapest, 2002/2003 őszi félév

Tartalomjegyzék

I. Előszó

„Azonban a próbálkozások 50 éves múltjából – a véleményéhez makacsul ragaszkodó néhány embert kivéve – mindenki számára világossá vált, hogy az általános intelligencia létrehozásának ez a módszere csődöt mondott.” (Hubert Dreyfus, 1992)

„*Ez olyan, mintha a filozófusok kikiáltanák magukat a színpadi bűvész trükkjei szakértőinek és akkor, amikor annak a magyarázatát kérnék tőlük, hogy a bűvész hogyan csinálja a ketté-fűrészelt-hölgy trükkjét, azt mondanák, hogy teljesen világos: a bűvész valójában nem fűrészeli ketté a hölgyet, csak ennek a benyomását kelti bennünk. „Na de ezt, hogy csinálja?” kérdezzük, „Az nem tartozik ránk” mondják a filozófusok*.” (Daniel Clement Dennett, 1972)

A huszadik század második felében nagy vitákat kavart a mesterséges intelligencia kérdése, s ennek a vitának néhány karakteres álláspontját fogjuk áttekinteni a következő dolgozatban. Mint a *filozófiatörténeti felvezetés* fejezetből kiderül, az MI kapcsán felmerülő – jórészt elmefilozófiai és nyelvfilozófiai – problémák nem új keletűek, gyökereik megtalálhatóak jóval korábban élt filozófusoknál. Sokan gondolták úgy, hogy az erős MI kísérlete kudarcot vallott (lásd Dreyfus fenti idézetét), mások az MI –t kritizáló filozófusokat bírálták (Dennett fenti idézete), annyi azonban bizonyos, hogy a gépek filozófiai értelmezése volt az a mozzanat, amely a hetvenes évek végétől nagyban hozzájárult az ún. „második kognitív forradalom” –hoz (Pléh Csaba, 1996). Ez egy szinttel absztraktabb kérdéseket vetett fel, mint a korábbi megközelítések: ideiglenesen még a pszichológusok is eltekintettek attól, hogy hogyan valósítja meg az elme a megismerést, a kognitív tudomány viszont általánosságban attól tekintett el, hogy miféle lény a megismerő. A pszichológusok számára azonban a gépek csak eszközök az emberre vonatkozó modellek explicitté tételéhez.

Csáji Balázs Csanád
Budapest, 2002. november 24.

II. Bevezetés – Mi az a mesterséges intelligencia?

Mielőtt elkezdenénk tárgyalni a mesterséges intelligenciával kapcsolatos különböző filozófiai problémákat, először tekintsük át, hogy mit is neveznek mesterséges intelligenciának a különböző szerzők. Az MI –nek négyfajta definíciója elterjedt, amiket két dimenzió mentén értelmezhetünk: (a) az egyik dimenzió, hogy a definíció a gondolkodást vagy a cselekvést célozza-e meg, míg a (b) másik dimenziója a felosztásnak az, hogy a sikert az emberi teljesítményhez mérjük vagy a siker mércéje az intelligencia egy idealizált koncepciója: a racionalitás. Ezek alapján a négyféle álláspont:

(1) Emberi módon gondolkodó rendszerek:
Amely elsősorban a kognitív vagy megismerés-tudományok megközelítése. Célja az emberi elme működését és megismerést modellező rendszerek kialakítása, hogy ezáltal is közelebb kerüljünk az elme megértéséhez.

(2) Emberi módon cselekvő rendszerek:
Ez a megközelítést Alan Turing nevéhez kötődik, akinek elhíresült Turing tesztje éppen az emberi viselkedést állította az intelligencia kritériumának, és így az elérendő célnak. Róla részletesebben lesz szó a későbbiekben.

(3) *Racionálisan gondolkodó rendszerek*:
Logicista megközelítés, mely az emberi gondolkodásnál valamilyen értelemben tökéletesebb, racionálisabb gépek / programok megalkotását tűzi ki célul.

(4) *Racionálisan cselekvő rendszerek*:
A modern informatikai tudományok / számítástudomány megközelítése, amely nem tűzi ki célul, hogy az így kialakult rendszerek valóban gondolkodjanak, azt sem, hogy közel hasonló módon működjenek mint az emberek, csak azt, hogy minél racionálisabban viselkedjenek. (pl.: előre jelezzék nekünk a földrengéseket, segítsenek diagnosztizálni betegségeket, stb.) A racionálisan viselkedő rendszereket *ágensek* nevezik.

John. R. Searle (Searle, 1980) bevezetett egy azóta elterjedt definíciót, mi szerint megkülönböztetjük a mesterséges intelligencia *gyenge* és *erős* változatát. Searle *gyenge MI –nek* nevezi azt az álláspontot, mely szerint ki lehet alakítani olyan rendszereket, amelyek úgy *cselekszenek* mintha intelligensek lennének, de a gyenge MI semmit nem mond arról, hogy egy ilyen gép valóban rendelkezik-e elmével vagy sem. Ezzel szemben *erős MI* –nek nevezett álláspont szerint olyan rendszerek is kialakíthatóak, melyek valóban *gondolkodnak,* tehát *elmének* tekinthetőek. Ez alapján az erős MI fő kérdése, hogy: *egy megfelelően programozott számítógép tekinthető-e elmének, abban az értelemben, hogy egy ilyen számítógép valóban megért dolgokat és egyéb kognitív állapotokkal rendelkezik?* Ez lesz a jelen dolgozat egyik fő kérdése, ami köré a különböző álláspontok és érvek csoportosulnak.

III. Filozófiatörténeti felvezetés

Mint a bevezetésben jeleztük, az a probléma, amivel ez dolgozat foglalkozik az (amit gyakran az MI *erős* megközelítésének is neveznek), hogy egy megfelelően programozott számítógép tekinthető-e *elménék*, abban az értelemben, hogy egy ilyen számítógép valóban *megért* dolgokat és egyéb *kognitív* állapotokkal rendelkezik. Több érv született *pro* és *kontra* – főleg a huszadik század második felében – és ezekben az érvekben kulcsszerepet játszik a gondolkodás és a nyelv viszonya, ezért itt álljon egy vázlatos filozófiatörténeti áttekintése ennek a problémának (a teljesség igénye nélkül) egészen Turing -ig, akivel részletesebben fogunk foglalkozni a következő fejezetben.

Sok más problémához hasonlóan, a gondolkodás és a nyelv kapcsolata már **Platón**nál is megjelenik. Számára úgy tűni, hogy a gondolkodás és a nyelv azonos egymással, mivel a beszéd és a gondolat ugyanannak az aktusnak a külső és belső oldala. Ezt olvashatjuk „A Szofistában" -ban:

„Nemde a gondolat és a beszéd azonos, csakhogy az egyik a léleknek bent, önmagával folytatott hangtalan beszélgetése: s ez az amit gondolkodásnak nevezünk." (Platón, Szofista)

Ez a megközelítés, mely szerint a gondolkodás és a nyelv természete között szoros kapcsolat van, sőt gondolkodás is nyelvi természetű, vagy éppen azonos a nyelvel mert a nyelvben megy végbe sok más gondolkodónál is megjelenik, lásd például: Hunboldt feltevéseit vagy Jerry Fodor gondolkodás nyelvére (mentáli) vonatkozó elméletét. Akadnak olyan filozófusok is, akik (Hubert Dreyfus, 1979) szerint nemcsak a nyelvfilozófia alapjai találhatóak meg már Platónnál, hanem az MI története jószerével i.e. 450 –ben is kezdődhetett volna, amikor egy Platón dialógusban ezt kérdezi Szókratész Euthyprosztól:

„Tudni szeretném, mi a jóság meghatározása, mitől lesz jó minden cselekedet… hogy mércének használhassam, mind a te mind más emberek cselekedeteinek megítélésénél."

Dreyfus szerint Szókratész itt egy *algoritmust* keres a jóság és annak hiányának megkülönböztetésére.

Arisztotelész volt az első aki megpróbálkozott a következtetés eljárásának általános tanulmányozásával. Arisztotelész számára a logika[1] tanulmányozása nem a szavak tanulmányozását jelenti, hanem a gondolkodását, amelynek a szavak a jelei. Megkísérelte a „helyes gondolkodás" módszereit szabályokba foglalni, s kialakította a híres szillogisztikáját, ami körülbelül két ezer évig volt uralkodó a logikával foglalkozók körében. Jelentőségét megvilágíthatjuk azzal amit Kant írt róla „*A tiszta ész kritikája*" –nak előszavába:

„A logika a legrégibb időktől kezdve ezen a biztos úton jár, ami abból látható, hogy Arisztotelész óta egyetlen lépést sem kellett hátrálnia." (Immanuel Kant, A tiszta ész kritikája, Bevezetés, 1781)

A mesterséges intelligencia kutatás logicista megközelítése azt reméli, hogy ilyen fajta (logikai következtetéseken alapuló) programokra alapozva intelligens rendszereket képes létrehozni. Már az 1960 –as években léteztek olyan programok,

[1] Arisztotelész nem használta a logika szót hanem *analitikának* nevezte az érvelések tanulmányozását.

amelyek – elegendő idő és memória esetén – képesek voltak a probléma logikai jellegű megfogalmazásából kiindulva, megadni a probléma megoldását, ha az létezett (ha pedig nem, a program soha nem hagyta abba a megoldás keresését)[2]. Természetesen egy ilyen megközelítéssel több probléma van, pl.: az informális tudásunkat formális logikai eszközökkel kifejezni nem mindig könnyű feladat, különösen, ha tudásunk nem abszolút bizonyos. Valamint minden logikai rendszerrel kapcsolatban felvethetőek bizonyos matematikai / logikai ellenérvek (pl: Gödel nem-teljességi tétele), melyekről később lesz szó.

Az első az erős MI problémájához szorosabban kapcsolható megközelítés **René Descartes** –nál található. Descartes kétféle szubsztanciát különböztet meg: *gondolkodó* és *kiterjedt* szubsztanciát. Ha úgy definiáljuk a gépeket, hogy olyan dolgok amelyek csak kiterjedt szubsztanciából állnak, akkor teljesen világos, hogy az lesz Descartes filozófiájának a következménye, hogy a gépek nem tudnak gondolkodni (mivel ha tudnának, akkor a gondolkodó szubsztanciáról kiderülne, hogy nem is valódi szubsztancia, mivel visszavezethető egy másik szubsztanciára). Ezen kívül Descartes „*Az értekezés a módszerről*” című művében (1637), két „biztos eszközt” is ad számunkra annak megállapítására, hogy az emberi cselekvést utánzó gépek mégsem emberek:

„S itt hosszasan elidőztem, hogy kimutassam, ha volnának olyan gépek, amelyek egy majom vagy más oktalan állat szerveivel és külső alakjával bírnának, semmiképp sem tudnók felismerni, hogy nem egyeznek meg mindenben ezekkel az állatokkal. Ellenben ha volnának olyan gépek, amelyek a mi testünkhöz hasonlítanának és a mi cselekedeteinket utánoznák, amennyire erkölcsileg csak lehetséges, akkor mégis volna két biztos eszközünk arra annak megállapítására, hogy azért mégsem igazi emberek. Az első az, hogy a gépek sohasem tudnának szavakat vagy másjeleket használni, mint mi tesszük, ti. hogy gondolatainkat közöljük másokkal. Mert nagyon jól el tudjuk képzelni, hogy egy gép úgy van alkotva, hogy szavakat mond ki, sőt, hogy egyes szavakat olyan testi cselekvések alkalmával mond ki, amelyek névi változást idéznek elő szerveiben. Pl. Ha az egyik helyen érintik azt kérdeti, hogy mit akarnak tőle; ha más helyen érintik, kiabál, hogy fáj neki, s több efféle. De azt nem tudjuk elképzelni, hogy a szavakat különféleképpen elrendezze s ezáltal értelmesen tudjon felelni arra, amit jelenlétében mondanak, amint ezt a legtompaeszűbb emberek is meg tudják tenni. A második az, hogy habár néhány dolgot éppoly jól vagy még jobban csinálnának, mint akármelyikünk, de okvetlenül csődöt mondanának másokban; ebből pedig megtudhatnánk, hogy nem tudatosan cselekszenek, hanem csak szerveinek elrendezésénél fogva. Míg ugyanis az ész egyetemes eszköz, mely minden esetben feltalálja magát, addig ezeknek a szerveknek minden különös tevékenység számára különös berendezésre van szükségük; ennélfogva erkölcsileg lehetetlen, hogy annyiféle berendezés legyen egy gépben, hogy az élet minden helyzetében úgy tudjon cselekedni, mint mi tudunk értelmünk segítségével.” (René Descartes: Értekezés a módszerről, 1608 - 1643)

Tehát:

(1) A gépek szerinte sohasem lesznek képesek a nyelvet úgy használni, ahogy mi tesszük, azaz a gondolataink kifejezésére.
(2) Noha elképzelhető, hogy ezek a gépek egyes dolgokat éppolyan jól, vagy még jobban csináljanak mint az emberek, Descartes szerint okvetlenül csődöt mondanának más dolgokban, mivel nem tudatosan cselekszenek, hanem a szerkezetük révén.

Megjegyezésre érdemes, hogy Descartes megkülönböztet *metafizikai* és *morális* bizonyosságot (a bizonyosság mértéke szerint), és ebben a kérdésben, mint ahogyan láttuk, csak morális bizonyossággal foglalt állást.

[2] A legelterjedtebb logikai programozási nyelv napjainkban a PROLOG.

Descartes ugyan az embereket, mint gondolkodó lényeket nem tekintette gépeknek, de az állatokat már igen. Azonban műve után 100 évvel (1748 –ben) **Julien Offray de la Mettrie** „*L'Homme Machine*" könyvében explicit módon kifejti, hogy az emberek (is) automaták.

A világ első számítógépeinek a tervezése **Charles Babbage** (1791 – 1871) nevéhez fűződik[3], az első programokat pedig **Augusta Ada Byron**[4] (1815 – 1852) tervezte. Pl.: javaslatot tett a Bernoulli számok[5] géppel való kiszámítására. Emlékirataiban Ada Byron foglalkozik MI –hez köthető kérdésekkel, melyekre Turing is kitér később elemzendő cikkében *Lady Lovelace ellenvetése* néven. Ada konkrétan a következőket írja a Babbage tervezte analitikus gépről:

„The Analytical Engine has no pretension to originate anything. It can do whatever we know how to order it to perform." (Lady Lovelace's memoir)

A következő fejezetben Turing híres 1950 –es „*Computing Machinery and Intelligence*" cikkével foglalkozunk, ami mint látni fogjuk bizonyos módon kapcsolódik Descartes megközelítéséhez. Természetesen az a történeti áttekintés amely itt szerepelt nem teljes, csak jelezni szerettem volna, hogy azok a problémák amelyek az 1950 –évektől oly éles vitákat váltottak ki, nem minden előzmény nélkül jelentek meg, hanem gyökereik már sokkal korábban is megtalálhatóak.

Ada Augusta Byron

René Descartes

[3] A „Differencia Gép" és az „Analitikus Gép". Fizikailag egyik sem készült el teljesen. Az analitikus gép egyes darabjai a Londoni Természettudományi Múzeumban találhatóak.

[4] Byron lánya.

[5] $B_0 = 1, \sum_{i=1}^{n-1} \binom{n}{i} B_i = 0$ alakú számok.

IV. Turing öröksége

Ha úgy interpretáljuk Descartes – az előző fejezetben idézett – érvét, hogy: a gépek nem képesek gondolkodni (mivel csak *kiterjedt* szubsztanciával rendelkeznek, és nem rendelkeznek *gondolkodó* szubsztanciával), ezért nem képesek a nyelvet értelmesen (tetszőleges szituációkban) használni, akkor ebből *kontrapozícióval* egy olyan állításhoz jutunk, hogy ha képesek lennének a nyelvet értelmesen használni, akkor abból tudnánk, hogy gondolkodni is tudnak. Hasonló álláspontot foglalt el **Alan Matheson Turing** (1912 – 1954) cambridgei matematikus is az 1950 –es években híres cikkében (Turing, 1950). A cikk legfőbb állítása – az én olvasatomban – az, hogy ha képesek lennénk egy olyan megfelelően programozott számítógépet létrehozni, amely tetszőleges témáról folytatott beszélgetésben úgy tudna részt venni, hogy nem tudnánk megállapítani, hogy egy emberrel vagy egy géppel beszélgetünk, akkor feltételeznünk kellene, hogy a gép rendelkezik kognitív állapotokkal, tehát azt, hogy a gép *valóban gondolkodik*. A gondolkodással rendelkezés ezen megállapítását Turing úgy képzelte, hogy egy kérdezőbiztos egy számítógép segítségével tartja a kapcsolatot egy szomszéd szobában lévő emberrel vagy egy programozott számítógéppel, és meg kell állapítania, hogy akivel beszélt, az egy ember volt, vagy egy gép. Ha olyan szintre fejlődnének a programok, hogy ennek megállapítása már nem volna lehetséges, akkor az azóta *Turing tesztnek* nevezett próbán átment programozott számítógépekről *feltételeznünk* kellene, hogy gondolkodnak. Hasonlóan ahhoz, ahogyan a világban észlelt emberekről is csak *feltételezzük*, hogy gondolkodnak (elmével rendelkeznek), mivel értelmes válaszokat adnak a kérdéseinkre, értelmes módon tudunk velük kommunikálni. Ezen feltételezés nélkül könnyen a szolipszizmus csapdájába eshetnénk, érvel Turing. Például ilyen beszélgetéseket képzel el Turing, miután a tesztalanyt (gép vagy ember) megkérték, hogy írjon egy szonettet:

„Kérdező:	*A szonetted első sorában azt olvashatjuk, hogy: „egy nyári naphoz vagy hasonló nékem”, nem lenne legalább ilyen jó vagy még jobb a „tavaszi nap”?*
Tesztalany:	*A vers üteme felborulna.*
Kérdező:	*És a „téli nap” –hoz mit szólsz? Akkor az ütemmel nem lenne gond.*
Tesztalany:	*Igen, de senki sem szereti, ha egy téli naphoz hasonlítják.*
Kérdező:	*De azt azért mondhatnád, hogy Pickwick úr a Karácsonyra emlékeztet téged?*
Tesztalany:	*Bizonyos értelemben igen.*
Kérdező:	*Nos, a karácsony is egy téli nap és nem hiszem, hogy Pickwick úr ellene lenne a hasonlatnak.*
Tesztalany:	*Ezt nem gondolhatod komolyan! Egy „téli nap” alatt egy tipikus téli napot szoktak érteni és nem egy olyan kivételes napot, mint amilyen a karácsony.”* (Turing, 1950)

Turing az emberi intelligencia egy operacionális vagy funkcionalista definícióját kívánta megadni, amikor azt javasolta, hogy tekintsünk intelligensnek minden olyan gépet, mely szöveges terminálon keresztül majdnem ugyanolyan eséllyel képes elhitetni kérdezőjével, hogy emberrel kommunikál, mint emberi ellenfele. Az érv amely szerint *ha* egy ilyen gép viselkedése megkülönböztethetetlen az emberétől *akkor feltételeznünk* kell, hogy a gép gondolkodik elég meggyőzőnek tűnik (legalábbis számomra), de Turing több az álláspontjával szemben felvethető

ellenérvet megvizsgál. Pl.: az un *alkalmatlansági argumentumok*at, melyek szerint a gépek sohasem lesznek képesek X –re, ahol X helyére sokféle dolgot lehet behelyettesíteni, pl.:

„Kedvesnek, leleményesnek, szépnek, barátságosnak, kezdeményezőnek lenni, humorérzékkel rendelkezni, hibázni, szerelembe esni, tejszínes epret élvezni, tapasztalatból tanulni, szavakat helyesen használni, saját gondolataiknak az alanya lenni, olyan változékony viselkedést felmutatni, mint egy ember, valami igazán újat megtenni, stb." (Turing, 1950)

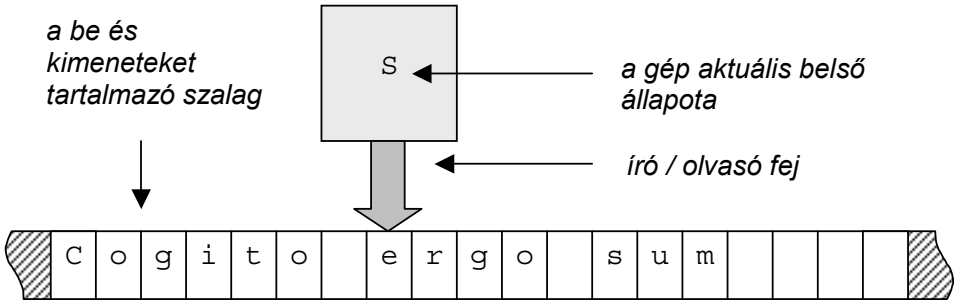
Turing rámutat, hogy ezen ellenvetéseket általában nem támasztják alá érvekkel, s inkább csak az előítéleteknek köszönhetőek. (Bár több ezek közül a gépi tudat kérdését feszegeti, amiről később részletesebben lesz szó.) Turing szerint az említett ellenvetésekből sugárzó kétely a gépeknek, mint olyan eszközöknek az eddigi tapasztalataiból erednek, melyek kevés észlelést és nulla következtetést igénylő ismétlő feladatokat végeznek. Turing emlékeztet, hogy a 40 –es évek végén a lakosság zöme nem hitte el, hogy a gépek találják majd meg egyenletek numerikus megoldását, vagy pedig megjósolnak ballisztikus röppályákat. Sőt ma is sok szakmailag olvasott ember nem hiszi el, hogy a gépek képesek tanulni.

Egy másik érdekes ellenvetés – amelyet már a bevezetésben is említettünk – Lady Lovelace ellenvetése, miszerint egy gép csak azt tudja csinálni, ami mi „megmondtunk" neki, hogy csináljon. Ez az ellenvetés véleményem szerint analóg Descartes második gépeket limitáló megállapításával, miszerint a gépek csak különös szerkezetük révén cselekszenek, és így nem lehetnek univerzális eszközök. Turing bevallja, hogy nincsenek meggyőző érvei ezen ellenvetések ellen, de valamiféle ötletekkel vagy analógiás érvekkel szolgál a számunkra. Kifejti, hogy a számítógépek olyan programozását, hogy azok intelligens viselkedést produkáljanak, nem direkt módon képzeli el, hanem olyan módon, hogy *tanulni* képes gépeket épít. Napjaink mesterséges intelligencia kutatása is nagymértékben foglalkozik a különböző gépi tanulási módszerekkel, és több bíztató eredmény született már, tehát Turing ötlete nagyon gyümölcsözőnek bizonyult. A másik analógiás érve, amit érdemes felidézni az un. *hagymahéj analógia*[6], amivel azt szeretné alátámasztani, hogy az elme mechanikus természetű (s így teljes mértékben az agy terméke). Analógiája alapján az agyat vizsgálva az elme bizonyos funkcióiról ki fog derülni, hogy mechanikus úton magyarázhatóak, s ezért (mivel így nem tartoznak a *valódi* elméhez), mint egy hagymahéjat leválaszthatjuk az elméről. A maradékot tovább vizsgálva újabb funkciókról derül ki, hogy mechanikus természetűek, és ezeket ismét leválaszthatjuk az elméről. Kérdés az, hogy marad-e valami az elméből a folyamat végére. Ha nem, akkor el kell ismernünk az elme mechanikus természetű.

Mielőtt rátérnénk más ellenvetésekre és Turing –nak az ezekre adott válaszára, lássuk, hogy mit is ért számítógépen. Turing elég pontosan definiálja, hogy ő mit ért számítógép alatt: még 1936 –ban bevezette az absztrakt gépek egy osztályát, amiket azóta az ő tiszteletére Turing gépeknek neveznek. Először röviden ismertetem a Turing gépek egy szemléletes definícióját, majd a formális matematikai definíciót. A gép rendelkezik *véges* számú belső állapottal. A gép minden időpillanatban pontosan egy belső állapotban van. Adott egy szalag, amely mindkét irányban potenciálisan végtelen, és cellákra van osztva. A szalag potenciálisan végtelenségén azt értjük, hogy miden időpillanatban véges hosszúságú ugyan, de balról is, jobbról is, új cellákkal

[6] Érdekes az a tény, hogy az első hagymahéj analógiás érvelés a buddhista tanításokban található, akik szintén a lélek semmisségének a kimutatására használták ezt az analógiát.

kiegészíthetjük. Adott a szalag szimbólumainak valamely *véges* halmaza, amelyet a gép jelkészletének (ábécéjének) nevezünk. A gépnek van egy író olvasó feje, amely adott időpillanatban a szalag pontosan egy celláján helyezkedik el. A gép *nem folytonosan* működik, hanem *diszkrét* időpillanatokban. A gép az éppen olvasott szimbólumtól és a belső állapotától függően ír egy szimbólumot a szalag aktuális cellájára vagy lép egyet jobbra vagy balra és megváltoztatja a belső állapotát. Ha egy szalagot ráillesztünk a Turing gépre, akkor a gép író-olvasó fejét a szalag egy cellájára helyezve, s a gép valamely belső állapotban van, akkor a gép elkezdi a működését a szalagon: az író-olvasó fej törli és írja a szimbólumokat, s balra vagy jobbra lépeget a szomszédos cellákra. Ha valamikor a gép megáll, a megálláskor a szalagon lévő jelsorozatot a számítás eredményének nevezzük.



egy Turing gép vázlatos ábrája

Definíció: Egy Turing gép formálisan egy $M = (K, \Sigma, \delta, s)$ rendezett négyes, ahol K az állapotok véges halmaza, $s \in K$ kezdőállapot, Σ pedig betűk véges halmaza (abc). $K \cap \Sigma = \varnothing$ valamint Σ tartalmazza a kezdet és az üres szimbólumokat. $\delta : K \times \Sigma \to (K \cup \{\uparrow, \downarrow, \hookleftarrow\}) \times \Sigma \times \{\leftarrow, \rightarrow, \_\}$ ez az ún. átmenetfüggvény.

A gép első ránézésre furcsának tűnhet, de rendkívül hatékony, Papadimitriu például ezt írja róla könyvében:

„Gyenge és ügyetlen megjelenése ellenére a Turing gép a hatékonyság számottevő romlása nélkül képes szimulálni bármely algoritmust. .... Bámulatos, milyen kevésre van szükségünk ahhoz, hogy mindenünk meglegyen!” (Papadimitriu, 1994)

Felmerülhet a kérdés, hogy milyen feladatokat lehet megoldani Turing géppel? Le lehet-e valóban írni egy Turing géppel bármely algoritmust, amit algoritmusnak gondolunk? A tapasztalat azt mutatja, hogy igen. Eddig még senki nem tudott olyan algoritmus adni, amit nem lehetett Turing géppel utánozni. Ezt matematikailag természetesen nem lehet bebizonyítani, hiszen az algoritmus heurisztikus fogalma nehezen fogható meg matematikailag. Ennek ellenére meg van fogalmazva, majdnem úgy, mint egy tétel. Úgy nevezik, hogy Church-tézis, amely szerint *minden algoritmikus feladat elvégezhető Turing géppel*[7].

[7] Természetesen a mai digitális számítógépek működése tökéletesen „belefér” a Turing gépek elméletébe (szimulálhatóak Turing gépekkel).

V. Matematikai ellenvetések

Az előző fejezetben láttuk, hogy milyen diszkrét állapotú gépekkel képzeli Turing kialakíthatónak az MI –t. Természetesen néhány matematikai / logikai ellenérv is tehető egy így értelmezett számítógéppel kapcsolatban, melyek némelyikével maga Turing is foglalkozik 1950 –es cikkében. Ilyen ellenvetés lehet az ún. *megállási probléma*. A megállási probléma annak az eldöntése, hogy egy program véget fog-e érni, vagy örökké fog futni. Turing bebizonyította, hogy bármely *H* algoritmus esetén létezik olyan *P* program, hogy *H P* megállási problémáját korrekt módon eldönteni nem lesz képes. Ez a tétel analóg **Kurt Gödel** híres nem-teljességi tételével, amely azt mondja ki, hogy egy megfelelő erősségű logikai rendszerben (amelye pl. tartalmazza a Peano aritmetika axiómáit) konstruálható olyan állítás, amely a rendszeren belül sem nem bizonyítható, sem nem cáfolható, hacsak a rendszer nem inkonzisztens. (Bővebben (Smullyan, 1999) –ben) Hasonló eredményei vannak Church –nek, Kleene –nek, Rosser –nek is. Természetesen Turing is tisztában volt az általa bevezetett diszkrét véges állapotú gépek korlátozottságával, de szerinte semmi okunk sincsen annak feltételezésére, hogy az emberi értelemnek nincsenek hasonló korlátai. Felmerülhet a kérdés, hogy egy ilyen korlátozottság nem segíthet-e egy Turing tesztben résztvevő kérdezőbiztosnak, hogy eldöntse, hogy emberrel vagy számítógéppel beszélget. Feltehetné mondjuk a miden géphez biztosan létező *megállási problémát* s ezáltal lebuktathatná a gépet. Azonban a kérdezőbiztosnak semmilyen adata nincs az adott gép működéséről, de ha még pontosan ismerné is a gépek szerkezetét az eldöntési problémához szükséges speciális programot nem tudná megkonstruálni. De ha még valamilyen véletlen folytán rendelkezésére állna is az eldöntési problémához szükséges program, és azt feladná a terminál másik végén lévő tesztalanynak (akiről nem tudja, hogy gép vagy ember) akkor sem lenne képes ez alapján kiszűrni a gépeket, ugyanis a program próbálkozhatna egy ideig, aztán adhatna egy olyan választ, hogy: „*Nem tudom, ez a probléma nagyon bonyolult*.” Ami teljesen emberi válasz lenne.

Hasonló választ fogalmaz meg **J. R. Lucas**, aki szerint az ember sem képes kimutatni egy nagy számítógépet vagy az agyát leíró rendszer konzisztenciáját és nem képes kimutatni a megfelelő Gödeli állítás igaz voltát. Következésképp az emberek ugyanolyan korlátozásnak vannak kitéve, mint a formális rendszerek. (Lucas, 1961)

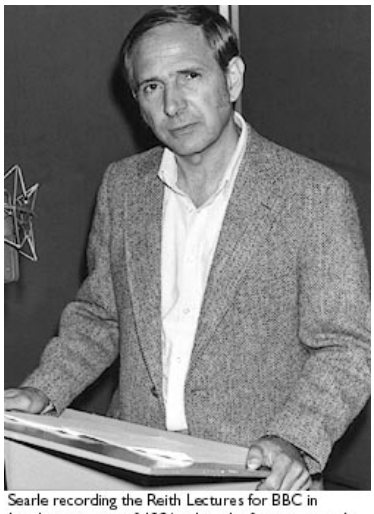
Ennek a kérdésnek volt további utóélete, pl: **Roger Penrose** is foglalkozik vele „*A császár új elméje*” című könyvében. Penrose azt állítja, hogy a matematikai „belátás” nem lehet algoritmikus:

„*Nos eme vélt „univerzális” rendszer, vagy algoritmus nem lehet egyike azoknak, amit mi matematikusok az igazság kiderítésére használunk. Mert ha ez így lenne, akkor meg tudnánk konstruálni a gödeli állítását és tudnánk, hogy ez egy igaz matematikai állítás. Így kénytelenek vagyunk arra következtetni, hogy az az algoritmus, amit a matematikusok a matematikai igazság eldöntéséhez használnak túlságosan bonyolult ahhoz, hogy az érvényességét valaha is megtudjuk*.” (Penrose, 1990)

Ugyan Turing gépekre vonatkozó javaslatának vannak matematikai nehézségei, azonban ezeknél súlyosabb ellenvetések is tehetőek ellene, a gép szimbólumainak szemantikai értékével kapcsolatban (szimbólum lehorgonyzás), amivel a következő fejezetben foglalkozunk részletesebben.

VI. A kínai szoba argumentum

Függetlenül a matematikai / logikai ellenérvtől létezik egy sokkal súlyosabb probléma is, amellyel szembe kell nézniük azoknak, akik szerint a diszkrét állapotú szimbolikus jelfeldolgozást végző számítógépek kognitív állapotokkal rendelkezhetnek: ez pedig a *szimbólum lehorgonyzás* problémája. **John R. Searle** „*Az elme, az agy és a programok világa*" című cikkében szereplő híres „kínai szoba" érve jól szemlélteti a probléma lényegét: Searle megkérdőjelezi a szimbolikus MI központi feltételezését, vagyis azt, hogy ha egy szimbólumrendszer képes olyan viselkedést generálni, amely megkülönböztethetetlen az emberétől, akkor kell, hogy legyen elméje. Searle demonstrációja azon alapul, hogy elképzeli magát, amint mindent megcsinál, amit egy számítógép megcsinál: egy szobában ül, van egy könyve (program), amiben különböző utasítások vannak kínai jelek feldolgozására vonatkozóan angolul (tegyük fel, hogy angolul tökéletesen beszél, kínaiul viszont egyáltalán nem), van potenciálisan végtelen papírja (memória) amit használhat a számítások elvégzésére. Egy ablakon keresztül szimbólum sorozatokat kap kínaiul és a könyvben lévő angol nyelvű utasításoknak megfelelően feldolgozza, majd egy eredményt egy másik ablakon kiadja. Ebben az esetben, még ha egy a kínai szobát kívülről szemlélő ember számára megkülönböztethetetlenek lennének is a szobából kijövő szövegek attól, amit mondjuk egy kínai anyanyelvű ember produkálna, mégsem mondhatjuk, hogy a gép (az ember a szobában) megértette a beszélgetést. A csak angolul tudó ember (Searle) aki a szobában ül és a jeleket dolgozza fel teljesen világos, hogy nem értene semmit a beszélgetésekből. Nem köt semmit a szimbólumokhoz, azokat csak a formájuk alapján dolgozza fel. (Searle, 1980) Searle megközelítéséhez hasonló probléma merül fel pl. akkor, amikor valaki kínaiul akar megtanulni egy kínai – kínai (értelmező szótár) segítségével. Noha minden szimbólumnak ott van a magyarázata, sohasem jutna odáig, hogy bármit is megértsen, mivel a szimbólumok más szimbólumokkal vannak csak magyarázva. A szimbólum lehorgonyzás problémájáról bővebben **Hernád István** cikkében olvashatunk (Hernád, 1992).



Searle recording the Reith Lectures for BBC in London, summer of 1984, where he first presented the "Chinese Room" argument. © BBC photo

Turinghoz hasonlóan Searle is foglalkozik különböző ellenvetésekkel cikkében. Például érvelhetnénk az argumentummal szemben úgy, hogy a kínai szobában ülő ember nem az egész számítógépnek felel meg, csak a *központi feldolgozó egységnek* (CPU) és mi nem azt szeretnénk állítani, hogy a CPU megért dolgokat vagy gondolkodik, hanem azt, hogy az egész számítógép, a rendszer ért meg dolgokat. Ezt az ellenvetést Searle *rendszerelvű válasznak* nevezi és meggyőző viszont válasza úgy szól, hogy képzeljük el, hogy az ember memorizálta a kínai jelek feldolgozásáról szóló könyvet és a számításokat is fejben végzi, sőt ilyen módon a kínai szoba falai is feleslegesek. Ott áll tehát áll előttünk egy ember, akinek bármilyen kérdést feltehetünk kínaiul (írásban) és ő válaszolni fog rá szintén kínaiul, miközben az ég világon nem ért semmit a beszélgetésekből.

Mint láttuk a fő probléma a számítógéppel az volt, hogy nem tudta semmihez sem kötni a szimbólumok jelentését, s éppen ezért nem értett semmit a beszélgetésekből. Érdekes elgondolkodni azonban azon, hogy nem menthető-e meg az erős MI a kínai szoba argumentumtól és egyéb a szimbólumok szemantikai értékét hiányoló argumentumoktól az *externalizmussal*. **Hilary Putnam** az externalizmusra vonatkozó tételét – „*The Meaning of Meaning*" –ben – eredetileg jelentésekre mondta ki, de később kiterjesztették gondolatokra is. Az externalizmus klasszikus megközelítése azt mondja, hogy az amit a szavakon értünk függ a külvilágtól, a kiterjesztés pedig azt állítja, hogy amit a gondolataink kifejeznek az szintén függ a külvilágtól. Ekkor a kínai szimbólumok jelentését az általuk jelölt a külvilágban létező dolgokhoz kötve talán megoldhatnánk a szimbólumok szemantikai értékének kérdését.

A MI erős megközelítésének a problémája első látásra úgy tűnik, hogy a klasszikus test – lélek vagy test – elme problémakörnek egy új formába öltöztetetett régi megoldása. Úgy tűnik, hogy Turing és a szimbolikus MI követői visszavezethetőnek tartják a gondolkodást anyagi természetű folyamatokra (mivel a Turing által leírt számítógépek fizikailag megvalósíthatóak), míg Searle cikkében ennek lehetőségét vitatta volna el. A helyzet ennél egy fokkal bonyolultabb, ugyanis, Searle csak a megismerés szabályalapú, szintaktikai szimbólum-feldolgozó koncepcióját kritizálja, de hisz a gondolkodás anyagi természetű voltában, s egy *biológiai naturalista* álláspontot foglal el. Ezt írja a cikke zárszavában:

„Tud egy gép gondolkodni? Nézetem szerint csak egy gép tud gondolkodni, méghozzá valójában csak nagyon különleges fajtájú gépek, vagyis az agy és olyan gépek amelyek ugyanazokkal az oki hatóerőkkel rendelkeznek, mint az agy. Ez a fő oka annak, hogy az MI erős verziója keveset tud mondani a gondolkodásról, mivel semmit sem tud mondani a gépekről. Saját definíciója szerint programokról szól és a programok nem gépek. Az intencionalitás, bármi is legyen az, biológiai jelenség, és mint ilyen okozatilag függ keletkezésének sajátos biokémiájától éppúgy mint a tejelválasztás, a fotoszintézis vagy más biológiai jelenségek. Senki sem feltételezné, hogy tejet vagy cukrot elő tudunk állítani azáltal, hogy lefuttatjuk a tejelválasztás vagy a fotoszintézis formális modelljének számítógépes szimulációját, amikor azonban az elméről van szó, sok ember a mély és tartós dualizmus következtében hajlamos hinni ilyen csodákban: úgy gondolják, hogy az elme formális folyamatok kérdése és független a meglehetősen sajátlagos anyagi okoktól, ellentétben a tejjel és a cukorral." (Searle, 1980)

Fontos észrevenni, hogy Searle kínai szoba érve használható mindenfajta funkcionális / operacionális megközelítés ellen, mivel nem csak (kínai) szimbólumok jöhetnek a szobába bemenetként, hanem valós számok, vagy akármilyen adat, amit önmagában a szobában ülő ember nem tud a külvilághoz kötni. Tehát a később ismertetendő konnekcionizmus is ki van téve Searle argumentumának, hiba nincs központi jelfeldolgozó egysége vagy szimbolikus szabályai. Ugyanis egy konnekcionista modellt is be lehet memorizálni, s így a rendszerelvű válaszra adott ellenvetés minden funkcionalista megközelítés ellen alkalmazható. Azonban a funkcionalizmusnak is van válasza Searle álláspontjával szemben, amit az *agyprotézis kísérlet* fejezetben ismertetünk.

VII. Konnekcionizmus

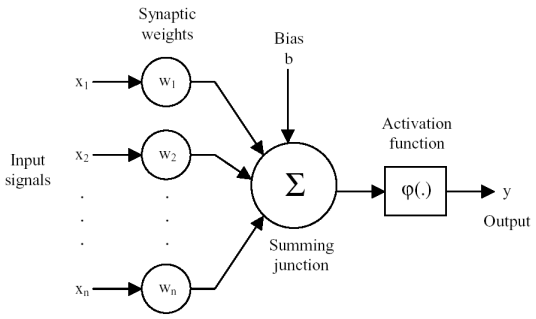
Az MI (és az elme modellezésének) szabályalapú, szintaktikai szimbólum feldolgozó koncepciója szerint (mint amilyen pl.: a Turing által javasolt modell) a „HA X AKKOR TEDD Y –t!” minden gondolkodási mechanizmus általános sémája. Az ilyen megközelítések ellen direkt módon alkalmazhatóak a szemantikai értékek hiányára épülő ellenérveket, mint amilyen Searle kínai szoba argumentuma. Azonban léteznek másfajta megközelítései is az elme modellezésének pl.: a *konnekcionizmus*, amely megkérdőjelezi a korábbi egyközpontú, szekvenciális logika-centrikus modellt. Egységes koncepció ez is, de az egységet egy sokkal „primitívebb” szinten találja meg. A konnekcionista felfogás nyíltan olyan modelleket szeretne kialakítani a megismerésről, amelyek az idegrendszer működéséből merítik analógiájukat vagy irányító metafórájukat. Ennek megfelelően a feldolgozás egysége náluk nem a szimbólum, hanem a neuronokra emlékeztető (elméleti idegsejtszerű) egységek puszta izgalmi mintázata. Szélsőséges megfogalmazásban: tudásunk nem más, mint neurális hálózatok izgalmi mintázata. Fontos megemlíteni, hogy a konnekcionizmus szemantikailag „nem áttetsző” típusú megismerés modellálást valósít meg:

„*Egy rendszert akkor nevezünk szemantikailag áttetszőnek, ha tiszta leképezés írható le a rendszer viselkedésének szimbolikus (fogalmi szintű) szemantikai leírása, valamint formális számítási tevékenységén belül reprezentált tárgyainak valamilyen kivetíthető szemantikai értelmezése között.* ” (Clark, 1999)

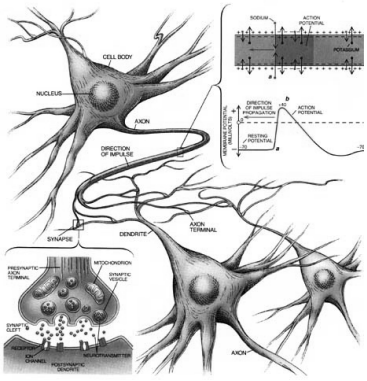
Áttetsző modellek például a Jerry Fodor vagy Daniel C. Denett által képviselt álláspontok, tehát a vélekedés-vágy koncepcióját használó megközelítések. A konnekcionizmus azonban tagadja, hogy a fejünkben kifejezések és attitűdök lennének, ezek legfeljebb a leíró tudós segédeszközei. Ennek megfelelően konnekcionista modellben nincsenek szimbólumok és szabályok. A modell decentralizált abban az értelemben, hogy nincs egy kitüntetett processzor vagy jelfeldolgozó egység (mint például a Turing gépnél), hanem párhuzamosan működő, egyenrangú egységek versengése jellemzi. A tudás a hálózatban résztvevő egységek (idegsejtek) kapcsolataival reprezentálható. Minden reprezentáció ezen „építőkövekből” épül fel. A tanulási folyamat a hálózati súlyok megváltozásával zajlik. A sémák-problémája úgy oldódik fel, hogy a sémák nem egyebek, mint korábbi tapasztalatból fakadó gyakori együtt járások. **Pléh Csaba** egy cikkében az alábbi állításokkal jellemzi a konnekcionista megközelítést (Pléh, 1996):

1. Idegrendszerszerű modellálás.
2. Minden tudásunk csomópontok izgalmával, s a köztük lévő kapcsolatokkal írható le.
3. A hálózatok címkézetlenek.
4. Nincsenek szimbólumok és szabályok.
5. Párhuzamos aktivitás és versengés jellemzi a rendszert.
6. A tudás a hálózat egy részének aktivált állapota.
7. Mikroreprezentáció: minden reprezentáció kicsiny építőkövekből áll össze.
8. A tanulás a hálózat súlyviszonyainak beállítása.
9. A sémák a konnekcionista rendszerben együtt aktivált hálózatrészek
10. A „neurális”, pszichológiai és számítástechnikai modellálás új szövetsége.

A konnekcionizmusnak és szimbólikus megközelítésnek természetesen kevert változatai is kialakultak, az ilyen megközelítések többsége feltesz egy külső vagy belső eredetű, nem transzparens rendet, és egy erre koherenciát és logikát ráépítő másodlagos szintet. Egy ilyen álláspontot képvisel például **Andy Clark** (Clark, 1999).



Egy neuron modellje (elméleti idegsejt)[8]



Valódi neuronok vázlatos ábrája

Searle a *kínai szoba* argumentumot bemutató cikke végén több az álláspontjával szemben felvethető ellenérvet is megvizsgál, például a konnekcionista választ is, amit ő *agyszimulátor válasznak* nevez. Ahogy az Searle válaszából kiderül ő minden funkcionalista megközelítést kritizál és nem csak a komputációs (szimbolikus / szabályfeldolgozó) modelleket. A konnekcionizmus ellen úgy módosítja az argumentumot, hogy a szobában lévő ember nem egy könyvből kiolvasott utasítások alapján dolgozik és papíron számol, hanem különböző csapokból és szelepekből álló finom rendszert működtet, amely pontosan szimulálja egy kínaiul tökéletesen beszélő ember idegsejtjeinek szinaptikus kapcsolatait. Ebben az esetben sem értene semmit a csapok és csővezetékek kezelője, és természetesen a csővezetékek sem. De ha még egy olyan álláspontra helyezkednénk is, hogy az ember és a csővezetékek *együtt* tudatosak, akkor itt ismét bevethetjük a *rendszerelvű válasz* ellen használt módszert. Tehát a csővezetékek és a csapok rendszerét kezelő ember (elméletileg) be tudja memorizálni az egész rendszert (és ezzel belsővé teheti), és fejben is elvégezheti a műveleteket (csapok nyitását és a folyadékok áramlását), s ebben az esetben sem értene semmit a kínaiul folyó beszélgetésekből.

---

[8] Ez egy egyszerű idegsejt modell, amit még McCulloc és Pitts vezetett be az ötvenes évek elején. Ma már sokkal bonyolultabb idegsejt modellek is léteznek. Az itt említett modell (amit a szakirodalomban *perceptronnak* neveznek) a bemenetekből (x_i) a következő módon számolja az idegsejt kimenetét (y):

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b), \qquad \varphi(x) = \frac{1}{1+e^{-x}}$$

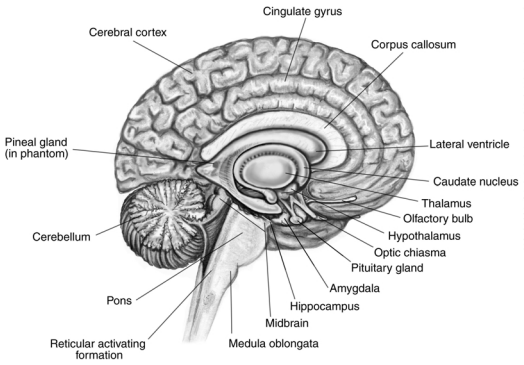
φ tipikusan egy nemlineáris függvény, mint amilyen a fent használt logisztikus függvény. Egy ilyen neuronmodellel könnyen lehet tanulási tulajdonságokat elérni, mint ahogyan azt Rosenblatt megmutatta 1958 –ban. Még az ilyen egyszerű mesterséges neuronokból felépülő hálózatok is rendkívül hatékony viselkedést mutatnak és univerzálisak abban az értelemben, hogy véges sok ilyen idegsejt képes bármilyen kis hibával approximálni (közelíteni) függvényeket. Ilyen módon talán matematikai bizonyossággal cáfolni lehet Descartes második kritériumát a gépekkel gondolkodásával szemben, miszerint a gépek nem lehetnek egyetemes (univerzális) eszközök.

A konnekcionista (s így funkcionalista) megközelítésekkel szemben – amelyek eltekintenek attól, hogy milyen közegben megy végbe a gondolkodás, ha az funkcionálisan megfelel a gondolkodó emberek idegsejtjeinek szinaptikus kapcsolatainak, más ellenérvek is tehetőek. Pl.: olyan dolgokat is elméknek kellene tekintenünk, ha azok funkcionálisan megfelelnek egy konnekcionista modellnek, amelyeket intuitívan nem szívesen neveznénk elmének. Például, egy nagyváros (mondjuk London) vízvezeték hálózata is megvalósíthatja (elméletileg) az idegsejtek szinaptikus kapcsolatát, vagy elképzelhetjük, amint Kína összes lakosa – több mint egymilliárd ember – összeáll, és mindegyikük szimulál egy darab idegsejtet, és az idegsejt tüzelési erősségét (a bemenetektől függően) továbbítja kijelölt szomszédainak. Ilyen módon a kínaiak hálózata megvalósíthatja funkcionálisan egy ember szinaptikus kapcsolatait, de sem London vízvezeték-hálózatát sem Kína egész lakosságának ily módon való együttműködését nem tekintenénk szívesen elmének (Clark, 1999). Ezek az ellenvetések abba az irányba mutatnak, hogy vagy intuíciónk téveszt meg minket vagy pedig valamilyen komoly baj van a funkcionalizmussal.

Searle alternatívaként a biológiai naturalizmust javasolja, amely álláspont szerint nem tekinthetünk el azoktól az anyagi hatóerőktől, amelyekben a gondolkodás végbemegy. Azonban Hans Moravec a funkcionalizmus védelmében megfogalmazott a következő fejezetben ismertetett érve szerint nem egyeztethető össze konzisztens módon a biológiai naturalizmus a funkcionalizmus tagadásával.

VIII. Az agyprotézis kísérlet

Mint láttuk a *biológiai naturalizmus* (pl: Searle) álláspontja az, hogy bizonyos folyamatok elválaszthatatlanok attól az anyagi közegtől, amiben végbemennek, s így a *funkcionalizmus* megközelítése, mely elvonatkoztat az adott funkciót megvalósító közegtől, hibás. Pl.: Searle olyan példákat hoz, hogy egy vihar számítógépes szimulációjától senki sem várja, hogy vizesek legyünk, ugyanígy az agy számítógépes szimulációjától sem szabad várni, hogy gondolkodjon. Ezzel az állásponttal szemben (a funkcionalizmus védelmében) hozza fel **Hans Moravec** az ún. agyprotézis kísérletét. (Moravec, 1988). Az érvelését a következőképpen foglalhatnánk össze: tegyük fel, hogy a funkcionális neurobiológia már odáig fejlődött, hogy az agy minden neuronja és összeköttetései tökéletesen ismertek. Tehát funkcionálisan az agy minden részéről és neuronjáról meg tudjuk mondani, hogy milyen bemenetre milyen kimenetet produkál (ennek lehetőségét Searle sem kérdőjelezi meg). Tegyük fel továbbá, hogy képesek vagyunk ezen viselkedéseket utánzó *mű* neuronokat építeni. Képzeljünk el egy olyan kísérletet melyben egy ember agyának egy neuronját kicseréljük egy funkcionálisan ugyanúgy működő *mű* neuronra. Mi történik ekkor? Változik valamit az ember viselkedése? Az argumentum úgy folyatódik, hogy szépen sorban cseréljük ki az ember agyának összes neuronját. Ekkor funkcionálisan az ember ugyanúgy fog működni, mint előtte. Kérdés az, hogy mi történt közben a tudatával? Eltűnt-e valamikor a kísérlet közben? Bár ezt nehéz lenne ellenőriznünk, azt azonban elvárjuk, hogy a kísérlet alanya beszámoljon nekünk a kísérlet közben végbemenő változásokról. Folyamatosan megfigyeljük a kísérlet közben, pl: folyamatosan egy Turing tesztnek tesszük ki. Moravec meg van győződve arról, hogy a kísérleti alanynak megmaradna a tudata, és így a neuronok funkcionális tulajdonságai az egyetlen lényegesek a tudat szempontjából.



Érdekes megjegyezni, hogy Searle is foglalkozik az agyprotézis kísérlettel, de ő arra a megállapításra jut, hogy a kísérlet közben eltűnne az tudat:

„*Meglepetésére ön azt fogja tapasztalni, hogy tényleg elveszti uralmát a külső viselkedése fölött. Tapasztalni fogja pl., hogy amikor a szemész a látását ellenőrzi és hallja a hangját: „Egy piros objektumot mutatok önnek, mit lát?" azt szeretné kiabálni, hogy „nem látok semmit, teljesen megvakultam!" Ehelyett azonban hallja a saját hangját, amint minden ellenőrzése alól kivonva magát azt mondja: „egy piros tárgyat látok magam előtt"… A tudatos tapasztalása lassan semmivé foszlik, miközben a külsőleg megfigyelhető viselkedése változatlan marad.* " (Searle, 1992)

Többen kritizálták Searle –t a túlzott intuitív és kellően alá nem támasztott érvelései miatt, s ebben az esetben argumentuma valóban túl intuitívnak tűnik. Mások (Stuart J. Russell) szerint olyan válaszokra számíthatunk attól, akinek az összes neuronját kicseréltük, és érdeklődünk a hogylétéről, hogy: „*Köszönöm, jól vagyok. Meg kell mondanom, meglep, de elhittem a kínai szoba érvet.*"

IX. Összefoglalás

Összefoglalva: a mesterséges intelligencia erős megközelítésének legfőbb problémája az, hogy lehetséges-e, hogy egy számítógép kognitív állapotokkal rendelkezzen, tehát tud-e egy számítógép gondolkodni. Mint láttuk, már a XVII. – XVIII. században is foglalkoztak ezzel a kérdéssel a filozófusok. Például Descartes a gépek gondolkodásával szemben, míg de la Metrie a gépek gondolkodása mellett foglalt állást, sőt azt állította, hogy az emberek is gépek. A huszadik század közepén Turing egy olyan álláspontot fogalmazott meg, amely szerint bizonyos körülmények között ésszerű azt a feltételezést megtennünk, hogy egy számítógépnek van tudata (ha egy számítógép átmegy a Turing teszten). Számítógépnek ő egy véges állapothalmazzal rendelkező szimbolikus jelfeldolgozó rendszert nevezett. Ez ellen a megközelítés ellen matematikai / logikai ellenérvek is tehetőek, mint amilyen pl: a megállási probléma vagy a Kurt Gödel féle nem-teljességi tétel. Ezeknél még erősebb ellenérvnek tűnik azonban a szimbólumok szemantikus értékét hiányoló argumentum, melyet először Searle fogalmazott meg a „kínai szoba" érvében. Searle gondolatmenete mindenfajta funkcionalista megközelítés számára kihívást jelent, így a később ismertetett konnekcionista álláspontnak is. A konnekcionalizmus a hagyományos centralizált (egyközpontú) szimbólum feldolgozáson alapuló megközelítés helyett egy elosztott / decentralizált, párhuzamos jelfeldolgozást végző rendszer felépítést javasol, ahol nincsenek szimbólumok és szabályok. Ez a megközelítés nyíltan az emberi agy idegsejt hálózatából meríti vezérlő analógiáját. A hagyományos szintaktikus szimbólumfeldolgozó rendszerek és a konnekcionizmus több keveréke is létezik, pl.: Andy Clark álláspontja. Mivel a konnekcionizmus is egy funkcionalista álláspont, így szembe kell néznie azokkal problémákkal amit – hogy Hernád István kifejezésével éljünk – a szimbólumok lehorgonyzása jelent. Searle is foglalkozik a konnekcionizmussal cikkében és a kínai szoba argumentumot kiterjeszti erre a típusú megközelítésre is. Felvethető az az álláspont, hogy talán az externalizmussal meg lehetne menteni a funkcionalista erős MI megközelítéseket. Azonban más ellenvetések is megfogalmazhatóak a konnekciozismussal kapcsolatban, pl.: ezen értelmezés szerinte egy nagyváros vízvezeték-hálózata vagy nagy mennyiségű ember együttműködése elmének tekinthető. Azonban a funcionalizmusnak is vannak ütőkártyái, mint amilyen pl.: Moravec agyprotézis kísérlete. Az agyprotézis kísérlet azt igyekszik kimutatni, hogy a Searle által is képviselt biológiai naturalizmus nem egyeztethető össze konzisztensen a funkcionalizmus tagadásával.

Az erős MI kérdése természetesen se pozitívan, se negatívan nincs eldöntve, de talán ezt nem is vártuk egy filozófia problémától, főleg nem, ha az olyan kérdéseket feszeget, mint az emberi gondolkodás és a tudatosság kérdése.

Végül csak annyit jeleznék, hogy más problémák is felmerülnek a számítógépek gondolkodásával kapcsolatban, pl.: érvelhetünk az MI ellen úgy is, hogy megkérdőjelezzük a gépek szabad akaratát, amelyet pedig az embereknél elfogadunk. Ezen kérdések tárgyalása azonban túl messzire vezetett volna (pl.: az etika területére), ezért nem tárgyaltuk őket ebben a dolgozatban.

Referenciák

1. Clark, Andy: *A megismerés építőkövei*. Osiris Kiadó, 1999. (Ford. Pléh Csaba)

2. Descartes, René: *Értekezés a módszerről*. (Ford: Szemere Samu, Boros Gábor), Ikon Kiadó, 1993.

3. Dreyfus, Hubert L.: *What Computers Can't Do: A Limits of Artificial Reason*. Hamper and Row, New York, 1979.

4. Dreyfus, Hubert L.: *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press, Cambridge, Massachusetts, 1992.

5. Farkas, Katalin; Kelemen, János: *Nyelvfilozófia*. Áron Kiadó, 2002.

6. Futó, Iván (szerk.): *Mesterséges Intelligencia*. Aula Kiadó, 1999.

7. Haykin, Simon: *Neural Networks, A Comprehensive Foundation*. 2nd edition. Prentice Hall, 1999.

8. Hernád, István: *A szimbólum lehorgonyzás problémája*. Magyar Pszichológiai Szemle, 1992 – 93, 335 –346. (Ford. Vinkler Zsuzsanna)

9. Kelemen, János: *A nyelvfilozófia rövid története*. Áron Kiadó, 2000.

10. Loewer, Barry: *A kognitív tudomány filozófiai kérdései*. (Ford: Farkas Katalin) Budapesti Könyvszemle, 2000.

11. Lucas, J. R.: *Minds, Machines and Gödel*. Philosophy, 36, 1961.

12. Moravec, Hans: *Mind Childeren: The Future of Robot and Human Intelligence*. Harvard University Press, Cambridge, Massachustetts, 1988.

13. Papadimitriu, Christos H.: *Computational Complexity*. Addison Wesley Publishing Company, 1994. Magyarul: *Számítási bonyolultság*. Novadat Kiadó, 1999.

14. Penrose, Roger: *The Emperor's New Mind*, 1990. Magyarul: *A császár új elméje*. Akadémiai Kiadó - Budapest, 1993.

15. Pinker, Steven: *The Language Instict*, 1994. (Ford. Bócz András) Magyarul: *A nyelvi ösztön*. Typotex Kiadó, 1999.

16. Platón: *A szofista*. (Ford: Kövendi Dénes), Platón Összes Művei. II. Gondolat Kiadó, 1989.

17. Pléh, Csaba (szerk.): *Agy és tudat*. BIP Kiadó, Budapest, 2002.

18. Pléh, Csaba (szerk.): *A megismerés-kutatás útjai*. Akadémia Kiadó, 2000.

19. Pléh, Csaba: *A modern kognitívizmus mozgalma és változásai*. Megtalálható: Kognitív tudomány, Osiris Kiadó és Láthatatlan Kollégium, Budapest, 1996.

20. Pléh, Csaba (szerk.): *Megismerés-tudomány és mesterséges intelligencia*. Akadémia Kiadó, 1998.

21. Russel, Stuar J.; Norvig, Peter: *Artificial Intelligence. A Modern Approach*. Prentice Hall, 1995. Magyarul: *Mesterséges Intelligencia*. *Modern megközelítésben*. Pantem Kiadó, 2000.

22. Schacter, Daniel L.: *Searching for Memory*, Basic Books, 1996. (Ford. Dankó Zoltán) Magyarul: *Emlékeink Nyomában.* Háttér Kiadó, Budapest, 1998.

23. Searle, John. R.: *Minds, Brains and Programs*. Behavioral and Brain Science, 1980, 3, 417 - 424. Magyarul: *Az elme, az agy és a programok világa*. Kognitív tudomány, Osiris Kiadó és Láthatatlan Kollégium, Budapest, 1996. (Ford: Thuma Orsolya)

24. Smullyan, Raymond: *Gödel nemteljességi tételei*. (Ford. Csaba Ferenc), Typotex kiadó, 1999.

25. Turing, Alan Matheson: *Computing Machinery and Intelligence*. Mind 59, no. 236 (1950), pp. 4 – 30.