SEARLE, John R.: Az elme, az agy és a programok világa

in.:

KOGNITÍV TUDOMÁNY

– p. 136–151.

Osiris Kiadó
Láthatatlan Kollégium

Budapest
1996

(Szemeszter)

JOHN R. SEARLE

AZ ELME, AZ AGY ÉS A PROGRAMOK VILÁGA*

Milyen pszichológiai és filozófiai jelentősége lehet azoknak a legújabb próbálkozásoknak, amelyek az emberi megismerőképességek számítógépes szimulációjával foglalkoznak? A kérdés megválaszolásakor hasznos, ha megkülönböztetjük a MI (Mesterséges Intelligencia) úgynevezett „erős" verzióját a MI „gyenge" vagy „óvatos" verziójától. A MI gyenge verziója szerint a számítógépnek az elme tanulmányozásában az a legfőbb értéke, hogy nagyon hatékony eszköz. Például lehetővé teszi a hipotézisek jóval pontosabb megfogalmazását és tesztelését. Ezzel szemben az MI erős verziója a számítógépet nem pusztán eszköznek tartja az elme tanulmányozásában, hanem a megfelelően programozott számítógépet valóban elménck tekinti abban az értelemben, hogy a megfelelő programok birtokában a számítógép szó szerint *megért* és egyéb kognitív állapotokkal rendelkezik. Mivel a MI erős verziója szerint a beprogramozott számítógépnek kognitív állapotai vannak, a programok nem pusztán olyan eszközök, amelyek a pszichológiai magyarázatok tesztelését teszik lehetővé, hanem sokkal inkább a programok maguk válnak a magyarázattá.

A MI gyenge verziójának állításaival szemben nincs ellenvetésem, legalábbis ami ezt a cikket illeti. A jelen cikk azoknak az állításoknak a megvitatására irányul, amelyeket a MI erős verziójának tulajdonítok, különös tekintettel arra az állításra, amely szerint a megfelelően beprogramozott számítógép rendelkezik szó szerinti értelemben vett kognitív állapotokkal, és amely szerint ezáltal a program képes az emberi kogníció magyarázatát adni. Mostantól, amikor a MI-ra hivatkozom, az előbbi két állítással jellemzett erős verzióra gondolok.

Roger Schank és a Yale Egyetemen dolgozó kollégái (Schank és Abelson, 1977) munkáját fogom áttekinteni, mivel az ő tevékenységüket jobban ismerem, mint más hasonló munkákat, és mivel nagyon világos példáját adják annak a fajta munkának, amelyet vizsgálni szeretnék. A további elemzés azonban független Schank programjának részleteitől. Ugyanezen érvek vonatkoznak Winograd SHRDLU (Winograd, 1972) és Weizenbaum ELIZA (Weizenbaum, 1965) nevű programjára, és valójában az emberi mentális jelenségek minden Turing-gépen történő szimulálására.

*Minds, brains, and programs. *Behavioral and Brain Sciences*, 1980, *3*, 417–424. Fordította: Thuma Orsolya.

Nagyon röviden és számos részlet mellőzésével, Schank programja a következőképpen írható le: a program célja az ember történetmegértő képességének a szimulációja. Az emberek történetmegértő képességének jellemző vonása az, hogy a történetre vonatkozó kérdéseket akkor is meg tudja válaszolni, ha az arra vonatkozó információ expliciten nem szerepelt egyáltalán a történetben. Így például, ha a következő történetet kapjuk: „Egy férfi bement egy étterembe és egy hamburgert rendelt. Mikor a hamburger megérkezett, keményre volt égetve, erre a férfi dühösen kiviharzott az étteremből anélkül, hogy a hamburgert kifizette volna vagy borravalót hagyott volna." Ha most azt kérdeznék: „A férfi megette a hamburgert?", valószínűleg azt válaszolnánk „Nem, nem ette meg." Hasonlóképpen, ha a következő történettel találkoznánk: „egy férfi bement egy étterembe és rendelt egy hamburgert, mikor a hamburger megérkezett, meg volt elégedve vele, és az étteremből kifelé menet a felszolgálónőnek nagy borravalót adott, mielőtt kifizette a számláját" majd ezután azt kérdeznék tőlünk: „megette a férfi a hamburgert?," valószínűleg azt válaszolnánk: „Igen, megette a hamburgert." Schank gépei ugyanígy meg tudnak válaszolni éttermekre vonatkozó hasonló kérdéseket. Ehhez rendelkezniük kell az emberek éttermekre vonatkozó ismereteihez hasonló információk „reprezentációjával", amely lehetővé teszi, hogy a fentiekhez hasonló kérdésekre válaszoljanak, amennyiben ilyen típusú történeteket kapnak. Amikor a gép megkapja a történetet és aztán a kérdést, olyan válaszokat nyomtat ki, amelyeket olyan emberektől várnánk, akik hasonló történeteket hallgattak. Az MI erős verziójának úttörői azt állítják, hogy a kérdések és válaszok ilyen sorozatakor a gép nemcsak szimulál egy emberi képességet, hanem

1. a szó szoros értelmében *megérti* a történetet és válaszol a kérdésre, továbbá
2. amit a gép és a programja csinál, az *megmagyarázza* azt az emberi képességet, amelynek a révén történeteket értünk meg és válaszolunk a rá vonatkozó kérdésekre.

Számomra, ahogy a következőkben megpróbálom bemutatni, egyik állítást sem támasztja teljes mértékben alá Schank munkája.

Az elmére vonatkozó elméletek tesztelésének egyik módja az, hogy megkérdezzük magunktól, milyen lenne, ha az elménk valóban azoknak az elveknek megfelelően működne, amelyeknek megfelelően a teória szerint minden elme működik. Alkalmazzuk ezt a tesztet a Schank-programra a következő gondolatkísérlet segítségével. Tételezzük fel, hogy be vagyok zárva egy szobába, és kapok egy nagy halom kínai írást. Továbbá tételezzük fel (ahogy valójában ez is a helyzet), hogy nem tudok kínaiul sem írni, sem olvasni, és hogy még abban sem vagyok bizonyos, hogy felismerem a kínai írást, mint olyat, ami eltér mondjuk a japán írástól vagy az értelmetlen krikszkrakszoktól. Számomra a kínai írás csak egy csomó értelmetlen krikszkraksz. Tételezzük fel továbbá, hogy az első halom kínai írást követően egy második halom kínai írást kapok egy olyan szabálykészlettel együtt, amely összekapcsolja az első halmot a második halommal. A szabályok angolul vannak és éppúgy megértem azokat, ahogy bármely angol anyanyelvű személy. A szabályok lehetővé teszik, hogy a formális szimbólumok egyik készletét összekapcsoljam a formális szimbólumok egy másik készletével. Amikor azt mondom, „formális", ezen azt értem, hogy a szimbólumokat kizárólag a formájuk alapján tudom azonosítani. Tételezzük most fel azt is, hogy kapok egy harmadik halom kínai írást instrukciókkal, szintén angolul, melyeknek révén össze tudom kapcsolni e harmadik halom elemeit az első két halommal, és ezek a szabályok arról is eligazítanak, hogyan adjak vissza bizonyos kínai

szimbólumokat bizonyos formában válaszként bizonyos formákra, amelyeket a harmadik halommal kaptam. Nincs tudomásom arról, hogy azok, akiktől a szimbólumokat kapom, az első halmot „forgatókönyvnek", a második halmot „történetnek", a harmadik halmot pedig „kérdéseknek" hívják. Továbbá arról sem tudok, hogy a szimbólumokat, amelyeket a harmadik halomra adok válaszként, „a kérdésekre adott válasznak" hívják, az angol nyelvű szabály-készletet pedig „programnak". Hogy a történetet egy kicsit tovább bonyolítsuk, képzeljük most el, hogy ezek az emberek angol nyelvű történeteket is adnak nekem, amelyeket megértek, és azután angolul kérdeznek engem a történetekről, és én angolul válaszolok rájuk. Tételezzük fel azt is, hogy egy idő után a kínai szimbólumkezelő utasításokat olyan jól követem és a programozók olyan jól írják a programokat, hogy a külső néző szempontjából – vagyis olyan személy szempontjából, aki a Kínai Szobán kívül tartózkodik – a kérdésekre adott válaszaim tökéletesen megkülönböztethetetlenek a kínai anyanyelvűek válaszaitól. Pusztán a válaszaimat tekintve senki sem tudja megállapítani, hogy nem beszélek kínaiul. Azt is tételezzük fel, hogy a válaszaim az angol kérdésekre, ahogy ez kétségtelenül így is van, megkülönböztethetetlenek más angol anyanyelvűek válaszaitól azon egyszerű oknál fogva, hogy én is angol anyanyelvű vagyok. Külső nézőpontból – annak a nézőpontjából, aki a válaszaimat olvassa – a kínai és az angol kérdésekre adott válaszok egyformán jók. De a kínai esetében, szemben az angollal, a válaszokat értelmezetlen, formális szimbólumokkal végzett műveletek eredményezik. Ami a kínait illeti, egyszerűen úgy viselkedem, mint egy számítógép: komputációs műveleteket végzek formálisan meghatározott elemeken. A kínai feladatokban egyszerűen a számítógépes program egy megvalósulása vagyok.

Tehát a MI erős verziójának hívei azt állítják, hogy a beprogramozott számítógép a történeteket megérti, és hogy a program bizonyos értelemben megmagyarázza az ember megértő folyamatait. Abban a helyzetben vagyunk, hogy gondolatkísérletünk fényében vizsgálhatjuk meg ezeket az állításokat.

1. Ami az első állítást illeti, számomra nyilvánvalónak tűnik a példából, hogy egy szót sem értek a kínai történetekből. Bár a bemenetem és a kimenetem megkülönböztethetetlen egy kínai anyanyelvűétől, és bármilyen tetszés szerinti formális programmal rendelkezhetem, mégsem értek belőle semmit. Ugyanezen okból Schank számítógépe sem ért semmit a történetekből sem kínaiul, sem angolul, sem más nyelven, mivel a kínai esetében én vagyok a számítógép, azokban az esetekben pedig, amikor a számítógép nem én vagyok, a gép semmivel sincs jobb helyzetben, mint amikor én nem értek egy szót sem.

2. Ami azt a második állítást illeti, hogy a program az emberi megértést magyarázza, azt látjuk, hogy a számítógép és programja nem adja elégséges feltételét a megértésnek, mivel a számítógép és a program minden megértés nélkül egyszerűen csak működik. Vajon egyáltalán szükséges feltétele a program a megértésnek, vagy jelentősen hozzájárul-e ahhoz? A MI erős verziójának egyik állítása az, hogy amikor pontosan megértek egy angol nyelvű történetet, ugyanazt teszem – vagy talán ugyanabból többet –, mint amikor a kínai szimbólumokkal tényekedem. Az angolt, amit megértek, egyszerűen a formális szimbólumkezelés többlete különbözteti meg a kínaitól, amelyet nem értek. Nem bizonyítottam be, hogy ez az állítás hibás, de a példa alapján bizonyára kevéssé hihető. Egy ilyen állítás elfogadása azon a feltevésen alapszik, hogy lehet olyan programot készíteni, amely pontosan olyan bemenettel és kimenettel szolgál, mint az anyanyelvűek, feltételezve

továbbá azt is, hogy ezek a személyek a leírás egy olyan szintjével rendelkeznek, amelyen ők a program megvalósulásai. E két feltevés alapján következtetünk arra, hogy még ha Schank programja nem mond is el mindent a megértésről, legalábbis egy része a történetnek. Empirikusan mindez lehetséges, de egyelőre a legkisebb okunk sincs arra, hogy igaznak higgyük, mivel, ahogy a példa utalt rá – bár csak implicite – a számítógépes program egyszerűen nem releváns az én történetmegértésem szempontjából. A kínai esetében rendelkezem mindennel, amivel csak a mesterséges intelligencia szolgálhat a program révén, mégsem értek semmit; az angol esetében mindent megértek, pedig egyelőre egyáltalán semmi okom nincs azt feltételezni, hogy megértésemnek bármi köze lenne számítógépes programokhoz, vagyis tisztán formálisan meghatározott elemeken végzett komputációs műveletekhez. Ha a programot úgy definiáljuk, mint formális elemeken végzett komputációs műveleteket, a példa azt mutatja, hogy ezek önmagukban nincsenek érdemleges kapcsolatban a megértéssel. Biztosan nem elégséges feltételei annak, és semmi okunk feltételezni, hogy szükséges feltételei lennének vagy egyáltalán hogy jelentősen hozzájárulnának a megértéshez. Vegyük észre, hogy az érv ereje nem pusztán abból ered, hogy a különböző gépek ugyanarra a bemenetre ugyanazt a kimenetet adhatják, miközben eltérő formális elvek szerint működnek – egyáltalán nem erről van szó. Sokkal inkább arról, hogy bármiféle pusztán formális elveket táplálunk is a számítógépbe, azok nem lesznek elégségesek a megértéshez, mivel az ember mindenfajta megértés nélkül is képes a formális elvek követésére. Nincsen semmi okunk azt feltételezni, hogy efféle elvek szükségesek lennének vagy egyáltalán hozzájárulnának a folyamathoz, mivel azt sincs okom feltételezni, hogy amikor angolul megértek valamit, bármifajta formális program szerint működnék.

Mi az tehát, amivel rendelkezem az angol mondatok esetében, de a kínai mondatoknál nem? A nyilvánvaló válasz az, hogy az előbbiről tudom, mit jelent, míg fogalmam sincs, hogy mit jelent az utóbbi. Mégis miben áll és miért nem adható át a gépnek, bármi legyen is az? Később vissza fogok térni erre a kérdésre, de előbb folytassuk a példát. Ezt a példát volt alkalmam megmutatni több mesterségesintelligencia-kutatónak és érdekes módon nincs egyetértés abban, hogy mi lenne a megfelelő válasz rá. Meglepően változatos válaszokat kapok, és az elkövetkezőkben ezek közül fogom a legáltalánosabbakat megvitatni (földrajzi eredetükkel egyetemben).

Először azonban szeretnék tisztázni néhány a „megértéssel" kapcsolatos gyakori félreértést: sok vitában a „megértés" szó pompás körülírásait lehet megtalálni. Kritikusaim azt emelik ki, hogy a megértésnek számos különböző fokozata van; hogy a „megértés" nem egy egyszerű két argumentumú predikátum; hogy a megértésnek különböző fajtái és szintjei vannak, és a kizárt harmadik elve gyakran nem alkalmazható közvetlenül az „x megérti y-t" formájú állításokra; hogy sok esetben elhatározás kérdése és nem ténykérdés, hogy „x megérti y-t" és így tovább. Mindezekre a megállapításokra a válaszom: persze, persze; de ezeknek semmi közük az itt tárgyalt érvekhez. Vannak világos esetek, ahol a „megértés" szó szerint értendő és világos esetek, ahol nem; csupán erre a két esetre van szükségem az érvelésemhez. Angol történeteket megértek, francia történeteket kevésbé értek, német történeteket még kevésbé, és kínaiakat egyáltalán nem. Másfelől az autóm és a számológépem semmit sem ért: nem ez a dolguk. „Megértést" és más kognitív predikátumokat gyakran tulajdonítunk metaforikusan vagy analógiásan autóknak, számológépeknek és egyéb mesterséges tárgyaknak, de az ilyenfajta tulajdonítások semmit sem bizonyítanak. Ilyesmiket mondunk, hogy „az ajtó tudja, mikor kell kinyílnia, mert fotocel-

lás", „a számológép tudja, hogyan kell (megérti, hogyan kell, képes rá) összeadni és kivonni, de nem tud osztani" és „a termosztát észleli a hőmérséklet változását". Az effajta tulajdonítások oka meglehetősen érdekes és azzal van kapcsolatban, hogy saját szándékrendszerünket kiterjesztjük a mesterséges tárgyakra; eszközeink a céljaink kiterjesztései, és így természetesnek találjuk, ha metaforikusan szándékot tulajdonítunk nekik, de gondolom, effajta példák nem okoznak földindulást a filozófiában. Egy automatikusan működő ajtó egyáltalán nem abban az értelemben „érti" meg a fotocella „utasításait", ahogyan én értek angolul. Ha a Schank-féle programmal ellátott számítógép történetmegértését metaforikus értelemben vennénk, úgy, ahogy az ajtó ért, és nem úgy, ahogy én értek angolul, a témát nem lenne értelme megtárgyalni. Newell és Simon (1963) azonban azt írja, hogy az a fajta kogníció, amelyet a számítógépnek tulajdonítanak, pontosan olyan, mint az emberi kogníció. Szeretem ezt az állítást az egyenességéért, és épp az ilyenfajta állításokat fogom megtárgyalni. Amellett fogok érvelni, hogy szó szerinti értelemben a beprogramozott számítógép pontosan annyit ért, mint az autó és a számológép, vagyis egyáltalán semmit. A számítógép nem részlegesen vagy hiányosan ért meg (ahogy én értek németül), hanem egyáltalán nem képes megérteni.

Most következzenek a válaszok:

I) A rendszerelvű válasz (Berkeley). „Miközben igaz, hogy a szobába zárt egyetlen személy nem érti meg a történetet, valójában a helyzet az, hogy ő csak része egy teljes rendszernek, és a rendszer az, amely a történetet megérti. A személy előtt egy nagy könyv van, benne a szabályokkal, van sok papírfecnije és ceruzái a számítások elvégzéséhez, »adatbankjai« vannak a kínai szimbólumrendszerről. A megértés nem magának a személynek tudható be, hanem inkább az egész rendszernek, amelynek része."

A válaszom a rendszerelvű elméletre meglehetősen egyszerű: a személy tegye belsővé a rendszer minden elemét. Memorizálja a könyvből a szabályokat és a kínai szimbólumok adatbankjait, továbbá minden számítást fejben végezzen el. Az egyén így megtestesíti a teljes rendszert. Semmi olyan nincs a rendszerben, amely a személyben ne lenne meg. Még a szobát is elhagyhatjuk, és feltételezhetjük, hogy a szabadban dolgozik. Ugyanúgy semmit sem ért az emberünk kínaiul, ebből következően pedig a rendszer sem, mert semmi sincs a rendszerben, ami ne lenne a személyben benne. Ha ő nem érti meg, akkor a rendszer sem értheti meg semmi módon, mert a rendszer csak egy része a személynek.

Valójában egy kissé zavarban vagyok, amikor most a rendszerelvű elméletre válaszolok, mivel az elmélet annyira valószínűtlennek tűnik a számomra. Az elképzelés szerint noha a személy nem ért kínaiul, valahogyan a személy és a papírdarabkák *összeköttetése* érthet kínaiul. Nehéz elképzelnem, hogy bárki, aki nem foglya ennek az ideológiának, egyáltalán elképzelhetőnek tekintsen ilyesmit. A MI erős ideológiájának sok elkötelezettje végül mégis hajlani fog arra, hogy valami ilyesmit mondjon, úgyhogy folytassuk egy kicsit tovább a kifejtést. Az elmélet egyik verziója szerint a rendszert belsővé tevő példában a személy ugyan nem ért kínaiul olyan értelemben, ahogy a kínai anyanyelvűek értenek (mert például nem tudják, hogy a történet éttermekről és hamburgerekről szól stb.), mégis „a személy, mint formális szimbólumkezelő rendszer" *valóban ért kínaiul.* A személy kínai nyelvű alrendszerét, vagyis formális szimbólumkezelő rendszerét ne tévesszük össze az angol nyelvű alrendszerével.

Így valójában két alrendszer van a személyben: az egyik ért angolul, a másik ért kínaiul, és „csak arról van szó, hogy a két rendszernek kevés köze van egymáshoz". Erre azt kell válaszolnom, hogy nemcsak kevés közük van egymáshoz, hanem mégcsak távolról sem hasonlítanak egymásra. Az az alrendszer, amely angolul ért (az „alrendszerek" zsargonjában beszélve), tudja, hogy a történetek éttermekről és hamburgerevésről szólnak; tudja, hogy az éttermekről kérdezik, és hogy legjobb tudása szerint válaszol, különböző következtetéseket vonva le a történet tartalmából és így tovább. A kínai rendszer azonban egyiket sem tudja ezek közül. Míg az angol alrendszer tudja, hogy a „hamburger" a hamburgerre vonatkozik, addig a kínai alrendszer csak azt tudja, hogy „kriksz, kriksz" után „kraksz, kraksz" következik. Mindössze annyit tud, hogy az egyik végén formális szimbólumok jönnek be és az angol nyelvű szabályoknak megfelelően kezelődnek, majd a másik végén más szimbólumok távoznak. Az egész eredeti példa lényege az az érv, hogy az ilyen szimbólumkezelés önmagában semmilyen szó szerinti értelemben nem elégséges a kínai megértéséhez, mert az ember leírhatja azt, hogy „kraksz, kraksz" azután, hogy „kriksz, kriksz" anélkül, hogy bármit is értene kínaiul. Az sem elegendő az érveléshez, ha a személyen belüli alrendszereket posztulálunk, mert az alrendszerek sem jutnak tovább, mint az előzőekben a személy; nincs bennük semmi még távolról hasonló sem ahhoz, amivel az angol nyelvű személyek (vagy alrendszerek) rendelkeznek. Valójában az előbb leírt esetben a kínai alrendszer egyszerűen csak része az angol alrendszernek, egy olyan része, amely értelmetlen szimbólumok kezelését végzi az angol nyelvű szabályoknak megfelelően.

Tegyük fel magunknak a kérdést, vajon mi motiválja elsősorban a rendszerelvű választ, vagyis mi lenne az a *független* alap, amely megköveteli, hogy az ágens alrendszerekkel rendelkezzen, amelyek szó szerint megértik a kínai történeteket? Amennyire meg tudom állapítani, az egyetlen alap az, hogy a példában a bemenet és a kimenet pontosan megegyezik az anyanyelvű kínaiakéval, és a program az egyiktől halad a másikig. A példák lényege azonban az volt, hogy megpróbáljam bemutatni, mennyire nem elegendő ez a megértéshez, abban az értelemben, ahogy a történeteket angolul értem meg, mert ugyan a személy, és ezáltal a személyt felépítő rendszerek összességének rendelkezésére áll a megfelelő bemenet, kimenet és program, mégsem ért meg semmit abban a releváns szó szerinti értelemben, ahogy én értem meg az angolt. Az egyetlen, ami arra a kijelentésre motiválhat, hogy *kell* legyen bennem egy olyan alrendszer, amely megérti a kínait, az, ha a program és én átmegyünk a Turing-teszten, ha meg tudok téveszteni kínai anyanyelvűeket. Az egyik kérdéses pont éppen a Turing-teszt alkalmassága. A példa azt mutatja, hogy lehet két olyan „rendszer", amely átmegy a Turing-teszten, de megérteni csak az egyik képes; és ezzel szemben nem érv az, hogy mivel mindkét rendszer átment a Turing-teszten, mindkettőnél fel *kell* tételeznünk a megértést, mivel ez az állítás nem veszi figyelembe azt az érvet, hogy az a rendszer bennem, amely angolul ért, sokkal többel rendelkezik, mint az a rendszer, amely pusztán feldolgozza a kínait. Röviden tehát a rendszerelvű válasz egyszerűen megkerüli a kérdést, amikor érvelés nélkül azt hangsúlyozza, hogy a rendszernek meg kell értenie a kínait.

Ezenfelül, úgy tűnik, a rendszerelvű válasz önmagában is abszurd következtetésekre vezet. Ha azt a következtetést, hogy kell hogy legyen bennem kogníció, azon az alapon vonjuk le, hogy bizonyos fajta bemenettel, kimenettel és egy köztük levő programmal rendelkezem, akkor úgy tűnik, mindenféle nem kognitív alrendszer is kognitívvé válik. A leírás egy szintjén például a gyomrom információt dolgoz fel, és akárhány számítógépes

program megvalósulása lehet, de feltehetőleg nem akarnánk azt állítani, hogy megértés jellemzi (vö. Pylyshyn, 1980). Ha azonban elfogadjuk a rendszerelvű választ, akkor is nehéz belátni, hogyan kerülhetnénk el, hogy a gyomrot, a szívet, a májat és egyebeket mind megértő rendszereknek nevezzük, mivel elvi alapon nem lehet megkülönböztetni azt, ami arra indíthat, hogy a kínai alrendszert megértőnek nevezzük attól, aminek alapján a gyomrot megértő rendszernek tekinthetjük. Mellesleg erre az érvre az sem válasz, hogy a bemenet és a kimenet a kínai rendszer számára az információ, míg a gyomor számára az étel és az élelmiszerek, mivel az ágens, vagyis az én nézőpontomból nincs információ sem az ételben, sem a kínai jelekben – a kínai pusztán egy csomó értelmetlen krikszkraksz. A kínai jelek esetében az információ kizárólag a programozó és az értelmező szemében létezik, és semmi nem tarthatja vissza őket attól, hogy az emésztőszerveim bemenetét és kimenetét információként kezeljék, ha úgy akarják.

Ez az utóbbi érv felveti a MI erős verziójának néhány ettől független problémáját, amelyeknél érdemes elidőznünk egy pillanatra, hogy világosabban lássuk őket. Ha a MI erős verziója a pszichológia egy ága szándékozik lenni, akkor meg kell tudnia különböztetni azokat a rendszereket, amelyek valóban mentálisak, azoktól, amelyek nem azok. Képesnek kell lennie arra, hogy megkülönböztesse azokat az elveket, amelyek szerint az elme működik, azoktól az elvektől, amelyek a nem mentális rendszereket működtetik, különben nem tudja majd megmagyarázni, hogy mi a speciálisan elmebeli az elmében. A mentális – nem mentális megkülönböztetése nem függhet pusztán a szemlélő látásmódjától, hanem a rendszer intrinzikus része kell legyen, különben csak a szemlélőn múlna, hogy az embereket nem mentálisnak, a hurrikánokat viszont mentálisnak tekinti-e, ha úgy kívánja. A MI-irodalomban mégis meglehetősen gyakran elmosódik ez a megkülönböztetés, olyannyira, hogy hosszú távon végzetesnek bizonyulhat arra az állításra nézve, hogy a MI kognitív kutatási terület. McCarthy például azt írja, hogy „olyan egyszerű gépekről, mint a termosztát, elmondható, hogy hiedelmeik vannak, és a hiedelmek megléte jellemző a legtöbb problémamegoldásra képes gépre" (McCarthy, 1979). Mindenkinek, aki esélyesnek tartja a MI erős verzióját egy elme-teóriaként, el kell gondolkoznia azon, hogy mit is implikál ez a kijelentés. Arra szólít fel, hogy a MI erős verziójának felfedezéseként fogadjuk el, hogy a fémdarab a falon, amelyet a hőmérséklet szabályozására használunk, ugyanolyan értelemben rendelkezik hiedelmekkel, mint ahogy magunknak, házastársunknak vagy gyermekeinknek vannak hiedelmei; továbbá, hogy a „legtöbb" egyéb gép a szobában – a telefon, a magnó, a számológép, az elektromos villanykapcsoló – szintén rendelkezik hiedelmekkel ebben a szó szerinti értelemben. A jelen cikknek nem célja, hogy érveket hozzon fel McCarthy megállapításával szemben, úgyhogy a következőkben minden érvelés nélkül egyszerűen kijelentéseket teszek. Az elme tanulmányozása annak a ténynek a megállapításával kezdődik, hogy míg az emberek rendelkeznek hiedelmekkel, addig a termosztátok, a telefonok és a számológépek nem. Ha olyan elmélethez jutunk, amely tagadja ezt a megállapítást, akkor sikerült felmutatnunk az elmélet ellenpéldáját, és az elmélet hamis. Az embernek az a benyomása, hogy azok a MI-hívők, akik ilyesmiket írnak, úgy vélik, hogy ezt megengedhetik maguknak, mivel nem veszik igazán komolyan és nem hiszik, hogy bárki is komolyan fogja azt venni. Azt javaslom, hogy legalább egy pillanatra vegyük komolyan. Egy percre gondoljunk csak bele jobban abba, mire lenne szükség annak megalapozásához, hogy az a fémdarab a falon valódi hiedelmekkel rendelkezik; olyan vélekedésekkel,

melyek irányultan illeszkednek, propozicionális tartalmuk van és kielégülési feltételekkel rendelkeznek; hiedelmekkel, melyek lehetnek erősek vagy gyengék; ideges, aggódó vagy biztos hiedelmekkel; dogmatikus, racionális vagy babonás hiedelmekkel; vak hittel vagy hezitáló gondolatokkal; bármifajta hiedelemmel. A termosztát erre nem pályázhat. A gyomor, a máj, a számológép és a telefon sem. Ugyanakkor, mivel az elképzelést komolyan vesszük, azt is vegyük észre, hogy ha igaz lenne, akkor arra az állítására nézve lenne végzetes, hogy a MI erős verziója elmetudomány, mivel így az elme mindenhol ott lenne. Amit tudni akartunk, éppen az volt, hogy mi különbözteti meg az elmét a termosztáttól és a májtól. Ha McCarthynak igaza lenne, a MI erős verziójának semmi reménye nem lenne arra, hogy ezt megválaszolja.

II. A robot-válasz (Yale). „Tételezzük fel, hogy Schank programjától eltérő programot írunk. Tételezzük fel, hogy egy számítógépet teszünk a robot belsejébe, és ez a számítógép nem csak formális szimbólumokat venne be bemenetként, és formális szimbólumot adna ki kimenetként, hanem ténylegesen oly módon működtetné a robotot, hogy amit a robot csinálna, az nagyon hasonló lenne az érzékeléshez, járáshoz, mozgáshoz, szögbeveréshez, evéshez, iváshoz – bármihez, amit csak akarunk. A robothoz például egy kamera lenne illesztve, amelynek a segítségével »látna«, lenne karja és lába, amellyel »cselekedhetne«, és mindezt a számítógépes »agya« irányítaná. Egy ilyen robot, szemben Schank számítógépével, rendelkezne valódi megértéssel és más mentális állapotokkal."

A robot-válasszal kapcsolatos első észrevételem az, hogy finoman elismeri, hogy a kogníció nem csupán formális szimbólumok kezelésének kérdése, mivel ez a válasz a külső világgal való oki kapcsolatok készletét hozza be (vö. Fodor, 1980). Mégis, a robotválaszra az a válaszom, hogy a „perceptuális" és „motoros" készségek ilyenfajta hozzáadása nem ad hozzá semmit sajátlagosan a megértés, illetve általánosságban az intencionalitás szempontjából Schank eredeti programjához. Hogy ezt belássuk, vegyük észre, hogy ugyanaz a gondolatkísérlet alkalmazható a robot esetére is. Tételezzük fel, hogy ahelyett, hogy számítógépet raknának a robotba, engem raknak be egy szobába, mint az eredeti Kínai Szoba esetében, további kínai szimbólumokat kapok, további angol nyelvű instrukciókkal, hogy hogyan feleltessek meg kínai szimbólumokat kínai szimbólumoknak, és hogyan adjak kínai visszajelzést a külvilág számára. Tételezzük fel, hogy a tudtom nélkül, néhány kínai szimbólum, amit kapok, a robothoz erősített kamerán keresztül érkezik hozzám, más kínai szimbólumok pedig, amelyeket én adok ki, arra szolgálnak, hogy meghajtsák a robotban levő motort, amely a robot karjait és a lábait mozgatja. Fontos hangsúlyoznom, hogy magam pusztán formális szimbólumokat kezelek: nem tudok semmit ezekről a további tényezőkről. „Információt" kapok a robot „perceptuális" berendezésétől, és „instrukciókat" adok motoros berendezésének anélkül, hogy ezek bármelyikéről is tudnék. A robot homunculusa vagyok, de a hagyományostól eltérő homunculus. Nem tudom, mi történik. Nem értek semmit, leszámítva a szimbólumkezelés szabályait. Ebben az esetben azt állítom, hogy a robot egyáltalán nem rendelkezik intencionális állapotokkal, egyszerűen csak mozgásokat végez az elektromos vezetékek és a program eredményeképpen. Továbbá a program megvalósulásaként én sem rendelkezem a releváns értelemben használt intencionális állapotokkal. Csupán annyit teszek, hogy formális szimbólumokat kezelek formális instrukciókat követve.

III. Az agyszimulátor-válasz (Berkeley és MIT). „Tételezzük fel, hogy olyan programot tervezünk, amely nem reprezentálja a világról meglévő információinkat, ahogyan Schank forgatókönyvei teszik, hanem a neuronkisülések tényleges sorozatát szimulálja a szinapszisoknál, ahogyan az az anyanyelvű kínai agyában zajlik, amikor történeteket ért meg kínaiul és válaszol azokra. A gép kínai történeteket és rájuk vonatkozó kérdéseket vesz fel bemenetként, a történet feldolgozásában a tényleges kínai agy formális struktúráját szimulálja, és kimenetként kínai válaszokat ad. Azt is elképzelhetjük, hogy a gép nem egyszerű szeriális programmal működik, hanem párhuzamosan működő programok egy egész készletével, ahogyan feltételezhezően a tényleges emberi agy működik, amikor természetes nyelven dolgoz fel. Nos ebben az esetben bizonyára azt kellene mondanunk, hogy a gép megértette a történeteket, és amennyiben ezt tagadjuk, vajon nem kellene-e azt is tagadnunk, hogy az anyanyelvű kínaiak megértették a történeteket? A szinapszisok szintjén miben különbözik vagy különbözhetne a számítógép programja és a kínai agy programja?"

Mielőtt szembeszállnék ezzel a válasszal, szeretnék visszatérni egy megjegyzés erejéig arra, hogy furcsa válasz ez a mesterséges intelligencia (vagy funkcionalizmus stb.) képviselőitől: úgy gondoltam, hogy a MI erős verziójának alapeszméje az, hogy nincs szükségünk arra, hogy tudjuk, hogyan működik az agy, ahhoz, hogy tudjuk, hogyan működik az elme. A kiinduló hipotézis, legalábbis én úgy gondoltam, az, hogy a formális elemeken végzett komputációs folyamatokból álló mentális működéseknek van egy olyan szintje, amely az elme lényegét alkotja, és amely a legkülönbözőbb agyi folyamatban valósulhat meg, ugyanúgy, ahogy bármely számítógépes program végrehajtható különböző számítógép hardware-eken. A MI erős verziójának feltételezése szerint az elme olyan az agynak, mint a program a hardware-nek, és ezért érthetjük meg az elmét anélkül, hogy neurofiziológiával foglalkoznánk. Ha tudnunk kellene, hogyan működik az agy, ahhoz, hogy MI-val foglalkozhassunk, nem érdekelne minket a MI. Ugyanakkor még ha megközelítenénk is az agy működését, az sem lenne elég a megértés létrehozásához. Ahhoz, hogy ezt belássuk, képzeljük el, hogy a szobában az egynyelvű, szimbólumkártyákat kevergető ember helyett egy olyan ember van, aki szelepekkel összekötött vízvezetékcsövek finom rendszerét működteti. Amikor megkapja a kínai szimbólumokat, az angol nyelvű programból kinézi, melyik szelepet kell be-, illetve kikapcsolnia. Minden vízvezeték-kapcsolat a kínai agy egy szinapszisának felel meg, és az egész rendszer úgy van megszerelve, hogy a megfelelő tüzelések után, vagyis a megfelelő csapok kinyitása után a csőrendszer kimeneti végén kínai nyelvű válaszok hullanak ki.

Vajon hol van a megértés ebben a rendszerben? Kínai nyelvű bemenetet vesz be, a kínai agy szinapszisainak formális rendszerét szimulálja, és kínai nyelvű kimenetet ad ki. Emberünk azonban természetesen nem ért kínaiul, és a vízvezeték sem, továbbá ha hajlandóak vagyunk elfogadni azt a szerintem abszurd elképzelést, hogy valami módon az ember *és* a vízvezeték *összeköttetése* ért meg, emlékezzünk arra, hogy emberünk elvileg belsővé teheti a vízvezetékek formális rendszerét, és gondolatban minden „neurontüzelést" elvégezhet. Az agy-szimulátorral az a probléma, hogy az aggyal kapcsolatban nem azt szimulálja, amit kellene. Amíg csak a neurontüzelések sorozatának formális struktúráját szimulálja, addig nem szimulálja azt, ami az aggyal kapcsolatban számít, vagyis az oki sajátságait, azt a képességét, hogy intencionális állapotokat hoz létre. A vízvezetékes példa azt mutatja, hogy a formális sajátságok nem elégségesek oki sajátságokként: minden formális sajátságot kifaraghatunk a releváns neurobiológiai oki sajátságokból.

IV. A kombinációs válasz (Berkeley és Stanford). „Míg az előző három válasz önmagában nem teljesen meggyőző a Kínai Szoba ellenpélda megcáfolására, addig ha egyszerre nézzük a hármat, együttesen sokkal inkább meggyőzőek és bizonyító erejűek. Képzeljünk el egy robotot, amelynek a koponyaüregébe egy agy formájú számítógépet helyeztek, képzeljük el, hogy a számítógép az emberi agy minden szinapszisát programozza, képzeljük el, hogy a robot egész viselkedése megkülönböztethetetlen az emberi viselkedéstől, és az egészet egy egységes rendszerként gondoljuk el, és nem pusztán egy bemenettel és kimenettel dolgozó számítógépként. Ebben az esetben a rendszernek minden bizonnyal intencionalitást kellene tulajdonítanunk."

Teljes mértékig egyetértek azzal, hogy ebben az esetben ésszerűnek és valóban elkerülhetetlennek találnánk annak a hipotézisnek az elfogadását, hogy a robot intencionalitással bír, ha semmi többet nem tudnánk róla. Valójában a küllem és a viselkedés mellett a kombináció többi eleme teljességgel irreleváns. Ha olyan robotot tudnánk építeni, amelynek viselkedése megkülönböztethetetlen lenne az emberi viselkedés széles skálájától, mindaddig intencionalitást tulajdonítanánk neki, ameddig okot nem ad az ellenkezőjére. Szükségtelen lenne előre tudnunk, hogy a robot számítógép-agya az emberi agy formális megfelelője.

Egyáltalán nem látom azonban, mennyiben segíti ez a MI erős verziójának állítását, mégpedig a következő oknál fogva: a MI erős verziója szerint egy formális program megvalósulása a megfelelő bemenettel és kimenettel az intencionalitás elégséges feltétele és tényleges alkotórésze. Newell (1980) megfogalmazásában az elmeműködés lényege egy fizikai szimbólumrendszer működése. A mi példánkban azonban az, hogy intencionalitást tulajdonítunk a robotnak, nincsen semmilyen kapcsolatban formális programokkal. Pusztán azon a feltevésen alapul, hogy ha a robot úgy néz ki és úgy viselkedik, mint mi, akkor feltehetjük, amíg az ellenkezője be nem bizonyosodik, hogy bizonyára a mienkhez hasonló mentális állapotokkal rendelkezik, amelyek viselkedésének okai és amelyek a viselkedésében nyilvánulnak meg, továbbá, hogy van egy olyan belső mechanizmusa, amely képes létrehozni ilyen mentális állapotokat. Ha ettől a feltételezéstől független magyarázatot tudnánk adni a viselkedésére, nem tulajdonítanánk intencionalitást a robotnak, különösen ha tudnánk, hogy egy formális program működteti. Pontosan ez a lényege a II. ellenvetésre adott korábbi válaszomnak.

Tegyük fel, tudjuk, hogy a robot viselkedését teljes mértékben az magyarázza, hogy egy ember a robot belsejében nem értelmezett formális szimbólumokat kapott a robot szenzoros receptoraitól, és nem értelmezett formális szimbólumokat küldött a motoros rendszeréhez és az ember a szimbólumokat egy szabályhalmaznak megfelelően kezelte. Tételezzük fel továbbá, hogy az ember semmit sem tud a robotról, csak azt tudja, melyik műveletet kell elvégeznie, milyen értelmetlen szimbólumon. Ebben az esetben a robotot egy eredeti mechanikus bábnak tekintenénk. Az a hipotézis, hogy ennek a bábnak elméje van, indokolatlan és szükségtelen lenne, mivel semmi okunk nem maradt arra, hogy intencionalitást tulajdonítsunk a robotnak vagy a rendszernek, amelynek része (kivéve persze emberünk intencionalitását, amikor a szimbólumokkal manipulál). A formális szimbólumok kezelése folytatódik, a bemenet és a kimenet helyesen kerül összeillesztésre, de az intencionalitás egyetlen valódi lokusza az ember, ő pedig semmit sem tud a releváns intencionális állapotokról. Nem *látja* például, hogy mi éri a robot szemét, nem *szándékozik* mozgatni a robot karját és nem *érti meg* a robotnak szánt és a robot

által tett megjegyzéseket. Az előzőekben kifejtett okok miatt az a rendszer, amelynek az ember és a robot a része, éppúgy nem képes erre.

Hogy megértsük a kérdést, vessük össze ezt az esetet azzal, amikor teljesen természetesnek tartjuk, hogy intencionalitást tulajdonítsunk bizonyos más főemlős fajok egyedeinek, mint például majmoknak, emberszabású majmoknak és a háziasított állatoknak, mint például a kutyának. Durván két oka van annak, hogy ezt természetesnek találjuk: intencionalitás feltételezése nélkül nem tudjuk értelmezni az állat viselkedését, illetve látjuk, hogy az állat hozzánk hasonló matériából áll – az ott egy szem, az ott egy orr, ez a bőre stb. Mivel az állat viselkedése összekapcsolódik azzal a feltételezéssel, hogy mögötte ugyanaz az oki matéria áll, egyszerre feltételezzük, hogy az állat viselkedése mögött bizonyára mentális állapotok állnak, és hogy mentális állapotait olyan mechanizmus hozza létre, amely a mi matériánkhoz hasonló matériából készült. Bizonyára hasonló feltételezéssel élnénk a robottal szemben, ha nem lenne okunk arra, hogy ne tegyük, de mindenesetre azonnal felhagynánk az intencionalitás feltételezésével, amint megtudnánk, hogy egy formális program eredményezi a robot viselkedését és hogy irrelevánsak a fizikai szubsztancia valós oki sajátságai (lásd Searle, 1978).

Két további gyakran felbukkanó (és ezért megtárgyalásra érdemes) válasz van a példámra, bár ezek valójában nem érintik a lényeget.

V) A más elmék válasz (Yale). „Honnan tudjuk, hogy más emberek megértik a kínait vagy bármi mást? Csak a viselkedésükből. A számítógép pedig ugyanúgy átmegy a viselkedési teszteken, mint ők (elvileg), így ha kogníciót tulajdonítunk más embereknek, elvileg a számítógépnek is azt kell tulajdonítanunk."

Ez az ellenvetés csupán rövid választ érdemel. Ebben a vitában a problémás kérdés nem az, hogy honnan tudom, hogy más emberek kognitív állapotokkal rendelkeznek, hanem sokkal inkább az, hogy mi is az, amit nekik tulajdonítok, amikor kognitív állapotokat rendelek hozzájuk. Az érv lényege, hogy ez nem lehet csupán komputációs folyamat és annak eredménye, mert a komputációs folyamatok és azok kimenetei kognitív állapot nélkül is lehetségesek. Nem válasz erre az érvre, ha érzéketlenséget tételezünk fel. A „kognitív tudományokban" ugyanúgy előfeltételezzük a mentális folyamatok valódiságát és megismerhetőségét, ahogy a fizikai tudományok előfeltételezik a fizikai tárgyak valódiságát és megismerhetőségét.

VI. A sok ház válasz (Berkeley). „Az ön egész érvelése azt előfeltételezi, hogy a MI erős verziója csak analóg és digitális számítógépekről szól. Ez azonban csak a technika jelen állapota. Legyenek bármik is azok az oki folyamatok, amelyeket alapvetőnek tekint az intencionalitáshoz (feltéve, hogy önnek igaza van), végül is képesek leszünk olyan eszközöket készíteni, amelyek rendelkeznek ezekkel az oki folyamatokkal és mesterséges intelligenciával. Tehát az ön érvelése egyáltalán nem a mesterséges intelligenciának arra a képességére irányul, hogy kogníciót hozzon létre és azt magyarázza."

Igazán semmi ellenvetésem nincs ezzel a válasszal szemben, leszámítva azt, hogy valójában triviálissá teszi a MI erős verziójának célkitűzését azáltal, hogy bármi olyan valamiként definiálja, ami mesterségesen hozza létre és magyarázza a kogníciót. A mesterséges intelligencia nevében tett eredeti állítás jelentősége éppen az, hogy precíz, jól meghatározott

tézis volt: a mentális folyamatok formálisan meghatározott elemeken végzett komputációs folyamatok. Ennek a tézisnek a megkérdőjelezésével próbálkozom. Ha az állítást úgy definiáljuk újra, hogy megszűnik tézis lenni, akkor az ellenvetéseim a továbbiakban nem állnak, mivel a továbbiakban nincs tesztelhető hipotézis, amelyre vonatkoznának.

Most térjünk rá arra a kérdésre, amelyet ígéretem szerint megpróbálok megválaszolni: feltéve, hogy az eredeti példámban megértem az angol nyelvet, és nem értem meg a kínait, és feltéve, hogy a gép nem érti meg sem az angolt, sem a kínait, mégis kell legyen valami bennem, ami azt teszi, hogy megértem az angolt, és egy ennek megfelelő valami hiányozzon kell belőlem, ami azt teszi, hogy a kínait nem értem meg. Végül is miért ne ruházhatnánk fel ezekkel a dolgokkal, bármik legyenek is azok, egy gépet is?

Nem látom az okát, hogy elvileg miért ne adhatnánk az angol vagy a kínai nyelv megértésének képességét egy gépnek, mivel a testünk az agyunkkal egyetemben lényegében pontosan ilyen gép. Nagyon is erős érveket látok azonban amellett, hogy ilyesmit nem adhatunk egy gépnek, ha a gép működését kizárólag formálisan meghatározott elemeken végzett komputációs folyamatokként definiáljuk, vagyis ha a gép működését úgy definiáljuk, mint egy számítógépes program megvalósulását. Nem azért vagyok képes megérteni az angolt és rendelkezem az intencionalitás egyéb formáival, mert egy számítógépes program megvalósulása vagyok (felteszem, számtalan számítógépes program megvalósulása vagyok), hanem, amennyire tudjuk, azért, mert bizonyos biológiai (pl. kémiai és fizikai) szerveződéssel rendelkező bizonyos fajta szervezet vagyok, és bizonyos feltételek mellett ez a szerveződés mellesleg képes a percepció, a cselekvés, a megértés, a tanulás és más intencionális jelenségek létrehozására. A jelen érv lényegéhez tartozik, hogy csak olyasmi lehet intencionális, ami rendelkezik ilyen oki hatóerőkkel. Talán más fizikai és kémiai folyamatok is okozhatnak ugyanilyen hatást, talán például marslakók is rendelkezhetnek intencionalitással, pedig az agyuk más matériából van gyúrva. Ez empirikus kérdés, hasonlóan ahhoz a kérdéshez, hogy vajon fotoszintézis történhet-e másképp, mint klorofillal.

A jelen érv lényege az, hogy nincs olyan pusztán formális modell, amely önmagában elégséges lenne az intencionalitáshoz, mert a formális sajátságok önmagukban nem hoznak létre intencionalitást és önmagukban nincsen oki hatóerejük, leszámítva azt az erőt, a program megvalósulásakor, amellyel a formalizmus következő lépését hozzák létre a gép futtatásakor. A formális modell egyedi megvalósulásának minden más oki sajátsága irreleváns a formális modellre nézve, mert bármikor áttehetjük ugyanazt a formális modellt egy másfajta megvalósulási helyzetbe, ahol ezek az oki sajátságok nyilvánvalóan hiányoznak. Még akkor is, ha valami csoda folytán a kínaiak tökéletesen megvalósítanák Schank programját, ugyanazt a programot áttehetnénk angolra, vízvezetékekre vagy számítógépekre, amelyek közül egyik sem ért kínaiul, a programról nem is beszélve.

Az agyműködésben nem az lesz érdekes, hogy a szinapszisok elrendeződése milyen formális árnyékot vet, hanem az elrendeződés tényleges sajátságai. A mesterséges intelligencia erős változatának minden eddig áttekintett érve arra törekedett, hogy azt az árnyékot határolja körül, amelyet a kogníció vet, és aztán azt állítsa, hogy az árnyékok valóságos dolgok.

Végezetül szeretnék rámutatni néhány általános érvényű filozófiai kérdésre, amelyet az érvelés implicit módon tartalmaz. Az érthetőség kedvéért megpróbálom kérdés-felelet formában kifejteni, és mindjárt a régi kemény dióval kezdem:

„Tud egy gép gondolkodni?"

A válasz, nyilvánvalóan, igen. Mi magunk pontosan ilyen gépek vagyunk.

„Igen, de tud egy mesterséges, ember készítette gép gondolkodni?"

Ha feltételezzük, hogy lehetséges mesterségesen olyan gépet létrehozni, amely rendelkezik idegrendszerrel, axon és dendrit végződésű idegsejtekkel és a mienkéhez kielégítő mértékben hasonló minden egyébbel, a kérdésre a válasz akkor is nyilvánvalóan igennek tűnik. Ha pontosan le tudnánk másolni az okokat, akkor le tudnánk másolni az okozatot is. Ténylegesen lehetségessé válna a tudat, az intencionalitás és a többi létrehozása az emberi lényekétől különböző kémiai elvek alkalmazásával. Ez, ahogy mondtam, empirikus kérdés.

„OK, de tud egy digitális számítógép gondolkodni?"

Ha a „digitális számítógépen" olyasmit értünk, ami rendelkezik a leírás egy olyan szintjével, amelyen egy számítógépes program megvalósulásaként korrekt módon leírható, akkor a válasz természetesen megintcsak igen, mivel mi magunk számtalan számítógépes program megvalósulásai vagyunk és ugyanakkor gondolkodunk.

„De lehetséges az, hogy valami gondolkodik, megért és a többi, *kizárólag* annak köszönhetően, hogy egy megfelelő fajta programmal ellátott számítógép? A megértés elégséges feltétele lehet-e önmagában egy program megvalósulása, a megfelelő programé természetesen?"

Azt hiszem, ez a megfelelő kérdés, bár gyakran összetévesztik a korábbi kérdések némelyikével, és a válasz erre az, hogy ez nem lehetséges.

„Miért nem?"

Mert a formális szimbólumok kezelése önmagában nem rendelkezik intencionalitással, meglehetősen értelmetlen, mégcsak nem is *szimbólum*-kezelés, mivel a szimbólumok nem szimbolizálnak semmit. Nyelvészeti szóhasználattal élve csak szintaxisuk van, szemantikájuk nincs. Az a fajta intencionalitás, amellyel, úgy tűnik, a számítógépek rendelkeznek, kizárólag azoknak a fejében létezik, akik programozzák és akik használják őket, akik a bemenetet adják és akik értelmezik a kimenetet.

A Kínai Szoba példája ennek a bemutatását célozta azzal, hogy megmutatta, amikor bejuttatunk a rendszerbe valamit, ami valódi intencionalitással bír (egy embert), és a formális programmal beprogramozzuk, azonnal látszik, hogy a formális program semmi további intencionalitásnak nem hordozója. Egy ember számára semmit sem ad például hozzá a kínai megértésének képességéhez.

A MI-nak pontosan az a jellemzője, amely olyan vonzónak tűnt – a program és a megvalósulás közötti különbségtétel – bizonyult végzetesnek arra az állításra vonatkozóan, hogy a szimuláció lehet másolás. A program és a hardware-en való megvalósulása közötti különbségtétel párhuzamba állítható a mentális műveletek és az agyi műveletek szintjének megkülönböztetésével. Ha formális programként le tudnánk írni a mentális műveletek szintjét, akkor, introspekciós pszichológia és az agy neurofiziológiája nélkül is, úgy tűnik, le tudnánk írni azt, ami az elme lényege. Az az egyenlet azonban, hogy „az elme az az agynak, ami a program a hardware-nek", több ponton felborul, többek között a következő három helyen:

Először, a program és a megvalósulása közötti különbségtétel azt eredményezi, hogy ugyanaz a program mindenféle őrült módon is megvalósulhat az intencionalitás formája nélkül. Weizenbaum (1976, 2. fejezet) például részletesen bemutatja, hogyan készíthetünk számítógépet egy guriga WC-papír és egy halom kavics felhasználásával. Hasonlóképpen a kínai történetet megértő program beprogramozható vízvezetékcsövek rendszerébe,

szélmalmok sorozatába vagy egy egynyelvű angol emberbe, amelyek közül egyik sem jut el ezáltal a kínai megértéséig. A kavics, a WC-papír, a szél és a vízvezetékcső nem megfelelő matéria az intencionalitás birtoklásához – csak olyasmi lehet intencionális, ami az agyhoz hasonló oki hatóerővel bír –, és bár az angolul beszélő ember megfelelő matéria az intencionalitás birtoklására, könnyű belátni, hogy semmi plusz intencionalitáshoz nem jut azáltal, ha memorizálja a programot, mivel a memorizálás nem tanítja meg őt kínaiul.

Másodszor, a program tisztán formális, de az intencionális állapotok nem ilyen módon formálisak. A tartalmuk és nem a formájuk szerint vannak definiálva. Az a hiedelem, hogy esik az eső, nem egyfajta formális alakként van definiálva, hanem olyan mentális tartalomként, amely kielégülési feltételekkel, irányultsággal (lásd Searle, 1979) és hasonlókkal rendelkezik. Valójában a hiedelem mint olyan még csak nem is rendelkezik formális alakkal ebben a szintaktikai értelemben, mivel egy és ugyanaz a hiedelem végtelen számú szintaktikai kifejezést kaphat különböző nyelvi rendszerekben.

Harmadszor, ahogy az előbb említettem, a mentális állapotok és események a szó szoros értelmében az agyi műveletek eredményei, de a program ilyen módon nem eredménye a számítógépnek.

„Ha a programok semmi módon nem alkotói a mentális folyamatoknak, miért hitte olyan sok ember éppen az ellenkezőjét? Ez legalábbis valamiféle magyarázatra szorul."

Nem igazán tudok erre válaszolni. Az az elképzelés, hogy a számítógépes szimuláció maga a valódi dolog, azonnal gyanúsnak kellett volna tűnjön, mivel a számítógép semmiképpen sem korlátozódik arra, hogy mentális műveleteket szimuláljon. Senki nem tételezi fel, hogy egy ötös erősségű tűzriadó számítógépes szimulációja porrá égeti a környéket vagy hogy egy számítógépen szimulált vihartól csuromvizesek leszünk. Hogy a csodában gondolná bárki is azt, hogy a megértés számítógépes szimulációja ténylegesen megért bármit is? Néha mondanak olyasmit, hogy borzasztóan nehéz lenne elérni, hogy a számítógépek fájdalmat érezzenek vagy szerelmesek legyenek, pedig a szerelem és a fájdalom nem nehezebb és nem is könnyebb a kogníciónál vagy bármi másnál. A szimulációhoz csak a megfelelő bemenetre és kimenetre van szükség, valamint a programra a kettő között, amely az előbbit az utóbbivá alakítja. Ez minden, amit a számítógép felhasznál a működéséhez. A szimuláció összetévesztése a másolással ugyanolyan hiba, akár fájdalomról, szerelemről vagy kogníciórol, akár tüzekről vagy viharokról van szó.

Mégis számos oka van annak, hogy úgy tűnt – és sok embernek talán még mindig úgy tűnik –, a MI valamiképpen reprodukálni és ezáltal magyarázni tudja a mentális jelenségeket, és azt hiszem, nem sikerül megszabadulnunk ettől az illúziótól addig, amíg teljességgel fel nem tárjuk ennek a kiváltó okait.

Az első és talán legfontosabb ezek közül az „információfeldolgozás" fogalma körüli zűrzavar: a kognitív tudományok területén sokan úgy gondolják, hogy az emberi agy az elmével egyetemben valami olyasmit végez, amit „információfeldolgozásnak" neveznek, és a számítógép a maga programjával ezzel analóg módon végez információfeldolgozást, másrészről viszont a tüzek és a viharok egyáltalán nem végeznek információfeldolgozást. Így aztán, bár a számítógép képes bármiféle folyamat formális jellemzőit szimulálni, különleges kapcsolatban áll az elmével és az aggyal, mert amikor a számítógép megfelelően van beprogramozva, ideális esetben ugyanazzal a programmal, mint az agy, az információfeldolgozás megegyezik a két esetben, és a mentális folyamatok lényege ez

az információfeldolgozás. A gond ezzel az érvvel az, hogy az „információ" fogalmának kétértelműségén alapszik. Abban az értelemben, ahogyan az emberek „dolgozzák fel az információt", amikor mondjuk számtani problémákon gondolkoznak, vagy amikor történetekre vonatkozó kérdéseket olvasnak el és válaszolnak meg, a beprogramozott számítógép nem végez „információfeldolgozást". Amit csinál, az sokkal inkább formális szimbólumok kezelése. Az, hogy a programozó és a számítógép kimenetének értelmezője a világ tárgyait helyettesítő szimbólumokat használ, teljességgel a számítógép hatókörén kívül áll. A számítógép, hogy megismételjük, rendelkezik szintaxissal, de szemantikával nem. Ha azt gépeljük a számítógépbe, hogy „kettő meg kettő egyenlő"?, azt fogja kiírni „4". Fogalma sincs azonban arról, hogy a „4" négyet jelent vagy hogy egyáltalán bármit is jelent. Nem az elsőrendű szimbólumok értelmezéséhez szükséges másodrendű információk hiányáról van szó, hanem inkább arról, hogy az elsőrendű szimbólumok nincsenek értelmezve, már ami a számítógépet illeti. A számítógép csupán további szimbólumokkal rendelkezik. Az „információfeldolgozás" fogalmának bevezetése így aztán dilemmát okoz: vagy olyan módon alkotjuk meg az „információfeldolgozás" fogalmát, hogy a folyamat magában foglalja az intencionalitást, vagy olyan módon, hogy nem foglalja magában. Az első esetben a programmal ellátott számítógép nem végez információfeldolgozást, csak formális szimbólumokat kezel. A másik esetben a számítógép információfeldolgozást végez ugyan, de csak olyan értelemben teszi ezt, ahogyan a számológép, az írógép, a gyomor, a termosztát, a vihar és a hurrikán végez információfeldolgozást, vagyis rendelkezik a leírás egy olyan szintjével, amelyen úgy írható le, mint ami információt vesz be az egyik végén, azt átalakítja és információt küld ki végeredményként. Ebben az esetben azonban a külső szemlélőn múlik, hogy a bemenetet és a kimenetet a hétköznapi értelemben vett információként értelmezi-e. Semmi hasonlóságot nem állapítottunk meg a számítógép és az agy között az információfeldolgozás hasonlóságának a szempontjából.

Másodszor, a MI területén gyakran felbukkannak a behaviorizmus vagy az operacionalizmus maradványai. Mivel a megfelelően programozott számítógép az emberhez hasonló bemeneti-kimeneti mintázattal rendelkezhet, hajlamosak vagyunk arra, hogy az ember mentális állapotaihoz hasonló mentális állapotokat posztuláljunk a számítógépnél. Amint belátjuk, hogy mind fogalmilag, mind empirikusan lehetséges, hogy egy rendszer emberi képességekkel rendelkezzen bizonyos területen anélkül, hogy egyáltalán bármiféle intencionalitással bírna, felülkerekedhetünk ezen a késztetésünkön. Az asztali számológépem rendelkezik számolási képességgel, de intencionalitással nem. A jelen cikkben éppen azt próbáltam bemutatni, hogy egy rendszer egy kínai anyanyelvű személy bemeneti és kimeneti képességeinek pontos másolatával rendelkezhet, és mégsem ért kínaiul, függetlenül attól, hogyan volt beprogramozva. A Turing-teszt, azzal, hogy minden szégyenérzet nélkül behaviorista és operacionalista, tipikus példája ennek a hagyománynak, és úgy gondolom, hogy ha a MI kutatói végérvényesen elvetnék a behaviorizmust és az operacionalizmust, a szimuláció és a másolás közötti zűrzavar nagy része kiküszöbölhető lenne.

Harmadszor, az operacionalizmus maradványai összekapcsolódnak a dualizmus maradványaival. Valójában a MI erős verziójának csak akkor van értelme, ha elfogadjuk azt a dualista feltételezést, hogy ahol az elméről van szó, ott az agy nem számít. A MI erős verziójában (a funkcionalizmusban úgyszintén) csak a programok számítanak, a programok pedig függetlenek gépi megvalósulásuktól; valójában, ami a MI erős verzióját illeti,

ugyanaz a program végrehajtható egy elektronikus gépen, a karteziánus mentális szubsztancián vagy a hegeli világszellemen. A legmeglepőbb felfedezés, ami ezeknek a kérdéseknek a megtárgyalásakor ért, az volt, hogy sok MI-kutatót megdöbbentett az az elgondolásom, hogy a valós emberi mentális jelenségek függhetnek a valós emberi agy valós fizikai-kémiai tulajdonságaitól. Ha egy percre belegondolunk, látható lesz, hogy nem kellett volna meglepődnöm, mivel ha nem fogadjuk el valamilyen formában a dualizmust, a MI erős verziójának célkitűzése teljesen esélytelenné válik. A célkitűzése pedig az, hogy a mentális folyamatokat programok tervezésével reprodukálja és magyarázza. Ha azonban az elme nemcsak fogalmilag, hanem empirikusan sem független az agytól, a célkitűzés kivitelezhetetlen, mivel a program teljesen független a végrehajtásától. Hacsak nem hisszük azt, hogy az elme az agytól fogalmilag és empirikusan is elválasztható – ami a dualizmus egy erős formája –, nem remélhetjük, hogy a mentális folyamatokat programok írásával és futtatásával reprodukálhatjuk, mivel a programok az agytól és minden más egyedi megvalósulási formától függetlenek kell hogy legyenek. Ha a mentális műveletek formális szimbólumokon végzett komputációs műveletekből állnak, ebből az következik, hogy nincsenek érdemleges kapcsolatban az aggyal, az egyetlen kapcsolat csupán az, hogy az agy történetesen a program megvalósítására képes végtelen fajtájú gépek egyike. A dualizmusnak ez a formája nem a hagyományos karteziánus változat, amely azt állítja, hogy kétféle *szubsztancia* van, hanem abban az értelemben karteziánus, hogy szerinte ami az elmében specifikusan mentális, az nincs intrinzikus kapcsolatban az agy valós tulajdonságaival. Ezt a mögöttes dualizmust az álcázza, hogy a MI-irodalom gyakran kirohan a „dualizmus" ellen; a szerzők, úgy tűnik, csak arról felejtkeznek meg, hogy az álláspontjuk előfeltételezi a dualizmus egy erős formáját.

„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 a 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 sajátlagos 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 menetének számítógépes szimulációját, amikor azonban az elméről van szó, sok ember a mély és tartós dualizmus következtében hajlamos hinni ilyen csodában: úgy gondolják, hogy az elme formális folyamatok kérdése és független a meglehetősen sajátlagos anyagi okoktól, ellentétben a tejjel és a cukorral.

A dualizmus védelmében gyakran az a remény fogalmazódik meg, hogy az agy egy digitális számítógép (a korai számítógépeket mellesleg gyakran „elektronikus agynak" nevezték). Ez azonban nem segít. Persze, hogy digitális számítógép az agy. Mivel minden digitális számítógép, az agy is az. A lényeg az, hogy az agy oki képessége az intencionalitás létrehozására nem állhat egy számítógépprogram megvalósulásából, mivel bármely tetszőleges programra igaz az, hogy megvalósíthatja valami oly módon, hogy közben mégsem rendelkezik mentális állapotokkal. Bármi legyen az, amivel az agy intencionalitást hoz létre, ez nem állhat egy program megvalósításából, mivel nincs olyan program, amely önmagában elegendő lenne az intencionalitáshoz.