Searle kínai szoba kísérlete egyszerű és elegáns, és azt bizonyítja, hogy egyetlen számítógép sem képes a természetes nyelv megértésére.

Stephen Law: Filozófia. M-ÉRTÉK Kiadó Kft., Budapest, 2008.

Searle azt kifogásolja, hogy a gépi megértés pusztán a helyes program megtalálásán múlik. Ezzel a funkcionalizmus szívébe döf.

Brighton, H._ Selina, H.: Mesterséges intelligencia másKÉPp. Edge 2000 Kft., Budapest, 2008.

John Searle elmefilozófiai gondolatkísérlete

Ismeretterjesztő filozófia könyvekben a denevér mellett gyakori illusztráció néhány kínai írásjel vagy egy-egy jellegzetes kínai figura. Ezek a szimbólumok John Searle amerikai filozófus sokat hivatkozott gondolatkísérletére utalnak, amely *A kínai szoba* megjelöléssel vált ismertté. Ez az eszmefuttatás először 1980-ban jelent meg, magyarul a *Kognitív tudomány* c. tanulmánykötetben olvasható, *Az elme, az agy és a programok világa* címen.[1]

A tanulmányban Searle nem tesz mást, mint hogy következetesen végiggondol egy, a kognitívizmus illetve a mesterséges intelligencia (MI) kutatás területén akkoriban egyeduralkodó felfogást, amely szerint az elme és az agy kapcsolata hasonlatos a számítógép szoftverének illetve hardverének viszonyához (számítógép analógia). Az elme komputációs elméletének nevezett elképzelés szerint az agy működésének tudatos szintjén olyan információk feldolgozása történik, amelyek szimbólumokban „testesülnek" meg, és a kogníció - tágabb értelemben a mentális működések - ezeknek a szimbólumoknak a manipulációját jelenti.[2] Ennek lényegét Steven Pinker, a teória egyik közismert proponense a következőképpen foglalja össze: „Az agy speciális helyzete abból a speciális dologból adódik, amit az agy csinál.... Ez a speciális dolog pedig az információfeldolgozás, azaz a komputáció. Az információ és a komputáció az adatok mintázatában és logikai kapcsolataiban rejlik, amely független az adatokat hordozó médiumtól. ..... A vágyak és a vélekedések információk, amelyek szimbólumok konfigurációiban testesülnek meg. A szimbólumok valamilyen anyag fizikai állapotai, olyan anyagé például, mint a számítógépek csipjei vagy az agybeli idegsejtek. ....Ha a szimbólumot alkotó anyagdarabok megfelelő módon botlanak olyan további anyagdarabokba, amelyek egy másik szimbólumot alkotnak, az egyik gondolathoz tartozó szimbólumok egy másik, új szimbólumot alkothatnak, amely logikai kapcsolatban áll az előző gondolattal."[3]

A komputációs elmélet szerint az elme működésének megismeréséhez elegendő a szimbólummanipuláció törvényeinek meghatározása és a működés ezeknek megfelelő leírása. Ez a funkcionalizmusnak nevezett felfogás magában rejti azt a hitet is, hogy amennyiben a leírás kellően pontos, úgy megértettük az elmét, és bármilyen fizikai rendszeren megvalósíthatjuk, feltéve hogy az rendelkezik az ehhez szükséges komplex strukturális jellemzőkkel.[4] Az emberi agy molekuláris finomszerkezete csupán egyike az ilyen rendszereknek. Az „implementáció függetlenség" hitéből táplálkozik a mesterséges intelligencia kutatóknak az a reménye, hogy a megfelelően programozott számítógép felveszi

---

[1] *Searle*, John. R. (1980) *Minds, brains, and programs*. Behavioral and Brain Sciences 3 (3): 417-457. Magyarul: *Az elme, az agy és a programok világa*, in: *Kognitív tudomány*, Szerk.: Pléh Csaba, Budapest, Osiris Kiadó, 1996. 136—151. o.

[2] Az MI kutatók fizikai szimbólumrendszer hipotézise (FSZH) szerint a mentális állapotokat absztrakt, funkcionális szerepük leírásával lehet megragadni, de az így - algoritmusok, programok formájában - meghatározott szimbólumfeldolgozó „gépezetnek" valamilyen fizikai rendszeren kell megvalósulnia.

[3] Pinker, S. (2002): Hogyan működik az elme. Osiris Kiadó, Budapest. 32-33- o.

[4] „...a gondolati és fizikai rendszerek között megvalósítási viszony van: egy gondolati rendszer többféle fizikai rendszerben is megvalósulhat." Pléh Csaba (2013): *A megismeréstudomány alapjai. Az embertől a gépig és vissza*. Typotex, Budapest. 31. o.

az elme tulajdonságait, azaz intelligens rendszerré válik. (Ez a felfogás az MI ún. erős verziója.)[5]

Gondolatkísérletében Searle a fentiekben körvonalazott elméletet teszteli. Felteszi a kérdést, mi lenne, ha az elménk valóban a szóban forgó teória szerint működne? A válaszadás eszközeként pedig kigondolja a *Kínai szoba* néven ismertté vált „forgatókönyvet". A történet a következő: Valaki (a tanulmányban Searle) be van zárva egy szobába, és a külvilággal történő kommunikációja csupán papírlapokra írt üzenetekkel történik egy résen keresztül. Tételezzük fel, hogy kérdéseket kap kínai nyelven, amelyekre kínaiul kell értelmes válaszokat adnia. Noha nem tud kínaiul, az ismeretlen írásjelek mellé részletes szabálykészletet és instrukciókat kap a saját anyanyelvén (angolul).[6] A jeleket ezeknek az instrukcióknak az alapján kombinálva elvileg képes olyan jelkombinációkat összerakni, amelyek megkülönbözhetetlenek egy kínai anyanyelvű válaszaitól. Mint látjuk, a Turing próbáról van szó, de annak eredeti értelmezésétől eltérően most nem az az érdekes, hogy a gép válasza megkülönbözhetetlen-e egy valódi emberétől, hanem az, hogy a válasz hogyan konstruálódik - és van-e mögötte megértés. A szobában serénykedő ember (Searle) úgy viselkedik, mint egy számítógép (illetve annak processzora): komputációs műveleteket végez formálisan meghatározott elemeken, azaz egy program aktuális megvalósulása. Bár helyes válaszokat produkál, vagyis bemenetei és kimenetei azonosak az anyanyelvi beszélőével, nyilvánvaló, hogy semmit sem ért a történetből. Searle szerint ebből kényszerítő logikával következik, hogy egyetlen, ezen az elven működő szerkezetben sem keletkezhet megértés ilyen módon - beleértve az agyat, mint sajátos biológiai gépet is.[7] „Ha a programot úgy definiáljuk, mint formális elemeken végzett komputációs műveleteket, a példa azt mutatja, hogy ezek önmagukban nincsenek érdemleges kapcsolatban a megértéssel."[8]

Searle argumentációnak meggyőző erejét növeli az, hogy könnyen elképzelhető és a laikus olvasó számára is azonnal megérthető szituációt gondolt ki. Telitalálat, hogy magára veszi a komputációs gépezet szerepét - így nem szükséges bizonygatnia, hogy a formális műveletekkel operáló rendszerben nincs megértés - hiszen evidencia, hogy ő nem tud kínaiul.[9] Ezen túlmenően nyilvánvaló, hogy bár van megértés és tudatosság a rendszerben (Searle az angol szabálykönyvet és az angolul feltett kérdéseket megérti), az nem formális szimbólumkezelés eredménye. [10] Az is kézenfekvő, hogy az angol szöveg megértése nem a formális szimbólumkezelés többletéből adódik. (Vegyük észre, hogy érvelésének ebben a

---

[5] „...az MI erős verziója a számítógépet nem csupán eszköznek tartja az elme tanulmányozásában, hanem a megfelelően programozott számítógépet valóban elmének tekinti abban az értelemben, hogy ...szó szerint megért és egyéb kognitív állapotokkal rendelkezik. .... a programok nem pusztán olyan eszközök, amelyek a pszichológiai magyarázatok tesztelését teszik lehetővé, hanem sokkal inkább a programok maguk válnak magyarázattá." Searle: *Az elme, az agy és a programok világa*, 136. o.

[6] „A szabályok lehetővé teszik, hogy a formális szimbólumok egyik készletét összekapcsoljam a formális szimbólumok egy másik készletével." Searle, i.m. 137. o.

[7] „ a számítógép és a program minden megértés nélkül egyszerűen csak működik." Searle: *Az elme, az agy és a programok világa*, 138. o.

[8] Searle, i.m. 139. o.

9 Egy interjúban ezt a következőképpen fogalmazta meg: „Take this Chinese Room argument of mine; I think a lot of the effectiveness of that has to do with not just the abstract structure – that syntax isn't sufficient for semantics – but derives from the fact that here's a simple example that anybody can understand: you're locked in a room with a bunch of Chinese symbols on cards and you have a program which tells you how to give them back through a slot in the wall in response to other cards coming in, and all the same you don't understand Chinese. Now any kid can understand that..... And the beauty of the example was I didn't have to consider consciousness, and secondly, I didn't have to ask the "How do you know?" question because I made it about my own case and it's obvious I don't know Chinese." Julian Moore: Interview with John Searle. In: Philosophy Now Volume 25, Winter 1999. p. 37-41

[10] „bármiféle pusztán formális elveket táplálunk a számítógépbe, azok nem lesznek elégségesek a megértéshez, mivel az ember mindenfajta megértés nélkül is képes a formális elvek követésére." Searle, i.m. 139. o.

részében Searle váltakozva két, a tesztelni kívánt teória szerint azonos elven működő komputációs rendszert idéz fel: a saját agyának folyamatait, illetve a kínai szobában lejátszódó, azzal analóg történéseket.) Argumentációja ezen a ponton fontos, továbbvezető kérdéseket vet fel: „Mi az tehát, amivel rendelkezem az angol mondatok esetében, de a kínai mondatoknál nem? ...Mégis, miben áll, és miért nem adható át a gépnek, bármi legyen is az?"[11]

Searle - nevezetesé vált tanulmányának megírását követően - folyamatosan továbbgondolta és továbbfejlesztette a gépi funkcionalizmus kritikájára irányuló argumentációját. Egy 1990-ben tartott előadásában kifejtette, hogy nem csupán arról van szó, hogy egy (bármilyen) program szintaxisa nem elegendő jelentés létrehozásához. Maga a szintaxis is megfigyelőt és értelmezőt feltételező mentális konstrukció, így nem képezi a fizikai világ részét.[12] A komputációs állapotok és az információfeldolgozás - akárcsak a digitális számítógép - nem mások, mint tudatos megfigyelőknek a fizikai világ működésére irányuló szubjektív interpretációi. Amikor ezt az értelmezési keretet használjuk az agyműködés leírására, elkerülhetetlen a felhasználó felbukkanása is, aki komputációsan értelmezi a folyamatokat (homunculus fallacy). Searle erre tipikus példaként David Marr közismert könyvét említi, amelyben a szerző a látást olyan komputációs folyamatként írja le, amely a retinán megjelenő kétdimenziós mintázatból a külső világ háromdimenziós leírását produkálja outputként.[13] A probléma az - írja Searle - hogy nem világos, tulajdonképpen ki olvassa ezt a leírást? Marr könyvét és más, a témában standard forrásként tekintett műveket olvasva olyan érzésünk támad - folytatja érvelését -, hogy az általuk leírt rendszerek automatikusan felidéznek egy homunkuluszt, hogy legyen, aki azok működését komputációs folyamatként értelmezi. (Érdemes felidézni, hogy Marr könyve vezető szaktekintélyek vélekedése szerint a 2. helyet foglalja el a kognitív tudomány 100 legfontosabbnak tartott mű között.)[14]

A téma újabb átfogó (lezáró?) összegzése, és a korábbinál is tágabb episztemológiai - tudományfilozófiai kontextusba helyezése Searle egy 2002-ben megjelent írásában (21 év a kínai szobában) olvasható.[15] A tanulmány bevezető soraiban leszögezi, hogy nevezetes gondolatkísérletének alaptézisét - miszerint bármely számítógép-program formális, absztrakt szintaktikai folyamatainak implementációja elégtelen az emberi megismeréssel velejáró mentális, szemantikai tartalmak előidézésére - a megfogalmazása óta eltelt 21 évben semmi és senki nem cáfolta meg. Amikor ezeket a sorokat írom, Searle argumentuma változatlanul érvényes. Ahogyan a fizikalista elmemagyarázatoknak szembe kell nézniük Nagel érveivel, úgy megkerülhetetlen vonatkoztatási rendszert jelent Searle gondolatkísérlete a funkcionalizmus különböző változatai számára. Ez a két filozófus olyan erős igazolási feltételeket fogalmazott meg az elme működését magyarázni kívánó koncepciók kigondolói számára, hogy felvethető a kérdés: van-e, lehetséges-e egyáltalán bármiféle tudományos magyarázat erre a problémára? Míg Nagel mostanáig a negáció térfelén maradt,[16] Searle kidolgozta saját tudatelméletét. A következő bejegyzésben ennek rövid bemutatására és értékelésére kerül sor.

---

[11] Searle, i.m. 139. o.

[12] „Syntax is not intrinsic to physics. The ascription of syntactical properties is always relative to an agent or observer who treats certain physical phenomena as syntactical." Searle, J. R.: Is the brain a digital computer? In: Philosophy in a New Century. Selected Essays. Cambridge University Press, Cambridge, 2008. p. 93.

[13] Marr, D. (1982): Vision. Freeman, San Francisco.

[14] Az információ forrása: Pléh Csaba (2013): *A megismeréstudomány alapjai. Az embertől a gépig és vissza*. Typotex, Budapest. 23. o.)

[15] Searle, J. R.: Twenty-one years in the Chinese Room. In: Philosophy in a New Century. Selected Essays. Cambridge University Press, Cambridge, 2008.

[16] Legutóbbi, tavaly megjelent könyve is ezt bizonyítja: Nagel, T.: Mind and Cosmos. Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False. Oxford University Press, Oxford, 2012.