sciendo

Műszaki Tudományos Közlemények vol. 11. (2019) 175–178.
https://doi.org/10.33895/mtk-2019.11.39
https://eda.eme.ro/handle/10598/31236
Angol nyelven: https://doi.org/10.33894/mtk-2019.11.39

EME MTSZ MTK

MESTERSÉGES TUDATOSSÁG ELMÉLETI MODELLJE

THEORETICAL FRAMEWORK OF ARTIFICIAL CONSCIOUSNESS

Szigeti Ferenc, ifj.

Budapesti Műszaki és Gazdaságtudományi Egyetem, Gépészmérnöki Kar, Budapest, Magyarország
szigetif.31@gmail.com

Abstract

Human consciousness is our most perplexing quality, still an adequate description of it's workings have not yet arrived. One of the most promising ways to solve this issue is to model consciousness with artificial intelligence (AI). This paper makes an attempt to do that on a theoretical level with the methods of philosophy. First I will review the relevant papers concerning human consciousness. Then considering the state of the field of AI at the moment, I will arrive at a model of artificial consciousness.

Keywords*: artificial intelligence, philosophy, consciousness, cognitive science.*

Összefoglalás

Az emberi tudatosság az egyik legérdekesebb tulajdonságunk, amelynek kielégítő leírása egyelőre még várat magára. Az egyik legcélszerűbb megközelítése a problémának a mesterséges intelligenciákkal (MI) történő modellezés. Az alábbi cikk erre tesz kísérletet elméleti szinten, a filozófia eszköztárával. Áttekintve a kognitív tudományok emberi tudatossággal kapcsolatos eredményeit, valamint a mesterséges intelligenciák jelenleg rendelkezésre álló módszereit figyelembe véve, felvázolok egy tudatos működésre képes MI modellt.

Kulcsszavak*: mesterséges intelligencia, filozófia, tudatosság, kognitív tudományok.*

1. Tudatosság problémája

Az elmefilozófia jelenleg legelfogadottabb és a legtöbb tudományos irányzattal kompatibilis tudatosságleírása a funkcionalizmus, ami szerint a mentális tulajdonságokat, állapotokat a betöltött funkcióikkal lehet leírni és azok megvalósíthatóak különböző fizikai rendszerekben (egyetlen hátránya, hogy a fenomenális tudatosságot nem tudja magyarázni, így ezzel a részével a tudatosság problémájának nem is foglalkozunk majd[1]).

Többféleképpen meg lehet fogalmazni, mi is az a bizonyos funkció, amelyet a tudatosság ellát, viszont ezek a megfogalmazások nem feltétlenül egyenértékűek egymással. Lehet ez a gondolkodás, a kreativitás, a probléma megoldás, a döntéshozatal, az önálló cselekvés, a figyelem irányítása, célok megfogalmazása és követése. Ezek a fogalmak mind olyanok, amelyeket legtöbbször azonosítunk vagy szoros kapcsolatba hozunk a tudatosság jelenségével. Egy ilyen jellegű definíció a tudatosságot tipikusan egy kognitív modulra vezeti vissza (ami legtöbbször hasonlóan nehezen megfogható). Így igazán egyik sem mutat rá egy fizikailag jól körülhatárolt vagy leírt jelenségre, amelyre alapozva egy valós MI modell elképzelhető lenne.

Az emberi agyról és elméről már nagyon sokat tudunk, kis túlzással akár elkezdhetnénk idegsejtről idegsejtre megvalósítani, viszont ez nem tűnik egy célszerű megoldásnak. Olyan komplexitást kellene lemásolnunk, amely szinte felfoghatatlan. Pontosan ezért van szükség olyan modellekre, amelyek tudnak ezen egyszerűsíteni és mégis képesek megtartani a tudatosság lényeges funkcióit.

[1] Ez a tudatosság nehéz problémája, amely értelmében nincs kielégítő fizikai alapú magyarázat arra, hogy miért van az érzékelésnek, gondolkodásnak egy átélési élménye, egy szubjektív minősége

2. Milyen funkciókkal bír a tudatos működés?

A kognitív tudományokban az elme feltérképezése közben a tudatosság jelenségével is foglalkoznak. A jelenség megfejtésére irányuló kísérletek eredményeiből levonhatók következtetések arra vonatkozólag, hogy pontosan milyen funkciókkal is bír a tudatosság. Ezek megismerhetőek a globális munkatér modell keretében [1]:

– a cselekvés szándékos irányítása;
– hosszútávú és explicit információ tárolás;
– új tervek készítése mentális folyamatok újszerű kombinálásával;
– a figyelem irányítása.

A cikkben következő modellnek a fentebbi négy jelenséget vagy funkciót kell elsősorban jól leírnia. A modell alapjául szolgáló kognitív tudományokbeli szakirodalmat ehhez röviden áttekintjük a modell érthetőbbé tétele érdekében.

3. Tudatossághoz alapvető neurális folyamatok áttekintése

Ebben a fejezetben minden bekezdés egy kognitív modellt foglal össze röviden, amelyeknek az elemeit majd valamilyen formában felhasználjuk a későbbiekben.

Az emberi tudatosságot jól leíró modell a globális neurális munkatér (global neuronal workspace theory) modell [2]. Az elmélet szerint a tudatosság az információ globális megosztottságát az egymáshoz kapcsolódó kognitív modulok között (piramidális neuronok útján[2]) jelenti. Ez a kapcsolódás különösen a magasabb kognitív funkciókat ellátó modulok között jelenik meg, így ezek összekapcsolva alkotják a munkateret, ahol megvalósul a tudatos működés.

A figyelem működése leírható négy folyamat egymásra hatásával [3]. A munkamemóriában (working memory) kerülnek az információk feldolgozás alá, viszont csak néhány másodpercre és ideiglenesen, mivel erősen korlátozott a kapacitása, nem tud egyszerre minden bejövő információt feldolgozni. A kompetitív kiválasztás (comptetitive selection) mechanizmusa fogja eldönteni, hogy a rendelkezésre álló számos információcsatorna közül pontosan melyik fog a munkamemóriához hozzáférni az adott pillanatban. A Top-down érzékenységi irányítás (top-down sensitivity control) felel a különböző információs csatornák jelerősségének beállításáért, ezen jelerősségek alapján dől el, hogy melyik fog helyet kapni a munkamemóriában a kompetitív kiválasztás útján. A kiugrás szűrő (salience filter), a haladéktalanul figyelmet kívánó információs csatornák jelerősségét emeli meg az adott helyzetben, ez természetesen evolúciósan tanult ingerekre működik, valamint a környezetétől nagyon eltérő ingerekre (például hajlamosak vagyunk egy zöld réten egy piros virágot nagyon hamar észrevenni). A modell szerint az első három folyamat (munkamemória, kompetitív kiválasztás, top-down érzékenységi irányítás) alkot egy körfolyamatot, amely lehetővé teszi a szándékosan és tudatosan irányított figyelmet.

A munkamemória leírható négy folyamat: a figyelem, a szenzoros reprezentációk (beérkező ingerek), a hosszú távú memória reprezentációi (emlékek) és az előrelátás termékeként. Ez nem egy modul, hanem mindig a feldolgozandó információhoz megfelelő módon szerveződik. Limitált[3] a kapacitása (3-4 elem). A hosszútávú memória aktiválása útján működik [4].

A munkamemória megtalálható az állatoknál is, a kísérleti eredmények nem teljesen világosak, de úgy tűnik, alapvetően még kisebb a kapacitásuk, könnyebben elterelődik a figyelmük és kizárólag az adott feladat megoldására tudják használni [5]. Viszont alapvető működése egyezhet a miénkkel, ha néhány tekintetben nem is úgy használják, mint mi.[4]

A figyelmi séma elmélet (attention schema theory) úgy magyarázza a szándékos figyelmet és a tudatosságot (itt awareness[5]), hogy az elme alkotott egy egyszerűsített figyelmi modellt (ahogyan egy egyszerűsített test modellünk is van), amely alkalmas a figyelem irányítására. Ezen belső modell alapján az agy úgy látja, hogy rendelkezik egy olyan képességgel (tudatosság az „awareness” értelemben), amely képes a figyelem irányítására bármilyen fizikai folyamat közreműködése nélkül (a belső modell egyszerűsít természetesen, így a fizikai részleteket már nem tartalmazza, ezért tűnhet úgy mintha nem fizikai folyamat lenne mögötte) [6].

Tovább is léphetünk a tényleges modellre, amely felhasználja az itt ismertett elméleteket, hogy a korábban felsorolt funkciókat magyarázni tudja.

4. Az emberi tudatosság egyszerűsített modellje

Az alábbi fejezetben áttekintjük a fejezetcímben megnevezett modellt, amely a kutatás jelenlegi

[2] Az ilyen típusú neuron nagyon hosszú axonnal és szer-teágazó dentrittel rendelkezik, ezzel lehetővé téve az egymástól fizikailag távol elhelyezkedő agyi területek összeköttetését [2].

[3] Ennek az oka az lehet, hogy az agy képtelen fenntartani túl sok agyi területen az egymástól független aktivitást a különböző reprezentációkhoz, minimális interferenciával [4].

[4] Az állatok között komoly különbségek vannak, a kísérletek általában majmokra, delfinekre, fókákra, rágcsálókra és esetleg varjúfélékre vonatkoznak.

[5] Magyar nyelven nincsen erre a fogalomra külön szó, de ez a tudatosságnak az a minősége, amely a szubjektivitást vagy önreflexiós képességet fejezi ki

eredménye. Ez egy kognitív modell, amely alapján neurális modell, majd implementációs modell készíthető mesterséges intelligenciákhoz. Az alfejezetekben sorban válaszokat fogunk kapni arra, hogy a modell hogyan magyarázza a tudatosság 2. fejezetben ismertetett négy funkcióját.

4.1. Hogyan működik az emberi agy?

A beérkező információkat (külső és belső ingerek) neuronok egy rendszere dolgozza fel, mindehhez energiát felvéve a szervezetből. Az információfeldolgozás végén pedig valamilyen cselekvés vagy mozgás fog megvalósulni. Fontos megérteni, hogy a neuronok akkor kapnak energiát, ha munkát végeznek, vagyis információt dolgoznak fel. Több energiát szeretnének, így folyamatosan keresik a lehetőséget arra, hogy dolgozhassanak.

4.2. Hogyan irányítható a figyelem tudatosan?

A figyelem, ahogyan korábban láttuk, nem egy modul, hanem egy folyamat, vagyis inkább jelenség. Egyszerűen azonosítható ez a fogalom a neuronok aktivitásával. Ott van a figyelem, ahol egy csoport neuron aktív, vagyis éppen információt dolgoz fel. Egyszerre számos agyi terület vagy neuron csoport lehet aktív, viszont ezek közül csak a legaktívabb néhány fog helyet kapni a munkamemóriában.

A munkamemória nem más, mint a belekerülő információt tartalmazó neuronok csoportjai. Azok az agyi területek, amelyek kellőképpen aktívak. Itt most úgy tűnhet, mintha egymással határoznánk meg a munkamemóriát és a figyelmet, ami bizonyos szempontból igaz is, mivel egyfajta körfolyamatot fogunk elérni. Ahelyett, hogy a korábban leírt négy folyamat termékeként írnánk le a munkamemóriát, azonosítsuk a figyelemmel és a hosszú távú memóriával. A hosszú távú memória jelenti a neuron hálózatunkban megmaradt emlékeket, amikor azok aktiválódnak, bekerülnek a munkamemóriába. Vagy inkább a munkamemória ott aktiválódik és ezzel együtt a figyelem is oda irányul.

A lényeges gondolat, hogy igazán mindhárom neuronok aktivitását jelenti. Jól láthatóan rengeteget egyszerűsítettünk a korábbi kognitív modelleken azzal, hogy ezeket a rendszereket azonosítottuk. Viszont ezzel még nem válaszoltuk meg az eredeti kérdést.

A figyelemnek mindig valamire irányulnia kell, ez lehet egy inger (belső vagy külső) vagy egy emlék. Az első egy közvetlen beérkező szenzoros információ, míg a második egy korábban történt hasonló információ (vagy már egy sokszorosan átdolgozott és más információkkal kombinált információ). A figyelem tudatos irányítása legalább két ilyen információ (amelyekből valamennyi mindenképpen emlék kell legyen) körfolyamatából alakul ki. Ezek közül mindig egyszerre egy van a figyelem középpontjában (magasabb az aktivitása, mint bármely más információnak). A körfolyamat akkor valósulhat meg, ha képes az információlánc önmagát erősíteni és így elérni, hogy felváltva legyenek a tagjai a legaktívabbak.

Feltehetően úgy tudják az információk egymást erősíteni, ha az őket tároló neuron csoportok kapcsolatban állnak egymással, hiszen akkor az áram ténylegesen körfolyamatként futhat végig a rendszeren. Ebből következően azok az agyi területek, amelyek jobb kapcsolódással rendelkeznek egymás felé (globális neurális munkatérben ismertetett munkatér vagyis a tudatosság színhelye vagy a magasabb kognitív modulok), könnyebben tudják megtartani a figyelmet.

A figyelem váltása tudatosan is jól értelmezhető ez alapján, az információs láncból átválthat a rendszer egy kapcsolódó másik információs láncba vagy esetleg csak egy része cserélődik a láncnak. Ilyen módon tudjuk például végigjárni az emlékeinket vagy végiggondolni a megtanult ismereteinket.

Felmerülhet a kérdés, ha az állatok (megint csak alapvetően magasabb szintűekre gondolva) is hasonló munkamemóriával rendelkeznek, akkor az eddigiek alapján miben térnek el tőlünk vagy mi nem vonatkozik rájuk? Egyelőre a vázolt modell szerint minőségileg semmilyen eltérés nincs, ugyanazok a folyamatok érvényesek rájuk is, csak esetleg érzékenyebbek a kiugrás szűrőre vagy kisebb az agyuk komplexitása, így kevesebb agyi területük lehet aktív egyszerre.

A komplexitáson túl is van azonban egy különlegessége az emberi agyi működésnek. Ehhez vesszük igénybe a figyelem séma elméletet. A figyelem séma is értelmezhető egy emlékként (ha nem közvetlen szenzoros információ, akkor emlék kell legyen, hogy hűek maradjunk a korábbi megkülönböztetésünkhöz), vagy emlékek egy halmazaként. Ezek lesznek azok az emlékek, amelyek lehetővé teszik az előzőleg ismertett kölcsönös erősítés hatékony működését. Kapcsolódnak egy kiterjedt rendszerként számos agyi területhez (munkatér a piramidális neuronokkal) és így képesek vagyunk a segítségükkel eljutni elménk bármelyik szegletébe, egyfajta autópályaként szolgálnak, amelyen végigfuthat a figyelem és elérhet bármilyen eltárolt információt.

A figyelmi sémával pedig bevezettük a rendszerbe a szubjektív élményt, amelyet tulajdonít az agy a legaktívabb információnak. Egyszerre számos informá-

ció aktív valamilyen szinten, viszont ebben az egyszerűsített figyelmi sémában egy bizonyos legerősebb fog igazán feltűnni. Az egymással „versenyző" számos aktivált hurok közül mindig csak néhány lesz (a legaktívabbak), amelyeket a figyelmi séma úgy érzékel, hogy a figyelem középpontjába kerültek.

Ezt a figyelmi sémát ne úgy képzeljük el, mint ami valahogyan „fölötte áll" minden agyi folyamatnak vagy az „igazi okozója" a tudatos működésnek. Nem, ez egy olyan rendszer, amely lehetővé teszi a figyelem hatékony irányítását, lehetővé teszi, hogy a figyelem vagyis igazából a neuronok aktivitása ott legyen magas, ahol szükség van rá, és a bejövő ingerek ne zavarják meg minden esetben (csak ha feltétlen szükséges) a koncentrált információ feldolgozást. Ez a figyelmi séma lehetővé teszi a magasabb szintű gondolkodást egyszerűen azzal, hogy képes az információ feldolgozást irányítani.

4.3. A cselekvés szándékos irányítása

A cselekvések indítása is egy bizonyos agyi területen található neuron csoport aktiválásával történik. Annyinak kell csak történnie, hogy a figyelem ráirányuljon, vagyis eljusson az aktivációs hurok addig az agyi területig. Erre alkalmas a jó összeköttetésekkel rendelkező figyelmi séma vagy munkatér, amely így képes a cselekvések indításáért felelős neuroncsoportokat aktiválni.

4.4. Hosszútávú és explicit információ tárolás

Akkor erősödnek a neurnok között a kapcsolatok, ha minél többször használják az adott kapcsolatot. A modell lehetővé teszi, hogy elegendő ideig aktív maradjon egy adott inger ahhoz, hogy emlékké válhasson, az inger hurokban tartása folytán.

A felejtés mechanizmusa pedig a neuronok egyéni viselkedésével magyarázható, igényük van több energiát kapni, így az olyan kapcsolódásaikat tartják meg, amelyeket használ a rendszer. Ha egy kapcsolatot nem használ, azt a neuronok leépítik és más irányba próbálnak kapcsolódni helyette. A korábban már eltárolt emlékek így elveszíthetik néhány vagy összes kapcsolatukat és előhívhatatlanná válnak.

4.5. Új tervek készítése mentális folyamatok újszerű kombinálásával

Ez a funkció a kreativitást és a gondolkodást jelenti. A modell szerint bármely kellőképpen kapcsolódó agyi terület egy másikkal kapcsolatba kerülhet. Természetesen ha jobb az összeköttetés, akkor jobban működik a figyelem megtartása is, mert közvetlenebb a kapcsolat és kevésbé halványul el a jel rövidebb idő alatt. Valamint szükséges ennek a funkciónak a működéséhez, hogy emlékek kerüljenek egymással hurokba, mivel azok tartalmazzák a korábbi mentális folyamatokat.

5. Mesterséges tudatosság modellje

Zárásként foglaljuk össze mi a modell lényege. A tudatos működést körfolyamatként írjuk le, ami ilyen módon önmagát képes irányítani bizonyos korlátokon belül. A munkamemória, hosszú távú memória, figyelem mind ugyanaz a mechanizmus, vagyis neuronok aktivitása.

A hurkok versenyeznek, hogy minél aktívabbak legyenek, mert akkor több ideig tudnak kapcsolódni vagyis több energiát kapnak a bennük található neuronok, amelyeknek végső soron ez az egyéni céljuk.

Az emlékemlék hurkokból kialakult összetettebb emlékek szolgálnak alapul a magasabb kognitív funkcióknak, viszont ezek egy része már a DNS-be íródva öröklődve jut el hozzánk és nem kell aktívan megtanulnunk. Egy ilyen komplex emlékhalmaz a figyelem séma, amely hatékonnyá teszi a tudatos működést.

A vázolt modell minden további nélkül megvalósíthatónak tűnik, miután egy neurális modell, majd egy implementációs modell is elkészül, amelyek egyre gyakorlatiasabb formába öntik a modellt. Ezek a kutatás következő lépései.

Köszönetnyilvánítás

A cikk megírását támogatta az Emberi Erőforrások Minisztériuma az Új Nemzeti Kiválóság Program (ÚNKP) keretében 2018/2019. tanévben elnyert Nemzeti Felsőoktatási Kiválóság Ösztöndíj – Felsőoktatási Alapképzés Hallgatói Kutatói Ösztöndíjjal. Ezentúl köszönöm a támogató munkáját Gyarmathy Ákosnak és konzulensemnek Dr. Héder Mihálynak!

Szakirodalmi hivatkozások

[1] Dehaene S., Naccache L.: *Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework*. Cognition 79/1-2. (2001) 1–37.

[2] Dehaene S., Changeux J. P., Naccache L.: *The global neuronal workspace model of conscious access: from neuronal architectures to clinical applications.* Characterizing consciousness: From cognition to the clinic?. Springer, Berlin, Heidelberg, 2011. 55–84.

[3] Knudsen E. I.: *Fundamental components of attention.* Annu. Rev. Neurosci. 30 (2007): 57–78.

[4] Eriksson J., et al.: *Neurocognitive architecture of working memory*. Neuron 88.1 (2015): 33–46.

[5] Carruthers P.: *Evolution of working memory*. Proceedings of the National Academy of Sciences 110. Supplement 2 (2013): 10371–10378.

[6] Graziano M. S. A., Webb T. W.: *The attention schema theory: a mechanistic account of subjective awareness.* Frontiers in psychology 6. (2015) 500.