A valószínűtlen mesterséges neuronoktól az idealizált kognitív modellekig: A mesterséges intelligencia filozófiájának újraindítása

Catherine Stinson

Megjelenés a *Tudományfilozófia* folyóiratban

Absztrakt

Az elmefilozófián belül hatalmas irodalom foglalkozik a mesterséges intelligenciával, de alig tesz említést a módszertani kérdésekről. A tudományfilozófián belül is egyre több munka foglalkozik a modellezés módszertanával, amely alig említ példákat a kognitív tudományokból. Itt kapcsolódnak össze ezek a viták. A tudományfilozófiai szakirodalomban az idealizálás fontosságáról kidolgozott meglátások módot adnak a konnekcionista hálózatok neurális valószínűtlenségének megértésére. A neurokognitív tudományból származó meglátások megvilágítják, hogyan választják ki a modellek és a célpontok közötti releváns hasonlóságokat, hogyan igazolják a modellezési következtetéseket, és a modellek metafizikai státuszát.

1. Bevezetés

A mesterséges intelligencia (AI) filozófiája körülbelül 20 éve az elmefilozófia egy meglehetősen poros sarkába szorult.[1] Annak ellenére, hogy a tudomány és a mérnöki tudomány számos ágában igen széles körben alkalmazzák az olyan módszereket, mint a támogató vektor gépek, a döntési fák, a főkomponens-elemzés és a neurális hálózatok, a kortárs mesterséges intelligencia nagyrészt elkerülte a tudományfilozófusok figyelmét.[2] Ugyanebben az időszakban a modellek és szimulációk hiánypótló témából a tudományfilozófián belül népszerű aldiszciplínává nőtték ki magukat, bár ezek a viták általában a modelleknek egy maroknyi területen (közgazdaságtan, klímatudomány, fizika, ökológia) való használatára összpontosítanak, amelybe a neurokognitív tudományok nem tartoznak bele.

Itt az ideje, hogy ezek az elhidegült rokonok újraegyesüljenek. A gépi tanulás beszivárgása az életünket közvetítő technológiák nagy részébe aligha lehetne aktuálisabb a mesterséges intelligencia módszereinek megértése. Hasonlóképpen, annak megértése,

---

hogy hogyan és mikor bízzunk meg a[1] Az újbóli aktivitás jeleiről lásd Buckner (2019).

[2] Godfrey-Smith (2006, 2009) nagyon röviden említi a neurális hálózatokat.

az éghajlati modellek előrejelzései sürgősen fontosak. Ez az írás megnyitja az utat a mesterséges intelligencia filozófiájának a tudományfilozófiába való visszaemeléséhez, és mindkét oldalról bemutat néhány lehetséges előnyt.

A mesterséges intelligencia tudományfilozófiába való visszahozásának természetes első lépése, hogy újra megvizsgáljunk egy problémát, amely központi helyet foglal el a mesterséges intelligencia módszertanáról szóló vitákban: miért tartják hasznosnak a neurális plauzibilitást a konnekcionista modellekben, amikor a modellekről tudjuk, hogy nem reálisak? Ezt a problémát soha nem oldották meg, de továbbra is releváns, legutóbb a mélytanulásról szóló vitákban, ahol az "ellenséges példák" feltárják a számítógépes és az emberi látás közötti különbségeket (lásd Han et al. 2019). A konnekcionizmus fénykorában hiányzott a filozófiai szókincs ennek a kérdésnek a megválaszolásához. Itt azt a problémát, hogy miért és hogyan kell a kognitív modelleknek neurálisan plauzibilisnek lenniük, a tudományos modellezés egy általánosabb kérdésének szemszögéből vizsgáljuk: miben kell a modelleknek hasonlítaniuk a célrendszerükre ahhoz, hogy releváns, általánosítható eredményeket produkáljanak? Ez hasznosnak bizonyul a konnekcionista modellek megértésében.

A neurokognitív tudományokból származó felismerések szintén fontos hiányosságokat tárnak fel a modellezés és szimuláció elméletében, amely a tudományterületek korlátozott köréből származó példákra támaszkodik. A neurokognitív tudományok kritikusabban szemlélik a reprezentációt, és mélyebbre ásnak az ok-okozati összefüggések és a modellek metafizikájának kérdéseiben. A konnekcionista modellezésben a következtetésnek a fajtákon alapuló elemzése kiterjeszthető a modellekre általánosabban is.

A 2. szakasz felszínre hozza a konnekcionizmus problémáját, amely egyszerre

támogatja és nem követi a neurális plauzibilitást. A 3. szakasz áttekinti a modellekkel és szimulációkkal kapcsolatos legújabb filozófiai munkákat, hogy megmutassa, hogyan lehet a konnekcionizmus

A módszertani rejtély megoldható, ha a konnekcionista modelleket a kognitív mechanizmusok idealizált modelljeiként értelmezzük. A 4. szakasz a neurokognitív tudományok szemszögéből vizsgálja a tudományos modellezés standard filozófiai beszámolóinak hiányosságait, és a konnekcionista modellek elemzéséből kiindulva felvázolja a modellek és a célok közötti kapcsolat újszerű bemutatását. Az 5. szakasz bemutatja, hogyan alkalmazható ez a számítás a konnekcionista modellek egy sor példáján.

2. A konnekcionista modellek neurális valószínűtlensége

Bár a konnekcionista modellezés iránti filozófiai érdeklődés hosszabb múltra tekint vissza, nagyrészt a Parallel Distributed Processing (PDP) kutatócsoportból ered, amelynek

kétkötetes "bibliája" (Rumelhart és McClelland 1986a; McClelland és Rumelhart 1986) vitát váltott ki a kognitív tudomány számítási módszereiről.[3] A standard konnekcionista hálózati architektúra egy háromrétegű, előrecsatolt hálózat egyszerű neuronszerű egységek, ahol minden egység kimenetet küld a következő magasabb réteg minden egységének. A kapcsolatok bármilyen mintázata lehetséges, beleértve a ritka, oldalsó, visszacsatolt vagy rekurrens kapcsolatokat is. A kortárs mély tanulási hálózatok több mint 3 réteget tartalmaznak, és gyakran kis szomszédságokban kapcsolódnak. A hálózat aktivitását az egyes egységek aktivációja, az egyes kapcsolatok súlya és az aktiválási függvény határozza meg, amellyel az egység kimenete a bemeneti aktivációk súlyozott összege alapján kerül kiszámításra. A súlyok beállítása egy olyan tanulási szabály segítségével történik, amelynek célja a teljes hiba minimalizálása.

Első pillantásra úgy tűnik, hogy a konnekcionista projekt neurálisan hihető mesterséges intelligencia modellek építéséről szól. A PDP biblia bevezetője szerint: "Az

egyik oka a vonzalomnak, hogy

[3] Marcus (2018) nyomán jelenleg újra érdeklődés mutatkozik e kérdések némelyike iránt.

a PDP-modellek nyilvánvalóan "fiziológiai" jellege: Úgy tűnik, sokkal szorosabban kötődnek az agy fiziológiájához, mint más típusú információfeldolgozási modellek" (McClelland és Rumelhart 1986, 10). Közelebbről megvizsgálva azonban mind a kijelentés, mind a projekt motivációi nehezebben értelmezhetőnek bizonyulnak. Mit értünk "ízlés" alatt?

Miért van a "fiziológiai" szó idézőjelben? Mi alapján vonzó, hogy "fiziológiai" íze van?

A PDP-csoportot az inspirálta, hogy a klasszikus mesterséges intelligencia modelljei alkalmatlannak tűntek bizonyos típusú számításokhoz:

> a biológiai hardver egyszerűen túl lassú ahhoz, hogy a mikrostruktúra szekvenciális modelljei hihető magyarázatot adjanak... Minden egyes további megkötés több időt igényel egy szekvenciális gépben, és ha a megkötések pontatlanok, a megkötések számítási robbanáshoz vezethetnek. Mégis az emberek gyorsabbak, nem pedig lassabbak lesznek, ha képesek kihasználni a további kényszereket. (McClelland és Rumelhart 1986, 12.)

Azt is érdemes megjegyezni, hogy a PDP csoport projektje nagyon is a klasszikus mesterséges intelligenciával volt összhangban, mivel olyan modelleket akartak létrehozni, amelyek a pszichológiai kísérletek eredményeinek megfelelő kimenetet produkálnak, és figyelmet fordítottak a reakcióidőre: ezek a lépések egyenesen a kognitív pszichológusok eszköztárából származnak.[4] De mivel a PDP-biblia a kognitív tudomány hagyományos megközelítéseitől való elfordulásnak tekintették, a biológiai hardverre való hivatkozással ellenvetéseket váltott ki a mesterséges intelligencia és a kognitív pszichológia főáramából. Ezeket az ellenvetéseket az alábbiakban négy probléma köré csoportosítjuk.

2.1 A szintek problémája

[4] Hinton, Rumelhart és McClelland mind pszichológusként kezdték.

Az első fő kritika arra vonatkozik, hogy a PDP-modellek milyen szinten kívánnak elhelyezkedni.

Broadbent szerint McClelland és Rumelhart (1985) helytelenül úgy állította be az elosztott memória rendszerét, mint amelynek "pszichológiai és nem csupán fiziológiai szintű következményei vannak" (Broadbent 1985). Broadbent a szintekre való hivatkozással Marr-ra (1982) utal, azzal az implikációval, hogy a megismerésnek függetlennek kellene lennie a megvalósítástól.

Fodor és Pylyshyn (1988) dilemmaként fogalmazza meg Broadbent kihívását: vagy a konnekcionista modellek a szimbolikus modellek "puszta implementációi", vagy nem képesek megfelelően megragadni a megismerést. Ha a PDP-modellek pszichológiai modellek, akkor az idegi részleteknek irrelevánsnak kellene lenniük, és nem nyújtanak előnyt. Ha a PDP-modellek implementációs szintű modellek, akkor az idegtudósok számára érdekesek lehetnek, de nem kognitív tudományok.

Sok oldalra lenne szükség ahhoz, hogy ennek a reakciónak az összes változatát felsoroljuk. Elég, ha csak annyit mondunk, hogy a *Stanford Encyclopedia of Philosophy a* közvélekedés szerint kétféle konnekcionista létezik: implementációs és radikális. Az implementációs konnekcionisták "azt vallják, hogy az agy hálója egy szimbolikus processzort valósít meg", míg a radikális konnekcionisták "azt állítják, hogy a szimbolikus feldolgozás egy rossz találgatás volt az elme működéséről" (Garson 2015). Egyes konnekcionista projektek, például Hinton (1990) cikkei azt mutatják, hogy a PDP-modellek képesek strukturált reprezentációkra és soros feldolgozásra, azaz implementációs konnekcionizmusra. Más konnektonista projektek, mint például Plaut (1995), azt mutatják, hogy ami a felszínen sorozatos feldolgozásnak tűnik, az jobban megmagyarázható a hálózati szintű részletekkel, azaz a radikális konnekcionizmus. A nem annyira radikális konnekcionizmus, amely azt állítja, hogy a szimbolikus

feldolgozás egy rossz tipp arra, hogyan működnek *egyes* mentális funkciók, közelebb áll ahhoz, amit a legtöbb konnekcionista hisz.

Akárhogy is, a konnekcionisták általában nem fogadják el, hogy modelljeik *pusztán* implementációk. Rumelhart és McClelland (1985) azt kifogásolja, hogy a kognitív pszichológusokat érintő kérdések nagy része inkább algoritmikus, mint számítási szinten történik.[5] Smolensky (1988, 1988a) a konnekcionista modelleket "szubszimbolikus szinten" lévőnek írja le, és azt mondja, hogy a konnekcionista kutatás célja egy "középút a szimbolikus számítás megvalósítása és a struktúra figyelmen kívül hagyása között" (Smolensky 1988a). Hogy pontosan mi ez a középút, az nem világos.

2.2 A neurális részletprobléma

Egy másik jól begyakorolt kihívás, hogy a konnekcionista modellek nem hasonlítanak az agyakra a részleteikben. A backpropagation algoritmus hírhedt arról, hogy neurálisan valószínűtlen; a hibajelek általában nem terjedhetnek visszafelé a neurális kapcsolatok hálózatán keresztül, ahogyan azt az algoritmus megköveteli. Hasonlóképpen, a konnekcionista modellek csomópontjai általában determinisztikus aktivációs függvényekkel rendelkeznek, míg a valódi akciós potenciálok sztochasztikusak.

A neurális részletesség problémájának egyik legfontosabb példája az egyes egységek értelmezésének rugalmassága. A lokális reprezentációkkal rendelkező hálózatokban az egységek meghatározott jelentést kapnak, mint például a Jets és a Sharks tagjainak neve, foglalkozása és életkora McClelland (1981) szerint. Az elosztott reprezentációjú hálózatokban "minden egységet egy sok számítóelemre elosztott tevékenységminta reprezentál, és minden egyes számítóelem sok különböző egység reprezentálásában vesz részt" (Hinton 1984).

A legtöbb konnekcionista modellben túl kevés egységet használnak ahhoz, hogy reális agymodellek legyenek. Egyes konnekcionista hálózatokban az egységek kifejezetten

neuronok egész populációit helyettesítik,

[5] A konnekcionista modellek és a Marr-féle szintek kapcsolatáról lásd Churchland és Sejnowski (1990).

az egység aktiválásával, amely egy populációs vektort képvisel.[6] Így jelentős eltérések vannak abban, hogy egy egység minek felel meg.

Az egyik legvitatottabb példa a múlt idejű tanulás (Rumelhart és McClelland 1986). Ez a hálózat angol igéket fogad el bemenetként, és megtanulja kiadni azok múlt idejét. Egy sor példán keresztül képzik ki, amelyek között vannak szabályos igék (add "ed") és szabálytalan igék (went, swam) is. A múlt idejű igék tanulásának sikere a múlt idejű igék konjugációjának megtanulásában a szabályos és szabálytalan igéket elválasztó explicit szabályrendszer nélkül, ahogy Boden fogalmaz, "elméleti dinamit" volt (Boden 2006, 956). Ugyanakkor a múlt idejű nyelvtanulót hevesen kritizálták azért is, mert nem szimulálta a fiziológiai részleteket. A bemeneti és kimeneti igék kódolása, mint fonetikus hármasok, úgynevezett "Wickelfeaturek", talán a legkevésbé hihető részlet.

A konnekcionizmus kritikusai hibaként kezelik ezeket a disanalógiákat, de a PDP csoport jól tudta, hogy a "fiziológiai" ízlés megállja a helyét a valóságos részletességben. A PDP-biblia 2. kötetének 20. fejezete leírja, hogy a mesterséges neurális hálózatok milyen módon nem olyanok, mint a valódi agyak. A bevezető is elhatárolódik attól, hogy a fiziológiai plauzibilitás a cél:

> "Bár a PDP-modellek vonzerejét határozottan növeli fiziológiai plauzibilitásuk és idegi inspirációjuk, nem ezek az elsődleges alapjai a számunkra való vonzerejüknek.... A PDP-modellek pszichológiai és számítási okokból vonzanak minket" (McClelland és Rumelhart 1986, 11).

[6] Wilson és Cowan (1972) olyan egyenleteket vezettek le a neuronpopulációk átlagos tüskefrekvenciájára, amelyek lehetővé teszik, hogy a véletlenszerű, sűrű kapcsolatokkal rendelkező neuronpopulációkat aggregátumként kezeljük, és ezek az egyenletek nagyban

megfelelnek a konnekcionista modellekben használtaknak.

A valósághű idegrendszeri részletek hiánya nyilvánvalóan tervezési jellemző volt.

Részben az a helyzet, hogy a gyakorlati szempontok megkövetelik, hogy a modellek ne legyenek túl bonyolultak. A konnekcionisták körében gyakori refrén, hogy hiba túl sok részletet beletenni egy modellbe: "Nem szükséges a konyhai mosogatót beletenni ahhoz, hogy betekintést nyerjünk... csak szimulálni a pokolba a populációkat mindenből, ami a modellben van, esztelen" (J. D. Cowan, idézi Anderson és Rosenfeld 2000). McClelland (2009) amellett érvel, hogy bár a modellezésben az egyszerűsítéseknek ára van, a megértés érdekében egyszerűsíteni kell. Úgy tűnik azonban, hogy a valószínűtlenség mélyebbre hatol, mint a pragmatizmus.

2.3 Az absztrakciós probléma

Egy másik rejtély, hogy a konnekcionisták néha matematikai kifejezésekkel írják le modelljeiket. Smolensky azt állítja, hogy a konnekcionizmus azt vizsgálja, hogy a folytonos (és nem a diszkrét) matematika mit tud felfedni a megismerés természetéről (1991).

Thomas és McClelland a konnekcionista modelleket "az univerzális függvények közelítésében részt vevő statisztikai modellek egy alosztályának" nevezi (2008).

Erre példa Touretzky és Hinton (1988) munkája, amely bemutatja, hogy az elosztott reprezentációk segítségével "olyan munkamemóriát lehet létrehozni, amely sokkal kevesebb egységet igényel, mint a potenciálisan tárolható különböző tények száma" (Touretzky és Hinton 1988). Itt nem tesznek erőfeszítést arra, hogy az általános szerkezeti jellemzőkön túlmenően neurális részleteket is újrateremtsenek. A lényeg egy olyan tulajdonság bemutatása, amellyel az ilyen hálózatok rendelkeznek, függetlenül attól, hogy az egységek mit képviselnek, ugyanakkor a modellt egyértelműen a munkamemória vizsgálatának szánják. Elgondolkodhatunk azon, hogyan képes

mindkettőre.

2.4 A magyarázat problémája

Az utolsó kihívás a PDP-modellek mint magyarázatok státuszát érinti. Green aggódik, hogy "ha a konnekcionista modelleket NEM tekinthetjük a megismerés elméletének, a szó hagyományos tudományos értelmében, akkor felmerül a kérdés, hogy pontosan mik is ezek, és miért kellene figyelmet fordítanunk rájuk" (Green 1998). Green szerint a konnekcionista hálózatok elméletként való értelmezése csak akkor lehetséges, ha azok "a megismerés alapjául szolgáló agyi tevékenység szó szerinti modelljei" (Green 1998). Ezt azonban aláássa a konnekcionista modellek valószínűtlensége.

A klasszikus mesterséges intelligenciában egy olyan számítógépes programot, amely egy kognitív feladatban az emberi teljesítményhez hasonló kimenetet produkál, az adott kognitív feladat elméletének tekintik. Amikor Newell és Simon (1961, 1976) elméleteknek nevezik programjaikat, a deduktív- nomológiai (DN) elméletre (Hempel 1958) gondolnak: "[Egy] elméletként használt számítógépes programnak ugyanaz az ismeretelméleti státusza, mint egy elméletként használt differenciálegyenlet- vagy differenciálegyenlet-halmaznak" (Newell és Simon 1961). A programban szereplő logikai kalkulusnak ugyanaz a státusza, mint a fizikai tudományokban elméletet alkotó törvény- és megfigyelési állításoknak.

A konnekcionista modellek nem elméletek a DN értelmében; nem következtetnek logikusan a viselkedésre, és nem kódolnak törvényszerű szabályszerűségeket. Az 1980-as évek végére a DN-elképzelés már nem volt a tudományos magyarázat elfogadott nézete, de a konszenzus hiánya arról, hogy mi lépjen a helyére, nyitva hagyta, hogy a konnekcionista modellek milyen magyarázatokat adnak.

3. A konnekcionista modellek mint a kognitív mechanizmusok idealizált modelljei

A tudományfilozófia legújabb fejleményei rávilágítanak a fenti problémákra.

3.1 Mechanisztikus magyarázat

A biológiai tudományokban a mechanisztikus magyarázati szemlélet nagyrészt kiszorította a DN-elméletet. A konnekcionisták által érintett szinteket mechanisztikus szinteknek lehet tekinteni (Craver 2007). A mechanisztikus magyarázat egy jelenséget a mechanizmusok többszintű rendszerében helyez el, ahol minden egyes szint korlátozza a szomszédos szinteket, és a szomszédos szintek korlátozzák őket. A mechanisztikus magyarázat magában foglalja annak bemutatását, hogy az alkotó egységek és tevékenységeik hogyan szerveződnek egy jelenség létrehozásához, és a mechanizmus szerepének azonosítását a magasabb szintű jelenségekben.

Az a felvetés, hogy a konnekcionista modellek mechanisztikus magyarázat szempontjából is értelmezhetők, Miłkowski (2013) felveti, és Stinson (2018) továbbfejleszti. Ez a felismerés összhangban van a PDP-csoport kinyilvánított motivációival. Ahelyett, hogy a fiziológiát és a megismerést függetlenként tekintenék, a konnekcionisták inkább azt vizsgálják, hogy a fiziológiai mikrostruktúra milyen módon korlátozza a megismerést. A PDP bibliája felsorolja azokat a korlátokat, amelyeket az idegtudományból vesznek át, többek között: "*Nagyon sok neuron van... A neuronok nagyszámú más neurontól kapnak bemenetet... A tanulás a kapcsolatok módosításával jár... A neuronok úgy kommunikálnak, hogy aktivációt vagy gátlást küldenek a kapcsolatokon keresztül...*" (Rumelhart és McClelland 1986b, 130-32).

Az, hogy a mechanisztikus magyarázatoknak nincs privilegizált szintje, segít megmagyarázni, hogy az egységek miért felelhetnek meg egyetlen neuronnak, neuronpopulációnak vagy magasabb szintű entitásoknak, például fonetikai reprezentációknak. A konnekcionista modellek a mechanizmusok rendszerében tetszőleges számú helyet vizsgálhatnak. Ahogy Churchland fogalmaz: "A hálózati

modellek... fontos módon függenek az elemzés minden szintjéről származó korlátozásoktól..... Mivel a hálózatok a szerveződés teljesen különböző szintjein érvényesülő elveket hivatottak tükrözni, megvalósításuk is különböző léptékű lesz az idegrendszerben" (Churchland és Sejnowski 1990).

3.2 Absztrakció

Az egyszerűség nemcsak a modellek működéséhez, hanem a magyarázathoz is elengedhetetlen. Az egyszerűsítés ára az, hogy amikor egy egyszerűsített modellből következtetést vonunk le, előfordulhat, hogy a modell érdekes tulajdonságai a modellnek a céltól eltérő aspektusaiból erednek, nem pedig abból, ami a modell és a cél közös. A konnekcionista modellek sok szempontból különböznek az agyaktól, ezért elvárható, hogy másképp viselkedjenek. Ez egy példa a tudományos modellekkel kapcsolatos nagyon általános aggodalomra, nevezetesen arra, hogy mely részleteket kell pontosan megragadni ahhoz, hogy a modell tájékoztasson minket a célrendszerről, és melyeket lehet biztonságosan megváltoztatni. Ez a probléma a tudományfilozófiában sok munkával foglalkozott.

Az egyszerűsítés egyik fajtája az, amit Cartwright (1989) absztrakciónak nevez. Az absztrakt modellek eltávolítják a részleteket, hogy a kisszámú változó hatása könnyebben vizsgálható legyen önmagában. A túl sok részlet elvonása hibához vezethet, ha a változók között összetett kapcsolatok vannak, így az egyes változók külön-külön történő vizsgálata nem egyértelműen informatív az összképet illetően. Mindazonáltal gyakran ez teljesen jogos. Összehasonlítható azzal, ahogyan a kísérleteknek kontrollálniuk kell a változókat ahhoz, hogy értelmezhetőek legyenek. Kompromisszumot kell kötni a sok kontrollálatlan változót tartalmazó naturalisztikus terepkísérletek és a könnyebben értelmezhető, de kevésbé külső érvényességű laboratóriumi kísérletek között.

A trükk az, hogy kitaláljuk, mely részletek számítanak. Morgan (2002, 2003) az anyagszerűség fontossága mellett érvel: az azonos anyagok megosztása a kísérleteket közelebb hozza a célpontokhoz, mint a modelleket, ami a kísérleti rendszereket nagyobb

valószínűséggel teszi a releváns tulajdonságok megosztására. Parker ellenzi Morgan értékelését a számítógépes

szimulációkat "csak matematikai modellezési gyakorlatoknak" (Parker 2009), azzal érvelve, hogy a szimulációk fizikai modellek. Rámutat arra, hogy az időjárás-előrejelzésben a modellezők jobban tudják beállítani a releváns kezdeti feltételeket egy számítógépes szimulációban, mint egy laboratóriumi modellben, amely ugyanazokat az anyagokat használja, mint a valós időjárási rendszerek. A szimulációk tehát jobban képesek az időjárás előrejelzésére, mint az "azonos anyagú" laboratóriumi modellek, és a kulcs az, hogy "a kísérleti és a célrendszerek valóban hasonlóak voltak-e azokban a vonatkozásokban, amelyek relevánsak, tekintettel a célrendszerrel kapcsolatos konkrét megválaszolandó kérdésre" (Parker 2009, 493). Egy másik meteorológiai modellben a légkör jelentős térfogatait egy rácsháló homogén pontjaiként kezelik, míg a rácsháló felbontásánál kisebb léptékű komplex dinamikára vonatkozó méréseket egyetlen paraméterértékkel közelítik. A finomabb felbontású ismert részletek figyelmen kívül hagyása pontosabb időjárás-előrejelzéshez vezet, mintha ezeket a részleteket a modellben szerepeltetnék (Norton és Suppes 2001, 95-96). (Lásd még Küppers és Lenhard 2004.)

Hasonlóképpen a kognitív modellezésben, amikor a cél egy kognitív ágens viselkedésének előrejelzése, az olyan pragmatikus szempontok, mint a pontosság maximalizálása, elsőbbséget élveznek a rendszer finomabb részleteinek modellezésével szemben. Ezt mutatja a támogató vektor gépek népszerűsége az ImageNET képfelismerési kihíváson (Russakovsky et al. 2015), amelyek alig vagy egyáltalán nem tesznek erőfeszítést az emberi vizuális feldolgozás utánzására.

Giere (2004) és Godfrey-Smith (2006) a modellek reprezentációs szerepére összpontosítanak, és hasonlóképpen azzal érvelnek, hogy a modellek és a célpontok a releváns tekintetben hasonlóak, ami igazolja az egyikből a másikra való következtetéseket. Az, hogy mely hasonlóságok relevánsak, a kontextustól függ: "a

tudósok folyamatos folyadékmodelleket használnak a víz reprezentálására a

a folyadékáramlás tanulmányozására, és molekuláris modelleket használnak a víz ábrázolására a Brown-mozgás tanulmányozásához" (Giere 2004).

Ha a múlt idejű tanulók célja az lett volna, hogy modellezzék, hogyan reprezentálódnak az igék az agyban, vagy hogy részletesen szimulálják a ragozást, akkor a bemeneti és kimeneti igék kódolásának módja lett volna releváns, és a Wickel-funkciók használata nem lett volna helyénvaló. Rumelhart és McClelland azonban azt szerette volna látni, hogy a strukturált szabálykövető viselkedésnek tűnő viselkedés megvalósítható-e anélkül, hogy ezt a struktúrát beépítenénk. Az igék reprezentálásának módját e cél szempontjából irrelevánsnak tartották.

Winsberg kiemeli az érvek fontosságát, annak bizonyítására, hogy a tudósok által "az adott eszközök manipulálásából kapott eredmények megfelelően bizonyító erejűek az őket érdeklő rendszerek osztályát illetően" (Winsberg 2009, 577).
Ezek az érvek nem csak a hasonlóságon alapulnak, hanem azon is, hogy ismerjük a jó modellek építésének módját, amely a múltbeli sikerekből származik, amikor ugyanazt a modellezési trükköket használtuk.

Batterman (2001, 2002) leírja, hogy az olyan matematikai eszközök, mint a renormalizációs csoportok használata hogyan függ attól, hogy az egyes problémákat minimális modellekre redukáljuk. A Batterman által "aszimptotikusnak" nevezett módszerek képesek megmagyarázni azokat az univerzális, stabil jelenségeket, amelyek közösek például a fázisátmenetek kritikus pontja közelében lévő mikroszerkezetileg különböző folyadékokban, valamint a ferro- és paramágneses állapotok között átmenő mágnesekben (Batterman 2001, 38). Ezek a módszerek nemcsak magyarázatot kínálnak ezekre az univerzális jelenségekre, hanem azzal, hogy "megmondják, hogy a különböző részletek milyen (és miért) irrelevánsak az érdeklődésre számot tartó viselkedés

szempontjából, ugyanez az elemzés azonosítja azokat a fizikai

olyan tulajdonságok, amelyek a vizsgált univerzális viselkedés szempontjából relevánsak" (Batterman 2001, 42).

Valami hasonló aszimptotikus magyarázat jelenik meg Fuhs és Touretzky (2006) modelljében a térbeli memória útvonal-integrációjáról. Modelljük arra keres magyarázatot, hogy a labirintusokban navigáló patkányok hogyan képesek hatékony utakat találni a célpontokhoz, függetlenül a korábban megtett utaktól, valamint a rácssejtek tüzelési mezőiben található sajátos hatszögletű mintázatoktól. A hexagonális mintázatok lehetséges magyarázataként Fuhs és Touretzky kimutatta, hogy "hexagonálisan elosztott aktivitásdudorok spontán módon keletkezhetnek a neuronok lapján egy spin glass típusú neurális hálózati modellben" (Fuhs és Touretzky 2006, 4266). A spin glass modellekben minden egység egy többdimenziós rácsban kapcsolódik a legközelebbi szomszédjaihoz. Ez a hálózati struktúra lazán az entorhinális kéreg helyi struktúráján alapul, ahol rácssejtek találhatók, abból a feltételezésből kiindulva, hogy a dendritek szorosan egymáshoz vannak csomagolva. Ha az egyenletes méretű körök vagy hengerek szorosan egymás mellé vannak pakolva, a legnagyobb sűrűségű elrendezés egy hatszögletű mintázat. Ez attól függetlenül igaz, hogy távközlési kábelekről vagy idegpályákon futó dendritekről van szó. Fuhs és Touretzky (2006) a szoros csomagolásra vonatkozó ezen tény alapján indokolja, hogy a rácssejtek konnekcionista modelljében az egységek hatszögletű mintázatban helyezkednek el, annak ellenére, hogy a valódi dendritek nem tökéletesen hengeresek és nem is egyenletes méretűek. Magyarázatuk azon múlik, hogy *nem* használtak-e pontosabb, részletesebb modellt, mert a feltételezés nélkül, hogy a dendritek egyenletes hengerek, a szoros csomagolásra vonatkozó geometriai tényt nem lehetett volna alkalmazni.

3.3 Idealizáció

Cartwright (1989) definíciója szerint az idealizálás olyan részleteket ad hozzá vagy változtat meg, hogy az idealizált modell olyan tulajdonságokkal rendelkezik, amelyek a célrendszerben nincsenek jelen. Laboratóriumi kísérletekben

az idealizációk esetleg kényelmesebb anyagokat helyettesítenek, vagy a számítások megkönnyítése érdekében valószínűtlen értékeket rendelnek a változókhoz. A konnekcionista modellezésben az idealizálás legegyértelműbb esetei a backpropagation és a determinisztikus aktiválási függvények. Első pillantásra úgy tűnik, hogy a rossz részletek beillesztése, szemben az irreleváns részletek egyszerű eltávolításával, rosszabb modellt eredményezne, de ez általában nem így van.

Számos szerző hasonlította az idealizációkat a fikciókhoz, és azt javasolta, hogy a modelleket ugyanúgy értelmezzük, mint az irodalmat. Mäki (2012) elemzése más irányba megy. Mäki amellett érvel, hogy a látszólag hamis idealizációkat többféleképpen is igazként lehet értelmezni. Egyes tényezők elhagyása "elhanyagolhatósági feltételezésnek" is tekinthető, vagyis annak, hogy ezeknek a tényezőknek elhanyagolható hatásuk van, tekintettel a modell tervezett céljaira és célközönségére (Mäki 2012, 222). Az "alkalmazhatósági feltételezések" a modell tervezett felhasználását olyan területekre korlátozzák, ahol a kihagyott tényezőknek elhanyagolható hatása van (Mäki 2012, 225). Másfajta feltételezések azzal az indokkal védhetik egy idealizált modell használatát, hogy az idealizálás révén a modell jobban követhetővé válik, vagy pedagógiai célokra alkalmasabbá válik (Mäki 2012, 228-230). Ezek a feltételezések nem mindig kerülnek explicit módon megfogalmazásra.

Bizonyos esetekben a backpropagation használata a konnekcionista modellekben a követhetőségi feltételezéssel volt igazolható, mivel egy ideig ez volt az egyetlen ismert módszer a súlyok frissítésére, amely garantáltan konvergált. Más esetekben a backpropagation elhanyagolhatósági feltételezéssel indokolható. Például a NETtalk (Sejnowski és Rosenberg 1986) esetében a backpropagation a modell célját tekintve nem problémás, mivel a céljuk az, hogy megmutassák, hogy egy angol szavak kiejtésére képes

rendszernek nem kell bonyolult szabályrendszert kódolnia. E célból jogosan feltételezhetjük egyszerűen, hogy a

az agynak van valamilyen módja a hibajelek továbbítására, anélkül, hogy aggódnánk amiatt, hogy ez pontosan hogyan történik. Az, hogy a hibajelek milyen útvonalakon haladnak, nem befolyásolja azt, hogy mit vizsgálnak. Ezzel szemben Suri és Schultz (2001) a tanulási mechanizmusok modellje, így a hibajelek terjedésének módja rendkívül fontos. Az ő modelljükben nem használnak backpropagationt; ehelyett a bazális ganglionok anatómiáját reprodukálják bizonyos részletességgel, és csak azokat az útvonalakat veszik figyelembe, amelyek léteznek az agyban, és amelyeken keresztül a visszajelzésekről ismert, hogy ténylegesen közlekednek.

A furcsa "'fiziológiai' ízű" kifejezést úgy is értelmezhetnénk, hogy a konnekcionista modellek a kognitív mechanizmusok idealizált modelljei. Ha a lényegtelen részleteket eltávolítjuk, másokat pedig idealizálunk, az agykéreg egyszerű tanulási egységek összekapcsolt hálózata.

3.4 Discovery

A modellek sokféle célt szolgálnak a tudományban, és sokféle stratégiát lehet alkalmazni a mechanizmusok keresése során. Anderson és Rosenfeld (2000) konnekcionizmus-története bemutatja, hogy a konnekcionista modellezők között mindig is nagyon eltérő megközelítések voltak a tekintetben, hogy mennyi fiziológiai részletet tartalmazzon, és mik a célok. E célok között mérnöki, matematikai, pszichológiai és idegtudományi kérdések is szerepelnek. A különböző ismeretelméleti szerepekre szánt modellek különböző jellemzőket igényelnek.

Steinle (1997, 2002) szerint a kutatási projekt különböző szakaszaiban végzett kísérleteknek általában különböző ismeretelméleti céljaik vannak, ami azt jelenti, hogy különböző típusú kísérleteket végeznek. Például a korábbi feltáró kísérletek általában a paraméterértékek sokkal több kombinációját próbálják ki a potenciálisan értelmes

paraméterek keresése során.

korrelációkat, míg a későbbi "elméletvezérelt" kísérletek nagy pontosságú berendezéseket használnak, és "jellemzően a különböző lehetséges kimenetelekre vonatkozó egészen konkrét elvárásokkal végzik" (Steinle 1997).

Steinle elemzése a modellekre is érvényes. A kutatási projekt különböző szakaszaiban használt modelleknek általában különböző ismeretelméleti céljaik vannak, és ennek megfelelően eltérőek lehetnek abban a tekintetben, hogy mennyire idealizáltak vagy specifikusak kell lenniük e célok elérése érdekében. Az ismeretelméleti célok közötti különbség tükröződik a NETtalkban használt tanulási mechanizmusok részletességében mutatkozó különbségekben a bazális ganglionok modelljeihez képest.

4 A modellek ismeretelmélete és metafizikája

Ez a szakasz Irvine (2014) szellemében folytatódik, ahol a kognitív idegtudományok számítógépes modellezési gyakorlatát figyelembe véve problematizálta és felülvizsgálta a modellek és szimulációk irodalmából származó állításokat. További példák a neurokognitív tudományok számítógépes modellezésével kapcsolatos munkákra: Kaplan (2011), amely amellett érvel, hogy az idegtudományban a számítógépes magyarázatok mechanisztikus magyarázatok; Chirimuuta (2018), amely amellett érvel, hogy a számítógépes idegtudományban "számos példája van a matematikai, nem oksági magyarázatnak"; és Stinson (2018), ahol amellett érvelek, hogy a konnekcionista kognitív modellek a tendenciák logikájával magyaráznak, szemben a klasszikus AI által a legjobb magyarázatra való következtetés használatával.

Ezekben a példákban közös, hogy az ok-okozati és ontológiai kérdésekkel foglalkoznak, ellentétben a modellek és szimulációk irodalmára jellemző reprezentációkra való összpontosítással (lásd Suárez 2003; Giere 2004; Weisberg 2012; Frigg and Nguyen 2016). Ennek talán az az oka, hogy a kognitív tudósok

megtanulták, hogy némi gyanakvással tekintsenek a reprezentációkhoz mint magyarázatokhoz való folyamodásokra, és inkább azon aggódnak, hogy

a feltételezett reprezentációk megszerzik és továbbítják tartalmukat. A reprezentációnak ez a kritikusabb felfogása és kauzális-mechanisztikus hajlama hasznos lehet a modellek és szimulációk irodalmában.

A szakirodalomban három nyitott kérdés az, hogy hogyan kell megítélni a releváns hasonlóságot, hogyan indokolt a modellekből a célokra való következtetés, és mi a modellek metafizikai státusza. Az első két kérdéssel kapcsolatban úgy tűnik, hogy a tudomány jelenlegi állása szerint esetről esetre kell eldönteni, hogy mely hasonlóságok relevánsak (Parker 2009), és hogy érvekkel kell igazolnunk a modellválasztásunkat (Winsberg 2009). Szükség van a modellezők által e döntések meghozatalakor alkalmazott kritériumok mélyebb elemzésére.

Godfrey-Smith (2009) a modellek metafizikai státuszának problémáját veti fel. Megjegyzi, hogy a modellrendszerek "bizonyos értelemben ugyanolyanok, mint a célrendszerek, amelyek megértéséhez a modellek segítségével segítséget nyújtanak" (Godfrey-Smith 2009), de elutasítja azt, amit a tudósok mondanak arról, hogy a bolygók, populációk vagy gazdaságok még laza értelemben is "a számítógépen belül vannak". Ellenáll annak, hogy a számítógépes modelleknek tárgyi mivoltot tulajdonítsunk, hogy azokat "árnyékos, kiegészítő, megragadható dolog[ok]nak" tekintsük, és helyesli a Deena Weisbergnek tulajdonított megjegyzést, miszerint a matematikusok platonizmusa "népi ontológia" (Godfrey-Smith 2009). Godfrey-Smith úgy írja le a platonista nézetet, hogy a modellt egy matematikailag vizsgálható absztrakt entitásnak tekinti, majd az absztrakt tulajdonságok leképezését követeli meg a célpont fizikai tulajdonságaira.

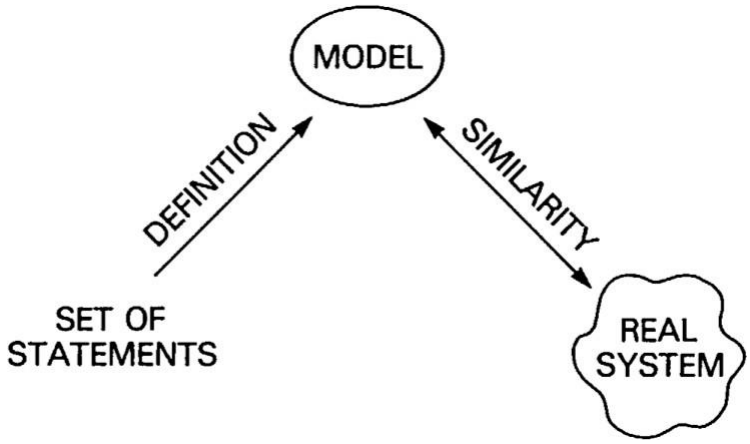
A platonista nézetet hallgatólagosan elvetik, mert a Harmadik Ember problémájába ütközik, mivel feltételezi az absztrakt tárgyak független valóságát. De ha a modellek nem olyan tárgyak, amelyek közvetlenül összehasonlíthatók a

céltárgyakkal, akkor továbbra is fennáll a kérdés, hogyan tájékoztathatnak bennünket a konkrét dolgokról. A reprezentációk és a fikciók túlságosan rugalmasak;

a fikcióban bármi megtörténhet, ezért nem korlátozza megfelelően a valós célokra való következtetéseket. Ahogy Frigg és Nguyen érvel: "Szinte bármit elképzelhetünk szinte bármilyen tárgyról, de ha nincsenek olyan kritériumok, amelyek megmondják, hogy ezek közül a képzelgések közül melyeket kell igaznak tekinteni a célpontra vonatkozóan, akkor ezek a képzelgések nem adnak engedélyt semmilyen helyettesítő érvelésre" (2016). A metafizikai probléma és a következtetési probléma tehát feszültségben áll egymással.

Winsberg (2010) hasonlóképpen megjegyzi, hogy "a szimuláció gyakorlói" azt az elképzelést támogatják, hogy a szimulációk szó szerint utánozzák a célrendszereket, így a folyadékdinamika szimulációja úgy tekinthető, mint egy kísérlet egy "virtuális szélcsatornában" (Winsberg 2010, 35). Winsberg azonban felveti azt a problémát, hogy "vajon egy szimuláció *megbízhatóan* utánozza-e, milyen mértékben és milyen körülmények között az érdeklődésre számot tartó fizikai rendszert" (Winsberg 2010, 37).

A modellek reprezentációjának általános feltételezése az, hogy a modell és a világ közötti kapcsolat a hasonlóságon alapul, Giere (1988) diagramját követve, amelyet itt az 1. ábrán mutatunk be. Winsberg megjegyzi, hogy a modellek és a célok közötti kapcsolatnak "sokkal bonyolultabbnak kell lennie, mint a mimikri" (Winsberg 2010, 39). Hogy mi lehet ez a bonyolultabb kapcsolat, az mindhárom kérdés szempontjából kulcsfontosságú.

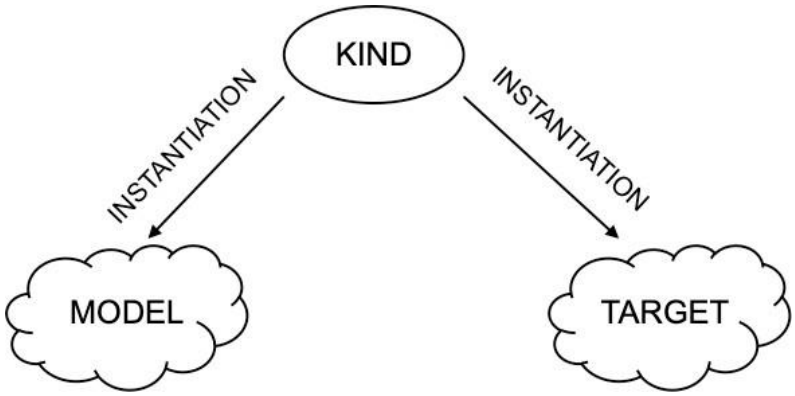


1. ábra. Az elmélet, a modell és a valóság közötti kapcsolat, Giere (1988, 83) alapján.

Frigg és Nguyen a DEKI beszámolójában egy bonyolultabb jelöltkapcsolatot kínál. A DEKI szerint a modelleket úgy értelmezik, mint amelyek az érdeklődésre számot tartó tulajdonságok egy halmazát példázzák, amelyeket a célpontnak tulajdonított tulajdonságokra képeznek le. A modell és a célpont közötti út mentén több állomás hozzáadása megoldja a közvetlen reprezentációs számlák számos problémáját (lásd Frigg and Nguyen 2016), de megtartja a reprezentációs számlák fő gyengeségét, mivel a tulajdonságkészletek közötti leképezés kevéssé biztosítja, hogy az egyik megbízhatóan utánozza a másikat.

4.1 A modellek reprezentációs elszámolásának alternatívája

A konnekcionista modellezésben való következtetésről szóló elemzésem Stinson (2018) alapján kiterjeszthető a modellek alternatív beszámolójára. Ez a számla szilárdabb kapcsolatot biztosít a modellek tulajdonságai és a célok között, és legitimálja a tudósok nézeteit a modellek ontológiájáról. Azt állítom, hogy a konnekcionista modellekből közvetve következtetéseket vonunk le a célpontjaikra olyan fajtákon keresztül, amelyeket mind a modell, mind a célpont példáz. Ezt a kapcsolatrendszert szemléltető diagramot a 2. ábra mutatja be.



2. ábra. A modell, a cél és a fajta közötti kapcsolat.

Eszerint a modellezők által a modellekből a célpontokra levont következtetések a következő szerkezetűek:

P1AT célrendszer a K típus egy példánya.

P2AM modella K egy példánya.

EzértT-nek hasonlónak kell lennie az M-hez.

A konnekcionista modellek esetében T a vizsgált agy vagy kognitív rendszer, M pedig a konnekcionista modell.

Ez az extra lépés, hogy a modellt először egy fajtához, majd aztán a fajtát a célponthoz társítjuk, több okból is ígéretes. Az egyik az, hogy jobban értelmezi az idealizálást, mint a modellezés reprezentációs beszámolói. Egy jó modell gyakran nagyon minimális, csak az érdekes tulajdonságokat tartalmazza, és kevés egyéb részletet. Ha a hasonlóság lenne a modell-cél relációk kritériuma, akkor a több részlet jobb lenne, nem pedig kevesebb. De ha egy faj jellemző tulajdonságainak megragadása a cél, akkor az idealizált modellek pontosan azok, amelyekre törekedni kell. A hasonlóság az, amit következtetésként le akarunk vonni, nem pedig az, amire a modellépítés során törekszünk.

A második, hogy szükséges útmutatást adhat arra vonatkozóan, hogy mely hasonlóságok a relevánsak a modellben való rögzítéshez. A modellnek azokat a tulajdonságokat kell utánoznia a céltárgyból, amelyek jellemzőek arra a fajtára, amelyhez mindketten tartoznak. Amíg van módunk a fajták kiválasztására, ez konkrét útmutatást ad arra vonatkozóan, hogy melyek a releváns hasonlóságok: ez az, amire a reprezentációs számlák nem voltak képesek.

A számla egyéb szép tulajdonságokkal is rendelkezik. Jól illeszkedik Godfrey-Smith megfigyeléséhez, miszerint a modellek néha egy céljelenség egy esetét írják le, majd csomópontként működnek, lehorgonyozva a "tényleges világban" előforduló összes esetet (Godfrey-Smith 2009). A csomópont az a fajta, amelyet a modell megragad. A céltárgy nélküli modellek hasonlóképpen úgy kerülnek figyelembe vételre, hogy a fajtát általánosított céltárgyként kezeljük.

Ami még hiányzik ebből a beszámolóból, az a fajták elhatárolásának módja, hogy megmondjuk, mely fajtákhoz tartozik egy célpont, és az univerzálisok problémájának megoldása. Az egyik hiányzó darabot Khalidi (1998, 2013) tágabb értelemben vett fajtákról alkotott nézete nyújthatja. Khalidi amellett érvel, hogy az olyan tudományos fajtákat, mint a parazita, a folyadék vagy a skizofrénia, "valódi fajtáknak" kell tekinteni[7] , mert "olyan dolgokat fedeztünk fel róluk, amelyeket semmiképpen sem feltételeztünk, amikor először bevezettük őket" (Khalidi 1998). A fajtáknak ez a leírása nem feltételez sem esszenciális természetet, sem szigorúan hierarchikus kapcsolatokat a fajták között. Mint ilyen, eléggé promiszkuzív ahhoz, hogy a legtöbb olyan jelenséget befogadja, amelyet modellezni szeretnénk. De mivel a fajok tagjainak nem önkényes közös tulajdonságaik vannak, a fajok alapot nyújtanak arra a következtetésre, hogy a tagok valószínűleg rendelkeznek a fajra jellemző tulajdonságokkal, amelyek közösek

egymással.

[7] Khalidi (2013) a "természetes fajok" kifejezést használja, de egy beszélgetésben azt mondja, hogy bárcsak "valódi fajoknak" nevezte volna őket.

Khalidi fajtái robusztusabbá tehetők, ha összekapcsoljuk őket Andersen (2017) Dennett (1991) "valódi minták" információelméleti frissítésével. A valódi minta olyan, amely "megbízhatóan kiválasztható és nyomon követhető az időben, és amely lehetővé teszi, hogy a véletlennél jobb előrejelzéseket tegyünk" (Andersen 2017). A jelenségek olyan gyűjteménye, amely Andersen szerint valódi mintázatot nyilvánít meg, Khalidi szerint egy valódi fajta tagjainak számítana.

Andersen megjegyzi a minták "pazarló" voltát, mondván, hogy "rengeteg különböző módja lehet az ilyen minták kiválasztásának, amelyek prediktív fogást adnak nekünk a rendszerről" (Andersen 2017), de ahogyan Khalidi fajtáinak promiszkuitása sem lehet aggasztó, mivel nem feltételeznek semmit az esszenciákról vagy a fajok hierarchiájáról, úgy Andersen mintáinak pazarlósága sem lehet aggasztó, mert "a realizmus mértéke nagyon-nagyon minimális" (Andersen 2017). Mind Khalidi, mind Andersen amellett érvel, hogy az az aggodalom, hogy ez túl sok fajt vagy mintát enged meg, túlzó. Azokat a kritériumokat, hogy a fajtákat vagy mintákat megbízhatóan ki lehet választani, nyomon lehet követni, és hasznos előrejelzéseket lehet készíteni, a jerry-rigged fajták nem teljesítik.

Azzal az aggodalommal szemben, hogy a minták epifenomenálisak, Andersen azt állítja: "A minták túlnyomó többsége ellenténYszerűen robusztus, mivel mikrofizikai részleteikben minden egyes token-instanciában különbözhettek volna anélkül, hogy ezáltal megváltoztatták volna a relátum oksági profilját" (2017). Az, hogy az oksági folyamatok idealizált és egyszerűsített modelljei gyakran a leghasznosabbak annak kitalálásához, hogy hogyan működnek ezek a folyamatok, lenne

rejtélyes, ha nem lenne az a helyzet, hogy ezek a minták valamilyen értelemben valóságosak, ami túlmutat a mikrofizikai részleteik valóságán.[8]

Mind Khalidi, mind Andersen legalább egy minimális valóságot állít az ő fajtáik/mintáik számára. Andersen szerint a minták az oksági nexus részét képezik, és "a magasabb szintű okok ugyanolyan valóságosak, mint az alacsonyabb szintű okok" (Andersen 2017). Az ő deke-je az univerzálisok problémájának megkerülésére az, hogy ami valóságos, az "az oksági nexus és az abban instanciált minták, amelyek információs struktúrájúak, de ahol maga az információ valami másnak a struktúrája, nem pedig egy reifikált extra szubsztancia" (Andersen 2017).

A tudósok népi ontológiáját támogatva egy lépéssel tovább is mehetünk, és úgy értelmezhetjük ezeket a valóságra vonatkozó állításokat, mint amelyek további árnyékos dolgokat implikálnak. Mielőtt azonban beindulna a térdreakció, hogy ez az univerzumok problémájába ütközik, nézzünk meg néhány újabb fejleményt a metafizikában, ahol tiszteletre méltó lehetőségek állnak rendelkezésre arra, hogy az univerzumokat bizonyos értelemben konkrétnak tekintsük.

Hennig (2014) Baxter (2001) alapján egy lehetséges megoldást kínál az univerzálisok problémájára, amely szerint a fajtáknak konkrét *aspektusaik* vannak, így bizonyos értelemben "ugyanolyanok", mint a fajta példányai. Hennig a következőképpen foglalja össze a számvetést: "az, hogy Szókratész instanciálja a fajta ülő dolgot, azt jelenti, hogy Szókratésznek van egy olyan aspektusa, amely egyben a fajta ülő dolog aspektusa is. Ez az aspektus kétféleképpen írható le: (1) mint Szókratész qua ülő dolog vagy (2) mint ülő dolog qua Szókratész által instantiált dolog" (Hennig, 2014). Hennig a következőképpen pontosít:

[8] Khalidi és Andersen egyaránt elutasítja azt a felvetést, hogy Kim kauzális kizárási érve problémákat okozhat a promiszkuitás/professzivitás fajtáinak vagy mintáinak valósága szempontjából.

> Az, hogy Szókratész az "ülő dolog" fajtát példázza, azt jelenti, hogy van egy ülő dolog, amely azonos Szókratésszel. Ez a dolog az egyik aspektusa. Szókratész az ülő dolog egy példánya, és az ülő dolog Szókratész egyik aspektusa. Az aspektus nem egy harmadik entitás, amely közvetít Szókratész és az egyetemes "ülő dolog" között; csak két dolog van: Szókratész és az aspektus. (Hennig, személyes közlés)

Az aspektusok, ellentétben a platóni univerzálékkal, konkrétak és a világban vannak. A "Szókratész qua ember" aspektusnak húsa és csontjai vannak.

Ez a rövid kitérő a kortárs metafizikába azt mutatja, hogy léteznek olyan tiszteletreméltó lehetőségek, amelyek lehetővé teszik számunkra, hogy komolyan vegyük a tudósok nézeteit a modellek valóságáról. Ha ezeket a darabokat összerakjuk, a fajta és a cél közötti kapcsolatot úgy értelmezhetjük, mint az aspektus és a példány közötti kapcsolatot, a modell és a cél közötti kapcsolatot pedig úgy, mint ugyanazon aspektus két példánya közötti kapcsolatot. Csábító úgy gondolkodni a modellekről, mintha kettő lenne belőlük, az ideális és a példányosított. Az ideális az, amiről a tudósok azt gondolják, hogy valóban a célrendszerben, és valóban a számítógépben van. Az instanciált modell az az eszköz, amelyet az ideális eléréséhez használunk. Az előbbi a szempont. Az utóbbi a példány.

A modellezés gyakorlata kiválasztja a célpont egy aspektusát, amelyet meg kell vizsgálni, majd ennek az aspektusnak egy olyan példányát építi fel, amely kényelmesen manipulálható. Egy olyan matematikai modell, mint a Hodkin-Huxley-egyenlet, elég közel áll ahhoz, hogy a vizsgált aspektus legyen, míg az olyan modellek, amelyek a célpont anyagát más anyaggal helyettesítik (mint az építészeti modellek, a modellorganizmusok és az analóg modellek), olyan példányok, amelyek további, a

kérdéses fajtára nem jellemző aspektusokkal rendelkeznek.

Most is ülök, miközben írok, így én is az "ülő dolog" egyik példánya vagyok. Ülő helyzetem révén a "Szókratész kvázi ülő dolog" modelljeként működhetek. E modell alapján feltételezhetném, hogy Szókratész lábujjai is hajlamosak lehettek elaludni, miután túl sokáig ült egy kemény széken. De vannak más aspektusaim is, amelyek nem közösek Szókratésszel. Nem lenne bölcs dolog arra következtetni, hogy Szókratész is általában teát iszik, és angol nyelvű beszélgetéseket hallgat angolul, miközben egy kemény széken ül. Az "ülő dolog" mellettem lévő példányai osztoznak ezek közül néhány aspektusban, de abban az értelemben osztoznak bennük, hogy a "Propeller Coffee-ban dolgozó személy" példányai. Az "ülő dolog" jobb modellje a zajoktól és forró italoktól elszigetelt lenne.

A számítási modellek helyzete egy kicsit árnyaltabb. Bizonyos szempontból olyanok, mint a matematikai modellek, mivel közel állnak ahhoz, hogy tiszta szempontok legyenek. De ahogy Parker állítja, a számítási modellek egyben fizikai modellek is, amelyeknek saját tulajdonságaik vannak (például, hogy tranzisztorokkal készülnek). Bizonyos körülmények között ezek a más szempontok lényegtelenné tehetők, de nagy mágnesek jelenlétében vagy vízbe merülve a számítási modellek elektronikus eszközként mutatják meg színüket. Egy számítási modell kvázi tranzisztorokkal készült, nem biztos, hogy informatív a megismerésről, de egy számítási modell kvázi konnekcionista hálózatnak annak kellene lennie.

A számítógépeket az teszi olyan hasznossá a modellezésben, hogy úgy tervezték őket, hogy képesek legyenek az általad választott szempontot feltárni, miközben elszigetelik azt a szempontot a többi aspektusuktól (tranzisztorokkal készültek, notebook méretűek, Kínában gyártják őket stb.). Ha csak bizonyos kimeneti folyamokat veszünk figyelembe, például képeket, nyomtatásokat vagy bizonyos fájlokat (szemben a CPU

hőmérsékletének mérésével, vagy azzal, hogy megnézzük, mi történik, ha kalapáccsal ütjük), és feltételezzük, hogy a fordítókód értelmezi ezt a kimenetet, egy programozónak

a számítógépet a legkülönfélébb szempontok egy példányává teheti. Másfajta modelleket, mint például a genetika területén a gyümölcslegyeket, szintén azért választják, mert egy adott szempont vizsgálatát könnyebben megvalósíthatóvá (gyorsabbá, olcsóbbá, etikusabbá) teszik, mintha az adott szempontot magán a célszemélyen vizsgálnák. Ahhoz, hogy egy modell minimálisan megfelelő legyen, a modellnek az érdeklődésünk tárgyát képező szempontnak egy példányának kell lennie. Az, hogy a modell és a céltárgy azonosak abban az értelemben, hogy egy aspektuson osztoznak, szankcionálja az egyikből a másikra való következtetéseket.

A modellépítés kiindulópontja annak a K fajtának a meghatározása, amelyhez a vizsgált rendszer tartozik, és amelynek vizsgálatára a modellt tervezik. A következtetés erősségét befolyásoló egyik tényező az, hogy K egy valódi, robusztus, általánosításokat fenntartani képes fajtáról van-e szó. A konnekcionista modellezésben a legáltalánosabb K-k esetében a modell kiválasztása egyenlő azzal a fogadással, hogy a megismerés szempontjából releváns általánosítások egy része a hálózat szintjén működik. Egy másik tényező az, hogy az M a K reprezentatív példánya-e. A K-kre nem jellemző tulajdonságokkal rendelkező modellek reprezentatívabbak. Végül, a következtetés attól függ, hogy T is K fajtájú-e. Ha K egy valódi fajta, és M K minimális példánya, akkor bármit is találunk K fajtájáról M vizsgálatával, annak - minden más esetben - igaznak kell lennie T-re is, feltéve, hogy T K-hoz tartozik. Még mindig előfordulhat, hogy T atipikus releváns szempontból, így nem rendelkezik ugyanazokkal a tulajdonságokkal, mint M, annak ellenére, hogy ugyanahhoz a K fajtához tartozik.

A modelleknek a neurokognitív tudományokban való figyelembevétele a modell és a céltárgy kapcsolatának összetettebb bemutatását, a releváns hasonlósággal kapcsolatos kérdésekre adott jobb válaszokat, valamint a modellek metafizikájának

részletesebb kifejtését tette szükségessé. A modellek újszerű episztemológiájának és metafizikájának puszta körvonalait itt a konnekcionista modellezés következtetéseinek elemzéséből vázoltuk fel.

5 Következtetések fajtákon keresztül a számítógépes kognitív tudományban

Nézzük meg, hogyan működik ez a számla a gyakorlatban, néhány példán keresztül. Marr (1969) kisagyról szóló elméletében a kiindulópontok néhány alapvető anatómiai ismeret a kisagyban található sejttípusokról és a közöttük lévő kapcsolatok mintázatáról és számáról; az a hipotézis, hogy a kisagy funkciója a motoros készségek tanulása; és a mesterséges intelligencia irodalmában akkoriban aktuális elképzelések a funkcióelemzésről (Marr 1969, 469). Felveti, hogy "a mohaszál-granuluszsejt-Purkinje-sejt elrendeződés mintafelismerő eszközként működhet", ahol a "mohaszál-granuluszsejt artikuláció lényegében mintaelválasztó" (Marr 1969, 440). Marr ezután matematikailag levezeti a kodonok méretére és más mértékekre vonatkozó korlátozásokat.

Ebben az esetben Marr a kisagy funkcionális anatómiájától absztrahál egy általános K-fajta, amelyet a sejttípusok közötti kapcsolatok száma és típusa határoz meg, a laza határokat meghatározó korlátozásokkal. A matematikai levezetések feltárják K tulajdonságait, és ezeket a tulajdonságokat a célrendszerre, a kisagyra vonatkozó hipotézisként alkalmazzák.

Hinton (1984) az elosztott reprezentáció egy másik korai tárgyalásában leírja, hogy a ritkán kódolt, elosztott reprezentációk milyen tulajdonságokkal rendelkeznek, mint például a hatékony adattárolás, a tartalomcímezhető memória és az automatikus általánosítás. Ezeket a tulajdonságokat mind formális levezetéseken, mind egyszerű konnekcionista modelleken keresztül állapítja meg. Hinton amellett érvel, hogy valahányszor "absztrakt modelleket valósítunk meg az agyban elosztott reprezentációk segítségével", ezeket a tulajdonságokat "primitív műveleteknek" tekinthetjük (Hinton 1984, 3).

Ebben az esetben K elosztott reprezentációk, M pedig K-t egy egyszerű hálózatban instanciálja, amely megtanulja a szóforma és a jelentés közötti asszociációkat. Itt M-et olyan példánynak választjuk, ahol az érdeklődésre számot tartó tulajdonságokat nehéz elérni: "Ez egy olyan eset, amelyben az elosztott reprezentációk sokkal kevésbé *tűnnek* alkalmasnak, mint a lokálisak, mivel az asszociációk tisztán önkényesek" (Hinton 1984, 3). Az érdeklődésre számot tartó tulajdonságok mégis megerősítést nyernek az M-ben, és a következtetés az, hogy ezek a tulajdonságok a K általános tulajdonságai. Ebben az esetben Hinton stratégiája az, hogy egy olyan modellt választ, amelyről úgy tűnik, hogy nem valószínű, hogy rendelkezik az érdeklődésre számot tartó tulajdonsággal, annak bizonyítására, hogy a tulajdonság általánosítható a fajtagok között.

Az elosztott reprezentációk ezen tulajdonságait az agykérgi rendszerek részletesebb modelljeiben is megerősítették. Babadi és Sompolinsky (2014) például elemzi a ritkaság (kevés neuron válaszol egy adott ingerre) és a kiterjedés (megnövekedett dimenzionalitás az agykérgi rétegben) számítási előnyeit "klaszterezett ingerek általános együtteseiben", "viszonylag egyszerű és biológiailag plauzibilis architektúrákra és dinamikára" összpontosítva (Babadi és Sompolinsky 2014, 1213). Következtetéseket vonnak le a szaglás és a vizuális feldolgozásra, valamint a kisagy mohaszálaira. Billings és munkatársai (2014) a kisagyban a ritka kódolást vizsgálják "a tüskés neuronok biológiailag részletes hálózati modelljei segítségével, amelyek paramétereit kísérleti mérésekkel korlátozták", hogy meghatározzák, hogy a szinaptikus összeköttetés milyen mértékben járul hozzá a hatékony mintaszeparációhoz.

Ezek az esetek egyre specifikusabb K-ket határoznak meg, amelyekhez kisagyi hálózatok tartoznak. Az ezekben a példákban szereplő modellek megerősítik a korábbi tanulmányokban vizsgált általánosabb típusok általános tulajdonságait, és árnyaltabb

képet alkotnak a

specifikusabb fajták, valamint azoknak a határeseteknek a vizsgálata, ahol a fajta tipikus tulajdonságai megszűnnek. Ahogy a modellek egyre részletesebbé és valósághűbbé válnak, úgy erősödnek a modellből a céltárgyra való következtetések, mivel a modell és a céltárgy több közös tulajdonsággal rendelkezik, de a következtetések hatóköre csökken, ahogy a fajta egyre specifikusabbá válik. Babadi és Sompolinsky következtetései a szagló- és látókéregre is vonatkoznak, míg Billings és munkatársai következtetései a kisagyra specifikusak.

Itt egy kontinuum van az általánosabb elméleti modellek és az egyes agyterületek specifikusabb modelljei között. A modellek bárhol elhelyezkedhetnek e két szélsőség között, a következtetés erőssége és az általánosíthatóság közötti kompromisszumokkal.

6. Következtetés

Sokat nyerhetünk azzal, ha a mesterséges intelligencia filozófiáját újra összekapcsoljuk a tudományfilozófiával. Elidegenedésük vákuumot hagyott ott, ahol a mesterséges intelligencia módszertani kritikájának kellene lennie. Nehéz lenne eléggé hangsúlyozni, hogy milyen sürgős szükség van erre a munkára. Hasonlóképpen, ha a tudományfilozófia hiteles támogatást akar nyújtani az éghajlati modellekkel szembeni támadásokkal szemben, akkor a számítási modellekről szóló beszámolóknak mélyebbre kell hatolniuk a fikciónál.

Ezeken az élet-halál motivációkon túl filozófiai és tudományos értéket is képvisel a mesterséges intelligencia tudományfilozófián belüli felkarolása.

A mesterséges intelligencia részéről választ kapunk arra a kérdésre, hogy a neurális hardverre vonatkozó korlátozások egy részét beépítve a konnekcionista modellekbe, miért képesek jobban megragadni a megismerést, még a valósághű részletességre való törekvés nélkül is. A tudományfilozófiából származó meglátások

segítenek megállapítani, hogy a konnekcionista modellek a megismerés alapjául szolgáló mechanizmusok idealizált, többszintű modelljeiként értelmezhetők. Ez lehetővé teszi erősségeik és gyengeségeik értékelését.

A tudományfilozófia részéről a kognitív tudományok számítási modelljeinek alapos vizsgálata megalapozza a modellek újszerű episztemológiáját és metafizikáját, ami segít megvilágítani a modellek és szimulációk szakirodalmában még fennálló problémákat. A modellek és a célok közötti közvetítés közvetve, a fajtákon keresztül lehetővé teszi, hogy konkrétabb válaszokat kapjunk azokra a kérdésekre, hogy hogyan válasszuk ki a releváns hasonlóságokat a modellépítés során, hogyan igazoljuk a modellekből a célokra való következtetéseket és a modellek metafizikai természetét. A mesterséges neurális hálózatok a valódi kognitív rendszerekről úgy mondanak el nekünk valamit, hogy bemutatják azon fajták tulajdonságait, amelyekhez mindketten tartoznak, vagy azokat a szempontokat, amelyek közösek bennük. Ez az elemzés más típusú modellek, köztük a modellorganizmusok és a matematikai modellek megértésében is hasznosnak bizonyulhat.

Hivatkozások

Andersen, Holly K. 2017. "Mintázatok, információ és ok-okozati összefüggések". *The Journal of Philosophy* 114(11):592-622.

Anderson, James A. és Edward Rosenfeld, szerkesztők. 2000. *Beszélő hálók: A neurális hálózatok szóbeli története*. Cambridge, MA: MIT Press.

Babadi, Baktash és Haim Sompolinsky. 2014. "Sparseness and Expansion in Sensory Representations" (Ritkaság és bővülés az érzékszervi reprezentációkban). *Neuron* 83(5): 1213-1226.

Batterman, Robert W. 2001. *Az ördög a részletekben: Aszimptotikus érvelés a magyarázatban, a redukcióban és az emergenciában*. New York: Oxford University Press.

--- 2002. "Aszimptotika és a minimális modellek szerepe". *The British Journal for the Philosophy of Science* 53:21-38.

Baxter, Donald. 2001. "Az instanciálás mint részleges identitás". *Australasian Journal of*

Filozófia 79(4): 449-64.

Billings, Guy, Eugenio Piasini, Lőrincz Andrea, Nusser Zoltán és R. Angus Silver. 2014. "Network Structure within the Cerebellar Input Layer Enables Lossless Sparse Encoding." *Neuron* 83(4): 960-974.

Boden, Margaret. 2006. *Az elme mint gép: A kognitív tudomány története*. Oxford: Clarendon Press.

Broadbent, Donald. 1985. "A szintek kérdése: McClelland és Rumelhart kommentárja". *Journal of Experimental Psychology: General* 114 (2): 189-92.

Buckner, Cameron. 2018. "Empirizmus varázslat nélkül: transzformációs absztrakció a mély konvolúciós neurális hálózatokban". *Synthese* 195(12), 5339-5372.

Cartwright, Nancy. 1989. "Kapacitások és absztrakciók". In: *Tudományos magyarázat*, szerk. Philip Kitcher és Wesley C. Salmon, 349-56. Minneapolis: University of Minnesota Press.

Chirimuuta, Mazviita. 2018. "Magyarázat a számítógépes idegtudományban: Causal and Non-causal". *The British Journal for the Philosophy of Science* 69(3):849-880.

Churchland, Patricia S. és Terrence J. Sejnowski. 1990. "Neural Representation and Neural Computation", *Philosophical Perspectives* 4: 343-382.

Craver, Carl F. 2007. *Az agy magyarázata: Mechanizmusok és az idegtudomány mozaikos egysége*. Oxford: Clarendon Press.

Dennett, Daniel C. 1991. "Valódi minták". *The Journal of Philosophy* 88(1): 27-51. Fodor, Jerry A. és Zenon W. Pylyshyn. 1988. "A konnekcionizmus és a kognitív Építészet: A Critical Analysis." *Cognition* 28: 3-71.

Frigg, Roman és James Nguyen. 2016. "A modellek fikciós szemlélete Reloaded". *The Monist* 99(3): 225-242.

Fuhs, Mark C. és David S. Touretzky. 2006. "A Spin Glass Model of Path Integration in Rat Medial Entorhinal Cortex". *Journal of Neuroscience* 26(16): 4266-4276.

Garson, James. 2015. "Connectionism." *Stanford Encyclopedia of Philosophy*.

Giere, Ronald N. 1988. *A tudomány magyarázata: A Cognitive Approach*. Chicago, London: University of Chicago Press.

--- 2004. "Hogyan használják a modelleket a valóság ábrázolására". *Tudományfilozófia* 71(5): 742-752.

Godfrey-Smith, Peter. 2006. "A modellalapú tudomány stratégiája". *Biology and Philosophy* 21: 725-40.

--- 2009. "Modellek és fikciók a tudományban". *Philosophical Studies* 143: 101-16.

Green, Christopher D. 1998. "Are Connectionist Models Theories of Cognition?" (A megismerés elméletei).

Psycoloquy 9(4).

Han, Chihye, Wonjun Yoon, Gihyun Kwon, Seungkyu Nam és Daeshik Kim. 2019. "Fehér és fekete dobozos ellenpéldák reprezentációja mély neurális hálózatokban és emberekben: A Functional Magnetic Resonance Imaging Study." *arXiv:1905.02422*.

Hempel, Carl G. 1958. "Az elméletalkotó dilemmája: Tanulmány az elméletalkotás logikájáról". In *Minnesota Studies in the Philosophy of Science, Vol II*, szerk. Herbert Feigl, Michael Scriven és Grover Maxwell. Minneapolis: University of Minnesota Press.

Hennig, Boris. 2015. "Az instancia az aspektus ellentéte". *Australasian Journal of Philosophy* 93:1, 3-20.

Hinton, Geoffrey E. 1984. "Elosztott reprezentációk." CMU-CS-84-157. Carnegie Mellon University, Computer Science Department.

---, szerk. 1990. *Mesterséges intelligencia: Special Issue on Connectionist Symbol Processing* 46(1-2).

Irvine, Elizabeth. 2014. "Modellalapú elméletalkotás a kognitív idegtudományban". *The British Journal for the Philosophy of Science* 67(1): 143-68.

Kaplan, David M. 2011. "Magyarázat és leírás a számítógépes idegtudományban". *Szintézis* 183(3): 339-373.

Khalidi, Muhammad Ali. 1998. "Természetes fajták és átívelő kategóriák". *The Journal of Philosophy* 95(1): 33-50.

--- 2013. *Természeti kategóriák és emberi fajták: Classification in the Natural and Social Sciences*. Cambridge University Press.

Küppers, Günter és Johannes Lenhard. 2004. "A szimulációk ellentmondásos státusza". In *Proceedings of the 18th European Simulation Conference,* 271- 75.

Mäki, Uskali. 2012 "A hamis idealizációk igazsága a modellezésben". In *Modellek, szimulációk és reprezentációk,* szerk. Paul Humphreys és Cyrille Imbert, 216-233. Routledge.

Marcus, G. 2018. "Deep Learning: A Critical Appraisal." *arXiv*:1801.00631 [cs.AI].

Marr, David. 1969. "A kisagyi kéreg elmélete". *The Journal of Physiology* 202(2): 437-470.

--- 1982. *Látomás: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W. H. Freeman & Company.

McClelland, James L. 1981. "Általános és specifikus információk kinyerése a tárolt specifikus ismeretekből". *Proceedings of the Third Annual Conference of the Cognitive Science Society (A Kognitív Tudományok Társasága harmadik éves konferenciájának jegyzőkönyve)*, 170-72.

--- 2009. "A modellezés helye a kognitív tudományban". *Topics in Cognitive Science* 1(1): 11-38.

McClelland, James L. és Rumelhart, David E. 1985. "Elosztott memória és az általános és specifikus információ reprezentációja". *Journal of Experimental Psychology: General* 114 (2): 159.

--- 1986. *Párhuzamos elosztott feldolgozás: Pszichológiai és biológiai modellek. 2. kötet*: *Explorations in the Microstructure of Cognition.* Cambridge MA: MIT Press.

Miłkowski, Marcin. 2013. *A számítógépes elme magyarázata*. Cambridge MA: MIT Press.

Morgan, Mary S. 2002. "Modellkísérletek és modellek a kísérletekben". In *Model- Based Reasoning: Science, Technology, Values*, szerk. Lorenzo Magnani és Nancy J. Nersessian, 41-58. New York: Kluwer Academic Publishers.

--- 2003. "Anyagi beavatkozás nélküli kísérletek: Kísérletek: Modellkísérletek, virtuális kísérletek és virtuálisan kísérletek." In *The Philosophy of Scientific Experimentation* (A *tudományos kísérletezés filozófiája)*, szerk. Hans Radder, 216-35. University of Pittsburgh Press.

Newell, Allen és Herbert A. Simon. 1961. "Az emberi gondolkodás számítógépes szimulációja".

Science 134(3495): 2011-7.

--- 1976. "A számítástechnika mint empirikus kutatás: Szimbólumok és keresés." *Communications of the ACM* 19(3): 113-26.

Norton, Stephen és Frederick Suppe. 2001. "Miért jó tudomány a légköri modellezés". In *Changing the Atmosphere*, szerk. Clark A. Miller és Paul N. Edwards, 67-105. Cambridge, MA: MIT Press.

Parker, Wendy. 2009. "Tényleg számít az anyag? Számítógépes szimulációk, kísérletek és az anyagiság." *Synthese* 169(3): 483-96.

Plaut, David C. 1995. "Kettős disszociáció modularitás nélkül: Evidence from Connectionist Neuropsychology (Bizonyítékok a konnekcionista neuropszichológiából)." *Journal of Clinical and Experimental Neuropsychology* 17(2): 291-321.

Rumelhart, David E. és James L. McClelland. 1985. "Valóban szintek! Válasz Broadbentre." *Journal of Experimental Psychology: General* 114 (2): 193-97.

--- 1986. "Az angol igék múlt idejének tanulásáról". In McClelland és Rumelhart 1986, 216-71.

--- 1986a. *Párhuzamos elosztott feldolgozás: A megismerés mikrostruktúrájának vizsgálata, 1. kötet: Alapok*. Cambridge, MA: MIT Press.

--- 1986b. "PDP-modellek és a kognitív tudomány általános kérdései". In Rumelhart és McClelland 1986a, 110-46.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg és Li Fei-Fei. 2015. "ImageNet Large Scale Visual Recognition Challenge". *International Journal of Computer Vision* 115(3):211-252.

Sejnowski, Terrence J. és Charles R. Rosenberg. 1986. "NETtalk: A Parallel Network that Learns to Read Aloud (Párhuzamos hálózat, amely megtanul hangosan olvasni)". Technikai jelentés JHU/EEC-86/01, Villamosmérnöki és Számítógép-tudományi Kar, Johns Hopkins Egyetem.

Smolensky, Paul. 1988. "A konnekcionizmus helyes kezeléséről". *Behavioral and Brain Sciences* 11: 1-74.

--- 1988a. "A konnekcionista mentális állapotok alkotó szerkezete: A Reply to Fodor and Pylyshyn." *Southern Journal of Philosophy* 26(S1): 137-61.

--- 1991. "A konnekcionizmus, a konstituencia és a gondolkodás nyelve". In *Meaning in Mind: Fodor and His Critics*, szerk. Barry M Loewer és Georges Rey, 201-27. Blackwell Publishing.

Steinle, Friedrich. 1997. "Új területekre való belépés: A kísérletezés feltáró felhasználása". *Tudományfilozófia* 64: S65-S74.

--- 2002. "Tudománytörténeti és tudományfilozófiai kísérletek". *Perspectives on Science* 10(4): 408-32.

Stinson, Catherine. 2018. "Magyarázat és konnekcionista modellek". In *The Routledge Handbook of the Computational Mind*, szerk. Matteo Colombo és Mark Sprevak, 120-133. Routledge.

Suárez, Mauricio. 2003. "Tudományos reprezentáció: A hasonlóság és az izomorfizmus ellen". *International Studies in the Philosophy of Science* 17(3): 225-244.

Suri, Roland E. és Wolfram Schultz. 2001. "Temporális differenciamodell reprodukálja az anticipációs neurális aktivitást". *Neural Computation* 13(4): 841-62.

Thomas, Michael S.C. és James L. McClelland. 2008. "A megismerés konnekcionista modelljei". In *Cambridge Handbook of Computational Psychology*, szerk. Ron Sun. 23-58.

Touretzky, David S. és Geoffrey E. Hinton. 1988. "Egy elosztott konnekcionista termelési rendszer". *Cognitive Science* 12(3): 423-66.

Weisberg, Michael. 2012. *Szimuláció és hasonlóság: A modellek felhasználása a világ megértéséhez*. Oxford University Press.

Wilson, Hugh R. és Jack D. Cowan. 1972. "Excitatorikus és gátló kölcsönhatások modellneuronok lokalizált populációiban". *Biophysical Journal* 12(1): 1-24.

Winsberg, Eric. 2009. "Két módszer története". *Synthese* 169(3): 575-92.

--- 2010. *Tudomány a számítógépes szimuláció korában*. University of Chicago Press.