ELŐNYOMTATÁS (25/05/2018)
PUBLIKÁLT VÁLTOZAT IN V.C. MÜLLER (SZERK.), *PHILOSOPHY AND THEORY OF ARTIFICIAL INTELLIGENCE*. SAPERE 44. KÖTET. BERLIN: SPRINGER NATURE. 3-18. OLDAL.

Mesterséges tudatosság: A lehetetlenségtől a sokféleségig

Chuanfei Chin

Filozófia Tanszék, Szingapúri Nemzeti Egyetem, Szingapúr 117570
phiccf@nus.edu.sg

Összefoglaló. Hogyan váltotta fel a sokféleség a lehetetlenséget a mesterséges tudattal kapcsolatos filozófiai kihívásokban? Értékelem a mesterséges tudatossággal kapcsolatos legújabb viták egy olyan irányvonalát, amelyben a tudatos gépek létrehozásának lehetőségével kapcsolatos metafizikai és magyarázó kihívások a "milyen" tudatos lénynek lenni vagy tudatos állapotban lenni sokféleségével kapcsolatos ismeretelméleti aggályokhoz vezetnek. Először is elemzem azokat a korábbi kihívásokat, amelyek azt állítják, hogy a fenomenális tudatosság nem jöhet létre, vagy nem építhető meg gépekben. Ezek Block Chinese Nation és Chalmers Hard Problem című műveire épülnek. Az ilyen kihívások hatástalanítása érdekében a mesterséges tudat elméletírói empirikus módszerekre és magyarázó modellekre hivatkozhatnak. Másodszor, elmagyarázom, hogy ez a naturalista megközelítés miért eredményez ismeretelméleti rejtélyt a biológiai tulajdonságok szerepéről a fenomenális tudatban. Sem a viselkedési tesztek, sem az elméleti következtetések nem látszanak eldönteni, hogy a gépeink tudatosak-e. Harmadszor, értékelem, hogy az új kihívás kezelhető-e a tudatos állapotok finomabb taxonómiájával. Ezt a stratégiát a biológiai fajok és az állati tudatosság hasonló taxonómiáinak kidolgozása támogatja. Bár értelmet ad a mesterséges tudatosság néhány jelenlegi modelljének, kérdéseket vet fel azok szubjektív és morális jelentőségével kapcsolatban.

Kulcsszavak: mesterséges tudat, gépi tudat, fenomenális tudat, tudományos taxonómia, szubjektivitás.

1 Bevezetés

Szeretném nyomon követni a mesterséges tudattal kapcsolatos legújabb filozófiai viták egyik irányvonalát. Ezen a pályán a tudatos gépek megalkotásának lehetőségével kapcsolatos metafizikai és magyarázó kihívásokat felváltják az azzal kapcsolatos ismeretelméleti aggályok, hogy milyen sokféleképpen lehet "milyen" tudatos élőlénynek lenni vagy tudatos állapotban lenni. A *mesterséges tudatosság* itt elsősorban a nem szerves anyagokból épített gépek fenomenális tudatosságára utal. Az általam tárgyalt filozófusok és tudósok többségéhez hasonlóan én is Blockot (1995) követem abban, hogy a fenomenális tudat fogalmát a szubjektív tapasztalatra használom. Block definíciója szerint egy állapot fenomenális tulajdonságainak összessége az, hogy milyen érzés az adott tudatállapotban lenni, és egy lény fenomenális állapotainak összessége az, hogy milyen érzés az adott tudatos lénynek lenni. Az ilyen tudatos állapotok paradigmái közé tartozik az érzékelések, érzések és észlelések megléte.

A mesterséges tudatosságról szóló számos felmérés hangsúlyozza, hogy a mesterséges intelligencia kutatásának ez az alterülete többféle érdekeltségű (Gamez 2008; Holland és Gamez 2009; Reggia 2013; Scheutz 2014). Kutatási programjainak célja olyan gépek megalkotása, amelyek a tudattal kapcsolatos viselkedést utánoznak, a tudat kognitív struktúrájával rendelkező gépek, vagy tudatos állapotokkal rendelkező gépek. Gyakran tesznek különbséget az *erős* mesterséges tudatosság, amely tudatos gépeket céloz meg, és a *gyenge* mesterséges tudatosság között, amely a tudatosság néhány jelentős korrelátumát szimuláló gépeket épít. Természetesen egy kutatási programban mind az erős, mind a gyenge mesterséges tudatosság iránt érdeklődhetünk; és ugyanazt a modellt használhatjuk mind az erős, mind a gyenge mesterséges tudatosság vizsgálatára.

Az erős mesterséges tudatossággal szembeni filozófiai kihívásokra fogok összpontosítani. Először is, a következő részben két korábbi kihívást elemzek, amelyek azt állítják, hogy a fenomenális tudat nem jöhet létre, vagy nem építhető ki a gépekben. Ezek Block Kínai nemzet és Chalmers Kemény probléma című művén alapulnak. Az ilyen kihívások hatástalanítására empirikus módszerekre és magyarázó modellekre hivatkozhatunk. Másodszor kifejtem, hogy ez a naturalista megközelítés miért vezet a biológiai tulajdonságoknak a fenomenális tudatban betöltött szerepével kapcsolatos ismeretelméleti rejtélyhez. Sem a viselkedési tesztek, sem az elméleti következtetések nem látszanak eldönteni, hogy gépeink tudatosak-e. Harmadszor, értékelni fogom, hogy az új kihívás kezelhető-e a tudatos állapotok finomabb taxonómiájával. Ezt a stratégiát a biológiai fajok és az állati tudatosság finomabb taxonómiáinak kidolgozása támogatja. Bár értelmet ad a mesterséges tudatosság néhány jelenlegi modelljének, kérdéseket vet fel azok szubjektív jelentésével és morális státuszával kapcsolatban.

2 A mesterséges tudat lehetetlensége

A mesterséges tudattal foglalkozó szakirodalom számos filozófiai kihívást tartalmaz a tudatos gépek megalkotásának lehetőségével kapcsolatban (Bishop 2009; Gamez 2008; McDermott 2007; Prinz 2003; Reggia 2013; Scheutz 2014). Ezek a kihívások a tudat természetével és a tudathoz való hozzáférésünkkel kapcsolatos filozófiai érvekre támaszkodnak. A kihívások egyik csoportja a mesterséges tudat *metafizikai lehetőségével* szemben szólal meg. Ezek Block (1978), Searle (1980) és Maudlin (1989) provokatív gondolatkísérletein alapulnak, amelyek azt sugallják, hogy a gépek, bármilyen kifinomult funkcionális vagy számítástechnikai szempontból is, nem lehetnek tudatosak. A kihívások másik csoportja a tudatos gépek megépítésének *gyakorlati* lehetőségére irányul. Ezek a McGinn (1991), Levine (1983) és Chalmers (1995) által megfogalmazott filozófiai állításokon alapulnak, amelyek arról szólnak, hogy nem tudjuk, hogyan keletkeznek a fizikai állapotokból a tudatos állapotok. E kihívások szerint aligha várhatjuk el, hogy tudatosságot hozzunk létre gépekben, ha nem tudjuk megmagyarázni azt az emberi agyakban.

A mesterséges tudatosság legtöbb teoretikusát nem zavarják az ilyen kihívások. Scheutz (2014) felmérésében két olyan attitűdöt ír le, amely ezt az álláspontot alátámasztja. Én így értelmezem őket. Először is, egyes teoretikusok *pragmatikusan* viszonyulnak a tudat fogalmához. Ők ezt a fogalmat operacionális módon határozzák meg, azon folyamatok és elvek szempontjából, amelyeket a pszichológusok a tudatosság alapjául vesznek. Céljuk, hogy ezeket a folyamatokat és elveket a gépek teljesítményének javítására használják. Nem akarják, hogy

a tudatosság megismétlésére, így nem kell aggódniuk amiatt, hogy a tudatosság létrejöhet-e, vagy előállítható-e a gépekben. Ez a hozzáállás különösen jól áll azoknak, akiknek a kutatásai a gyenge mesterséges tudatosságra irányulnak. Másodszor, más teoretikusok *revíziós hozzáállást* képviselnek. Ők a tudat fogalmát a pszichológusok által azonosított mögöttes folyamatok és elvek empirikus vizsgálatával akarják finomítani vagy helyettesíteni. Ezzel mind a pszichológiához, mind a filozófiához hozzá kívánnak járulni. A vonatkozó folyamatok és elvek modelljei például új pszichológiai kísérleteket tehetnek lehetővé, és új tudatelméleteket hozhatnak létre. Ezek pedig befolyásolhatják a tudatról alkotott filozófiai intuíciókat és nézeteket.

Ez utóbbit úgy értelmezem, hogy az erős mesterséges tudatosság empirikus kutatását nem kell megállítani a filozófiában uralkodó intuíciók és nézetek miatt. Ennek demonstrálására megmutatom, hogy a mesterséges tudat elméletírói az empirikus módszerekre és magyarázó modellekre hivatkozva hatástalaníthatnak néhány filozófiai kihívást. Különösen azt fogom megvizsgálni, hogyan válaszolhatunk a tudatos gépek megalkotásának lehetőségével kapcsolatos két kihívásra - az egyik Block Kínai nemzet gondolatkísérletén, a másik Chalmers A tudat nehéz probléma című művén alapul.[1] Még azok az elméletalkotók is, akik kevésbé hajlanak arra, hogy komolyan vegyék a filozófiai kihívásokat, tisztázhatják módszertani elkötelezettségüket e válaszok mérlegelésével. Sőt, a következő szakaszban megmutatom, hogy az e válaszok mögött meghúzódó elkötelezettségek miért vezetnek egy olyan ismeretelméleti rejtélyhez, amely a mesterséges tudat minden teoretikusát érdekelheti.

(a) Az első kihívás a tudat természetével kapcsolatos. Azt sugallja, hogy tudatos gépeket nem lehet építeni, mivel a gépek nem lehetnek tudatosak. Pontosabban azt sugallja, hogy a gépek által megvalósítható funkcionális tulajdonságok nem elegendőek a tudatossághoz. Block gondolatkísérletében Kínában egymilliárd embert utasítanak arra, hogy másolja le a mentális állapotok funkcionális szerveződését egy emberi elmében. Rádiókapcsolatokon és műholdas kijelzőkön keresztül úgy irányítanak egy mesterséges testet, ahogyan a neuronok irányítják az emberi testet. A testbe érkező különböző érzékszervi bemenetekre megfelelő viselkedési kimenetekkel reagálnak. Block (1978) szerint azonban nem szívesen tulajdonítunk tudatosságot ennek a rendszernek: "prima facie kétséges, hogy egyáltalán rendelkezik-e mentális állapotokkal - különösen, hogy rendelkezik-e azzal, amit a filozófusok különbözőképpen "minőségi állapotoknak", "nyers érzéseknek" vagy "közvetlen fenomenológiai minőségeknek" neveznek" (73). Ha a Kínai Nemzettel kapcsolatos intuíciónk helytálló, akkor a tudatossághoz többre van szükség, mint a pszichológiában felfedezett funkcionális tulajdonságok. Ha ez így van, akkor a csak ezeket a funkcionális tulajdonságokat megvalósító gépek nem lehetnek tudatosak.

Nem hiszem, hogy a kínai nemzetről szóló intuíciónak kellene engednünk. Inkább empirikus módszerekkel kellene többet megtudnunk a tudat természetéről. A legjobb kutatásaink - a pszichológia, az idegtudományok és a mesterséges tudatosság területén - talán megállapítják, hogy a tudatossághoz elegendőek a durva szemcséjű pszichológiai szintű funkcionális tulajdonságok. Vagy azt is megállapíthatja, hogy a finomabb szemcséjű neurológiai szintű funkcionális tulajdonságok is szükségesek. Az, hogy a megfelelő tulajdonságok megvalósíthatók-e a gépeinkben, egy másik kérdés, amelyet szintén empirikus vizsgálattal kell eldönteni. E kutatásokat nem szabad *eleve* megelőznie annak, hogy az intuíciónknak

[1] Különösen a Prinz (2003) és Gamez (2008) által kínált válaszokból tanultam. A Searle kínai szobás gondolatkísérletén alapuló kihívásokat félretettem: ezeket a mesterséges tudatossággal foglalkozó irodalomban kimerítően elemzik, és úgy tűnik, hogy egyre csökkenő haszonnal. Az ezekre a kihívásokra adott egyik válasz az a) pontban adott válaszom mintájára modellezhető.

mondja egy gondolatkísérletben, és hogy ez mit jelent a tudatos gépek lehetőségéről.

Ebben a módszertani kérdésben még Block is egyetértene. Megjegyzi, hogy intuitív módon az emberi agy sem tűnik megfelelő rendszernek ahhoz, hogy rendelkezzen az általa "kvaliának" nevezett tapasztalatokkal, a tapasztalat szubjektív aspektusával. Tehát az intuíciónkra önmagában nem lehet hagyatkozni annak megítélésében, hogy melyik rendszer rendelkezik vagy nem rendelkezik qualia-val. Block szerint felülbírálhatjuk az intuíciót, ha független okunk van azt hinni, hogy egy rendszer rendelkezik qualia-val, és ha meg tudjuk magyarázni az ebben való hit látszólagos abszurditását. Itt a kínai nemzethez hasonló rendszerrel kapcsolatos aggályai főként azon alapulnak, hogy nincs elméleti alapunk arra, hogy elhiggyük, hogy van qualia. Úgy tűnik, egyetlen általa figyelembe vett pszichológiai elmélet sem magyarázza meg a qualia létezését. Ezért ragaszkodik a rendszerrel kapcsolatban: "minden kétség, hogy kvaliákkal rendelkezik, annak kétségbe vonása, hogy a kvaliák a pszichológia területéhez tartoznak" (84). Ahhoz, hogy ezt a kételyt eloszlassuk, olyan empirikus tudatelméletet kell alkotnunk, amely megmagyarázza a kvaliákat, és értékeli, hogy a kínai nemzetek, gépek és más rendszerek rendelkeznek-e velük.

(b) A második kihívás közvetlenül a tudatosság magyarázatát érinti. Azt sugallja, hogy nem tudunk olyan gépeket építeni, amelyek tudatosak, még akkor sem, ha a gépek tudatosak lehetnek. Chalmers (1995) szerint az előttünk álló nehéz probléma az, hogy megmagyarázzuk, hogyan keletkeznek a tudatos tapasztalatok az agyban zajló fizikai folyamatokból és mechanizmusokból. Megkülönbözteti ezt a könnyű problémáktól, amelyekben különböző pszichológiai funkciókat és viselkedéseket kell megmagyaráznunk számítási vagy neurális mechanizmusokkal. A Nehéz Problémát még nem sikerült megoldanunk, mert nem tudjuk, hogyan keletkezik a tudat az emberi agyban. De amíg ezt nem tudjuk, addig nem tudunk tudatot előállítani egy gépben, csak véletlenül. Gamez (2008) így foglalja össze ezt a tudatlanságunkon alapuló érvelést: "ha nem értjük, hogyan jön létre az emberi tudat, akkor nincs sok értelme annak, hogy megpróbáljunk egy robotot fenomenálisan tudatossá tenni" (892).

Két kapcsolódó okot találok arra, hogy elutasítsam ezt a kihívást. Először is, a tudat keletkezése nem feltétlenül igényel magyarázatot. Empirikus vizsgálattal talán képesek leszünk a tudatosságot anélkül is előállítani, hogy azt az agyban zajló fizikai folyamatokkal és mechanizmusokkal magyaráznánk. Ha ez így van, akkor elegendő, ha a gépekben megteremtjük azokat a feltételeket, amelyek az emberekben tudatosságot eredményeznek; nem kell filozófiailag kielégítő módon megértenünk, hogy a feltételek hogyan teszik ezt. A tudatosság gépekben való létrehozására irányuló kutatásaink ezután segíthetik az emberi tudatosság magyarázatára irányuló kutatásainkat. Ez a kutatási programok közötti kölcsönös megtermékenyítés összhangban lenne a Scheutz által kiemelt revíziós hozzáállással.

Másodszor, még ha szükségünk is van valamilyen magyarázatra a termelés lehetővé tételéhez, a tudat empirikus magyarázata nem feltétlenül igényli a Nehéz Probléma megoldását. Empirikus elméleteik révén a tudósok nem arra törekednek, hogy valamilyen metafizikailag érthető módon megmagyarázzák, hogyan "keletkeznek" egy jelenség tulajdonságai más, alacsonyabb szinteken lévő tulajdonságokból. Ehelyett arra törekednek, hogy a jelenség elméleti azonosságát a mögöttes tulajdonságai alapján állapítsák meg (Block & Stalnaker 1999; McLaughlin 2003; Prinz 2003; Shea & Bayne 2010). (Arról, hogy ez hogyan vonatkozik a tudatosságtudományra, a következő részben szólok bővebben). Elméleteik felépítéséhez a tudósok összefüggéseket vonnak le a szintek között, összekötve néhány magasabb és alacsonyabb szintű tulajdonságot. A biológiai és pszichológiai tudományokban az, hogy mi igényel ilyen jellegű magyarázatot a szintek között, a kontextustól függ: gyakran az határozza meg, hogy a magasabb vagy alacsonyabb szinteken milyen tulajdonságok tűnnek anomáliának (Wimsatt 1976; Craver 2009, 6. k.; Prinz 2012, 287-8). Ezek a gyakorlatok azt sugallják, hogy egy empirikusan sikeres elmélet

a tudatosságnak nem kell kitöltenie a fenomenális és a fizikai tulajdonságok közötti űrt - legalábbis nem a Chalmers-féle Kemény Probléma által meghatározott feltételek szerint.

3 A sokféleség a fenomenális tudatban

Megmutattam, hogy az empirikus módszerek és a magyarázó modellek miként tudják hatástalanítani a mesterséges tudat lehetőségével kapcsolatos filozófiai kihívásokat. Lehetővé teszik számunkra, hogy megcáfoljuk a tudat természetére vonatkozó gondolatkísérletekből származó intuíciókat, és aláássuk azokat az érveket, amelyek abból a tudatlanságunkból erednek, hogy a tudatos állapotok hogyan keletkeznek a fizikai állapotokból. Ezekre az empirikus módszerekre és modellekre hivatkozva naturalista megközelítést alkalmazunk a mesterséges tudat tanulmányozásában. Amennyire lehetséges, empirikus módszereket használunk a tudat természetével és a tudathoz való hozzáférésünkkel kapcsolatos kérdések megválaszolására. Ezáltal lehetővé tesszük, hogy a fenomenális tudattal kapcsolatos empirikus felfedezéseket felhasználjuk a mesterséges tudat fogalmi megértéséhez. Ez a naturalista megközelítés azonban egy másik filozófiai kihívást eredményez, amely abból fakad, hogy felfedezzük a tudat mögött rejlő sokféleséget. A mesterséges tudatossággal szembeni új kihívás ismeretelméleti jellegű: azt sugallja, hogy még ha tudatos gépeket is tudunk építeni, nem tudjuk megmondani, hogy a gépek tudatosak-e.

A kihívás abban rejlik, hogy nehezen tudjuk meghatározni a biológiai tulajdonságok szerepét a fenomenális tudatosságban. Amíg nem határozzuk meg szerepüket, nem tudjuk kideríteni, hogy az e tulajdonságok legalább egy részének hiányával rendelkező gépeink tudatosak-e. Ezt a nehézséget több filozófus is elemzi (Block 2002; Papineau 2002, 7. k.; Prinz 2003, 2005; Tye 2016, 10. k.). Érveiket azonban a mesterséges tudatosság teoretikusai nagyrészt figyelmen kívül hagyják. Én Prinz érveire fogok koncentrálni - mivel azok természetesen a tudat empirikus elméletével kapcsolatos munkájából erednek, és közvetlenül a mesterséges tudat teoretikusainak szólnak.

Prinz azzal kezdi, hogy pszichológiai szinten elemzi a tudatos állapotaink *tartalmát* és azokat a *feltételeket, amelyek mellett* tudatossá válnak. Nagel nyomán úgy véli, hogy a perspektíva birtoklása alapvető fontosságú a tudatosság szempontjából: "Nem lehet tudatos tapasztalatunk egy semmiből való kilátás" (2003, 118). Elemzése szerint az emberek érzékszerveinken keresztül "egy bizonyos nézőpontból" tapasztalják meg a világot. Tudatunk tartalma tehát egyszerre perceptuális és perspektivikus. Ezek a tartalmak akkor válnak tudatossá, amikor figyelünk. Amikor ezek a tartalmak elérhetővé válnak a mérlegelésünk és a cselekvés tudatos irányítása számára, lehetővé teszik a világra adott rugalmas válaszaink adását. Ezeket a hipotéziseket összefűzve Prinz azt javasolja, hogy a tudatosság akkor keletkezik az emberekben, amikor olyan jelenségekre figyelünk, hogy perspektivikus perceptuális állapotaink elérhetővé válnak a mérlegelés és a cselekvés szándékos irányítása számára.

Ezután empirikus tanulmányokra támaszkodva Prinz a tudatos állapotok tartalmát és feltételeit a számítási és neurális szintekre helyezi. Az információfeldolgozásban a tudatosság tartalma látszólag a köztes szinten helyezkedik el. A köztes szintű reprezentációink "előnypont-specifikusak és koherensek" (2003, 119). Ezek különböznek a magasabb szintű reprezentációktól, amelyek túl absztraktak ahhoz, hogy megőrizzék a perspektívát, és az alacsonyabb szintű reprezentációktól, amelyek túl lokálisak ahhoz, hogy koherensek legyenek. A megismerés számítógépes modelljeiben a figyelem egy olyan folyamat, amely a reprezentációkat a következő szakaszba szűri, míg a tudatos kontrollt a munkamemória, egy rövid távú

hosszú távú tárolókapacitás és végrehajtó képességek. Az emberi agyban ezeket a számítási folyamatokat a temporális kéregben lévő perceptuális központok, a parietális kéregben lévő figyelemközpontok és a frontális kéregben lévő munkamemória-központok közötti neurális áramkör valósítja meg (2003, 119; 2005, 388). Prinz (2012) több olyan bizonyítékot is idéz, amely arra utal, hogy a gammavektorhullámok döntő szerepet játszanak ezekben az agyi régiókban. Legújabb elmélete szerint tehát a tudat akkor és csak akkor keletkezik bennünk, "amikor és csakis akkor, amikor a középszintű reprezentációkat megvalósító vektorhullámok a gamma-tartományban tüzelnek, és ezáltal elérhetővé válnak a munkamemória számára" (293). Ez empirikusan nézve jó jelölt az emberi tudatosság neurofunkcionális alapjaira.

E fejlődés ellenére Prinz (2003, 2005) kiemel egy ismeretelméleti korlátot, amely független attól, hogy milyen empirikus tudatelmélet mellett döntünk. Azt állítja, hogy nem tudjuk meghatározni, hogy biológiai tulajdonságaink konstitutívak-e a tudatosság szempontjából. Így azt sem tudjuk felfedezni, hogy a gépeink, amelyekből legalább néhány ilyen tulajdonság hiányzik majd, tudatosak-e. Ez az alapja az erős mesterséges tudatosság kutatásával kapcsolatos pesszimizmusának: "Egyszerűen nem az a helyzet, hogy a tudat természetének tudományos vizsgálata eltünteti a gépi tudatosság kérdéseit. Még ha a tudat tudományos elméletei a saját mércéjük szerint sikeresek is lesznek, akkor is agnosztikusnak kell maradnunk a mesterséges tapasztalatokkal kapcsolatban" (117).

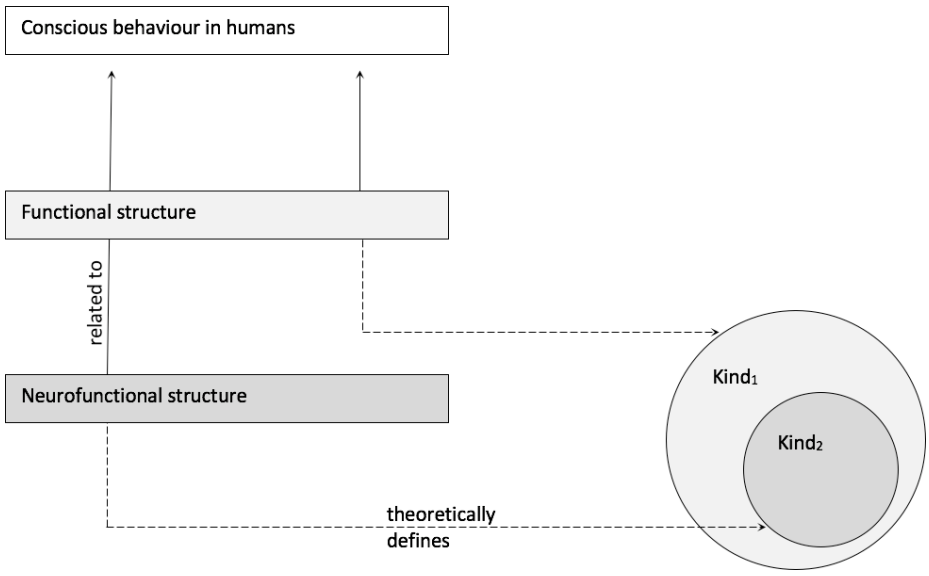
Mint mások, akik osztják pesszimizmusát, Prinz is arra hivatkozik, hogy a viselkedési tesztek elvileg nem képesek eldönteni ezeket a kérdéseket (Prinz 2003, IV; Block 2002; Papineau 2002, 7. fejezet, 2003). Hogyan találjuk meg a tudat konstitutív tulajdonságait? A szokásos módszer az, hogy teszteljük azt, amit Prinz "különbség-alkotóknak" nevez (121). Ez magában foglalja a folyamatok megváltoztatását egy vizsgált szinten, miközben más szinteken a folyamatokat változatlanul hagyjuk. Ha ez a változás különbséget tesz az emberek tudatos viselkedésében, akkor ezen a vizsgált szinten bizonyos tulajdonságok a tudat konstitutív tulajdonságai. Tegyük fel, hogy technikailag lehetséges az emberi agy neuronjait szilíciumchipekkel helyettesíteni. És tegyük fel, hogy ez nomológiailag lehetséges, miközben a pszichológiai és számítási szintek releváns folyamatai állandóak maradnak.[2] Ez a műtétileg megváltoztatott személy "egy normális ember funkcionális másolatává válik, normális aggyal" (123). A funkcionális másolat a terv szerint pontosan úgy fog viselkedni, mint a tudatos emberek - fájdalmat fog jelenteni, a harag jeleit fogja mutatni, látszólag "naplementét fog látni és rózsát fog szagolni". A tudatosság jelenlegi tesztjei mégis a viselkedésre összpontosítanak. Tehát nincs valódi tesztünk a tudatosságra a másolatban. Ezekkel a tesztekkel nem tudjuk megmondani, hogy a biológiai szintű tulajdonságaink konstitutívak-e a tudatosság szempontjából.

Egyetértek Prinzzel (2003) abban, hogy ez a gondolatkísérlet rávilágít egy "komoly ismeretelméleti problémára" (130). Valójában úgy vélem, hogy ő és mások alábecsülik a probléma mélységét. Ők arra összpontosítanak, hogy a viselkedéses tesztek kudarcot vallottak annak kiderítésében, hogy a biológiai tulajdonságok különbséget tesznek-e a tudatos állapotok között. Prinz azt állítja, hogy ez a "különbségtételes módszer tűnik az egyetlen módszernek arra, hogy kiderítsük, milyen szintek számítanak".
(130). Mégis, mint más filozófusok, ő is azt ajánlja, hogy használjuk következtetés a

[2] Ez egy gyakori idealizáció a gondolatkísérletben. A valóságban egynél több pszichológiai és egynél több számítási szinttel találkozunk (Prinz 2003, 120-1). A chipcsere során inkább a kevésbé finomra szabott pszichológiai és számítási szinteken tartjuk állandónak a folyamatokat. A tesztelés ismeretelméleti nehézsége megmarad, bár bonyolultabbá válik. Másutt, Chin (2016) című tanulmányomban a többféle probléma bonyolultabb változatait elemzem a tudattudományban; lásd még Irvine (2013) 6. fejezetét.

a legjobb magyarázat a tudat elméleti azonosságának megteremtésére (116).[3] Nem magyarázza meg, hogy ez az elméleti következtetés miért nem képes tisztázni a biológiai tulajdonságok szerepét a tudatban, és ezáltal javítani a tudat jelenlegi tesztjeit.

Hadd tegyem ezeket az összefüggéseket egyértelművé az 1. ábrán látható többféle problémán keresztül. Ahogy a gondolatkísérlet sugallja, legalább két funkcionális struktúrát fogunk felfedezni, amelyek az emberekben a tudatos viselkedésért felelősek. Az egyik egy neurofunkcionális struktúra, mint amilyet Prinz elméletében azonosítottunk. A másik egy olyan funkcionális struktúra, amely elvonatkoztat a neurofunkcionális struktúra egyes biológiai mechanizmusaitól. Ezért a neurofunkcionális struktúra által meghatározott fajta (fajta2) az absztraktabb funkcionális struktúra (fajta1) által meghatározott fajtába ágyazódik. A kind1 magában foglalja a tudatos embereket és funkcionális hasonmásainkat, míg a kind2 kizárja a funkcionális hasonmásokat. Melyik tehát a tudatosság struktúrája? Melyik struktúra határozza meg az összes és csakis tudatos lények által alkotott fajt?



1. ábra. A fenomenális tudat többféle fajtája

Prinz érvelése azt mutatja, hogy a jelenlegi, viselkedésen alapuló tesztek nem képesek megoldani ezt a többféle problémát. Ezt az érvet szeretném kiterjeszteni, hogy megmutassam, miért nem segít a legjobb magyarázatra való következtetés. Mind a neurofunkcionális struktúra, mind az absztraktabb struktúra összefügg a tudattal az emberben. Mindkettő szisztematikusan összefügg az emberek tudatos viselkedésével is. Ha a neurofunkcionális struktúra, az emberek tudata és tudatos viselkedése közötti szisztematikus kapcsolatokra összpontosítunk, akkor alátámaszthatjuk a tudat és a neurofunkcionális struktúra közötti azonosságot. Ez a lépés azonban *ad hoc jellegű*, a funkcionális másolatainkat végérvényesen nem tudatosnak minősíti. Másfelől, ha az ugyancsak szisztematikus kapcsolatokra összpontosítunk a

[3] További filozófusok: Block és Stalnaker (1999), McLaughlin (2003), Shea és Bayne (2010), valamint Allen és Trestman (2016), 4.3. §.

az absztraktabb funkcionális struktúrát, az emberek tudatát és tudatos viselkedését, akkor alátámaszthatjuk a tudat és az említett struktúra közötti azonosságot. Ez azonban ugyanúgy *ad hoc*, a duplikációkat önkényesen átminősítve tudatosnak.

Egyik hipotézis sem kínál egyszerűbb magyarázatot. Akár a neurofunkcionális struktúrával, akár az absztraktabb struktúrával azonosítjuk a tudatot, a teljes magyarázathoz mindkét struktúrára hivatkoznunk kell. Ha a tudatot a neurofunkcionális struktúrával azonosítjuk, akkor az absztraktabb struktúrát kell használnunk annak megmagyarázására, hogy a másolatok miért viselkednek ugyanúgy, mint az emberek, annak ellenére, hogy a másolatoknak nincs emberi agyuk. Ha a tudatot az absztraktabb struktúrával azonosítjuk, akkor a neurofunkcionális struktúrával kell megmagyaráznunk, hogy az absztraktabb struktúra hogyan valósul meg másképp a tudatos emberekben és a másolataikban. Az első hipotézis a tudatot az absztraktabb struktúra egyetlen megvalósításaként értelmezi, míg a második a neurofunkcionális struktúrát a tudat egyetlen megvalósításaként értelmezi. A magyarázó egyszerűség ismert normái tehát nem segítenek választani e hipotézisek között. Ezért tűnik megoldhatatlannak a többféle probléma. Ha nem tudjuk megoldani ezt a problémát, akkor nem tudjuk megmondani, hogy azok a biológiai tulajdonságok, amelyek a gépeinkből hiányoznak, konstitutívak-e a tudatosság szempontjából. És ezért nem tudjuk megmondani, hogy a gépeink tudatosak-e.

4 A tudományos taxonómiák fejlesztése

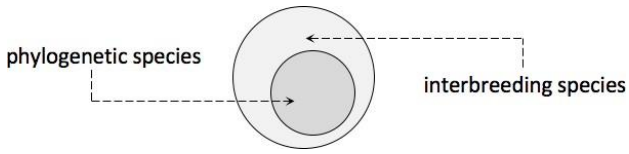
Megmutattam, hogy a mesterséges tudattal kapcsolatos korábbi filozófiai kihívásokat elhárító naturalista megközelítés miért okoz ismeretelméleti fejtörést a biológiai tulajdonságok tudatban betöltött szerepével kapcsolatban. Empirikus vizsgálatok révén több funkcionális struktúrát fedezünk fel, amelyek az emberi tudatosság alapjául szolgálnak. Sem a viselkedési tesztek, sem az elméleti következtetések nem képesek ezek közül egy struktúrát kiemelni, hogy meghatározzunk egy olyan fajt, amelyet minden és csakis tudatos lény alkot. Amíg nem oldjuk meg ezt a többfajta problémát, nem tudjuk meghatározni, hogy azok a biológiai tulajdonságok, amelyek a gépeinkből hiányoznak, konstitutívak-e a tudatosság szempontjából. Ebben a fejezetben azt szeretném megvizsgálni, hogy más tudósok hogyan alakítanak ki finomabb taxonómiákat a többféle fajtájuk kezelésére. Ezután értékelni fogom, hogy a mesterséges tudat elméletírói hogyan használhatják ezt a rendszertani stratégiát.

Hogyan merül fel máshol a többféle probléma? Az egyik kiemelkedő példa az, amit a biológusok "faji problémának" neveznek.[4] Amikor a biológusok megpróbálják az élőlényeket fajokba sorolni, a biológiai sokféleség hátterében többszörös struktúrákat fedeznek fel. E struktúrák középpontjában a kereszteződés, a genetikai vagy fenotípusos hasonlóság, az ökológiai fülke, az evolúciós tendencia vagy a filogenezis áll. Ezek egymásnak ellentmondó meghatározásokhoz vezetnek a faj fogalmának meghatározásához. A különböző struktúrák egymást átfedő, különböző organizmuspopulációkból álló fajokat határoznak meg. Coyne és Orr (2004) biológusok szerint legalább kilenc fajdefiníció marad "komoly versenytárs". Ezek közül hármat gyakran említenek a

[4] Ezt a problémát biológusok és filozófusok egyaránt elemzik: lásd a Coyne és Orr (2004); Cracraft (2000); Ereshefsky (2010, 2017); és Richards (2010) tanulmányait. Én is tanultam LaPorte (2004) elemzéséből, bár eltérő következtetésekre jutunk. Richards (2010) szerint a probléma a darwinizmus előtti időkre nyúlik vissza: Darwin maga is szembesült "a fajfogalmak sokaságával" (75).

filozófiai irodalom: a biológiai fajfogalom (BSC), a filogenetikai fajfogalom (PSC) és az ökológiai fajfogalom (ESC).[5] Ezek az evolúcióban részt vevő három elsődleges folyamatra összpontosítanak: a szexuális szaporodásra, a közös ősöktől való leszármazásra és a környezeti szelekciós nyomásra. A három közül melyik határozza meg a fajok természetét?

A BSC, a PSC és az ESC hívei néha azt állítják, hogy az ő fajmeghatározásuk a "legjobb".[6] A gyakorlatban azonban a biológusok empirikus érdekeiknek megfelelően választanak e definíciók között. Ahogy de Queiroz (1999) kifejti, "eltérnek a vonalszakaszok általuk legfontosabbnak tartott tulajdonságai tekintetében, ami a faji kritériumokkal kapcsolatos preferenciáikban tükröződik" (65). A BSC, a PSC vagy az ESC választása lehetővé teszi számukra, hogy a szexuális szaporodáshoz, a közös őstől való leszármazáshoz vagy az ökológiai fülkéhez kapcsolódó szélesebb körű magyarázó struktúrákat vizsgálják. Az élettörténet iránt érdeklődők például a PSC-t részesítik előnyben a BSC-vel szemben, mert úgy vélik, hogy a reproduktív izoláció "nagyrészt irreleváns a történelem rekonstruálása szempontjából" (Coyne és Orr 2004, 281). Azok, akik a biológiai sokféleség magyarázata iránt érdeklődnek, elutasítják a PSC-t, mert szerintük a filogenezis "nagyrészt irreleváns a természet különállóságának megértéséhez". Ehelyett a BSC-t használják a szexuálisan szaporodó populációk tanulmányozására, vagy az ESC-t az adaptív zónák tanulmányozására az ökológiában.



2. ábra. Két egymást átfedő biológiai faj

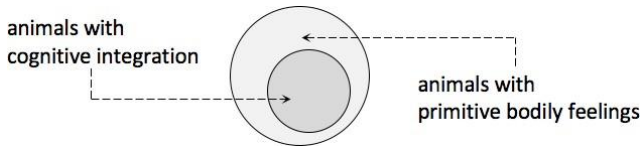
Az eredmény a fajok finomabb taxonómiája, amely felhasználható a biológiai sokféleségen belül található többféle faj kezelésére. A biológusok ma már különbséget tesznek a fajok között, amelyek kereszteződésből, a filogenetikai kapcsolatból és a környezeti szelekcióból erednek (Ereshefsky 2010). Amint a 2. ábra mutatja, a BSC és a PSC általában átfedő populációfajtákat határoz meg. Ha a genealógiailag különböző populációk képesek egymással szaporodni, akkor a filogenetikai faj populációi egy kereszteződő faj populációin belül fészkelnek be. Taxonómiájuk révén a biológusok tisztázhatják az e fajok közötti kapcsolatokat, és elhatárolhatják az e fajokat magában foglaló magyarázó struktúrákat.

---

[5] A BSC a fajokat úgy határozza meg, mint "a természetes populációk kereszteződő, más ilyen csoportoktól reproduktívan elszigetelt csoportjait" (Mayr 1969). A PSC meghatározása szerint ezek "az egyedi szervezetek legkisebb diagnosztizálható csoportja, amelyen belül a származás és a leszármazás szülői mintázata létezik" (Cracraft 1983). Az ESC meghatározása szerint "olyan vonal (vagy egymással szorosan rokon vonalcsoport), amely egy olyan adaptív zónát foglal el, amely minimálisan különbözik bármely más vonalcsoport adaptív zónájától, és amely elkülönülten fejlődik minden, az elterjedési területén kívül eső vonalcsoporttól" (Van Valen 1976).

[6] Ahogy Cracraft (2000) figyelmeztet, "a "legjobb" fogalma mindig relatív" (10). Arra ösztönöz, hogy "alaposan vizsgáljuk meg, hogy mit jelenthet *a legjobb*", beleértve azt is, hogy mennyire általános a definíció alkalmazása, és hogy egy általánosabb definíció mindig hasznosabb-e.

A finomabb taxonómia bevezetésével a biológiai magyarázat szempontjából nem az számít, hogy a BSC vagy a PSC kínálja-e a fajok "legjobb" meghatározását. A biológusoknak inkább azt kell biztosítaniuk, hogy azok, akiket a fajok kereszteződése érdekel, ne keverjék össze az osztályozást azokkal, akiket a filogenetikai fajok érdekelnek. Egy olyan környezetben, ahol közösek az érdekek, ilyen félreértés valószínűleg nem fog előfordulni. Például a szexuális szaporodás és annak hatásai iránt érdeklődő biológusok többsége a kereszteződő fajokra összpontosít. Érdeklődésük már eleve ezeket a releváns fajokat választja ki a szexuális szaporodással, az ősöktől való leszármazással és a környezeti szelekciós nyomással kapcsolatos, egymást átfedő fajok közül. Az egymással versengő érdekek kontextusában a biológusok elkerülhetik a félreértéseket, ha kifejezetten utalnak a kereszteződő fajokra, a filogenetikai fajokra vagy az ökológiai fajokra. Egyes általános összefüggésekben azonban a biológusoknak nem kell meghatározniuk, hogy milyen fajokra hivatkoznak. Előfordulhat, hogy a biológia különböző ágaira vonatkozó általánosításokat kívánnak tenni (Brigandt 2003). Így állításaik egységesen vonatkoznak a kereszteződő fajokra, a filogenetikai fajokra és az ökológiai fajokra. A több faj problémája az állati tudattal kapcsolatos vitákat is sújtja. Itt közelebb áll a mesterséges tudattal kapcsolatos ismeretelméleti rejtélyünkhöz. Az állati tudatosság esetében a probléma azért merül fel, mert az embernél legalább két kognitív struktúrát fedezünk fel a tudatosság alapjául. Mindkét struktúra különböző módon felelős az emberek tudatos viselkedéséért. Követni fogom, hogy Godfrey-Smith (2016a, b) hogyan különbözteti meg ezeket a struktúrákat. Az első a fájdalommal és más primitív testi érzésekkel, például a szomjúsággal és a légszomjjal kapcsolatos egyszerű információfeldolgozási módokat foglalja magában. Ez a struktúra lehetővé teszi, hogy a tényleges és potenciális sérülésekre rugalmas, nem reflexív viselkedéssel reagáljunk. A második struktúra az információfeldolgozás kifinomultabb módozatait foglalja magában, amelyek a különböző érzékszervekből és testi érzésekből származó információkat integrálják a memória, a figyelem és a végrehajtó kontroll segítségével. A megismerés egyes elméletei szerint ez a struktúra lehetővé teszi számunkra, hogy modellezzük a világot, mielőtt reagálunk rá.



3. ábra. Két egymást átfedő állatfaj

A 3. ábra azt mutatja, hogy ez a két kognitív struktúra két egymást átfedő állatfajtát határoz meg. A kognitív integrációval rendelkező állatfaj a primitív testi érzésekkel rendelkező állatfajon belül helyezkedik el, mivel a kognitív integráció több gépezetet igényel, például memóriát, figyelmet és végrehajtó kontrollt. Melyik tehát a tudat kognitív struktúrája? Melyik struktúra határozza meg az összes és csakis tudatos állatok által alkotott fajt? Ha a tudatossághoz kognitív integrációra van szükség, akkor csak a memóriával, figyelemmel és végrehajtó kontrollal rendelkező állatok számítanak tudatosnak. Ha azonban a tudatossághoz elegendőek a primitív testi érzések, akkor sokkal több állat számít tudatosnak, amennyiben rendelkezik a primitív testi érzésekhez kapcsolódó szenzomotoros képességekkel.

Godfrey-Smith (2016a, b) ezzel a többféle problémával szembesülve az állatok szubjektív tapasztalatainak finomabb taxonómiáját javasolja. Legalább kétféle szubjektív élmény létezik. Az alaptípus, amely először fejlődött ki, a fájdalom és más primitív testi érzések élményeiből áll; az összetett típus, amely később fejlődött ki, a különböző érzékszervekből és testi érzésekből származó információkat integráló élményekből áll. A szubjektív élmények mindkét fajtája megtalálható a tudatos emberekben: "Az emberi tapasztalatok nagy része valóban a különböző érzékek integrációját, az érzékek és az emlékezet integrációját stb. foglalja magában, de folyamatos szerepe van a tapasztalatok olyan, látszólag régi formáinak is, amelyek a feldolgozás szervezettebb fajtáinak behatolásaként jelennek meg" (2016b, 500). Taxonómiáján keresztül tisztázhatjuk a kétféle tapasztalat közötti kapcsolatokat, és elhatárolhatjuk a mindkét fajtát magában foglaló magyarázó struktúrákat.

A finomabb taxonómiát alkalmazva láthatjuk, hogy az állati viselkedés magyarázatában nem az számít, hogy a szubjektív élmények alapvető vagy összetett fajtája számít-e tudatosnak. Az állati tudatosság elméletírói inkább az empirikus érdekeiknek megfelelően bármelyik tapasztalati fajtára összpontosíthatnak, amennyiben terminológiájuk nem fedi el a két fajta közötti különbségeket. Godfrey-Smith például csak a kognitív integrációval járó tapasztalatokat minősíti tudatosnak: ""A tudatosság" valami, ami túlmutat a puszta szubjektív tapasztalaton, valami gazdagabb vagy kifinomultabb" (2016a, 53). A fájdalmat és más primitív testi érzéseket megtapasztaló állatoknak vannak kvalitásai; valaminek érzik magukat. De kognitív integráció nélkül nem számítanak számára tudatosnak: "Azon tűnődöm, vajon a tintahalak éreznek-e fájdalmat, vajon a sérülés érződik-e számukra valaminek, de ezt nem tekintem úgy, mint azon tűnődést, hogy a tintahalak tudatosak-e" (2016b, 484). Mint elismeri, más, eltérő érdeklődésű teoretikusok hajlamosak a qualia-t a fenomenális tudattal egyenlővé tenni: "Ha van valami, amit úgy *érez, mintha* egy rendszer *lenne,* akkor a rendszerről azt mondják, hogy egyfajta tudattal rendelkezik" (483- 4). Ezeknek az elméletalkotóknak viszont meg kell különböztetniük a fenomenális tudatot más, kifinomultabb, kognitív integrációt igénylő tudattípusoktól.

Hogyan oldhatná meg ez a rendszertani stratégia a mesterséges tudatossággal kapcsolatos ismeretelméleti rejtélyt? Kialakíthatjuk a tudati állapotok finomabb taxonómiáját, hogy kezelni tudjuk a mesterséges tudat teoretikusait zavaró sokféleséget. Ha Prinznek igaza van, akkor legalább kétféle állapotot kell megkülönböztetnünk. Az elsőt az olyan neurofunkcionális állapotok alkotják, mint amilyeneket az ő tudatelméletében meghatározott. Funkcionális másolataink nem rendelkeznek ilyen típusú állapotokkal. A második olyan funkcionális állapotokból áll, amelyek elvonatkoztatnak a neurofunkcionális állapotok bizonyos biológiai mechanizmusaitól; mind az emberek, mind a hasonmások osztoznak az ilyen típusú állapotokon. Ezzel a taxonómiával tisztázhatjuk a neurofunkcionális és a funkcionális állapotok közötti kapcsolatokat, majd elhatárolhatjuk a kétféle állapotot magában foglaló magyarázó struktúrákat. Az emberek és a hasonmások magyarázatában nem az számít, hogy a neurofunkcionális vagy a funkcionális állapotok számítanak-e tudatosnak. A tudat elméletírói inkább az állapotok bármelyik fajtájára összpontosíthatnak empirikus érdekeiknek megfelelően, mindaddig, amíg terminológiájuk nem homályosítja el a kétféle állapot közötti különbségeket. Azoknak, akik csak a neurofunkcionális állapotokat minősítik tudatosnak, még mindig el kell ismerniük a funkcionális állapotok szerepét, amelyek megmagyarázzák, hogy a másolatok miért viselkednek olyan módon, amely az embernél tudatosságra utal. Azoknak, akik a funkcionális állapotokat minősítik tudatosnak, még mindig el kell ismerniük a neurofunkcionális állapotok szerepét; ezek magyarázzák, hogy a funkcionális állapotok hogyan valósulnak meg az emberekben.

Ez az elemzés ismeretelméleti különbséget mutat a biológiai fajok és a mesterséges tudatosság esete között. A biológusok ma már biztosak abban, hogy a kereszteződő fajok, a filogenetikai fajok és az ökológiai fajok jelentős magyarázó szerepet játszanak. Tudják, hogy a BSC-hez, a PSC-hez és az ESC-hez kapcsolódó fajok különböző magyarázó struktúrákban vesznek részt, amelyek a szexuális szaporodáshoz, az ősi leszármazáshoz és az ökológiai fülkéhez kapcsolódnak. Ezzel szemben még nem ismerjük pontosan azokat az állapotokat, amelyek a mesterséges tudatosság kutatásában jelentős magyarázó szerepet fognak játszani. Ez a különbség azonban nem teszi érvénytelenné a taxonómiai stratégia alkalmazását. Csak a tudati állapotok ideiglenes taxonómiájával kell kezdenünk, hogy feltárjuk a minket érdeklő különböző magyarázó struktúrákat. Ahogy egyre többet tudunk meg ezekről a magyarázó struktúrákról, úgy finomíthatjuk a taxonómiát, hogy az pontosabban tükrözze a magyarázatainkban hivatkozott számítási és biológiai folyamatokat. Ez hasonló ahhoz, ahogyan a biológusok kidolgozták a fajok taxonómiáját.

Ez a taxonómiai stratégia máris értelmet adhat a mesterséges tudatosság néhány jelenlegi modelljének. Egyes teoretikusok szerint a megfelelő számítási folyamatok beépítése a gépekbe elegendő ahhoz, hogy azok tudatosak legyenek. Dehaene, Lau és Kouider (2017) például azt javasolják, hogy a gépek akkor tudatosak, ha képesek globális sugárzásra szelektálni az információt, rugalmasan elérhetővé téve azt a számítások számára, és ha képesek e számítások önellenőrzésére. Javaslatuk alátámasztására azt állítják, hogy egy olyan gép, amely mindkét számítási folyamattal rendelkezik, úgy fog viselkedni, "mintha tudatos lenne".
(492). Arra utaló bizonyítékokat is idéznek, hogy az emberek szubjektív tapasztalata "úgy tűnik, hogy összefügg" a globális műsorszórással és az önellenőrzéssel (492). Más teoretikusok úgy vélik, hogy a megfelelő biológiai folyamatok beépítése a gépekbe szükséges ahhoz, hogy azok tudatosak legyenek. Haladjian és Montemayor (2016) a tudatot az emberekben zajló biológiai folyamatokhoz kapcsolja, amelyek érzelmekkel és empátiával ruházzák fel őket. Tehát véleményük szerint a pusztán számításra tervezett, mesterséges intelligenciával rendelkező gépeknek nem lesznek szubjektív élményeik. Godfrey-Smith (2016b) szerint a gépeknek csak akkor lehetnek szubjektív élményeik, ha rendelkeznek bizonyos funkcionális tulajdonságokkal, amelyek az "élő tevékenységhez" (505) kapcsolódnak. Számára ezek közé a tulajdonságok közé tartozik az emberek komplex biológiai rendszereire jellemző robusztusság és alkalmazkodóképesség.

A mi szemszögünkből nézve a mesterséges tudatosság ezen modelljeinek nem kell konfliktusba kerülniük. Inkább úgy tekinthetünk rájuk, mint amelyek együttesen tisztázzák a mesterséges tudat kutatásához szükséges tudati állapotok finomabb taxonómiáját. Egyrészt Dehaene, Lau és Kouider (2017) olyan állapotokat vizsgál, amelyek tisztán számítási szempontból, biológiai mechanizmusokra való hivatkozás nélkül definiálhatók; különösen a globális műsorszórással és az önellenőrzéssel kapcsolatos magyarázó struktúrák érdeklik őket. Másrészt Haladjian és Montemayor (2016), valamint Godfrey- Smith (2016b) egy másfajta, részben biológiai fogalmakkal definiált állapotok iránt érdeklődnek; ők különböző nehézségeket vetnek fel az ilyen állapotok gépekben való megvalósításával kapcsolatban.

5 Következtetés

Ebben a tanulmányban egy olyan pályát értékeltem, amelyben a mesterséges tudatossággal kapcsolatos filozófiai kihívásokban a sokféleség a lehetetlenség helyébe lépett. Először is, két korábbi kihívással foglalkoztam, amelyek azt állítják, hogy a fenomenális tudatosság nem jöhet létre, vagy nem építhető ki a gépekben. Az első kihívás, a tudat természetéből fakadóan, a Block-féle

Kínai nemzet gondolatkísérlet. A második kihívás, a tudatosság magyarázatából, Chalmers Kemény problémáján alapul. Megmutattam, hogy a naturalista megközelítés, amely az empirikus módszerekre és a magyarázó modellekre apellál, hogyan tudja hatástalanítani ezeket a kihívásokat. Annak kiderítéséhez, hogy a gépek lehetnek-e tudatosak, empirikus módszerekkel kidolgozott tudatelméletekre kell támaszkodnunk, nem pedig a gondolatkísérletek által kiváltott intuíciókra a tudatról. A tudat empirikus magyarázatához nem kell filozófiailag kielégítő magyarázatot adnunk arra vonatkozóan, hogy a fizikai tulajdonságokból hogyan keletkeznek a fenomenális tulajdonságok.

Másodszor, elmagyaráztam, hogy ez a naturalista megközelítés miért vezet a biológiai tulajdonságoknak a fenomenális tudatban betöltött szerepével kapcsolatos ismeretelméleti rejtélyhez. Empirikus vizsgálat révén több funkcionális struktúrát fedezünk fel, amelyek az emberi tudatosság hátterében állnak. Amint több filozófus is érvelt, a viselkedési tesztek nem tudnak ezek közül egy struktúrát kiemelni, hogy meghatározzák az összes és csakis tudatos lények által alkotott fajt. Azzal érveltem, hogy a legjobb magyarázatra való következtetés sem segíthet. Ha nem tudjuk megoldani ezt a többfajta problémát, akkor nem tudjuk meghatározni, hogy azok a biológiai tulajdonságok, amelyek a gépeinkből hiányoznak, konstitutívak-e a tudatosság szempontjából. Azt sem tudjuk meghatározni, hogy ezek a gépek tudatosak-e.

Harmadszor, értékeltem, hogy egy más tudományokban használt rendszertani stratégia képes-e kezelni ezt az új kihívást. A biológiai fajok és az állati tudatosság teoretikusai a magyarázatokban hivatkozott, egymást átfedő fajok kezelésére finomabb taxonómiákat dolgoznak ki. Azzal érveltem, hogy hasonlóképpen a mesterséges tudatosság teoretikusai is kidolgozhatják a tudati állapotok finomabb szemléletű taxonómiáját, amely különbséget tesz a tudatosság empirikus elméletében meghatározott neurofunkcionális állapotok és a funkcionális állapotok között, amelyek elvonatkoztatnak a neurofunkcionális állapotok egyes biológiai mechanizmusaitól. Egy ilyen taxonómia lehetővé teszi számunkra, hogy tisztázzuk a kétféle állapot közötti kapcsolatokat, és elhatároljuk a mindkét fajtát magában foglaló magyarázó struktúrákat. Ezen túlmenően azzal érveltem, hogy ez a taxonómiai stratégia segít értelmet adni a mesterséges tudatosság jelenlegi modelljeinek, beleértve azokat is, amelyek csak számítási állapotokat igényelnek, és azokat is, amelyek részben biológiai állapotokat igényelnek. Ezeket úgy értelmezhetjük, mint a különböző típusú tudati állapotok vizsgálatára szolgáló modelleket.

Ez a stratégia három, egymással összefüggő kihívás elé állít bennünket, amelyek az új taxonómiában szereplő fajták magyarázó, szubjektív és erkölcsi jelentőségére vonatkoznak. Először is, meg kell állapítanunk, hogy ezek az állapotfajták jelentős magyarázó szerepet játszanak a mesterséges tudatossággal kapcsolatos kutatásokban. Ez elsősorban empirikus kihívás, a mesterséges tudat elméletíróitól függ, hogy feltárják a különböző, minket érdeklő magyarázó struktúrákat. Másodszor, meg kell vizsgálnunk az ilyen típusú állapotok szubjektív jelentőségét. Eddig úgy értelmeztük egy tudati állapot fenomenális tulajdonságait, hogy azok azt ragadják meg, "milyen érzés" az adott állapotban lenni. Ez az értelmezés azonban nem segít abban, hogy megkülönböztessük, mit jelentenek a többféle állapotfajták szubjektív szempontból. Ezt úgy tehetjük meg, ha megvizsgáljuk azokat a képességeket és kölcsönhatásokat, amelyeket az ezeket a fajtákat meghatározó mögöttes struktúrák tesznek lehetővé. Például egyes alapstruktúrák támogathatják azt, hogy milyen érzés mesterséges betegnek lenni, míg mások azt, hogy milyen érzés mesterséges ágensnek lenni. Harmadszor, fel kell tárnunk az ilyen típusú állapotok morális jelentőségét. Milyen módon számítanak a mesterséges betegek olyan morális betegeknek, akiknek a szenvedésén enyhítenünk kell? Milyen módon számítanak a mesterséges ágensek olyan erkölcsi ágenseknek, akiknek az életére oda kell figyelnünk?

Köszönetnyilvánítás

Köszönöm Arzu Gokmen, Michael Prinzing és Kaine Yeo javaslatát. Abhishek Mishra, Susan Schneider, Paul Schweitzer és Alexandra Serrenti kommentálta az előadást. Ezt a kutatást az NUS Early Career Award (NUS korai karrierdíj) támogatta.

Hivatkozások

Allen, C., & Trestman, M. (2016). Animal Consciousness. In E. N. Zalta (szerk.) *The Stanford Encyclopedia of Philosophy*, Winter 2016. https://plato.stanford.edu/archives/win2016/entries/consciousness-animal/. Hozzáférés 2018. január 10.

Bishop, J.M. (2009). Miért nem érzik a számítógépek a fájdalmat. *Minds and Machines* 19(4): 507-516. Block, N. (1978). Bajok a funkcionalizmussal. In N. Block, *Consciousness, Function, and Rep-*

bemutató: Összegyűjtött tanulmányok, 1. kötet (2007), 63-101. Cambridge, MA: MIT Press.

Block, N. (1995). A tudat funkciójával kapcsolatos zűrzavarról. In N. Block, *Conscious- ness, Function, and Representation: Collected Papers, 1. kötet* (2007), 159-213. Cambridge, MA: MIT Press.

Block, N. (2002). "A tudatosság keményebb problémája". In N. Block, *Consciousness, Function, and Representation: Collected Papers, 1. kötet* (2007), 397-433. Cambridge, MA: MIT Press.

Block, N., & Stalnaker, R. (1999). Fogalmi elemzés, dualizmus és a magyarázó szakadék. *Philosophical Review* 108(1): 1-46.

Brigandt, I. (2003). A fajpluralizmus nem jelenti a fajeliminativizmust. *Philosophy of Science* 70(5): 1305-1316.

Chalmers, D.J. (1995). Szembenézés a tudatosság problémájával. *Journal of Consciousness Studies* 2(3): 200-219.

Chin, C. (2016). *Határeset-tudatosság, fenomenális tudatosság és mesterséges tudatosság: A Unified Approach.* University of Oxford DPhil disszertáció.

Coyne, J.A., & Orr, H.A. (2004). Specifikáció: A Species Concepts katalógusa és kritikája. In A. Rosenberg & R. Arp (szerk.), *Philosophy of Biology: An Anthology*, 272-92. Oxford: Wiley-Blackwell.

Cracraft, J. (1983). Fajfogalmak és fajelemzés. In R.F. Johnston (szerk.), *Current Ornithology*, 159-87. New York: Springer.

Cracraft, J. (2000). Fajfogalmak az elméleti és alkalmazott biológiában: A Systematic Debate with Consequences. In Q.D. Wheeler & R. Meier, *Species Concepts and Phylogenetic Theory: A Debate*, 3-14. New York: Columbia University Press.

Craver, C.F. (2009). *Az agy magyarázata*. Oxford: Oxford University Press.

Dehaene, S., Lau, H., Kouider, S. (2017). Mi a tudatosság, és lehet-e a gépeknek tudata? *Science* 358(6362): 486-92.

Ereshefsky, M. (2010). Fajok, rendszertan és rendszertan. In A. Rosenberg & R. Arp (szerk.), *A biológia filozófiája: An Anthology*, 255-71. Oxford: Wiley-Blackwell.

Ereshefsky, M. (2017). Fajok. In E. N. Zalta (szerk.), *The Stanford Encyclopedia of Philosophy*, 2017 ősz. https://plato.stanford.edu/archives/fall2017/entries/species/. Hozzáférés: 10 Jan. 2018.

Gamez, D. (2008). Haladás a gépi tudatosságban. *Consciousness and Cognition* 17(3): 887-910.

Godfrey-Smith, P. (2016a). Az állati evolúció és a tapasztalat eredete. In D. L. Smith (szerk.), *Hogyan alakítja a biológia a filozófiát: New Foundations for Naturalism (A naturalizmus új alapjai)*, 51-71. Cambridge: Cambridge University Press.

Godfrey-Smith, P. (2016b). Elme, anyag és anyagcsere. *Journal of Philosophy* 113(10): 481- 506.
Haladjian, H.H., & Montemayor, C. (2016). Mesterséges tudatosság és a tudat-figyelem disszociáció. *Tudat és megismerés* 45(október): 210-25.
Holland, O., & Gamez, D. (2009). Mesterséges intelligencia és tudatosság. In W.P. Banks (szerk.), *Encyclopedia of Consciousness*, 37-45. Oxford: Academic Press.
Irvine, E. (2013). *A tudat mint tudományos fogalom: A Tudományfilozófia perspektívája.* Dordrecht: Springer.
LaPorte, J. 2004. *Természetes fajták és fogalmi változás.* Cambridge: Cambridge University Press.
Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64(October): 354-61.
Maudlin, T. (1989). Számítás és tudatosság. *Journal of Philosophy* 86(8): 407-32. McDermott,
D. (2007). Mesterséges intelligencia és tudatosság. In P. D. Zelazo, M. Moscovitch, & E. Thompson (szerk.), *The Cambridge Handbook of Consciousness*, 117-50. Cambridge: Cambridge University Press.
McGinn, C. (1991). *A tudatosság problémája: Essays Towards a Resolution*. Oxford: Blackwell.
McLaughlin, B.P. (2003). Egy naturalista-fenomenális realista válasz Block keményebb problémájára. *Philosophical Issues* 13(1): 163-204.
Papineau, D. (2002). *Gondolkodás a tudatosságról*. Oxford: Clarendon Press.
Prinz, J. (2003). Szint-fejű misztériánizmus és mesterséges tapasztalat. In O. Holland (szerk.), *Gépi tudatosság*, 111-32. Exeter: Imprint Academic.
Prinz, J. (2005). A tudat neurofunkcionális elmélete. In A. Brook & K. Akins (szerk.), *Cognition and the Brain: A filozófia és az idegtudományok mozgalma*, 381-96. Cambridge: Cambridge University Press.
Prinz, J. (2012). *A tudatos agy: Hogyan váltja ki a figyelem a tapasztalatot*. Oxford: Oxford University Press.
Queiroz, K. de. (1999). A fajok általános vonalas fogalma és a fajkategória meghatározó tulajdonságai. In R. A. Wilson (szerk.), *Species: New Interdisciplinary Essays (Új interdiszciplináris esszék)*, 49-89. MIT Press.
Reggia, J.A. (2013). A gépi tudatosság felemelkedése: Studying Consciousness with Computational Models. *Neural Networks* 44(August): 112-31.
Richards, R.A. (2010). *A fajok problémája: filozófiai elemzés*. Cambridge: Cambridge University Press.
Scheutz, M. (2014). Mesterséges érzelmek és gépi tudatosság. In K. Frankish & W. M. Ramsey (szerk.), *The Cambridge Handbook of Artificial Intelligence*, 247-66. Cambridge: Cambridge University Press.
Searle, J.R. (1980). Elmék, agyak és programok. *Behavioral and Brain Sciences* 3(3): 417-57.
Shea, N., & Bayne, T. (2010). A vegetatív állapot és a tudatosság tudománya. *British Journal for the Philosophy of Science* 61 (3): 459-84.
Tye, M. (2016). *Feszült méhek és páncélsokkos rákok: Tudatosak az állatok?* New York: Oxford University Press.
Valen, L.V. (1976). Ökológiai fajok, többfajúak és tölgyek. *Taxon* 25(2/3): 233-39. Wimsatt,
W.C. (1976). Redukcionizmus, szerveződési szintek és a test-lélek probléma. In G.G. Globus, G. Maxwell, & I. Savodnik (szerk.), *Consciousness and the Brain*, 205-67. Dordrecht: Springer.