Tento text patrí do voľne prepojenej série materiálov o umelej inteligencii na stránke http://marcelkvassay.sk

Stroj, inteligencia, vedomie

Marcel Kvassay

Čo je vedomie? Môže ho mať aj stroj? Medzi výskumníkmi tieto otázky stále vyvolávajú búrlivé diskusie. Aby zviditeľnili „zarážajúcu absenciu konsenzu“, Sloman a Chrisley v [10] zozbierali kolekciu predbežných definícií. Niektorí výskumníci boli presvedčení, že vedomie je „lokalizovateľné v špecifických oblastiach či procesoch mozgu“, no iní zasa tvrdili, že „hovoriť o lokalizovateľnosti vedomia by bolo ‚kategoriálnou chybou' “. A kým niektorí boli toho názoru, že počas spánku nie sme vedomí, ďalší im oponovali, že sme vedomí, keď snívame, a tak ďalej. Táto chaotická spleť protichodných názorov signalizuje, že naše bežné predstavy o vedomí často nepresahujú zárodočnú predvedeckú úroveň.

Cieľom tohto článku je neformálne porovnať dva teoretické rámce pre vedecké štúdium vedomia: „funkcionalizmus virtuálnych strojov“, ktorý sformulovali Sloman a Chrisley v [10], a „nereduktívny funkcionalizmus“, ktorý navrhol David Chalmers v [2, 3].

Kľúčové slová: *umelá inteligencia, filozofia mysle, kognitívna veda, strojové vedomie, umelé vedomie*

Keywords: *artificial intelligence, philosophy of mind, cognitive science, machine consciousness, artificial consciousness*

Toto pojednanie sa zrodilo šťastnou náhodou vďaka prieniku mojich osobných a pracovných záujmov. Relatívne nedávno som vstúpil do jednej z oblastí aplikovanej informatiky s názvom „inteligentné a znalostné technológie“. V snahe udomácniť sa v novom prostredí začal som skúmať jej prepojenia s umelou inteligenciou. Zameral som sa na pojem „inteligencie“, aby som zistil, či ho obe disciplíny chápu rovnako. Medzi článkami, ktoré som našiel na internete, jeden ma fascinoval názvom: „Virtuálne stroje a vedomie“ [10]. Autori začali nemenej pútavo:

> *Štúdium vedomia, etablovaná súčasť filozofie mysle a metafyziky, bolo na dlhé roky vykázané z experimentálnej vedy, no nedávno sa do nej opäť vrátilo (niektorí hovoria, že zadnými vrátkami, ktoré mali zostať zamknuté). Väčšina výskumníkov umelej inteligencie otázku vedomia ignoruje, na rozdiel od ľudí, ktorí diskutujú o jej možnostiach a hraniciach. Tvrdíme, že veľká časť týchto diskusií je zmätočná, pretože nie je jasné, čo sa pod pojmom „vedomie“ myslí. Čiastočne je to spôsobené tým, že „vedomie“ je zhlukový pojem („a cluster concept“). ([10], str. 3)*

„*Zhlukový pojem*“, vysvetľujú autori,

> *je tvrdší oriešok než vágny pojem („a vague concept“), napríklad „veľký“, „žltý“, ktorému nemožno presne určiť hranice pozdĺž určitého kontínua. Je to dokonca horší prípad, než skrížený pojem („a mongrel concept“), ktorý chybne kombinuje súvisiace, no pritom jednotlivo korektne definované pojmy, napríklad „fenomenálne vedomie“ („phenomenal consciousness“) a „prístupové vedomie“*

(„access consciousness“). Je pravda, že aj pri vágnych, skrížených, ba aj korektne definovaných pojmoch sa ľudia často sporia o tom, ako ich aplikovať na konkrétne prípady. No pri pojme „vedomie“ je zvlášť problematický ten fakt, že nezhody nie sú len empirické, ale aj hlboko konceptuálne, pretože diskutéri sa často nedokážu zhodnúť ani len na tom, aký typ dôkazu či experimentu by túto otázku definitívne rozhodol. ([10], str. 4-5)

Tieto pasáže vo mne hlboko zarezonovali; otázka vedomia ma totiž fascinovala už od strednej školy. V čase, keď väčšina mojich spolužiakov o nej vôbec nepočula, moje nečakané stretnutie s jogou rozmetalo na kúsky môj rodiaci sa svetonázor. Ak mám byť úprimný, ten svetonázor pozostával zväčša z balastu, ktorý oficiálna propaganda pumpovala do mladých hláv vo vtedy ešte (zdanlivo) pevnom komunistickom režime. Celkom prirodzene som začal inklinovať k indickej duchovnosti, až ma časom úplne pohltila. Vtedy som si ešte neuvedomoval, že aj učenie jogy a iných duchovných smerov možno zneužiť na vymývanie mozgov. Komunistický režim sa k joge staval podozrievavo a toleroval iba niektoré jej druhy, napríklad hathajogu a radžajogu. Tieto bývali prezentované ako metódy na zlepšenie zdravia a duševnej výkonnosti. Len v ojedinelých prípadoch si lektori mohli dovoliť naznačiť či nenápadne pošepkať do ucha, že jogou možno aktivovať aj hlbšie a mocnejšie úrovne vedomia a nadvedomia. Celkovo ich prezentácia pôsobila dojmom, že ide o otvorený repertoár techník, ktoré adepti môžu používať samostatne a podľa vlastného uváženia – samozrejme po náležitom zaškolení kompetentným inštruktorom, tak ako chodíme do autoškoly, aby sme získali vodičské oprávnenie. Po páde komunizmu sa však postupne začal vynárať iný obraz jogy. Bez ohľadu na to, či a nakoľko je spomenutý liberálny prístup efektívny, rozhodne to nebol spôsob, akým sa joga tradične praktizovala v Indii. Na jogu som však ani vtedy nezanevrel, pretože som si zvolil systém, ktorý sa v tomto ohľade – pre mňa navýsosť dôležitom – odchyľoval od zabehanej tradície.

Moja túžba po duchovnosti ma po čase doviedla do Indie, kde som pobudol niekoľko rokov. Bolo to užitočné a podnetné obdobie, no nakoniec skončilo dezilúziou. V tejto chvíli nie je dôležité, nakoľko to spôsobili moje vlastné chyby a nedostatky, a nakoľko nedostatky komunity, v ktorej som žil. Oba prvky nepochybne spolupôsobili. Podstatné je, že hoci moje nadšenie pre duchovné komunity vyprchalo, stále považujem duchovnosť za dôležitú, no teraz ju vnímam hlavne ako individuálnu psychologickú disciplínu. Za jej najdôležitejší aspekt považujem ideu odstupňovaných úrovní vedomia, ktoré zahŕňajú naše bežné bdelé vedomie, no nezačínajú ním, ani ním nekončia. Stručný popis týchto úrovní v slovenčine možno nájsť napríklad v [11].

Netreba hádam zdôrazňovať, že duchovne orientovaní ľudia majú sklon zavrhnúť myšlienku, že by stroj mohol mať vedomie, ako hlúposť. Aj ja som tak zareagoval na Slomanov a Chrisleyho článok po prvom, povrchnom prečítaní. Zdalo sa mi totiž, akoby tvrdili, že keď napríklad kalkulačka dokáže narábať s číslami, tak si ich musí byť istým spôsobom „vedomá“. Keď som si ale článok prečítal pozornejšie, a najmä keď som ho porovnal s inými, uvedomil som si niekoľko subtílnych, ale dôležitých aspektov, vďaka čomu som svoje pôvodné stanovisko prehodnotil. Pokúsim sa teraz tieto aspekty objasniť a tiež vysvetliť, prečo považujem apriórne odmietanie reduktívnych prístupov k vedomiu za chybu. Za tým účelom porovnám Slomanov a Chrisleyho prístup prezentovaný v [10] s „minimalistickou“ nereduktívnou alternatívou navrhnutou Davidom Chalmersom v [2, 3].

* * *

Úvodné pasáže Slomanovho a Chrisleyho článku sa venujú protichodným názorom na povahu vedomia. Aby zviditeľnili názorový chaos medzi výskumníkmi, vyzbierali od nich kolekciu predbežných definícií a prezentovali ju formou dvojstĺpcovej tabuľky. Každý riadok predstavuje jeden sporný bod, napríklad lokalizovateľnosť vedomia, zatiaľ čo stĺpce prezentujú protichodné názory na daný aspekt. Tak napríklad niektorí výskumníci tvrdili, že vedomie možno „lokalizovať v špecifických oblastiach alebo procesoch mozgu“, kým pre iných bola snaha lokalizovať vedomie „kategoriálnou chybou“. Podobne jedni tvrdili, že počas spánku vedomie chýba, no iní považovali snívanie za formu vedomia. Táto zmäť protichodných názorov podľa autorov indikuje, že naše bežné predstavy o vedomí nepresahujú zárodočnú predvedeckú úroveň.

Sloman a Chrisley navrhujú prístup založený na biologicky inšpirovanej architektúre, o ktorom veria, že im umožní nahradiť tieto nejasné predstavy presnejšími a empiricky overiteľnými pojmami. Vychádzajú z predbežnej domnienky či hypotézy, že aj keď termín „vedomie“ zatiaľ nemá presne definovaný význam, označujeme ním rôzne aspekty procesov a mechanizmov, ktorými ľudia a zvieratá spracúvajú informácie. Ich základný pracovný predpoklad je:

> *Javy, ktoré označujeme ako „vedomé“, neobsahujú nič magického; sú výsledkom činnosti veľmi zložitých biologických mechanizmov na spracovanie informácií, ktorým zatiaľ nerozumieme. ([10], p. 9)*

Autori pripúšťajú, že kým prvá časť predpokladu je samozrejmá pre každého naturalistu, jeho druhá časť je notoricky kontroverzná. Zvyšok ich článku je v podstate obhajobou tejto základnej tézy.

Keďže sa autori snažia vysvetliť vedomie ako biologický fenomén, ich prvým krokom je preskúmať existujúce biologické architektúry na spracovanie informácií. Začínajú reaktívnymi architektúrami, najjednoduchšou a vývojovo najstaršou skupinou. Reaktívne mechanizmy, píšu, „produkujú výstup alebo menia svoj vnútorný stav ... bez toho, aby ... explicitne reprezentovali a porovnávali alternatívy. “ Bolo by však chybou ich podceňovať. Vedia sa prispôsobiť aj učiť (napríklad zmenou váh v neurónových sieťach) a umožnili vznik mimoriadne robustných a úspešných biologických komunít:

> *Niektoré čisto reaktívne biologické druhy majú sociálnu architektúru, ktorá umožňuje mase čisto reaktívnych jedincov vytvoriť dojem pozoruhodnej inteligencie, napríklad termitom stavajúcim „katedrály“. Hlavnou črtou reaktívnych systémov je, že im chýba kľúčová schopnosť deliberatívnych systémov ... menovite schopnosť reprezentovať a uvažovať o [v danej chvíli] neexistujúcich či neviditeľných javoch, napríklad o možných budúcich akciách alebo skrytých objektoch. ([10], p. 22)*

„Teoreticky“, poznamenávajú autori, „reaktívny systém vie produkovať rovnako široké spektrum vonkajšieho správania ako zložitejšie systémy. Aby to však dokázal aj prakticky, môže potrebovať väčšiu pamäť na uloženie preddefinovaných vonkajších reakcií, než je celý vesmír.“

Ďalším stupňom hierarchie sú deliberatívne architektúry: tieto dokážu „reprezentovať možnosti (napr. možné alternatívne akcie alebo vysvetlenia nejasných zmyslových vnemov) v nejakej explicitnej forme, vďaka ktorej ich možno porovnať a zvoliť jednu z nich.“ Príklady umelých deliberatívnych systémov zahŕňajú „počítačové programy pre automatické dokazovanie matematických viet, plánovanie, stolné hry, spracovanie prirodzeného jazyka a rôzne typy expertných systémov.“ Tu sa mi dostalo nepriamej odpovede na moju pôvodnú otázku: sú to hlavne deliberatívne algoritmy, ktoré príbuzné odbory ako „inteligentné a znalostné technológie“ preberajú z umelej inteligencie a prispôsobujú si ich na svoje účely.

V ranej fáze umelej inteligencie sa zdalo, že deliberatívne algoritmy sú kľúčom k tomu, aby počítače a roboty inteligenciou dobehli ľudí. Nakoniec sa však ukázalo, že ľudský „zdravý rozum“ (*„common sense“*) je mimo dosahu deliberatívnych systémov. Bolo obrovským prekvapením – a zároveň sklamaním – že naprogramovať ekvivalent bežného zdravého rozumu je oveľa ťažšie, než napríklad automatické dokazovanie matematických viet. Istý čas trvalo, kým výskumníci tento výsledok strávili, a ešte aj dnes mnohí ľudia považujú matematické dôkazy za zložitejšie, než svoje každodenné fungovanie na základe „bežného rozumu“. No výskum umelej inteligencie ukázal, že tento pocit je klamný: zložitosť mozgových procesov, vďaka ktorým vieme interpretovať svet okolo nás a vykonávať aj tie najtriviálnejšie akcie, nám spravidla celkom uniká.

Hlavný zádrheľ v prvých inteligentných systémoch, vysvetľujú autori, spočíval v ich tendencii „zaseknúť sa v nekonečnej slučke identických neúspešných pokusov o riešenie nejakého podproblému.“ Postupne začali výskumníci takým situáciám čeliť pridaním „paralelne bežiaceho podsystému, ktorý monitoruje a vyhodnocuje deliberatívne procesy. Keď zistí, že sa deje niečo zlé,

môže spracovanie prerušiť a zamerať ho nádejnejším smerom.“ Autori tento monitorovací podsystém označujú termínom *meta-manažment*.

Architektúry s meta-manažmentom (teda s reflexiou a sebareflexiou) tvoria najvyšší stupeň biologickej hierarchie. „Ich najbohatšie formy sa vyvinuli relatívne nedávno a zdajú sa byť vlastné iba človeku,“ poznamenávajú autori, „hoci isté formy sebauvedomenia sa potvrdili aj u iných primátov.“ Autori veria, že tento typ architektúr postačuje na vytvorenie robotov vybavených „zdravým rozumom“ a inteligenciou zhruba na ľudskej úrovni, a prezentujú aj vlastného kandidáta: „human-like architecture for cognition and affect“ (skratka „H-CogAff“), čo by sa dalo približne preložiť ako „kognitívna a afektívna architektúra ľudského typu“.

Autori sa napokon dostávajú k bodu, ktorý je považovaný za hlavný kameň úrazu na ceste k robotom obdareným nielen inteligenciou, ale aj vedomím: k problému jedinečnosti a nevyjadriteľnosti našich subjektívnych zážitkov, niekedy označovanému aj ako „problém kválií“ („*the problem of qualia*“). Autori postupujú opatrne, až okľukou, a majú na to pádny dôvod. Dokonca hneď v úvode čosi letmo naznačujú, aj keď predpokladám, že to väčšina čitateľov (podobne ako ja) pri prvom prečítaní prehliadne:

> *Konkrétne sa pokúsime objasniť, ako záujem o otázky vedomia všeobecne a o kváliá osobitne vzniká celkom prirodzene v inteligentných strojoch s istým typom architektúry a meta-manažmentu. Objasnenie možnosti (ba až nevyhnutnosti) vzniku takýchto otázok ilustruje koncepciu „architektúrou motivovaných“ („architecture-driven“) pojmov (teda pojmov, ktoré v danej architektúre pravdepodobne vzniknú) a poskytuje nové podnetné vysvetlenia ľudských filozofických otázok, a tiež zmätkov, týkajúcich sa vedomia. ([10], pp.3-4)*

Takéto slová obvykle signalizujú extrémne reduktívny pohľad na povahu vedomia. Pokúsim sa ukázať, že to neplatí nevyhnutne pre rámec, ktorý Sloman a Chrisley navrhli, minimálne v tom zmysle, že ho možno výhodne skombinovať s menej reduktívnymi prístupmi. Za tým účelom najprv predstavím minimalistickú nereduktívnu alternatívu, ktorú navrhol David Chalmers v [2, 3].

* * *

Kým Aaron Sloman je už viac ako štyri dekády etablovaným odborníkom na umelú inteligenciu (a pôvodne filozof), David Chalmers vstúpil na pole filozofie mysle oveľa neskôr (v deväťdesiatych rokoch) a pôvodným vzdelaním je matematik. Ich profesijné trajektórie sú teda takmer „inverzné“, z čoho môže plynúť zvláštna komplementárnosť, ktorú pociťujem v ich názoroch. Hoci prvý sa prikláňa k reduktívnemu a druhý k nereduktívnemu náhľadu na vedomie, majú veľa spoločného. Prvá podobnosť vlastne ani neprekvapí. Berúc do úvahy rozmanitosť prístupov k štúdiu vedomia, je len prirodzené, že aj Chalmers vo svojom podnetnom článku [2] začína uznaním principiálnej neurčitosti úlohy:

> *Nestojí pred nami iba jeden problém vedomia. „Vedomie“ je nejasný pojem, ktorým označujeme mnoho rozličných javov. Každý z nich treba vysvetliť, no nie všetky sú rovnako obťažné. Na začiatok bude výhodné rozdeliť ich na „ťažké“ a „ľahké“. Medzi „ľahké“ problémy patria tie, na ktoré možno priamo aplikovať štandardné postupy kognitívnej vedy, čím ich objasníme v pojmoch výpočtových a neurónových mechanizmov. „Ťažké“ problémy sú tie, ktoré takýmto postupom odolávajú. [2]*

Medzi problémy, ktoré Chalmers označuje ako „ľahké“, patrí schopnosť kognitívnych systémov rozoznávať podnety z prostredia, kategorizovať ich a reagovať na ne, schopnosť integrovať informácie z rôznych zmyslov, ako aj schopnosť vnímať svoj vnútorný stav a ovládať svoje správanie na základe uvažovania. „Niet sporu o tom, že *tieto* javy možno vysvetliť,“ píše. Tým, že ich nazýva „ľahkými“ však nechce naznačiť, že sú triviálne: „Môže nám trvať sto aj dvesto rokov, než ich detailne objasníme. No v princípe máme všetky dôvody veriť, že metódy kognitívnej vedy a neurovedy nakoniec uspejú.“ Za „naozaj ťažký“ považuje problém subjektívneho zážitku („*experience*“):

> *Keď myslíme a vnímame, tak prebieha nielen spracovanie informácií, ale je prítomný aj istý subjektívny aspekt. Ako to kedysi vyjadril Nagel (1974), tento aspekt vystihuje, aké je to byť vedomým organizmom. Keď napríklad niečo vidíme, prežívame zároveň subjektívne vizuálne pocity a dojmy: špecifickú „červenosť" danej farby, jej „tmavosť" či „svetlosť", ale aj kvalitu „hĺbky" nášho zrakového poľa. Obdobné subjektívne dojmy sprevádzajú iné zmyslové modality: zvuk klarinetu, pach naftalínu. Potom tu máme telesné pocity, od bolesti až po orgazmus, čisto vnútorné myšlienkové predstavy, preciťovanú kvalitu emócií a tiež zážitok neutíchajúceho prúdu našich myšlienok. Všetky tieto stavy majú spoločné to, že sa môžeme pýtať: „Aké je to, byť v danom stave?" Všetky sú subjektívne prežívanými stavmi. [2]*

„Ťažký problém je práve preto ťažký," pokračuje, „lebo nie je otázkou fungovania mechanizmov":

> *Aj keď vyčerpávajúcim spôsobom vysvetlíme fungovanie všetkých subjektívne prežívaných kognitívnych a behaviorálnych mechanizmov—zmyslového rozlišovania, kategorizácie, prístupu k vlastnému vnútornému stavu a schopnosti verbálne ho popísať—stále zostane nezodpovedaná otázka: prečo je vykonávanie týchto funkcií spojené so subjektívnym prežívaním? Priamočiare vysvetlenie ich fungovania ponecháva túto otázku nezodpovedanú. [2]*

Pre Chalmersa je práve to kľúčovou otázkou vo vzťahu k vedomiu: „Prečo toto spracovanie informácií neprebieha 'v tme', bez vedomého subjektívneho prežívania?"

> *Vieme, že vykonávanie týchto funkcií je spojené so subjektívnym prežívaním, no práve táto skutočnosť je hlavnou záhadou. Kognitívne mechanizmy a ich subjektívne prežívanie oddeľuje „pojmová explanatórna priepasť" („explanatory gap"), ktorú treba premostiť. Čisto technický popis mechanizmov ostáva na jednej strane priepasti, materiál na premostenie musíme preto hľadať inde. [2]*

Chalmers uznáva, že pozoruhodný počet javov sa podarilo bezo zvyšku vysvetliť reduktívne prostredníctvom jednoduchších pojmov a entít, no upozorňuje, že to tak nebolo vždy. Prirovnáva situáciu k fyzike, kde sa občas stáva, že isté nové entity treba akceptovať ako základné. Takéto entity

> *nie sú vysvetlené v jednoduchších pojmoch. Práve naopak, sú považované za základné, a teória popisuje ich vzťahy k ostatným entitám. Napríklad, v devätnástom storočí sa ukázalo, že elektromagnetické procesy nemožno úplne vysvetliť v pojmoch mechanických procesov, na ktorých stáli dovtedajšie fyzikálne teórie, a tak Maxwell a ďalší zaviedli pojmy elektromagnetického náboja a elektromagnetických síl ako nové základné prvky. Aby sme mohli vysvetliť elektromagnetizmus, bolo treba rozšíriť ontológiu fyziky. Na uspokojivé vysvetlenie pozorovaných javov boli potrebné nové základné vlastnosti a zákony. [2]*

Skutočnosť, že sa nepokúšame vysvetliť entity ako hmotnosť, náboj a časopriestor v jednoduchších pojmoch, dôvodí Chalmers,

> *nevylučuje existenciu teórie hmotnosti alebo časopriestoru. Existuje predsa komplikovaná teória, ako tieto základné prvky navzájom súvisia a do akých základných zákonov vstupujú. Tieto základné princípy sú následne využité na vysvetlenie pozorovaných javov týkajúcich sa hmotnosti, priestoru a času na vyššej úrovni. [2]*

Chalmers navrhuje, aby sme akceptovali subjektívne prežívanie („*experience*") ako základný pojem. „Je zrejmé, že teória vedomia si vyžaduje pridať nové základné prvky do našej ontológie," tvrdí,

> *keďže terajšia fyzikálna teória je plne kompatibilné s absenciou vedomia. Mohli by sme pridať nejaký úplne nový nefyzikálny prvok, z ktorého by sa dali subjektívne zážitky odvodiť, no ťažko si predstaviť,*

aký prvok by to mohol byť. Je preto pravdepodobnejšie, že akceptujeme priamo subjektívne prežívanie ako jednu zo základných čŕt sveta, popri hmotnosti, náboji a časopriestore. Potom môžeme začať budovať teóriu subjektívneho prežívania („theory of experience“). [2]

„Nereduktívna teória subjektívneho prežívania,“ objasňuje ďalej,

pridá nové základné princípy k existujúcim prírodným zákonom. Z nich sa budú odvíjať vysvetlenia príslušných javov. Tak ako vysvetľujeme javy týkajúce sa hmotnosti pomocou základných princípov zahŕňajúcich hmotnosť a ďalšie veličiny, budeme aj javy týkajúce sa subjektívneho prežívania vysvetľovať pomocou základných princípov zahŕňajúcich subjektívne prežívanie a ďalšie veličiny. [2]

Chalmers tieto nové princípy nazýva „psychofyzikálnymi“, keďže ich úlohou je preklenúť „pojmovú explanatórnu priepasť“ („*explanatory gap*“) medzi kognitívnymi mechanizmami a ich subjektívnym prežívaním. Z toho zároveň vyplýva, že by nemali narúšať fyzikálne zákony, keďže tie, ako sa zdá, už tvoria uzavretý a sebestačný systém. Nové princípy tak budú skôr doplnkom fyzikálnej teórie. Chalmers uznáva, že jeho pozíciu možno klasifikovať ako dualizmus, keďže postuluje nové základné atribúty nad rámec toho, čo aktuálne potrebuje a používa fyzika. Napriek tomu tvrdí, že je to „nevinná“ forma dualizmu, plne kompatibilná s vedeckým náhľadom na svet:

Nič v tomto náhľade neprotirečí fyzikálnej teórii; len pridávame nové premosťujúce princípy, ktoré vysvetlia, ako sa subjektívne prežívanie objavuje či vynára z istých fyzikálnych procesov. Na tejto teórii nie je nič mystického -- formou sa blíži fyzikálnej teórii, pretože v nej vystupuje malý počet základných entít prepojených istými základnými zákonmi. Nepochybne musíme mierne rozšíriť ontológiu fyziky, no to isté predsa urobil aj Maxwell. Celková štruktúra tejto pozície je čisto naturalistická, pretože vníma vesmír ako pradivo základných entít prepojených jednoduchými základnými zákonmi, a zároveň umožňuje v tomto rámci vybudovať teóriu vedomia. Ak túto pozíciu treba nejako nazvať, dobrou voľbou by bol „naturalistický dualizmus“ („naturalistic dualism“). [2]

Uverejnenie tohto článku v roku 1995 vyvolalo medzi výskumníkmi búrlivú diskusiu. Chalmers reagoval na kritiky o dva roky neskôr v [3]. Vo svojej analýze sa najprv venoval reduktívnej, resp. „deflačnej“ kritike a rozlíšil dva typy materializmu:

Materialista prvého typu („type-A materialist“) popiera, že by nejaký „ťažký problém“ vedomia odlišný od „ľahkých“ vôbec jestvoval; materialista druhého typu („type-B materialist“) jeho existenciu explicitne alebo implicitne pripúšťa, no tvrdí, že sa dá vyriešiť v rámci materializmu. [3]

Materializmus prvého typu, objasňuje Chalmers,

netvrdí len toľko, že vedomie je identické s istou kognitívnou funkciou, alebo že hrá funkčnú rolu, alebo že keď vysvetlíme kognitívne funkcie, pomôže nám to vysvetliť vedomie. Je to oveľa extrémnejší postoj, že otázka vedomia vlastne vôbec nejestvuje: keď objasníme funkcie systému, dozvieme sa o ňom všetko, čo má zmysel vedieť. [3]

Toto stanovisko je v príkrom rozpore s bežnou intuíciou, poznamenáva Chalmers:

Na prvý pohľad akoby popieralo očividný fakt našej vedomej existencie. No zasluhuje si, aby sme ho brali vážne: nebola by to prvá ani posledná vedecká či filozofická teória, čo sa prieči zdravému rozumu. Na druhej strane, takéto teórie vyžadujú vecné a pádne argumenty. A v tomto konkrétnom prípade -- že v kognitívnych systémoch okrem funkcií skutočne niet čo skúmať -- máme právo požadovať mimoriadne silné argumenty. Aké argumenty teda proponenti tejto tézy ponúkajú? [3]

Hádam najčastejšou stratégiou materialistov prvého typu, pokračuje,

je redukcia „ťažkého problému“ pomocou analógie s inými oblasťami, kde taký pojem nie je potrebný. Tak napríklad Dennett si predstavuje vitalistu boriaceho sa s ťažkým problémom „života“ či

neurovedca skúmajúceho ťažký problém „vnímania“. Podobne Paul Churchland (1996) ponúka obraz filozofa devätnásteho storočia, ktorého trápi ťažký problém „svetla“, a Patricia Churchland analógiu s problémom „tepla“. Vo všetkých týchto prípadoch si ľudia v istej fáze mohli myslieť, že treba vysvetliť viac, než len štruktúru a funkcie, no veda ich nakoniec vždy usvedčila z omylu. Ťažký problém „vedomia“ teda tiež môže byť len zdanlivý. [3]

Chalmersa tieto argumenty nepresvedčili: „Medzi problémom vedomia a problémami v iných oblastiach nie je analógia,“ vraví, keďže v iných oblastiach „je viac či menej *zrejmé*, že vysvetliť treba práve štruktúru a funkcie, prinajmenšom keď abstrahujeme od aspektov subjektívneho prežívania“:

Keď sa napríklad zamyslíme nad problémom života, je hneď zrejmé, že potrebujeme vysvetliť štruktúru a funkcie: Ako sa živý systém samoorganizuje? Ako sa prispôsobuje prostrediu? Ako sa rozmnožuje? Aj vitalisti uznávali tento kľúčový bod: ich hlavnou motiváciou bola vždy otázka, „Ako by mohol čisto fyzikálny systém zvládať také zložité funkcie?“ a nie, „Prečo sa život objavuje pri vykonávaní týchto funkcií?“ Nie náhodou je teda Dennettova verzia vitalistu čisto imaginárna: nejestvuje žiadny „ťažký problém“ života, a nejestvoval ani pre samotných vitalistov. [3]

V prípade vedomia však medzi pozorované javy, ktoré treba vysvetliť, patrí nielen štruktúra a funkcie, ale aj subjektívne prežívanie. Preto je podľa Chalmersa táto analógia nepoužiteľná. To isté platí o „ľubovoľnom jave, ktorý pozorujeme v okolitom svete“:

Keď pozorujeme vonkajšie objekty, vnímame ich štruktúru a fungovanie, a to je všetko. Takéto pozorovania nedávajú žiaden dôvod postulovať existenciu novej triedy vlastností nad rámec tých, ktorými vysvetľujeme štruktúru a funkcie. Nikde mimo nás teda nenájdeme obdobu „ťažkého problému“ vedomia. Aj keby totiž vonkajšie objekty mali také doplnkové vlastnosti, tie by pre nás boli nedosiahnuteľné, keďže objekty pozorujeme „zvonka“ a sprostredkovane. Všetky takéto vlastnosti by ležali na opačnej strane neprekročiteľnej poznávacej bariéry. Vedomie sa tejto situácii jedinečným spôsobom vymyká, pretože neleží mimo nás, ale priamo v centre nášho epistemického vesmíru. V tomto jedinom prípade máme prístup aj k dačomu inému, než je štruktúra a funkcie. [3]

Ak chcú teda materialisti prvého typu presvedčiť ostatných, musia explicitne obhájiť tézu, že v prípade vedomia, rovnako ako v prípade života, netreba okrem štruktúry a funkcií vysvetliť nič iné. Namiesto toho,

proponenti často len vyhlásia, že vysvetliť funkcie stačí, alebo argumentujú spôsobom, ktorý v istom bode túto tézu nenápadne akceptuje ako samozrejmú. Ale to zjavne nepostačuje. Evidentne máme dobrý dôvod veriť, že teória vedomia musí vysvetliť nielen rozlišovanie a integráciu zmyslových podnetov, verbálny popis vnútorného stavu a podobné funkcie, ale aj subjektívne prežívanie, a rovnako evidentne sa otázka subjektívneho prežívania javí odlišná od otázky fungovania kognitívnych mechanizmov. Je pravda, že takéto „evidentné“ intuície sa môžu nakoniec ukázať ako nesprávne, no na ich vyvrátenie treba mimoriadne silné a vecné argumenty. [3]

A také argumenty, poznamenáva Chalmers, veľmi ťažko nájsť. V prvej vlne reakcií v časopise *Journal of Consciousness Studies* túto tézu otvorene hájil iba Daniel Dennett. Kľúč k pochopeniu Dennettovej pozície, domnieva sa Chalmers,

leží v tom, čo Dennett inde charakterizoval ako základ svojej filozofie: „absolutizmus pohľadu tretej osoby“ („third-person absolutism“). Ak sa človek pozrie sám na seba z hľadiska tretej osoby -- pozerajúc na seba takpovediac zvonka -- tieto reakcie a schopnosti sú nepochybne to hlavné, čo uvidí. Lenže ťažký problém spočíva v snahe vysvetliť, ako vnímame sami seba bezprostredne -- z pohľadu prvej osoby. Takže zmeniť perspektívu takýmto spôsobom a skúmať náš bezprostredný pohľad na seba z pozície tretej osoby -- čo je Dennettov obľúbený manéver -- opäť neznamená nič iné, ako

vopred predpokladať, že stačí vysvetliť kognitívne reakcie a verbálne popisy, čiže je to zasa pokus o dôkaz v bludnom kruhu. [3]

Chalmersa nezaskočil ani Dennettov protiargument, „odmyslite si funkcie a nezostane nič“:

Použijem analógiu, ktorú sformuloval Gregg Rosenberg. Farbu definujú tri atribúty: odtieň, sýtosť a jas. Keď si z farby odmyslíme odtieň, pravdepodobne z nej nezostane nič pozorovateľné, no to iste neznamená, že by farbu definoval iba odtieň. Aj keby teda Dennett dokázal obhájiť tézu, že kognitívne funkcie sú z nejakého dôvodu nutné pre subjektívne prežívanie (tak ako je pre farbu nutný odtieň), to by ani zďaleka nestačilo na dôkaz tézy, že v prípade vedomia okrem funkcií netreba vysvetliť nič iné. [3]

Na výzvu, aby „predložil ‘nezávislé’ dôkazy ... oprávňujúce ho postulovať subjektívne prežívanie“, Chalmers odpovedá: „Ale to je predsa nepochopenie problému: subjektívne prežívanie nepostulujeme, aby sme ním vysvetlili niečo iné; ono samo je tým javom, ktorý treba vysvetliť.“ A položartom dodáva:

Rád by som videl Dennettovu verziu „nezávislých“ dôkazov oprávňujúcich fyzikov zaviesť základné kategórie priestoru a času. Podľa mňa sú všetky také dôkazy skrz-naskrz časopriestorové, a rovnako aj „dôkaz“ subjektívneho prežívania nemôže nebyť skrz-naskrz subjektívne prežívaný. [3]

Zástancovia materializmu prvého typu radi naznačujú, že výsledky modernej vedy podporujú ich stanovisko, poznamenáva Chalmers, „ale empirická veda, nakoľko ju poznám, je v tomto ohľade celkom neutrálna: ešte nikdy som nevidel experimentálny výsledok, z ktorého by vyplývalo, že v prípade vedomia okrem funkcií netreba vysvetliť nič iné.“

Vzhľadom na celkové vyznenie Chalmersovho výkladu ma jeho záverečné slová prekvapili:

Tým nechcem povedať, že by bol materializmus prvého typu principiálne neobhájiteľný. V literatúre možno nájsť niekoľko sofistikovaných prác na jeho obranu (napríklad Shoemaker 1975 a White 1986), no aj tie musia nakoniec zvažovať alternatívy a zaoberať sa problémami, ktoré nastanú, keď prijmeme, že vedomie je svojbytný fenomén, ktorý treba vysvetliť. Tieto problémy (ontologické i epistemologické) sú nepochybne veľké; život by bol oveľa jednoduchší, keby ťažký problém vedomia nejestvoval. Myslím si však, že tieto problémy sa dajú prekonať; v každom prípade pokračovať v zarytom odmietaní existencie problému len preto, lebo je ťažký, má príchuť svojvoľného rozhodnutia z pozície sily. [3]

Chalmers si nemyslí, že by výskumné výsledky materialistov prvého typu boli bezcenné, alebo že by sa tento typ materializmu chystal v dohľadnej dobe zaniknúť:

Pravdepodobne si budeme musieť zvyknúť, že výskumná komunita je v tomto ohľade rozštiepená na dva základné tábory: na tých, čo sú presvedčení, že existujú len „ľahké“ problémy vedomia, a tých, čo akceptujú, že treba vysvetliť aj subjektívne prežívanie. Môžeme teda očakávať dva veľmi rozdielne druhy teórií vedomia: tie prvé vysvetlia funkcie a dodajú, „To je všetko,“ kým tie druhé akceptujú aj dodatočné explanatórne bremeno. V konečnom dôsledku hlavný pokrok príde skôr zvnútra jedného či druhého tábora, než z nekonečných sporov medzi nimi. Od istého bodu nemá zmysel ďalej sa škriepiť, a teoretici oboch táborov urobia lepšie, ak sa zhodnú aspoň v tom, že sa nezhodli, a zvyšnú energiu venujú rozvoju vlastných výskumných programov. Tak sa veci aspoň pohnú vpred. [3]

Chalmers následne pokračuje rovnako podnetnými analýzami materializmu druhého typu a nereduktívnych kritík, ku ktorým sa vrátim neskôr.

* * *

Slomanove a Chrisleyho názory sa zjavne blížia Dennettovým. Konštatuje to aj sám Sloman v krátkej reakcii [9] na Dennettovu knihu *The Intentional Stance*, i keď v nej zároveň argumentuje, že

intencionálne hľadisko samo osebe nestačí: skutočný vhľad do povahy inteligencie získame až z pohľadu dizajnéra -- teda ak sa ju pokúsime zostrojiť a overiť v praxi. To bol hlavný dôvod, prečo sa Sloman presunul z akademickej filozofie do oblasti umelej inteligencie. Pokúsim sa teraz stručne zhrnúť vybrané výsledky a dôsledky tejto voľby.

Slomanov a Chrisleyho „pohľad dizajnéra“ vychádza z nového typu funkcionálnej analýzy mentálnych pojmov, ktorý sami navrhli. Nazývajú ho „funkcionalizmus virtuálnych strojov“ („*virtual machine functionalism*“ -- skratka VMF) a tvrdia, že je imúnny voči mnohým štandardným filozofickým námietkam. „Väčšina filozofov a kognitívnych vedcov,“ píšu, „nараba s pojmom ‘funkcionalizmus’, akoby bol presne definovaný a všeobecne známy“. Ako príklad uvádzajú prácu [1] filozofa Neda Blocka, kde píše:

> *Podľa funkcionalistov majú mentálne stavy rovnakú povahu ako stavy automatu: definuje ich vzťah k vstupom, výstupom a iným stavom. Mentálny stav S1 teda neznamená nič iné, než produkciu istých špecifikovaných výstupov, ak v tomto stave systém dostane príslušné vstupy. Pre funkcionalistov stav bolesti spočíva iba v tom, že máme sklon povedať „Au, to bolí!“, zamýšľať sa, či nie sme chorí, nedokážeme sa sústrediť, atď. [1]*

Blockovo zhrnutie má podľa Slomana a Chrisleyho najmenej dve interpretácie. Tú prvú, pri ktorej môže mať entita v danej chvíli iba jeden nedeliteľný mentálny stav, navrhujú nazvať „funkcionalizmom atomických stavov“ („*atomic state functionalism*“), a vzápätí ju zavrhujú ako nepostačujúcu, keďže ľudské stavy ako hlad, smäd, údiv či hnev môžu navzájom koexistovať a objaviť sa i zmiznúť nezávisle. Pri druhej interpretácii môže mať entita naraz viac koexistujúcich, nezávisle sa meniacich a navzájom interagujúcich stavov:

> *Je možné, že Block si neuvedomil, že jeho príklady, ako ich bežne chápeme, boli tohto druhého typu: tá istá bolesť nás totiž môže súčasne rozptyľovať od práce i doviesť k úvahám o našom zdraví, takže pociťovanie bolesti, zamýšľanie sa nad zdravím, snaha dokončiť začatú prácu a neschopnosť sústrediť sa na ňu sú štyri koexistujúce stavy, ktoré nemusia začať ani skončiť súčasne. Keď bolesť ustúpi, môžeme aj naďalej premýšľať o svojom zdraví, a kým dokončujeme začatú prácu (napríklad rýľovanie v záhrade), môžeme sa rozhodnúť neskôr navštíviť lekára. Koexistencia interagujúcich podstavov je súčasťou našej bežnej predstavy o mentálnych stavoch, napríklad keď rozprávame o svojich protichodných túžbach či postojoch. ([10], str. 16)*

Práve túto „paralelizovanú“ verziu funkcionalizmu autori nazývajú funkcionalizmom virtuálnych strojov (VMF). Zároveň rozlišujú dve formy: *obmedzený* VMF, pri ktorom musí byť každý podstav priamo alebo nepriamo kauzálne prepojený so vstupmi a výstupmi celého systému, a *neobmedzený* VMF bez takej podmienky. Tento rozdiel je dôležitý. Všeobecne sa uznáva, že priestor možných návrhov inteligentných strojov je natoľko rozsiahly, až je nezvládnuteľný, a potrebujeme zmysluplné kritériá, ktoré by ho zmenšili. Biologicky inšpirované prístupy sa často odvolávajú na evolúciu ako filter: v ideálnom prípade by mal systém obsahovať iba také črty a funkcie, o ktorých možno vierohodne preukázať, ako mohli štruktúrne podobným organizmom pomôcť prežiť alebo získať evolučnú výhodu. Keďže VMF je biologicky inšpirovaný, musí obstáť aj v tomto smere. Kým *obmedzený* VMF si tu vedie dobre, *neobmedzený* VMF je diskutabilný. Nechcem sa tu púšťať do detailov; sústredím sa radšej na to, *prečo* sa autori snažia legitimizovať *neobmedzený* VMF. Dôvod je jednoduchý: ukazuje sa, že *neobmedzený* VMF dokáže elegantne objasniť isté prchavé a ťažko definovateľné črty ľudskej povahy, napríklad jedinečnosť a nevyjadriteľnosť našich subjektívnych zážitkov („*the problem of qualia*“). Autori to demonštrujú na sérii odstupňovaných príkladov:

> *Neobmedzený VMF pripúšťa, aby nejaký podstav S alebo bežiaci podproces bol neustále modifikovaný inými podstavmi prepojenými s vonkajším prostredím, hoci žiadna zo zmien stavu S neovplyvní nič, čo by mohlo v konečnom dôsledku spätne ovplyvniť okolie. Príkladom v počítači by mohol byť proces zbierajúci štatistiky o udalostiach odohrávajúcich sa v iných procesoch, ktorý však svoje výsledky*

> *neposiela ani nesprístupňuje iným častiam systému. Neobmedzený VMF dokonca pripúšťa, aby sa podsystémy na neurčito odpojili a pracovali izolovane. Napríklad nejaký subsystém sa môže rozhodnúť odohrať sám so sebou nekonečnú sériu šachových partií, alebo bez časového obmedzenia hľadať dôkaz Goldbachovej hypotézy. ([10], p.18)*

Tieto úvodné príklady asi nikoho neohúria. Kým štatistiky môžu mať zmysel pre *koncového používateľa*, je ťažké predstaviť si systémového administrátora, ktorý by ihneď nezlikvidoval všetky izolované a nekomunikujúce procesy, na ktoré naďabí. A vo väčšine prípadov by to iste bolo správne rozhodnutie, neplatí to však pre autonómne sa vyvíjajúce systémy, ktorých procesy potrebujú slobodu odpájať a pripájať sa k hlavnému procesu na základe vlastného rozhodnutia.

Prvý náznak skutočného zámeru autorov nájdeme o čosi ďalej pri zaujímavom type čiastočne odpojeného procesu,

> *ktorý kauzálne interaguje s inými procesmi, napríklad tým, že im posiela inštrukcie alebo odpovede na otázky, ale ktorého vnútorné detaily iné procesy neovplyvňujú. Môže im napríklad posielať svoje závery, ale nie dôvody, ktoré ho k nim doviedli. Iné časti systému potom vedia, čo bolo vyvodené, no nevedia prečo. ([10], pp. 18-19)*

Táto kategória zahŕňa dôležitý prípad čiastočne odpojeného procesu na úrovni meta-manažmentu, ktorý sleduje a vyhodnocuje iné procesy:

> *Tento proces vnútorného sebapozorovania nemusí mať žiadne kauzálne prepojenia s motorickými časťami systému, takže informácie v ňom zhromaždené nemožno verbálne reprodukovať ani inak sprístupniť navonok. Ak tento proces ovplyvňuje pozorované procesy ... potom môže mať pozorovateľné vonkajšie prejavy. Môže sa však stať, že vnútorné stavy pozorovacieho procesu sú príliš zložité alebo sa príliš rýchlo menia, než aby sa dali úplne zrekonštruovať z vonkajších prejavov -- čosi ako obmedzená kapacita prenosového kanálu. Pre taký systém môže byť jeho vnímanie seba samého („experience“) čiastočne nevyjadriteľné („partly ineffable“). ([10], p.19)*

Autori zjavne inklinujú k stanovisku, že čiastočná nevyjadriteľnosť na úrovni meta-manažmentu znamená, že sa nám podarilo v stroji reprodukovať prinajmenšom zárodok problému nevyjadriteľnosti subjektívneho prežívania („*the problem of qualia*“). Zároveň argumentujú aj podrobnou analýzou tvorby pojmov v samoorganizujúcich sa systémoch, ktoré -- ako ukazujú -- poskytujú rámec pre pokročilejšie formy subjektívneho prežívania. Ich analýza je náročná, preto sa vyhnem detailom. Musím však priznať, že ma presvedčili, že inteligentné stroje môžu disponovať rôznymi typmi nevyjadriteľného subjektívneho prežívania (presnejšie povedané, skôr technickými predpokladmi preň, pričom na „pravé“ subjektívne prežívanie sa podľa môjho názoru tieto predpoklady menia až v prítomnosti vedomia). Odteraz už teda nemôžem považovať púhu existenciu jedinečného a nevyjadriteľného ľudského subjektívneho prežívania za neprekonateľnú bariéru na ceste k strojovému vedomiu. Keď uvážime, že až donedávna som bol zarytým odporcom tézy, že raz vytvoríme stroj či robota obdareného vedomím, je to zo strany autorov obdivuhodný výkon. Ešte stále sa môže ukázať, že existujú isté typy subjektívneho prežívania vlastné iba ľuďom, ale to je už celkom iný problém. A zástancovia strojového vedomia sa môžu oprávnene pýtať: „Mohli by ste presnejšie špecifikovať, aké formy subjektívneho prežívania máte na mysli?“ Priznávam, že na túto otázku zatiaľ nedokážem uspokojivo odpovedať.

Zároveň mám voči Slomanovej a Chrisleyho argumentácii aj isté výhrady, hlavne pre jej „Dennettovské“ smerovanie, čiže snahu vyhnúť sa pojmu „vedomie“ v nereduktívnom zmysle. Bez neho si totiž nedokážem predstaviť, ako by sa mohli technické predpoklady, ktoré tak vynaliezavo sformulovali, premeniť na ozajstné „jedinečné a nevyjadriteľné subjektívne prežívanie“ inteligentného stroja obdareného vedomím. Zdá sa však, že podľa ich názoru žiadne také „plnohodnotné“ subjektívne prežívanie, ako si ho predstavujú neredukcionisti, nejestvuje, a práve to zrejme pokladajú za hlavný omyl neredukcionistov, čo dokladá aj ich zábavná úvodná kolekcia

„predvedeckých“ definícií vedomia. No aj keď pripustím, že zatiaľ nejestvuje žiadna všeobecne prijímaná definícia vedomia a že väčšina ľudí tento pojem používa chybne, stále nevidím, ako z toho vyplýva neexistencia vedomia v nereduktívnom zmysle, keď problém môže jednoducho spočívať v tom, že vzhľadom na jeho jedinečné vlastnosti je mimoriadne ťažké konzistentne ho popísať pojmami ľudského intelektu.

Ponaučenie, ktoré som si z tohto všetkého vzal, je v súlade s Chalmersovým záverom ohľadne materializmu prvého typu: kedykoľvek máme do činenia s hypotetickou entitou, ktorú sa nám z principiálnych dôvodov nedarí adekvátne popísať, je celkom normálne a prirodzené, že vedľa seba existujú a pôsobia dva tábory: tí, ktorí ju považujú za reálnu a pokúšajú sa o jej adekvátne pojmové vyjadrenie, a tí, ktorí o jej realite pochybujú a snažia sa vybudovať alternatívnu teóriu bez nej. V prípade vedomia obe línie zjavne produkujú cenné poznatky a navzájom sa motivujú k intenzívnejšiemu úsiliu o nájdenie definitívneho riešenia. Tento aspekt je podľa môjho názoru najcennejší a nestratí hodnotu, ani keby sa nakoniec ukázalo, že ani jedna z týchto línií nedokáže vytvoriť definitívny pojmový rámec pre adekvátny popis vedomia a subjektívneho prežívania.

* * *

Po prvom prečítaní Chalmersovho článku [2] som mal dojem, že jeho „naturalistický dualizmus“ principiálne vylučuje možnosť vytvoriť stroj obdarený vedomím, a je teda protipólom Slomanovej a Chrisleyho pozície. Iste si dokážete predstaviť môj šok, keď som neskôr na anglickej verzii Wikipédie našiel článok o strojovom vedomí (https://en.wikipedia.org/wiki/Artificial_consciousness), ktorý za najotvorenejšieho obhajcu jeho realizovateľnosti označil práve Chalmersa! „Ako je to možné?“ pýtal som sa sám seba. Wikipédia odkazovala na Chalmersov článok [5] z roku 2011, ktorý vychádzal z jeho nepublikovaného rukopisu z roku 1993. Keď som ho študoval, mal som dojem takmer úplnej zhody s názormi Slomana a Chrisleyho. „Hm, zdá sa, že to napísal, keď bol ešte materialistom,“ povedal som si. Nasvedčovali tomu aj Chalmersove vyjadrenia v biografickom článku [4], kde hovorí, že počas práce na knihe v polovici deväťdesiatych rokov dospel k záveru -- v rozpore so svojou pôvodnou mienkou -- že problém vedomia nemožno vyriešiť z pozície materializmu. A dodáva:

> *Nemyslím si, že veda o vedomí môže uspieť, ak bude čisto reduktívna, budovaná iba v pojmoch neurovedy či výpočtových algoritmov. Skôr si myslím, že to bude veda nereduktívna, ktorá sa neusiluje redukovať vedomie na nejaký fyzikálny proces, ale bude k nemu pristupovať ako k svojbytnej entite a hľadať spojivá medzi ním a mozgom, správaním a ďalšími kognitívnymi procesmi. [4]*

Na druhej strane, v relatívne nedávnej úvodnej poznámke k tomuto rukopisu v [5] Chalmers hovorí, že v globále stále „sympatizuje“ s jeho tézami, i keď sa už nestotožňuje s každým detailom. „Verí teda, že vedomie možno reprodukovať na počítači, alebo nie?“ mrmlal som si pod nos, kým som prehľadával jeho články v snahe nájsť dáku stopu. Nakoniec som si uvedomil, že som urobil podobnú chybu ako pri Slomanovom a Chrisleyho článku. Chalmers totiž v [2] poznamenáva, že už nie je materialistom, no zároveň svojím myšlienkovým experimentom s „prepínaním subjektívneho prežívania“ („*dancing qualia argument*“) obhajuje tézu „vypočítateľnosti vedomia“ („*computability of mind*“). Táto netypická kombinácia bola hlavným zdrojom môjho zmätku: dovtedy som totiž predpokladal, že téza vypočítateľnosti vedomia znamená funkcionalizmus a redukcionizmus zároveň, čiže prinajlepšom materializmus druhého typu. No Chalmers v tom istom článku redukcionizmus výslovne odmietol!

Ďalšou stopou boli Chalmersove úvodné slová v sekcii venovanej nereduktívnym kritikám jeho názorov v [3], kde poznamenal, že v predchádzajúcom článku [2] sa snažil o “zlatú strednú cestu”:

> *Umiernený charakter mojej pozície zrejme vyplýva z môjho sklonu k jednoduchosti a príklonu k vede. Reduktívny materializmus ponúka jednoduchý a v mnohých ohľadoch presvedčivý obraz sveta, a aj keď v prípade vedomia nefunguje, pokúsil som sa aspoň zachovať čo najviac z jeho príťažlivých čŕt. [3]*

Poslednou stopou bola skutočnosť, že vedomie má dve stránky: zážitkovú („*phenomenal*“), t.j. aké je to niečím byť alebo dačo zažívať, a funkčnú, t.j. účasť vedomia na mentálnym procesoch ako je uvažovanie, rozlišovanie a kategorizácia, verbálne popisy vnútorného stavu, atď. Zhruba by sa dalo povedať, že zážitkový aspekt vystihuje, čo vedomie *je*, kým funkčný aspekt, čo vedomie *robí*. Prvý sa v angličtine zvyčajne označuje ako „*phenomenal consciousness*“, čo by sa dalo preložiť ako „zážitkové vedomie“, resp. „subjektívne prežívanie“, kým druhý ako „*access consciousness*“, teda „prístupové vedomie“. Chalmers navrhuje mierne odlišnú terminológiu a označuje prvý aspekt ako „vedomie“ („*consciousness*“), kým druhý ako „uvedomovanie si“ („*awareness*“).

Pomerne dlho mi unikal zásadný význam poslednej stopy, a tak som zotrvával v chybnom presvedčení, že odklon od redukcionizmu nutne znamená aj úplné zavrhnutie tézy o vypočítateľnosti vedomia, teda o jeho možnej (re)produkcii v počítači či inom inteligentnom stroji. Až dodatočne som si uvedomil, že je možný aj polkrok, teda upustenie od vypočítateľnosti pre prvý aspekt vedomia pri jej súčasnom podržaní pre druhý aspekt. Toto sa nakoniec javí ako Chalmersova skutočná pozícia, a možno ju zhrnúť asi takto: „Nevieme síce vypočítať, čo vedomie *je*, ale vieme vypočítať, čo *robí*.“ Prvá časť zaručuje, že je to pozícia nereduktívna, kým druhá ju udržuje v rámci širšie chápaného funkcionalizmu. Chalmers sám ju nazýva aj „nereduktívnym funkcionalizmom“.

Je tento postoj filozoficky obhájiteľný? Chalmers verí, že áno, a predkladá aj kandidáta -- rámec inšpirovaný Russellovským monizmom. Opiera sa o skutočnosť, že

> *fyzika charakterizuje svoje základné entity iba zvonka („extrinsically“), teda v pojmoch príčin a následkov, a ponecháva ich vnútornú, intrinzickú povahu nešpecifikovanú. Aj keď vezmeme do úvahy všetko, čo nám fyzika hovorí napríklad o nejakej častici, stále je to len akési klbko kauzálnych tendencií („causal dispositions“); nedozvieme sa nič o entite, ktorá je ich nosičom. To isté platí o základných atribútoch, ako je hmotnosť či náboj: v konečnom dôsledku predstavujú aj ony iba zložité kauzálne tendencie (mať hmotnosť znamená vzdorovať zrýchleniu určitým spôsobom, atď.). No vždy, keď narazíme na nejakú kauzálnu tendenciu, môžeme pátrať po jej kategoriálnom základe, čiže pýtať sa: aká entita je jej zdrojom a nosičom? [3]*

Ak sa tejto otázke vyhneme postulátom, že svet pozostáva iba z kauzálnych tendencií, poznamenáva Chalmers, zostane nám len

> *obrovská pavučina príčin a následkov bez akýchkoľvek entít, ktoré by prostredníctvom nej vstupovali do vzájomných vzťahov! Základné častice fyziky a ich atribúty by sa premenili na prázdne menovky („placeholders“) ... a svet by razom stratil všetku svoju substanciu. [3]*

Idea „čisto štruktúrneho“ či „čisto kauzálneho“ sveta je svojím spôsobom príťažlivá, uznáva Chalmers, no nie je celkom isté, či je aj vnútorne súdržná („*coherent*“). To ho vedie k nastoleniu dvoch otázok:

> *(1) Aká je vnútorná, intrinzická povaha fyzikálnej reality? (2) A aké miesto v prirodzenom usporiadaní sveta patrí intrinzickým vlastnostiam subjektívneho prežívania („experience“)? Russell si ako prvý povšimol ... že tieto dve otázky môžu súvisieť. Možno intrinzické vlastnosti, nosiče pozorovateľných fyzikálnych tendencií, priamo predstavujú subjektívne prežívanie („are themselves experiential properties“), alebo sú aspoň jeho zárodkami („some sort of proto-experiential properties“), pričom subjektívny zážitok vzniká nejakou ich kombináciou. Subjektívne prežívanie by tak získalo miesto vo vnútri siete príčin a následkov, ktorú popisuje fyzika; nebolo by k nej viac prilepené zvonka ako dáky nepotrebný prívesok. Zároveň by tým prevzalo rolu, o ktorej možno povedať, že ju už bolo načase vyplniť. A čo je najdôležitejšie, nenarušilo by tým kauzálnu uzavretosť fyzického sveta: sieť fyzikálnych príčin a následkov by mala rovnaký tvar ako predtým; iba by sme vyfarbili jej uzly. [3]*

Zaujímavú úvahu o tom, prečo by intrinzické vlastnosti fyzickej reality mali mať vôbec dačo spoločné s vedomím, možno nájsť v článku G. Rosenberga [8]. S tým úzko súvisí aj otázka, či Russellovský monizmus prijatím kauzálnej uzavretosti fyzického sveta nevedie nutne k

epifenomenalizmu, teda k názoru, že vedomie nedokáže fyzický svet nijako ovplyvniť. V tomto ohľade Chalmers tvrdí, že umiestnením subjektívneho prežívania do uzlov kauzálnej siete mu priraďujeme kauzálnu rolu:

> *Zárodky subjektívneho prežívania („proto-experiences“) sa tak stanú základom kauzality na najnižších úrovniach. Môžeme teda očakávať, že vyššie formy subjektívneho prežívania, napríklad ľudské, zdedia kauzálnu relevantnosť elementárnych zážitkov či proto-zážitkov („proto-experiences“), z ktorých vzišli. Získame tak oveľa ucelenejší obraz o mieste vedomia v prirodzenom usporiadaní sveta. [3]*

Samozrejme, aj Russellovský monizmus má svoje problémy. Chalmers sám spomína „hrozbu panpsychizmu“ a kompozičný problém, teda hypotetický spôsob, akým by postulované elementárne subjektívne zážitky či proto-zážitky na mikroskopickej úrovni mali spolu konštituovať či inak vytvárať komplexné a ucelené subjektívne prežívanie vlastné ľuďom. K tomu treba pridať aj výhrady viacerých výskumníkov pristupujúcich k vedomiu z nereduktívnych pozícií, že Chalmers predsa len zostal príliš blízko funkcionalizmu. Aby som priblížil ich názory, stručne predstavím kritiku Chalmersovho článku [2] z pera E.J. Loweho [7].

Lowe oceňuje, že Chalmers odmieta „spohodlnelé predpoklady reduktívneho fyzikalizmu“, no obáva sa, že Chalmersov prístup aj tak „hrá fyzikalistom do karát tým, že budí dojem, akoby jediným problémom funkcionalizmu bola jeho neschopnosť objasniť subjektívne prežívanie (*‘qualia’*).“ Chalmersovo poňatie vedomia a subjektívneho prežívania má podľa neho vážne slabiny: nezohľadňuje napríklad, ako hlboko a neoddeliteľne je intencionálny (t.j. reprezentačný, sémantický a vzťahový) obsah vnímania ukotvený v jeho zážitkovej (*‘phenomenal’*) stránke. Namieta tiež proti použitiu Shannonovskej definície informácie, ktorú považuje za absolútne nevhodnú pre popis ľudských kognitívnych stavov. Nepáči sa mu ani Chalmersov terminologický návrh ohľadne pojmov vedomia (*‘consciousness’*) a uvedomovania si (*‘awareness’*):

> *Zdá sa mi, že podľa Chalmersovej definície „uvedomovania si“ by v princípe nebolo nič zlé na tvrdení, že počítač, ba aj termostat si „uvedomujú“ isté veci, no keby sme následne začali tvrdiť, že ľudia si tiež „uvedomujú“ veci iba v tomto oslabenom zmysle slova, úplne by sme tým skreslili povahu schopností, vďaka ktorým si „uvedomujeme“ sami seba, svoje myšlienky i subjektívne zážitky. [7]*

Hlavná Loweho výhrada smeruje proti Chalmersovej téze, že ľudské myslenie (a kognícia všeobecne) spočíva len v spracovaní informácií, ktoré môže v princípe rovnako dobre prebiehať aj v počítači. To podľa Loweho zvádza k dojmu,

> *že jedinou skutočne odlišnou črtou vedomia je jeho kvalitatívna či zážitková („phenomenal“) stránka („aké to je, byť niekým či niečím“, resp. „vnútorný, subjektívny pocit“). A vtedy sa začne javiť ako dáka čudná záhada, ako zvláštny vrtoch evolúcie, že bytosti ako my sú obdarené práve týmto typom vedomia na dôvažok ku všetkým našim intelektuálnym a kognitívnym schopnostiam -- berúc do úvahy, že podľa Chalmersa sú to všetko len schopnosti spracovať a uchovať informácie vhodným spôsobom. [7]*

Loweho riešenie tejto „čudnej záhady“ je,

> *že vedomie dostal do tejto prekérnej situácie práve Chalmers (a, ak mám byť spravodlivý, aj mnohí iní) tým, že mu chybne odmietol prisúdiť aktívnu rolu vo svojej koncepcii ľudského myslenia a chápania. Je to skrátka reduktívny a zhola nedostatočný „informačno-procesný“ pohľad na ľudskú kogníciu, ktorý vyvoláva klamný dojem, že „vedomie“ (vo forme subjektívneho prežívania a pod.) hrá púhu epifenomenálnu rolu podivnej doplnkovej črty ľudskej psychiky, ktorá je pre naše základné intelektuálne schopnosti prakticky bezvýznamná. [7]*

Chalmers v odpovedi na Loweho kritiku uviedol, že v článku [2] nebolo jeho úmyslom tvrdiť, že by si ľudia „uvedomovali“ veci iba v rovnakom (oslabenom) zmysle ako stroje. Podobne nechcel

rozoberať „presnú povahu vzťahu medzi vedomím a ‘intencionálnymi’ (čiže sémantickými) mentálnymi stavmi“, pretože to sú „veľmi hlboké a subtílne otázky“, ktoré by šli nad rámec uvedeného článku. „Čo sa týka intencionality,“ píše Chalmers, „som rozpoltený“:

> *Na jednej strane si uvedomujem dôležitosť jej fenomenálnej, zážitkovej zložky, no na druhej strane ma lákajú aj možnosti funkčných analýz špecifických intencionálnych obsahov.... Čím ďalej, tým viac sa kloním k [myšlienke]..., že vedomie je primárnym zdrojom významu, či ako to hovorí Lowe, že intencionálny obsah vedomia je ukotvený v obsahu fenomenálnom. [3]*

Táto Chalmersova odpoveď, stará dvadsať rokov, akoby naznačovala možnú modifikáciu jeho tézy o vypočítateľnosti „uvedomovania si“, teda funkčného aspektu vedomia. Samozrejme, vzhľadom na jeho nedávnu zmienku v [5], že „v globále stále sympatizuje s tézami“ svojho raného rukopisu, nemožno očakávať radikálnu zmenu v jeho názoroch. Ďalšie detaily sa pravdepodobne dozvieme, až keď sa skončí diskusia o jeho rukopise v časopise *Journal of Cognitive Science*.

Dovolím si uzavrieť tento článok osobnou poznámkou. Nie som v pozícii, aby som mohol o diskutovaných témach vynášať autoritatívne súdy, no to mi nebránilo aktívne formulovať vlastné hypotézy a sledovať, ako sa vysporiadajú s predkladanými otázkami a problémami. Aj keď moje hypotézy zväčša neobstáli, vďaka svojmu úsiliu teraz oveľa lepšie chápem problémy späté s pojmom „vedomie“. Zrejme nikoho neprekvapím, keď ako osoba s duchovnými sklonmi prezradím, že najhlbšie vo mne rezonovali názory E.J. Loweho. Zároveň však považujem Chalmersov „nereduktívny funkcionalizmus“ za neoceniteľný príspevok, za akúsi „minimalistickú“ verziu nereduktívneho prístupu k vedomiu a nádejný most, po ktorom raz môžu obojsmerne prúdiť výskumné výsledky medzi reduktívnym a nereduktívnym výskumným táborom.

Z môjho pohľadu sa dá Chalmersov najdôležitejší výsledok v tých článkoch, ktorými som sa dosiaľ zaoberal, vyjadriť tézou: „Aj keby sme pripustili, že funkčné aspekty ľudského vedomia sú len špecifickou formou strojového spracovania informácií, jeho zážitkový, fenomenálny aspekt aj tak zostane neredukovateľný na štruktúru a funkcie.“ Z tohto hľadiska je Chalmersovo „zapredanie sa funkcionalizmu“, za ktoré ho kritizoval Lowe a ďalší, vlastne jeho najsilnejšou zbraňou. Tak ako sú v matematike vety s najslabšími predpokladmi najširšie použiteľné, tak je aj Chalmersov výsledok mimoriadne pôsobivý pre tých, ktorí stoja na rozhraní medzi reduktívnym a nereduktívnym poňatím vedomia.

V neposlednom rade by som chcel zdôrazniť, že Slomanove a Chrisleyho myšlienky o architektúre inteligentných strojov považujem za mimoriadne prínosné, a paradoxne práve tie nakoniec možno aj budem prakticky využívať vo svojej práci na inteligentných a znalostných technológiách. Hoci sa Sloman a Chrisley názorovo klonia k Dennettovi, moje pôvodné chápanie ich základnej tézy -- že napríklad kalkulačka už tým, že vie narábať s číslami, musí si ich istým spôsobom uvedomovať -- bolo jednoznačne chybné. Sloman a Chrisley to tvrdia len o vysoko inteligentných strojoch so sofistikovanou vrstvou meta-manažmentu, ktoré sú schopné sebareflexie, čiže dokážu skúmať vlastné fungovanie. To, že neredukcionisti s týmto názorom nesúhlasia, nijako neznižuje praktickú hodnotu takých strojových architektúr. Ak máme v dohľadnej dobe vyvinúť roboty s inteligenciou blížiacou sa ľudskej (a ja osobne verím, že je to možné), potom je podľa mňa viac než pravdepodobné, že ich dizajn bude inšpirovaný výsledkami, ktoré dosiahol Sloman a jeho kolegovia od konca sedemdesiatych rokov minulého storočia. Loweho námietka (zdieľaná aj ďalšími nereduktívnymi výskumníkmi), že ľudská kognícia funguje celkom inak, podľa mňa nijako neznižuje praktickú užitočnosť týchto architektúr pre tvorbu inteligentných strojov. Navyše som videl niekoľko pozoruhodných obrán funkcionalizmu i vo vzťahu k ľudskej kognícii, takže táto diskusia ešte zďaleka nie je uzavretá.

A čo je ešte dôležitejšie, Slomanove a Chrisleyho architektúrne idey sa dajú priamo včleniť do Russellovského nereduktívneho rámca propagovaného Chalmersom. Tak by mohli získať chýbajúci zárodok subjektívneho prežívania nutný na to, aby sa ich funkčné predpoklady pre „nevyjadriteľnosť“ vnútorného stavu stroja premenili na opravdivé subjektívne prežívanie. V tejto

kombinácii ide o veľmi silný argument v prospech realizovateľnosti strojového či umelého vedomia, teda idey, ktorú som až dosiaľ vehementne odmietal. Teraz sa mi javí ako takmer nespochybniteľná. Jej praktické uskutočnenie však nebude ľahké. Nestačí prosto vyhlásiť: „Tu máme robota, ktorého inteligencia sa blíži ľudskej, a keďže sme sa zhodli, že intrinzické vlastnosti hmoty zahŕňajú aj zárodky subjektívneho prežívania, nie je najvyšší čas uznať, že tento a podobné roboty majú vedomie?“ Také ľahké to nie je, lebo nič nám nezaručuje, že keď budeme konštruovať stále inteligentnejšie roboty, zárodky subjektívneho prežívania prislúchajúce ich stavebným častiam sa pritom budú automaticky zoskupovať do vyspelejších foriem opravdivého subjektívneho prežívania. Vôbec to tak nemusí byť, a potom získame len inteligentných robotov bez vedomia, či presnejšie len s takými zárodkami subjektívneho prežívania, aké Russellovský monizmus priznáva hocakému kusu amorfnej hmoty. Otázka vedomia je pre mňa niečo iné než otázka inteligencie, a celkom ma nepresvedčil ani Chalmersov myšlienkový experiment s „prepínaním subjektívneho prežívania“ („*dancing qualia argument*“), lebo ten robí z funkčnej špecifikácie systému kľúčový prvok, ktorý bezo zvyšku určuje aj jeho subjektívne prežívanie. Moje pochybnosti pramenia z toho, že funkčné špecifikácie sú vyjadrené vo vonkajších, formálnych (extrinzických) pojmoch, kým intrinzické vlastnosti majú byť bohatšie a *niesť* tie extrinzické. Je teda otázne, nakoľko funkčná špecifikácia postačí pre strojovú reprodukciu vedomia (na rozdiel od inteligencie). Chalmers si je týchto problémov vedomý a uznáva, že proces vzniku subjektívneho prežívania z jeho zárodkov pravdepodobne nepodlieha pravidlám fyzickej kompozície, aj keď zároveň dúfa, že by mohol podliehať pravidlám *informačnej* kompozície. To sú zložité otázky, do ktorých sa teraz nemôžem púšťať, obmedzím sa teda na to hlavné: aj keď je momentálne cesta k realizácii umelého vedomia nejasná, Russellovský monizmus takú možnosť zjavne pripúšťa, aspoň v princípe.

Môj definitívny záver je teda nasledovný: Chalmers podľa mňa presvedčivo obhájil neredukovateľnosť ľudského vedomia, no z jeho argumentov zároveň vyplynulo aj dačo, čoho sa neredukcionisti spravidla štítia: že je možná aj strojová reprodukcia vedomia, aspoň v princípe. Je tu teda prvok prekvapenia pre oba tábory, čo považujem za znak opravdivého a významného pokroku na ceste k hlbšiemu pochopeniu ľudského vedomia.

Skončím tým, že sa ešte raz vrátim k svojmu východiskovému bodu: tento článok vznikol ako neformálne „nakuknutie“ do fascinujúcej oblasti na priesečníku mojich osobných a profesionálnych záujmov. Snažil som sa opierať o verejne dostupné zdroje a v referenciách uvádzať aktuálne webové linky na citované články. Čo sa týka reakcií na Chalmersov prvý článok [2], tie sa pôvodne objavili v časopise *Journal of Consciousness Studies*, ktorý nanešťastie nie je voľne dostupný, no napríklad Loweho reakcia [7] bola dodatočne publikovaná aj v otvorenom internetovom časopise *Antimatters*, odkiaľ ju stále možno stiahnuť. Niektoré ďalšie reakcie sú tiež voľne dostupné; linky na ne možno nájsť v Chalmersovej súhrnnej odpovedi na kritiku a komentáre [3]. V slovenčine je s podobnou tematikou dostupný napríklad zborník *Myseľ, inteligencia a život*. Teším sa na príležitosť oboznámiť sa s následným vývojom a diskusiami v tomto priesečníku umelej inteligencie, kognitívnej vedy a filozofie mysle. Pevne verím, že nový materiál a diskusie ma obohatia aspoň tak, ako idey prezentované v tomto článku.

Referencie:

[1] Block, N. (1996). What is functionalism? http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/functionalism.html (Originally in *The Encyclopedia of Philosophy Supplement*, Macmillan, 1996)

[2] Chalmers, D. J. (1995). Facing up to the problem of consciousness. http://consc.net/papers/facing.html (Originally in the *Journal of Consciousness Studies* 2:200-19)

[3] Chalmers, D. J. (1997). Moving forward on the problem of consciousness. http://consc.net/papers/moving.html (Originally in the *Journal of Consciousness Studies* 4:3-46)

[4] Chalmers, D. J. (2009). Mind and Consciousness: Five Questions. http://consc.net/papers/five.pdf (Originally in Patrick Grim, ed. *Mind and Consciousness: Five Questions*. Automatic Press, 2009.)

[5] Chalmers, D. J. (2011). A Computational Foundation for the Study of Cognition. http://consc.net/papers/computation.html (Originally in the *Journal of Cognitive Science* 12:323-57.)

[6] Kvassay, M. (2011). Psychological Foundations of Sri Aurobindo's Philosophy and His Approach to the Problem of Evil. http://marcelkvassay.net/article.php?id=psychological

[7] Lowe, E. J. (1995). There are no easy problems of consciousness. http://anti-matters.org/articles/46/public/46-41-1-PB.pdf (Originally in the *Journal of Consciousness Studies* 2:266-71.)

[8] Rosenberg, G. H. (1999). On the Intrinsic Nature of the Physical. http://www.newdualism.org/papers/G.Rosenberg/1998_Tucson_Paper.htm (Originally in *Toward a Science of Consciousness III, The Third Tucson Discussions and Debates*, 1999.)

[9] Sloman, A. (1988). Why Philosophers Should be Designers. http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-dennett-bbs-1987.pdf (Commentary submitted to *The Behavioral and Brain Sciences Journal*.)

[10] Sloman, A. & Chrisley, R. (2003). Virtual Machines and Consciousness. http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-chrisley-jcs.pdf (Originally in *Journal of Consciousness Studies*, 10:113–172)

[11] Šrí Aurobindova integrálna joga a mapa nadvedomia. http://aurobindo.sk/article.php?id=sk-SA02