Rozhovor

O vědomí, jazykových modelech, virtuální realitě a odvoditelnosti[1]

David J. Chalmers
New York University
chalmers@nyu.edu

Jakub Mihálik
Filosofický ústav AV ČR, v. v. i., Praha
mihalik@flu.cas.cz

Tento rozhovor vznikl v rámci výzkumného pobytu, který jsem jako stipendista Fulbrightova programu absolvoval pod vedením profesora Chalmerse v Centru pro mysl, mozek a vědomí *na Newyorské univerzitě (NYU) v akademickém roce 2022/2023. Toto výzkumné centrum, zaměřené na filosofii mysli, vede tento původem australský filosof společně s prof. Nedem Blockem. Rozhovor jsme museli několikrát odkládat vzhledem k bleskovým záplavám, které v červenci 2023 postihly údolí řeky Hudson, kde profesor Chalmers trávil část své letní dovolené. J.M.*

Jakub Mihálik (J. M.): *Mluvíme spolu krátce poté, co proběhly dvě velké konference věnované tématu vědomí, jedna na Sicílii a druhá – kterou jste spolu-organizoval – na NYU v New Yorku, přičemž na každé z nich vystoupily stovky přednášejících. Není, myslím, nijak přehnané říci, že Vaše filosofická práce inspirovala toto systematické a interdisciplinární úsilí o porozumění vědomí. Je snad možné říci, že jste přispěl k tomu, že vědomí už není považováno za obskurní badatelské téma. Rád bych se proto na úvod zeptal, jak se díváte na aktuální stav výzkumu vědomí. Která témata nyní považujete v rámci tohoto výzkumného úsilí za klíčová?*

1 Rozhovor vznikl v rámci výzkumné činnosti Oddělení analytické filosofie Filosofického ústavu AV ČR, v. v. i., v Praze, s podporou Fulbrightova programu. **DOI:** https://doi.org/10.46854/fc.2024.1r.105. Pozn. red.

David Chalmers (D. Ch.): Tato výzkumná oblast se postupem času stala velice mnohotvárnou. Její kořeny vyrůstají z řady různých vědeckých oborů, jako jsou psychologie a neurověda, do jisté míry také informatika, fyzika a některé oblasti chemie, přičemž zároveň velmi významně souvisí s filosofií. Věda o vědomí má ovšem dlouhou tradici – v 19. století byla reprezentována významnými osobnostmi, z nichž mnohé byly filosofy a zároveň psychology. Za všechny je možno uvést například Williama Jamese. Ve dvacátém století však nastalo dlouhé období, kdy problematika vědomí nebyla ve vědě příliš uznávaná a ve filosofii se jí dostávalo jen občasné pozornosti. Hodně se to změnilo až v 90. letech, ale neměl jsem na tom, myslím, tak úplně jednoznačnou zásluhu. V roce 1990 Francis Crick a Christof Koch napsali článek, v němž prohlásili, že vědomí se opět stalo významným tématem neurovědy.[2] Ve filosofii jsme zase měli Dana Dennetta, který v téže době psal knihu *Consciousness Explained*, jež vyšla v roce 1991 – ta byla opravdu velice důležitá. Významně přispěli také další lidé, například Bernie Baars v psychologii, Roger Penrose ve fyzice a mnozí další odborníci, včetně mnoha filosofů. Během devadesátých let se všechny tyto proudy postupně propojily v jeden obor a já měl to štěstí, že jsem se zúčastnil některých raných konferencí, kde k tomu propojování postupně docházelo. Úplně první konference v Tucsonu v roce 1994 svedla dohromady badatele z různých oborů a panovala zde opravdu tvůrčí atmosféra. Pak byla založena Asociace pro vědecký výzkum vědomí (*Association for the Scientific Study of Consciousness*, zkratka *ASSC*), která po roce 1997 začala pořádat každoroční setkání. Byla to velice vzrušující doba a bylo skvělé být součástí toho všeho.

Během posledních téměř třiceti let, jež uběhly od první tucsonské konference, jsme svědky toho, jak náš obor roste, rozšiřuje se a v mnohém dozrává. Nemyslím si, že mnoho problémů spojených s vědomím bylo již jednoznačně vyřešeno – možná dokonce žádný z nich. Filosofické otázky týkající se vědomí jsou stále velmi obtížně řešitelné, ale řekl bych, že nyní alespoň lépe chápeme, jaké možnosti pro filosofické přístupy k vědomí se nám otevírají. Co se týče vědy, máme přinejmenším kvalifikovanější poznatky o neurálních korelátech vědomí a o vztahu mezi vědomými a nevědomými procesy. Také jsme začali budovat spojení s mnoha klinickými obory, jako jsou neurologie, anesteziologie atd., přičemž někdy se při tom používají nástroje pocházející z vědy o vědomí. Řekl bych tedy, že náš obor během posledních třiceti let opravdu hodně vyzrál. Zároveň mám ale za to, že se věda o vědomí stále ještě nachází v raném stadiu, má před sebou ještě dlouhou cestu a za sto let

2 Crick, F. – Koch, Ch., Towards a Neurobiological Theory of Consciousness. *Seminars in the Neurosciences*, 2, 1990, s. 263–275.
– Autorem této i všech následující poznámek pod čarou je Jakub Mihálik. Pozn. red.

se pravděpodobně budeme za touto dobou ohlížet jako za érou, kdy jsme se v dané oblasti teprve snažili zorientovat.

J. M.: *To mě vede k otázce, jak vidíte současnou roli filosofie v mezioborovém výzkumu vědomí? Proměňuje se podle Vás ona role? Myslíte si, že se možná vědomí stane za nějakou dobu čistě vědeckým problémem, nebo bude hlas filosofie v souvislosti s ním vždy významný?*

D. Ch.: Osobně v našem oboru nevedu žádné silné rozlišení mezi filosofií a vědou. Je zjevné, že některé výzkumy jsou převážně empirické, zahrnují experimenty atd., zatímco jiné aktivity jsou spíše teoretické. Ale dost často se stává, že teoretický příspěvek k poznání vědomí může zformulovat vědec – a naopak filosofové se, alespoň v některých případech, sami podílejí na experimentech. Filosofové se každopádně velmi vážně zabývají experimentálními otázkami. Shrneme-li to, pak filosofové podle mě ve výzkumu vědomí hrají několik rolí. Jednou z nich je samozřejmě formulování problémů a vznášení otázek, které mají být zodpovězeny. Podle mě je obecně klíčovou rolí filosofů formulovat výzvy pro vědu.

Další významná interakce pak spočívá v interpretaci vědy. Když je proveden nějaký experiment, může se objevit vědec, který z něj bude chtít vyvodit dalekosáhlý závěr typu „tohle prokazuje tohle o povaze vědomí nebo o neurálních korelátech vědomí". Filosofové pak mohou říci: „zpomalte, tohle z toho neplyne, tady vycházíte z určitého sporného filosofického předpokladu." O takovéto sporné filosofické předpoklady se vědci zkoumající vědomí opírají často a podle mě je důležitým úkolem filosofů je odhalovat.

Jsou tu ale také určité projekty, v jejichž rámci filosofové a vědci spolupracují na společných úkolech a snaží se vědomí pochopit. Já jsem například nyní členem skupiny lidí, kteří uvažují o tom, jak vytvořit testy vědomí – a to je samozřejmě jak hluboce filosofický problém, tak zároveň problém, který vyžaduje zohlednění mnoha vědeckých poznatků. Zdá se mi nicméně – a jsem rád, že tomu tak je –, že filosofové a vědci nyní stále častěji spojují síly v takovýchto projektech.

Další role filosofů pak samozřejmě spočívá v analýze problémů souvisejících s tradičními filosofickými aspekty vědomí – metafyzikou, epistemologií atd. Myslím si, že filosofické zkoumání vědomí nemusí být navázáno na vědu, je sice skvělé, když navázáno je, ale máme také řadu velmi dobrých prací o vědomí, které nejsou nijak zvlášť zakotveny ve vědeckém výzkumu. Pokud pracujete na tradiční metafyzické otázce vztahu mysli a těla, bude empirický výzkum někdy relevantní a jindy nikoli, ale tato práce bude tak jako tak důležitá.

J. M.: *Je tomu zhruba 30 let od té doby, kdy jste zformuloval „těžký problém" vědomí – tedy otázku, jak mozek utváří vědomé stavy, u nichž je pro daný subjekt nějaké* (there is something it's like) *je zakoušet, neboli otázku, proč se všechno to dění v mozku neodehrává tak řečeno „ve tmě". Domníváte se, že jsme významně pokročili v řešení tohoto obtížného problému během tří dekád od doby, kdy jste jej zformuloval?*

D. Ch.: Rád bych řekl, že podle mého názoru formulace těžkého problému vědomí sama o sobě nebyla nějakým významně originálním příspěvkem. Řada lidí v té době už chápala, že ten problém existuje a že je velmi obtížně řešitelný. Myslím si, že jsem jen víceméně zvolil příhodné a užitečné označení, které pravděpodobně pomohlo ten problém zvýraznit a učinilo obtížnějším se mu vyhnout. Avšak sám ten problém má velmi dlouhou historii. Verze těžkého problému lze jistě najít u Leibnize, u Huxleyho,[3] podle některých odborníků i u Isaaca Newtona. Je ale potěšující pozorovat v posledních třiceti letech, že mnozí – určitě ne všichni, ale je jich docela hodně – vědci, kteří se zabývají tématem vědomí, chápou těžký problém nejen jako výzvu pro filosofii, ale i jako výzvu pro sebe samé. Často lze vypozorovat, že se vědci tomuto problému přímo věnují, a tak tomu dříve nebývalo. Existují přitom tisíce různých přístupů k řešení tohoto problému: někteří vědci zastávají silně realistický, nereduktivní přístup k vědomí, který nacházíme například v Tononiho teorii integrované informace (*integrated information theory*), další hájí široce iluzionistický přístup, jaký kupříkladu nacházíme v teorii schématu pozornosti (*attention schema theory*) Michaela Graziana, jiní pak zastávají v širokém smyslu komputační či funkcionalistický přístup, jaký nalezneme u teorie globálního pracovního prostoru (*global workspace theory*) Stana Dehaena, další zase mají k těžkému problému deflacionistický přístup atd. Bylo velmi zajímavé podrobně sledovat během uplynulých 30 let všechny tyto pokusy a závěry, k nimž dospěly.

Nikdo dosud podle mě nenalezl „svatý grál", kterým by bylo vysvětlení vědomí v jazyce fyziky – já sám mám za to, že něco takového není možné –, ale bylo zajímavé sledovat, jak se o to mnozí pokoušejí. Zároveň si myslím, že filosofové také dokázali k tomuto problému mnohé říci: někteří například navrhli fyzikalistické strategie řešení těžkého problému, zatímco jiní se zaměřili na méně ortodoxní přístupy, jako jsou panpsychismus, různé formy dualismu, ale i formy iluzionismu.

Ve své vlastní práci jsem v posledních letech zkoumal, nakolik si různé konstruktivní přístupy k těžkému problému dokáží poradit s obtížemi, kterým každý z nich čelí – panpsychismus čelí problému kombinace, dualismus

3 Jde o britského biologa T. H. Huxleyho (1825–1895).

čelí problému interakce a tak dále. Pokoušel jsem se vybrat nejlepší varianty těchto stanovisek a zjistit, jak daleko s nimi lze následně postoupit při řešení dané obtíže. Prozatímní závěr přitom obvykle zní, že jsme daný problém dosud uspokojivě nevyřešili, každé stanovisko čelí obtížím, které dosud nebyly adekvátně překonány. Bylo nicméně skvělé vidět, jak se tolik vědců a filosofů během posledních dekád ve své práci na těžký problém zaměřuje.

J. M.: *Máme tu tedy těžký problém, ale pak také problém relativně jednodušší: identifikovat neurální základ vědomí, tedy „neurální koreláty vědomí". Je obecně známo, že na již zmíněné newyorské konferenci ASSC došlo k vypořádání sázky, kterou jste před 25 lety uzavřel s neurovědcem Christofem Kochem. Sázka se týkala toho, zda budeme dnes, tedy po 25 letech od jejího uzavření, znát neurální koreláty vědomí. Mohl byste mi říci něco o výsledku té sázky a o aktuálním výzkumu neurálních korelátů vědomí?*

D. Ch.: Problém neurálních korelátů vědomí je skutečným středobodem vědy o vědomí už od doby, kdy jej Francis Crick a Christof Koch kolem roku 1990 nastolili jakožto hlavní téma této disciplíny. Jeho krása spočívá částečně v tom, že nevyžaduje, abychom našli řešení těžkého problému dříve, než najdeme neurální koreláty vědomí. Dosud jsme sice nedokázali vysvětlit vědomí, ale objasnění korelace je možná jednodušší než toto vysvětlení. Problematika korelace je navíc relativně neutrální projekt, do jehož řešení se může zapojit každý.

V roce 1998 proběhla druhá konference ASSC, tentokrát na téma neurálních korelátů vědomí. Christof a já jsme na ní oba byli a oba jsme mluvili o tomto tématu. A jednoho večera Christof řekl: „Myslím, že tam dospějeme! Myslím, že nejsme daleko! My ten problém vyřešíme – alespoň neurální koreláty vědomí do 25 let najdeme!" Já jsem odvětil: „To si nemyslím – možná, že dokážeme nalézt neurální koreláty vědomí, ale je to podle mě moc složité na to, abychom to stihli za 25 let." Skončilo to tím, že jsme uzavřeli sázku o několik lahví vína, která měla být rozhodnuta po 25 letech, konkrétně v roce 2023. To je samozřejmě letos, takže jsme na konferenci ASSC uspořádali večer, v jehož rámci byla sázka vypořádána. Při tom jsme se zároveň podívali nejen na vývoj v našem oboru v uplynulém mezidobí, ale podrobně také na nejnovější výsledky, které vzešly z projektu spolupráce soupeřů (*adversarial collaboration*), tedy zastánců konkurenčních teorií neurálního základu vědomí, kteří společně pracovali na empirických testech, které by tyto teorie mohly potvrdit či vyvrátit.[4] Konkrétně šlo o experiment, který měl posoudit predikce Tononiho teorie integrované informace a Dehaenovy teo-

4 Jde o projekt COGITATE – vice info lze nalézt na: URL www.arc-cogitate.com; [cit. 10. 1. 2024].

rie globálního pracovního prostoru. Když jsme porovnali výsledky, dospěli jsme k závěru, že ani jedné straně nedaly tak zcela zapravdu. U každé z konkurenčních stran došlo ke zpochybnění některých predikcí a celkový závěr, k němuž jsme dospěli, je takový, že ohledně neurálních korelátů vědomí stále neexistuje shoda. To je, zhruba řečeno, to, co jsem vlastně celou dobu očekával. Ve světle toho všeho Christof uznal porážku a zapili jsme to dobrým vínem. Abych k němu byl ale spravedlivý, byl to on, kdo v této sázce zariskoval. Já jsem nemusel dělat nic, stačilo jen říci „Ne", takže z mé strany tomu není tak, že bych něčeho dosáhl. Byl to nicméně pěkný způsob, jak představit některé pokroky v oblasti hledání neurálních korelátů vědomí, které se objevily v posledních 25 letech.

J. M.: *Rád bych se teď posunul k Vaší nedávné práci na otázce, zda tzv. „velké jazykové modely", jako je Chat GPT, mohou být vědomé či cítící. Tyto AI (artificial intelligence, tj. umělá inteligence) systémy, primárně trénované k predikci textu, projevily v posledních letech schopnosti překvapivě podobné těm lidským. Dokáží psát básně, shrnutí, možná i studentské eseje a sofistikovaně s námi konverzovat. Běžně přitom máme za to, že takovéto dovednosti jsou vyhrazeny lidem, tedy vědomým bytostem. Proč myslíte, že jsou mnozí z nás i přesto zdrženliví v přisuzování vědomí velkým jazykovým modelům a někteří, jako například Emily Benderová a její kolegové, dokonce nazvali tyto modely „stochastickými papoušky".*

D. Ch.: Celá otázka vědomí u AI systémů je samozřejmě velmi obtížně zodpověditelná a vysoce kontroverzní. Mnoho lidí má výhrady i k předpokladu, že by vůbec nějaký AI systém mohl být vědomý. V jedné z našich anket mezi filosofy se ukázalo, že možná jen polovina dotazovaných filosofů je otevřena závěru, že by AI mohla být vědomá, zatímco druhá polovina to odmítá.[5] Možná mají za to, že vědomí je nutně biologické, nebo že je naopak nutně nefyzické. Osobně ale nenahlížím, proč by křemík jakožto základ vědomí měl být v principu horší než jiné prvky, a tak jsem vždy stál na straně možnosti vědomí AI.

Až donedávna nicméně na ničem z toho z praktického hlediska moc nezáleželo, protože jsme měli k dispozici jen velice primitivní AI systémy, které nevykazují příliš známek vědomí. Nové jazykové modely, které se objevily v poslední době, jsou ale najednou velice sofistikované. O mnoha otázkách uvažují sofistikovanými způsoby, dokáží s vámi mluvit, vést rozhovor. Najednou vykazují chování, které bychom, alespoň pokud by jeho původcem byl člověk, chápali jako známku vědomí. Tohle tedy představuje přinejmenším

5 Viz Bourget, D. – Chalmers, D. J., Philosophers on Philosophy: The 2020 PhilPapers Survey. *Philosophers' Imprint*, 23, 2023, No. 11. DOI: https://doi.org/10.3998/phimp.2109.

velkou výzvu a je velmi zajímavé o tom uvažovat. Zároveň je samozřejmě pravda, že tyto jazykové modely jsou trénovány tím, že imitují velké množství lidského textu, což vede Emily Benderovou a další k tomu, že říkají, že jde o stochastické papoušky. Ale já nevím – pouhý fakt, že jsou trénovány na lidském textu, ještě nevylučuje, že vykazují velmi silné schopnosti uvažovat, tedy schopnosti, které by mohly být spojeny s vědomím. Mohlo by se například ukázat, že daný AI systém dokáže nejlépe predikovat lidský text, pokud má zakódovaný velmi hluboký model světa, podobný tomu jaký mají lidé, nebo pokud je dokonce vědomý. Podle mě tedy určitě není vyloučeno, že jsou tyto modely vědomé. Možná to není pravděpodobné, aktuálně tu stále ještě existuje mnoho překážek. Ale v článku, který jsem o tomto problému napsal, jsem se pokusil ukázat, že – platí-li některé předpoklady – existuje alespoň cca 25 % pravděpodobnost, že by tyto jazykové modely mohly být do deseti let vědomé.[6] Uvážíme-li navíc současnou rychlost jejich vývoje, možná je tato předpověď dokonce příliš konzervativní.

J. M.: *Ano, rychlost jejich vývoje je ohromující. Vraťme se ale ještě ke zmíněným překážkám. Ve zmíněném článku je označujete jako výzvy* (challenges) *z hlediska možnosti vědomí u AI systémů. Jde o to: měl-li by systém být vědomý, musel by podle řady badatelů disponovat globálním pracovištěm, světo-modely* (world--models) *a sebe-modely* (self-models)*, skutečnou pamětí atd. Které z těchto výzev chápete jako relativně snadno zvládnutelné a které naopak jako obtížnější?*

D. Ch.: Ve zmíněném článku je jich uvedeno šest a jsou to zároveň vlastně i důvody, proč je možné se domnívat, že současné AI systémy vědomé nejsou. Jedním z nich je biologie – podle některých odborníků vědomí vyžaduje biologický základ. Pokud by tomu tak bylo, pak by se AI systémy nikdy nemohly stát vědomými, takže to ponechávám stranou. Dalším z nich jsou atributy smyslů a těla. Paradigmatické jazykové modely jako GPT3 nemají ani smysly, ani tělo. Je sice sporné, zda jsou nutnou podmínkou vědomí, ale každopádně lidé aktuálně stále intenzivněji pracují na multimodálních jazykových modelech, které disponují zpracováním obrazu a zvuku, tedy jistým druhem smyslových procesů, a dokáží řídit tělo. Myslím si tedy, že možná jde jen o dočasnou překážku. Někteří odborníci poukazují na potřebu světo-modelů a sebe-modelů. Je velmi obtížné zjistit, zda AI systémy mají světo-modely a sebe-modely, ale myslím si, že máme celkem dobré důvody domnívat se, že alespoň pokud jde o světo-modely, mají k nim tyto systémy už výrazně nakročeno.

6 Chalmers, D. J., Could a Large Language Model be Conscious? *Boston Review*. Dostupné na: https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/; [cit. 10. 1. 2024].

Dále se tu vynořuje otázka, zda AI systémy disponují rekurentním zpracováním informací (*recurrent processing*) – zpětnovazebnými smyčkami (*feedback loops*), které by se mohly ukázat jako nutné pro ty formy paměti, které by zase mohly být nutné pro vědomí. Ve většině z těchto systémů probíhá zpracování informací směrem dopředu (*feed-forward*). Už se nicméně rozbíhá i projekt stavby rekurentních jazykových modelů, takže i tato překážka je možná jen dočasná. Totéž platí pro globální pracovní prostor – podle některých odborníků je globální pracovní prostor, tedy centrální oblast umožňující šíření informací do mnoha modulů, pro vědomí nutný. Není sice zjevné, že jazykové modely tento prostor mají, nicméně už se stavějí takové verze jazykových modelů, které globální pracovní prostor mají. Takže jde opět o něco, co je potenciálně jen dočasné.

Poslední překážka, která mi dělá starosti, je odpověď na otázku, zda jsou jazykové modely jednotnými původci jednání, což by muselo být nutně podmíněno tím, že by disponovaly jednotnými sadami přesvědčení a přání toho typu, který umožňuje vznik jednotných motivací, jaké mají lidé. Tyto modely bývají ovšem velice vrtkavé – jsou tak trochu jako chameleoni –, velice snadno zamění jednu osobnostní charakteristiku za jinou a nezdá se, že by byly jednotnými původci jednání v uvedeném smyslu. Na druhé straně se ale stále více objevují projekty, ve kterých jde o vytvoření tzv. konatelské AI (*agent AI*) – tedy takových jazykových modelů, které se drží jedné osobnostní charakteristiky a mají i přesvědčení a přání, třebaže stále ještě primitivní.

Uvažujeme-li tedy o dnešku, dost možná v některých kritériích selháváme. Počkejme ale třeba do roku 2032 – myslím si, že je docela dobře možné, že většina ze zmíněných překážek bude do té doby už překonána. Možná budou dokonce existovat sofistikované AI systémy, které budou mít všechny tyto atributy. I v té chvíli jistě někteří řeknou, že tyto systémy nejsou vědomé, ale pak se tedy budeme muset ptát, co jim chybí.

J. M.: *Ve světle toho všeho se zdá být důležité uvažovat o možnostech testování AI vědomí. Jaký typ testu považujete za nejspolehlivější?*

D. Ch.: Vytváření testů je samozřejmě velmi obtížné. Každý test je založen na dalekosáhlých filosofických předpokladech, a my nemáme k dispozici žádný nezávislý způsob, jakým je možné měřit vědomí, a tím platnost našich testů ověřit. U lidí spoléháme především na slovní výpovědi: když někdo řekne, že si je něčeho vědom, není důvod myslet si něco jiného, a tak to přijmeme. U jazykových modelů AI se to ale zdá být méně spolehlivé. Je snadné přimět je k tomu, aby řekly, že jsou vědomé, nicméně ony jsou přece trénovány k tomu, aby imitovaly lidi, kteří běžně říkají, že mají vědomí. Není tedy jasné, zda tomu lze přikládat velkou váhu. Klasickým testem AI je samozřejmě

Turingův test, u něhož jde o to, zda je takový systém neodlišitelný od člověka. Podle Turinga právě tato charakteristika stačí. Co se týče jazykových modelů, myslím si, že zatím Turingovým testem tak docela neprošly, ale mají k tomu už velice blízko. Jistě dokáží konverzovat způsobem, který se velmi podobá lidskému. Nyní však někteří odborníci začali vyjadřovat obavy ohledně toho, zda je Turingův test validní. Možná jím nějaký systém může projít, aniž by byl vědomý. A zároveň je to na druhé straně velice úzký test, takže mnohé AI systémy, které Turingovým testem sice neprojdou, by přesto mohly být vědomé.

Nyní jsme tedy podle mě v situaci, kdy v případě umělé inteligence nemáme k dispozici žádné velmi dobré testy prokazující vědomí. Máme teorie vědomí, které možná můžeme na AI systémy vztáhnout, například teorii globálního pracovního prostoru a komputační teorii. Ty nám mohou poskytnout určité důvody vedoucí k přisouzení vědomí, ale žádná z těchto teorií samozřejmě není prosta kontroverzí. Máme několik teorií, které si konkurují, a žádný konsensus. To nejlepší, co asi můžeme udělat, je uvážit různé teorie a ukazatele (*markers*) vědomí a podívat se, jak je na tom v těchto ohledech umělá inteligence. Myslím si ale, že ani pak si nebudeme zcela jisti, zda jsme dosáhli pravdivého poznání o těchto věcech. Možná až tehdy, když tyto systémy budou mezi námi, budou mluvit jako lidé a zcela běžně vypovídat o svých vědomých stavech, tak snad v určité chvíli nabudeme přesvědčení, že jsou vědomé, ale celá otázka prokazování vědomí je velice obtížná.

J. M.: *Jedna zásadní otázka, která se v této souvislosti nabízí, zní, zda bychom se měli snažit vytvořit vědomou AI nebo zda jde o něco, čemu bychom se naopak měli snažit vyhnout. Zastáváte ohledně tohoto nějaké stanovisko?*

D. Ch.: Vím, že je to velmi komplexní otázka. I když to chápeme jen jako vědecký a technologický úkol, vytvořit AI a umělé vědomí představuje fascinující cíl, který podle mě představuje jeden z největších problémů vědy. Zároveň to s sebou samozřejmě nese řadu nebezpečí, rizik a etických otázek. Pokud vědomá AI bude zároveň inteligentní AI, a to na lidské nebo vyšší úrovni inteligence, pak to povede k otázkám stran bezpečnosti: jak například můžeme tyto bytosti kontrolovat? Zároveň, pokud budou tyto bytosti vědomé, povede to k otázkám týkajícím se etiky: zaslouží si tyto bytosti zákonná a morální práva? Pokud je totiž určitý systém vědomý, pak si zasluhuje, aby na něj byly brány morální ohledy – a v té chvíli se musíme začít obávat, že když vytvoříme AI systém, o němž se domníváme, že by mohl být vědomý, pak může trpět, může zakoušet pozitivní a negativní stavy apod. Náhle o něm musíme začít uvažovat také z morálního hlediska a myslím si, že právě to je ten moment, kdy pravděpodobně musíme zpomalit a postupovat velmi opatrně. Nemůžeme bez přemýšlení o těchto důsledcích AI systémy vytvářet a pak

s nimi zacházet zcela podle své libosti, musíme brát v úvahu rovněž to, jaká by mohla být subjektivní zkušenost takového AI systému. V opačném případě bychom mohli postupně dospět až k morální katastrofě.

J. M.: *Běžná námitka proti tezi, že AI systémy mohou být vědomé, je založena na předpokladu, že vědomí nemůže existovat bez specifického druhu biologického substrátu – a vzhledem k tomu, že AI systémům tato biologická rovina chybí, nemohou být vědomé. Platí koneckonců, že všechny systémy, kterým běžně přisuzujeme vědomí, jsou biologické. Co si o této námitce myslíte?*

D. Ch.: Proti tomuto postoji jsem se vymezil už ve své první knize *The Conscious Mind*,[7] v níž jsem představil myšlenkový experiment, v jehož rámci postupně nahrazujeme neurony křemíkovými čipy. Snažil jsem se doložit, že máme docela dobrý důvod si myslet, že je-li tento experiment proveden správně, je výsledný křemíkový systém funkčním izomorfem. Pokud je systém, který podrobujeme tomuto experimentu, vědomý na začátku, bude vědomý i na konci. Předpoklad, že se v rámci tohoto experimentu vědomí postupně vytratí nebo náhle zmizí, je, jak jsem se snažil ukázat, takřka nepřijatelný. Domnívám se tedy, že se mi podařilo zformulovat argument – a pevně doufám, že platí –, že vědomí může existovat i bez biologie. Pokud tomu tak skutečně je, vyvrací to onu velmi obecnou námitku, o níž jste se zmínil ve své otázce. Každý, kdo má za to, že se vědomí neobejde bez určitého biologického základu, bude muset nyní říci, kde je v mé argumentaci chyba a co by se stalo, kdybyste neurony nahradil funkčně ekvivalentními křemíkovými čipy. Je to nemožné? Vytratilo by se vědomí postupně? Zmizelo by náhle? Takto jsem vymezil výzvy, jimž ono „biologické" stanovisko musí nyní čelit.

J. M.: *Váš zájem o vědomí a jeho „těžký problém" Vás přivedl až ke zkoumání takových teorií vědomí, jež se vymykají fyzikalistickému redukcionismu, který představoval v posledních dekádách takřka ortodoxii. Příklady takových nereduktivních teorií jsou například dualismus vlastností, panpsychismus či idealismus. Chtěl jsem se zeptat, kterou z těchto teorií aktuálně vidíte jako nejslibnější a proč.*

D. Ch.: Dlouho jsem svou důvěru vkládal do dualismu a panpsychismu a stanovisek, která jsou jim příbuzná. Shrneme-li to: zaměřoval jsem se na dualismus na jedné straně a russelliánský monismus na straně druhé. Každé z těchto stanovisek mi připadá odlišným způsobem přitažlivé.

7 Chalmers, D. J., *The Conscious Mind. In Search of a Fundamental Theory*. New York, Oxford University Press 1996, kapitola 7.

Dualismus však jen s obtížemi čelí problému interakce, tedy otázce, zda má vědomí vliv. Pokud ho nemá, platí epifenomenalismus a vědomí nehraje žádnou roli, což mnohým připadá neatraktivní. Alternativou je pak ovšem interakcionismus, kde vědomí hraje kauzální roli – jak to ale funguje? V některých svých pracích jsem se pokusil ukázat, že možná stojí za to alespoň zkoumat mezery v této úvaze a zaměřit se při tom především na kolaps vlnové funkce v kvantové mechanice. Právě ten totiž možná ukazuje, že vědomí může hrát kauzální roli. S kolegou Kelvinem McQueenem se nyní pokoušíme zkoumat, zda by za tohoto předpokladu mohl dualismus fungovat, ale bohužel stále ještě narážíme na mnoho obtíží a nezodpovězených otázek.[8]

Zároveň platí, že panpsychismus je velmi atraktivní teorie. Zdá se mi, že by se úspěšně mohla vyhnout ontologickým excesům dualismu, sjednotit mentálno a fyzično a zasadit vědomí hluboko do nitra přírodních procesů. Čelí nicméně obtížně řešitelnému *problému kombinace*: jak se naše vědomí vztahuje k malým vědomím na fundamentální úrovni fyziky a jak by se tato vědomí mohla kombinovat a vytvořit naše vědomí? Pokud se nekombinují, vypadá to, že naše vědomí možná bude opět epifenomenální, a jsme tedy následně konfrontováni s problémem interakce. Pokud se kombinují – pokud je naše vědomí nějak utvářeno, konstituováno těmito mikro-vědomími –, jsme naopak konfrontováni s problémem kombinace: jak by se miliardy nebo bilióny malých vědomí mohly spojit a společně utvořit mé vědomí? Nakolik je mi známo, nikdo dosud nedospěl ke skutečně pozitivnímu řešení tohoto problému.

Shrnu-li to: mám rád obě tato stanoviska, dualismus i panpsychismus, ale zároveň také vidím, že se potýkají s vážnými problémy. Více či méně zásadním problémům ale podle mě čelí i materialismus.

Vážně beru také iluzionismus, podle jehož zastánců je vědomí jen iluze. Připadá mi sice zhola nemožné tomuto stanovisku věřit, nicméně tento přístup ve své nejlepší podobě může předpovídat i to, že mi bude připadat nemožné mu věřit. A právě to mu tedy lze přičíst k dobru.

Toto jsou tedy přístupy, které nyní beru nejvážněji, ale v současné době jsem, řekl bych, nejspíše agnostik – a nemám tedy žádný nejoblíbenější přístup.

J. M.: *Zmíněné nereduktivní teorie vědomí – pomineme-li nyní iluzionismus – jsou někdy označovány za přehnaně spekulativní nebo nedostatečně vědecké. Jaká je Vaše odpověď na takovéto námitky?*

8 Viz Chalmers, D. J. – McQueen, K. J., Consciousness and the Collapse of the Wave Function. In: Shan Gao (ed.), *Consciousness and Quantum Mechanics*. New York, Oxford University Press 2023. DOI: https://doi.org/10.1093/oso/9780197501665.003.0002.

D. Ch.: Předně: jsou to metafyzické, nikoli vědecké teorie. Panpsychismus nebo dualismus samy o sobě určitě nepředstavují zplna vypracované teorie vědomí, jsou to spíše jen určitá stanoviska, teze. Každé z těchto stanovisek by muselo být nejprve doplněno celou řadou detailů, aby z něj mohla vzniknout seriózní ucelená teorie – a dojde-li k tomu, pak při tom budou hrát nesmírně relevantní roli vědecké poznatky a přístupy. Takže: dualismus ve své obecné podobě není založen na vědě. Nicméně kvantově mechanický dualismus je určitým způsobem provázaný s poznatky týkajícími se kolapsu vlnové funkce atd. – a je tedy jistě přinejmenším dosti založen na vědeckém výzkumu. Čím propracovanější se jakákoliv ze zmíněných teorií bude postupně stávat, tím relevantnější – a o tom jsem pevně přesvědčen – pro ni bude věda.

Nemyslím si, že filosofické stanovisko musí být nutně podpořeno vědeckými přístupy. Může být i bez nich zcela přesvědčivé. Pokud se ovšem ukáže, že tomuto stanovisku odporují vědecké důkazy, pak to jistě představuje obtíž. Co se týče dualismu, mnozí kolegové se domnívají, že existují vědecké důkazy, které podporují kauzální uzávěru fyzična, a lze je tedy chápat jako důkazy vyvracející přinejmenším interakcionistický dualismus. Tím se dostáváme k otázce, zda je kvantová mechanika slučitelná s popřením kauzální uzávěry fyzična. Každopádně si myslím, že když někdo tvrdí, že existují vědecké důkazy proti určitému filosofickému stanovisku, představuje to pro ono stanovisko vážný problém, nebo přinejmenším výzvu, na kterou je třeba odpovědět. Existují vědecké důkazy proti panpsychismu? Ne, já sám jsem přesvědčen, že vůbec žádné. Neexistuje nicméně ani moc vědeckých důkazů, které by panpsychismus podporovaly. V této souvislosti zastávám názor, že naše vědecké poznatky jsou do značné míry neutrální – a obecně si myslím, že tak je tomu s vědeckými závěry a filosofickými tezemi často. Jsem ale jednoznačně pro to, abychom, nakolik je to jen možné, provázali filosofické teze s vědeckými přístupy. Uvedu příklad: na Tononiho práci zabývající se teorií integrované informace lze nahlížet jako na pokus o sjednocení panpsychismu s myšlenkami, které nacházejí své předpoklady v teorii informace a neurovědě. Také některé verze iluzionismu lze úspěšně provázat s vědou. Myslím si tedy, že dost často nacházíme prostor pro interakci mezi vědci a filosofy právě při rozvíjení a hodnocení teoretických stanovisek týkajících se vědomí – a že tyto tendence ke spolupráci je třeba vždy podporovat.

J. M.: *Mluvíme-li o tezích, které určitým způsobem jdou „za" vědecké poznatky, ve své nejnovější knize* Reality+*, kterou jste věnoval tématu virtuální reality, uvažujete o hypotéze simulace* (simulation hypothesis)*, tedy o myšlence, že ži-*

jeme – a lidé vždy žili – v uměle vytvořené počítačové simulaci světa.[9] *Proč podle Vás tato hypotéza, která některým může připadat tak trochu přitažená za vlasy, zasluhuje vážnou filosofickou pozornost?*

D. Ch.: Já jsem filosof, kterému nevadí za vlasy přitažené myšlenkové experimenty – myslím si naopak, že je často užitečné o nich přemýšlet. Hypotéza simulace představuje přinejmenším docela dobrý způsob, jakým lze reformulovat Descartovu výzvu týkající se vnějšího světa. Otázka „Jak víte, že nežijete v simulaci?" je moderní verzí otázky: „Jak víte, že nejste klamáni zlotřilým démonem?" Mimo jiné jde tedy o uvedení do zásadních epistemologických problémů. V některých souvislostech má ale tato hypotéza ještě větší sílu a dopad – domnívám se, že před nás staví například velmi zajímavé metafyzické problémy. Může se například zdát, že pokud se nacházíme v simulaci, pak nic není skutečné. Ve své knize jsem se oproti tomu pokusil obhájit metafyzické stanovisko, podle něhož uvnitř simulace věci skutečné jsou, což má důležité důsledky z hlediska uvedených metafyzických a epistemologických otázek.

Někteří kolegové se přitom domnívají, že sama hypotéza simulace se liší od hypotézy zlotřilého démona. Mají za to, že jde o jinou hypotézu, kterou bychom měli brát vážně především vzhledem k nástupu jí relevantní technologie – dost brzy totiž budeme disponovat simulační technologií. Během několika desetiletí budeme dost možná mít k dispozici virtuální realitu, která bude nerozlišitelná od reality fyzické, což před nás ještě naléhavěji postaví otázku: „Jak víme, že v takové virtuální realitě nežijeme teď?" Podle Nicka Bostroma lze vzhledem k nástupu simulační technologie formulovat statistický argument, podle kterého je dokonce pravděpodobné, že v dějinách vesmíru existuje mnoho simulací, a pouze jeden nesimulovaný vesmír.[10] Nahlížíme-li na výše uvedené otázky z této perspektivy, může se tedy zdát dokonce pravděpodobnější, že skutečně žijeme v simulaci.

To vše je samozřejmě extrémně spekulativní, ale hodně z toho mi připadá filosoficky fascinující. Tyto otázky si zaslouží pozornost samy o sobě, ale zároveň představují i uvedení do úvah o nejzákladnějších filosofických problémech, což je důvod, proč jsem napsal knihu *Reality+*, v níž zkoumám pomezí mezi virtuální realitou, hypotézou simulace a velkými tradičními problémy filosofie. Podle mě mezi nimi probíhá obousměrná interakce, kterou nazý-

9 Chalmers, D. J., *Reality+. Virtual Worlds and the Problems of Philosophy*. New York, W. W. Norton 2022.

10 Viz Bostrom, N., Are We Living in a Computer Simulation? *Philosophical Quarterly*, 53, 2003, No. 211, s. 243–255.

vám *technofilosofii* a v jejímž rámci filosofie pomáhá osvětlit technologii – a zároveň také technologie pomáhá osvětlit filosofii.

J. M.: *Ano, to je podle mě jedna z mnoha podnětných koncepcí, které lze ve Vaší knize nalézt. Obhajujete v ní také myšlenku, že hypotézu simulace nelze vyloučit – a ostatně jste se k ní opět explicitně přihlásil i v odpovědi na mou předchozí otázku. Tento závěr mohou někteří lidé chápat i tak, že vede přímo ke skeptickému stanovisku, podle něhož o vnějším světě nemůžeme nic vědět. Ve své knize nicméně tento závěr striktně odmítáte a tím zároveň nabízíte i určitý typ odpovědi na tuto variantu karteziánské skepse. Proč podle Vás z naší neschopnosti vyloučit hypotézu simulace* nevyplývá *skepse ohledně vnějšího světa?*

D. Ch.: Domnívám se, že skeptická argumentace, která vychází z teze „nevíme, zda nežijeme v simulaci" a vede k závěru „nevíme nic o vnějším světě", předpokládá, že uvnitř simulace není nic skutečné, žádné z našich přesvědčení není pravdivé. To je stanovisko, které nazývám *simulační irealismus.* Já sám jsem však ve své knize obhajoval takový *simulační realismus,* podle něhož, i *pokud* žijeme v simulaci, tak většina z našich běžných přesvědčení je stále pravdivá. Pokud jsem přesvědčen, že je přede mnou stůl, skutečně je přede mnou stůl – je to sice digitální stůl, ale přesto stůl, a má přesvědčení o něm jsou tedy pravdivá. Zastánci standardního stanoviska by řekli „Nevím, zda nežiji v simulaci, ale pokud ano, pak tu stůl není. Proto nevím, zda tu je stůl." Já mám ale za to, že i když žiji v simulaci, stůl tu je, takže nelze vyvodit „Nevím, zda je tu stůl". Existuje samozřejmě ještě řada dalších variant skepse, které nás mohou trápit, a já netvrdím, že mé stanovisko představuje definitivní odpověď. Myslím si nicméně, že má argumentace přinejmenším nabízí zajímavý přístup k některým velmi tradičním otázkám týkajícím se skepse, který je zároveň novým způsobem provázán s moderními technologiemi.

J. M.: *Napadá mě, že se lze přirozeně domnívat, že ve filmu* Matrix *postava Nea, poté co spolkne červenou pilulku, získá nové důležité poznatky o tom, jaký je svět. Někdo by tedy mohl reagovat na Vaši odpověď na skeptický problém tak, že by řekl, že i kdybychom přijali, že červená pilulka, kterou Neo spolkne, a další věci v matrixu existují jakožto digitální objekty, přesto v určitém důležitém smyslu je Neo, předtím než pilulku spolkne, vážně pomýlený, a že jeho pomýlení zmizí teprve poté, co pilulku spolkne.*

D. Ch.: Ano, myslím si, že existuje něco velmi důležitého, co Neo předtím nevěděl a co se po spolknutí pilulky doví: je to fakt, že žil v simulaci. To je samozřejmě velmi důležitý poznatek. Předtím Neo nevěděl, že žije v simulaci, teď ví, že tomu tak je. Zjistil, že žije v digitálním světě vytvořeném simuláto-

rem – to vše je velice důležité. Je to obrovský objev, který učinil tím, že si vzal červenou pilulku. Skeptik by mohl říci, že předtím nevěděl, zda je tam a tam stůl, a teď možná ví, že tam stůl není – já bych tohle neřekl. Předtím byl Neo přesvědčen, že je tam stůl, a měl pravdu – a i poté tam stůl stále je. Jeho epistemický pokrok spočívá v tom, že se doví něco hlubokého o povaze své reality – žije v simulaci, svět je digitální, existují tvůrci, kteří nás do něj umístili atd. Lidé dnes občas používají termín „červená pilulka", zmiňují-li myšlenku, že je člověk klamán těmi, kdo jsou momentálně u moci, a osobně si myslím, že tohle pro Nea stále platí – stroje způsobily, že se domnívá, že žije v bazální realitě, přičemž on ve skutečnosti žije v simulaci.

J. M.: *Tím, že Neo spolkne červenou pilulku, tedy v jistém smyslu získává* metafyzické *poznání…*

D. Ch.: Ano, poznání povahy své reality. Dověděl se, že realita, ve které žije, je ve skutečnosti simulací, že je založena na digitálních procesech. Je to poznatek, který je důležitý – nicméně nepředstavuje rozdíl mezi stavem, kdy nevíme nic, a stavem, kdy něco víme.

J. M.: *Mluvili jsme o jazykových modelech a o virtuální realitě – a ve světle těchto souvislostí bych se nyní rád přesunul k další technologii, která sehrála ve Vaší práci významnou roli, konkrétně ke kosmoskopu. Tento imaginární přístroj se objevil ve Vaší knize* Constructing the World.[11] *Co je kosmoskop a jakou roli hraje v argumentaci této knihy?*

D. Ch.: V *Constructing the World* jsem se snažil ukázat, že můžete vše o světě vědět na základě relativně omezené třídy informací, tedy že jistá malá množina pravd by vás mohla přivést až k tomu, že poznáte všechny pravdy o světě. Podle mého názoru do této malé množiny patří fyzikální pravdy, ale také pravdy fenomenální, také jsou potřeba určité indexické pravdy o vašem umístění ve světě a určitá „celkovostní pravda", která říká „Tohle je vše, co je ve světě". Snažil jsem se ukázat, že z něčeho takového, co jsem nazval PQTI,[12] je v zásadě možné vyvodit všechny pravdy o světě. Pro ilustraci jsem jako součást obhajoby tohoto tvrzení využil imaginární přístroj zvaný kosmoskop, který poskytuje přístup ke všem pravdám tvořícím PQTI. Kosmoskop člověku zobrazuje aktuální stav fyzikální konfigurace všeho na světě, hustoty hmoty v prostoru a čase atd. Člověk v rámci tohoto zobrazení vstupuje do jis-

11 Chalmers, D. J., *Constructing the World*. New York, Oxford University Press 2012.

12 Tj. P – fyzikální [physical] pravdy, Q – fenomenální pravdy, T – celkovostní [totality] pravda, I – indexické [indexical] pravdy.

tého druhu virtuální reality, která mu poskytuje poznání všech vědomých prožitků všech lidí na světě. Kosmoskop vám dále ukazuje, kde se nacházíte, a také meze světa.

Kosmoskop je smyšlený přístroj, ale pokud by existoval, pak byste jeho prostřednictvím mohl zjistit o světě vlastně cokoli. Týkalo by se to přitom nejen faktů zahrnutých v PQTI, ale také celé řady dalších věcí, jako například toho, kdo byl Jack Rozparovač, nebo kdo spáchal určitý zločin, jaká je fundamentální vědecká povaha mozku nebo cokoli jiného. Snažil jsem se ukázat, že pomocí kosmoskopu byste se mohl dovědět vlastně všechny pravdy o světě nebo alespoň všechny pravdy patřící do určité třídy. Dovozoval jsem, že vzhledem k tomu, že kosmoskop nám vlastně poskytuje všechny pravdy v PQTI, lze z PQTI zjistit všechny pravdy o světě.

Argumentace, kterou jsem nastínil, ovšem ještě definitivně neprokazuje všechny ústřední *teze odvoditelnosti* (*scrutability theses*), které v knize *Constructing the World* k obhajobě svých závěrů potřebuji. Například neprokazuje, že ona inference má apriorní povahu, třebaže nás má uvedená argumentace k tomu nasměrovat. Kosmoskop jsem použil pro ilustraci jako přístroj, který měl přiblížit, že z PQTI bychom se mohli dovědět většinu pravd o světě. Následně v knize tuto tezi rozšiřuji tak, aby zahrnula veškeré pravdy, odpovídám na různé námitky, dovozuji, že zmíněná inference má apriorní povahu atd.

Někde doma ale mám křišťálovou kouli, kterou jsem jednou objevil v obchodě *Metaphysics' World*, kde mi řekli: „Tohle je kosmoskop!" Nyní mě tedy těší, že mohu říci, že kosmoskop vlastním. Nejsem si jist, zda funguje až tak dobře jako ten, který jsem popsal ve své knize, ale pracuji na tom.

J. M.: *To je skvělé! Váš projekt v knize* Constructing the World *je blízký Carnapovu projektu v* Der logische Aufbau der Welt *(1928). Carnap tvrdí, že všechny pravdy o světě jsou derivovatelné z malé množiny základních pravd, a jeho kniha bývá chápána jako návrh jistého druhu fundacionalistické epistemologie. Ve své knize se snažíte ukázat, že Carnapův projekt, který byl obecně velmi silně kritizován, je v modifikované podobě obhajitelný. Mohl byste se prosím pokusit říci, co na Carnapově projektu vidíte jako nejcennější a v čem je naopak třeba ho zásadněji revidovat?*

D. Ch.: Myslím si, že kritika Carnapova projektu se týkala hlavně jeho fenomenalismu – tedy myšlenky vyvodit všechny pravdy o světě z pravd o *smyslových datech*. Můj vlastní přístup však fenomenalismus nezahrnuje – součástí mého základu odvoditelnosti (*scrutability base*) je sice fenomenalita, ale také fyzika. Například Quine Carnapa kritizoval za to, že se spoléhá na rozlišení mezi analytickým a syntetickým. Já sám ale toto rozlišení mezi analytickým

a syntetickým nebo mezi a priori a a posteriori neodmítám, protože se mj. domnívám, že Quineovy argumenty proti tomuto rozlišení nejsou moc přesvědčivé. Jiní zase Carnapa kritizovali za to, že se spoléhá na definice klíčových termínů, přičemž ve filosofii panuje dosti obecná shoda na tom, že je velice těžké definice většiny běžných termínů nalézt. Ve své knize se každopádně pokouším ukázat, že projekt tohoto typu definice nevyžaduje – že to, co nazývám *projekt odvoditelnosti*, v podstatě zahrnuje jistý druh pojmové analýzy jednotlivých případů – a že jednotlivé případy dokážeme posuzovat i bez definic.

Vzdávám se v podstatě dvou Carnapových klíčových předpokladů, fenomenalismu a, chcete-li, „definicionalismu", a pokouším se ukázat, že i bez nich je možné realizovat jistou verzi jeho projektu. Stále tu totiž existuje omezená třída pravd, z nichž všechny ostatní pravdy o světě alespoň a priori vyplývají. Vyžaduje to sice určitou idealizaci a výsledek je jistým způsobem nejistý a omezený, já se ale přesto právě v knize *Constructing the World* snažím ukázat, že i navzdory tomu všemu mají tyto závěry dopady na mnoho věcí. Například to, co nazývám *tezí odvoditelnosti*, lze využít k definování významů našich termínů, které se – alespoň v jistých ohledech – chovají velice podobně jako fregovské *smysly,* takže můžete obhájit fregovské chápání významu. Dále se pokouším ukázat, že tyto úvahy nám mohou pomoci alespoň začít vytvářet základy některých metafyzických argumentů, které jsem představil ve svých dřívějších pracích a které propojují myslitelnost, možnost a tak dále.[13] A možná se ukáže, že jsou také epistemologicky relevantní v úvahách o některých otázkách týkajících se filosofické skepse. Pokusil jsem se tedy ukázat, že tento teoretický rámec by mohl vrhnout nové světlo na množství různých filosofických problémů a že mu navíc nechybí ani jistá zajímavost.

J. M.: *Těší mě, že to říkáte, protože mi vždycky připadalo, že Vaše obrana aprority v té knize představuje důležitý základ, z něhož vyrůstá to, co jinde říkáte o myslitelnosti – zde mám na mysli například Váš argument se zombie. Tím, že berete aprioritu vážně, brojíte proti současnému quineovskému proudu ve filosofii, který má patrně kořeny v Quineových* Dvou dogmatech empirismu. *V této souvislosti jsem se chtěl zeptat, v čem vidíte hlavní problémy, pokud jde o argumenty proti analytičnosti a možná i proti aprioritě, které Quine v této studii předkládá.*

13 Viz např. Chalmers, D. J., The Two-Dimensional Argument Against Materialism. In: týž, *The Character of Consciousness*. New York, Oxford University Press 2010, s. 141–205.

D. Ch.: Dnes už samozřejmě není nic radikálního na tom jít právě v těchto otázkách proti Quineovi. V anketě portálu PhilPapers se zcela nedávno přibližně 70 % filosofů nebo možná ještě více přihlásilo k názoru, že existuje *a priori* a distinkce mezi analytickým a syntetickým.[14] Quineovy argumenty ve „Dvou dogmatech"[15] jsou zajímavé, ale myslím si, že hodně kolegů je nyní nahlíží jako slabší, než se původně předpokládalo. Ve většině článku, zhruba v prvních čtyřech částech, Quine zvažuje různé definice a snaží se vysvětlit analytičnost pomocí významu, který je vysvětlen pomocí definice, která je vysvětlena pomocí synonymie, která je vysvětlena pomocí analytičnosti – a aha, máme kruh! Mnozí kolegové si však už povšimli toho, že s obdobnými tautologiemi se setkáváme u mnoha důležitých filosofických pojmů, takže ani podle jejich, a ani mého názoru není zcela evidentní, že by tu šlo o skutečně velký problém, který by byl pro analytičnost specifický.

Možná nejvlivnější a nejzajímavější částí Quineova článku je velice stručná závěrečná pasáž, v níž tvrdí, že jakékoli tvrzení by mohlo být drženo jako pravdivé za všech okolností a že žádné tvrzení není imunní vůči revizi. Jeho stoupenci se domnívali, že fakt, že žádné tvrzení není imunní vůči revizi, může představovat problém pro rozlišení mezi analytickým a syntetickým nebo mezi a priori a a posteriori, jež podle nich vycházejí z předpokladu, že některé pravdy jsou vůči revizi imunní. Je pravda, že na první pohled Quineovy argumenty zdánlivě přinejmenším ohrožují apriornost a analytičnost. V knize a v článku, které jsem na toto téma napsal,[16] jsem se však snažil ukázat, že tyto argumenty ve skutečnosti nefungují, pokud zkoumáme situace, kdy je něco drženo jako pravdivé, a situace, kdy něco není imunní vůči revizi. Vždy je možné revidovat věty – například jsem dříve přijímal určitou větu, ale už ji nepřijímám. Hlavní otázka však zní: Kdy, pokud toto uděláme, skutečně dojde ke změně významu? Quineův odpůrce říká: „Revidovat můžeme, pouze pokud dojde ke změně významu." Quine pak odpovídá: „Co je to ale změna významu? Teď znovu předpokládáte ten problematický pojem. Jak může existovat principiální rozlišení mezi revizemi, které mění význam, a těmi, které nikoli?"

Ve svých příspěvcích věnovaných této problematice jsem se snažil ukázat, že tu ve skutečnosti lze vést principiální rozlišení, a pokusil jsem se jej učinit s využitím bayesovkých pojmů. Předložil jsem argument obhajující tezi,

14 Viz Bourget, D. – Chalmers, D. J., Philosophers on Philosophy: The 2020 PhilPapers Survey, c.d.

15 Quine, W. V. O., Dvě dogmata empirismu. Přel. P. Sousedík. In: Quine, W. V. O., *Vybrané články k ontologii a epistemologii*. Ed. L. Dostálová – T. Marvan. Plzeň, Západočeská univerzita 2006, s. 79–99.

16 Jde o knihu Chalmers, D. J., *Constructing the World*, c.d.; a článek Chalmers, D. J., Revisability and Conceptual Change in "Two Dogmas of Empiricism". *The Journal of Philosophy*, 108, 2011, No. 8, s. 387–415.

podle níž existuje principiální rozlišení mezi relevantními druhy revize přesvědčení: mezi těmi, které zahrnují změnu významu, a těmi, které nikoli. Neřekl bych, že z tohoto automaticky vyplývá analytičnost a apriorita, ale myslím si, že je to dostačující na to, abychom dokázali odpovědět na Quineovy argumenty proti analytickému a apriornímu. Připadalo mi proto nejen zajímavé, ale i velmi užitečné o tom přemýšlet.

J. M.: *Stojíte za řadou vlivných a často průlomových příspěvků v mnoha oblastech filosofie. Vaše práce se navíc těší – navzdory obtížným tématům, kterými se zabývá – značnému vlivu a zájmu i mimo akademickou sféru. Je něco, čeho byste ve filosofii ještě rád dosáhl? Jaké jsou Vaše plány a aspirace do budoucna?*

D. Ch.: Je toho hodně, doufám, že jsem teprve na začátku! Rád bych pokročil v analýze všech problémů, o nichž jsem přemýšlel – u žádného z nich jsem se dosud nedobral jádra pudla, definitivního řešení! Jednoho dne bych chtěl proniknout k samé podstatě problému vztahu mysli a těla. Připadá mi, že jsme k ní zatím nedospěli, ale rád bych si myslel, že je taková věc možná. Rád bych plně porozuměl významu, rád bych plně porozuměl našemu vztahu k vnějšímu světu. Fascinují mě ale také nové technologie a domnívám se, že umělá inteligence a virtuální realita nás staví před celou řadu velmi praktických problémů, se kterými se budeme muset nějak vypořádat. Právě při řešení těchto otázek mohou sehrát klíčovou roli filosofové. Připadá mi, že teprve teď začínám rozumět úžasným pokrokům v oblasti umělé inteligence, o nichž bych rád v nadcházejících letech a dekádách systematicky přemýšlel.

Mám ostatně dojem, že nové problémy a úkoly vyvstávají neustále – dokonce jsem nedávno napsal text o etice: „Sentientism and Moral Status". Nikdy jsem si nemyslel, že bych právě k etice měl co říci, ale ukázalo se, že když člověk přemýšlí o jednom filosofickém tématu, je propojeno s celou řadou dalších, takže se najednou otevře celá krajina problémů a úkolů. To však není nic nového, stávalo se mi to vlastně v celé mé dosavadní kariéře – začal jsem od problému vědomí, a ten mě přímo zavedl k metafyzickým problémům a problémům významu, k otázkám modality, k tematice poznání a problémům vědy, ba možná i k problematice etiky a nových technologií. Je to jako neustále se otevírající spirála vedoucí permanentně se zvětšující krajinou problémů a úkolů k přemýšlení – a snad k tomu bude obdobným způsobem docházet bez ustání i nadále! Moc bych si to přál. Filosofie je téměř nevyčerpatelná sféra fascinujících témat a já jen doufám, že tu budu dost dlouho na to, abych mohl pokračovat v přemýšlení o všech.

J. M.: *To také doufám a těším se na Vaše budoucí pokusy zodpovědět fundamentální otázky. Moc Vám děkuji za Váš čas, bylo skvělé mít příležitost s Vámi mluvit o těchto obtížných otázkách.*

D. Ch.: Díky, Jakube, moc mě těšilo s Vámi mluvit o tom všem. A děkuji i všem čtenářům *Filosofického časopisu*, kteří náš rozhovor dočetli až sem. Díky za jejich pozornost a za to, že o těchto tématech přemýšlejí.[17]

17 Další informace o práci prof. Chalmerse lze nalézt na jeho webových stránkách: URL<https://consc.net/>; [cit. 24. 1. 2024].