Západočeská univerzita v Plzni

Fakulta filozofická

Bakalářská práce

Kritéria vědomí u člověka a umělé inteligence

Denisa Turečková

Plzeň 2024

Západočeská univerzita v Plzni

Fakulta filozofická

Katedra filozofie

Studijní program Humanitní studia

Studijní obor Humanitní studia

Bakalářská práce

Kritéria vědomí u člověka a umělé inteligence

Denisa Turečková

Vedoucí práce:

Mgr. Michal Polák, Ph.D.

Katedra filozofie

Fakulta filozofická Západočeské univerzity v Plzni

Plzeň 2024

Prohlašuji, že jsem práci zpracoval(a) samostatně a použil(a) jen uvedených pramenů a literatury.

Plzeň, duben 2024 ……

Poděkování

Chtěla bych poděkovat Mgr. Michalu Polákovi, Ph.D. za vedení práce, trpělivost a cenné připomínky.

Obsah

Úvod....1
1. Vymezení pojmu vědomí....3
2. Teorie vědomí....5
2.1. Teorie globálního pracovního prostoru....5
2.2. Informačně-integrační teorie....6
2.3. Teorie rekurentního zpracování....7
2.4. Teorie vyššího řádu....9
3. Kritéria vědomí u člověka....11
3.1. Neurální kritéria....11
3.1.1. Neurofyziologie mozku....11
3.1.2. Navrhovaná neurální kritéria....12
3.1.3. Neurální koreláty vědomí....14
3.2. Behaviorální kritéria....18
3.2.1. Odpovídající reakce....19
3.2.2. Schopnost učení....21
3.3. Subjektivní kritéria....22
3.3.1. Introspekce jako metoda posuzování vědomých stavů....22
3.3.2. Subjektivita a kvália jako kritérium vědomí....25
4. Kritéria a jejich využití v případě umělé inteligence....29
4.1. Neurální kritéria a umělá inteligence....29
4.2. Behaviorální kritéria a umělá inteligence....31
4.2.1. Argument čínského pokoje....32
4.2.2. Turingův test....34
4.3. Subjektivní kritéria a umělá inteligence....35
Závěr....37
Zdroje:....39

Úvod

V době prudce rostoucího technologického pokroku je zkoumání povahy vědomí úkolem, který je řešen na poli filozofie, kognitivní vědy, a nově také umělé inteligence. S tím, jak se schopnosti systémů umělé inteligence neustále vyvíjejí, což je doprovázeno snahou o sestrojení inteligence podobné té lidské, vyvstává otázka: Jaká kritéria určují existenci vědomí u lidí? A dají se tato kritéria aplikovat i v případě umělých agentů? Cílem mé práce je tedy představení a analýza kritérií, která vymezují vědomí u lidí, a možnost jejich aplikace v oblasti umělé inteligence.

V první části se zabývám vědomím a jednotlivými přístupy vědomí. Jelikož je pojem vědomí předmětem stále probíhajících debat, zaměřím se hned zezačátku na vymezení a definici pojmu vědomí pro tuto práci. Poté navazují teorie vědomí, přičemž k představení jsou vybrány čtyři konkrétní přístupy, které jsou v současnosti vnímány jako nejvhodnější kandidáti na „onu pravou" teorii vědomí. Jsou jimi: teorie globálního pracovního prostoru, integrovaná informační teorie, teorie rekurentního zpracování a teorie vyššího řádu. Každou teorii charakterizuji a popíši, jakým způsobem je vědomí v kontextu dané teorie chápáno.

V druhé části přecházím k analýze kritérií využívaných pro posuzování přítomnosti vědomí u člověka. Tato klíčová část je rozdělena na tři oblasti, ve kterých jsou kritéria charakterizována. Nejprve rozebírám oblast neurálních kritérií, kde je v rámci kontextu práce představena stručná neurofyziologie mozku. Poté vymezím ta neurální kritéria, která byla navržena jako ukazatele přítomnosti vědomí a vysvětlím, jak téma neurálních kritérií souvisí s neurálními koreláty vědomí. Po neurálních kritériích přecházím k behaviorálním, kde budu mluvit o *odpovídající reakci* jako o hlavním indikátoru přítomnosti vědomí. Pokusím se také vystihnout hlavní nevýhody spojené s behaviorálními kritérii. Jako poslední budou rozebrána kritéria subjektivní. Uvedu introspektivní metody, které jsou užívány v rámci posuzování vědomých stavů a poté představím subjektivitu jako kritérium vědomí.

Třetí část otevře téma umělé inteligence. Definuji samotný pojem umělé inteligence a poté rozvinu jednotlivé oblasti kritérií. Ve stejném pořadí (tedy neurální, behaviorální, subjektivní) projdu tyto oblasti a podívám se, jak se umělé systémy při hodnocení vědomí liší od lidských subjektů. U neurálních kritérií uvedu rozdíly a podobnosti neuronů,

perceptronů, neurálních sítí a umělých sítí. V kapitole o behaviorálních kritériích poukáži na problémy hodnocení behaviorálních reakcí a poté se podívám na myšlenkové experimenty *čínský pokoj* a *Turingův test*, které do problematiky chování spojeného s vědomím přináší dva rozdílné pohledy. V části o subjektivních kritériích poukáži na to, jak je obtížné řešit otázky týkající se (sebe)vědomí v oblasti umělých agentů.

Při zpracovávání tématu se tedy zaměřím na analýzu jednotlivých kritérií vědomí člověka a umělé inteligence. Budu se v obou případech držet tří vymezených oblastí kritérií (neurální, behaviorální, subjektivní) a pokusím se za pomoci komparace zjistit, zda lze kritéria vědomí člověka aplikovat i v případě umělých agentů. Při úvodní charakteristice teorií vědomí budu využívat knihu *Vědomí a jeho teorie* od Tomáše Marvana a Michala Poláka. Navržením kritérií vědomí člověka se zabývá článek *Criteria for Consciousness in Humans and Other Mammals*, který napsali Anil Seth, Bernard Baars a David Edelman. Kritéria vědomí u umělé inteligence poté řeší Raúl Arrabales, Agapito Ledezma a Araceli Sanchis v jejich publikaci s názvem *Criteria for Consciousness in Artificial Intelligent Agents*. Obě díla budu pro svoji práci proto také využívat.

1. Vymezení pojmu vědomí

Abychom se později mohli zabývat analýzou kritérií vědomí u různých entit, je důležité hned zezačátku definovat pojem vědomí, neboť by posléze mohlo docházet k různým interpretacím tohoto pojmu. Při jeho vyslovení si jistě každý dokáže vybavit alespoň nějakou představu o významu tohoto pojmu. Vědomí je neoddělitelnou součástí světa každého z nás. Jedná se o zcela subjektivní jev, jenž nám umožňuje unikátně prožívat vše, co se kolem nás děje, a právě díky vědomí jsme tím, kým jsme.[1] Nicméně je něco jiného vědomí vnímat z pohledu první osoby a poté se ho snažit vyložit, nebo dokonce zkoumat z pohledu třetí osoby. A ke složitosti této záležitosti přispívá i fakt, že každý obor napříč vědeckou komunitou k pojmu přistupuje jinak. Definice by měla vycházet z výzkumu, ale právě samotný výzkum vědomí je problematickou záležitostí. Pokud bychom se snažili o introspektivní zkoumání vědomí, narazili bychom brzy na problém s ověřováním takových subjektivních dat, navíc by také mohlo docházet k chybnému pozorování vlastních vědomých stavů. Naštěstí v současné době máme již velkou řadu technologií, které nám pomáhají s výzkumem vědomí z jiného než subjektivního pohledu.[2]

Thomas Nagel se ve své práci s názvem *What Is It Like to Be a Bat?* zaobírá otázkou definice vědomí a navrhuje svoji vlastní odpověď. Zastává názor, že vědomí je přítomno napříč různými druhy organismů, ačkoli prokázat to nelze nijak snadno. Podstatné ale je, že mít vědomí znamená mít možnost být *někým*, nebo *něčím*. Na toto téma Nagel rozvíjí myšlenku subjektivity vědomí a zdůrazňuje, že i přes veškerou snahu naše výzkumy cizího vědomí budou vždy ovlivňovány subjektivní zkušeností nás samých.[3] Pro naši práci je ale důležitý autorův popis vědomí. Pokud můžeme říci, že je něco jako *být tímto organismem,* pak tvrdíme, že onen organismus má vědomí. Není totiž možné představit si, jaké to je být malířským stojanem, a proto usuzujeme, že malířský stojan vlastní vědomí nemá.[4]

Studiu vědomí se, jak již bylo zmíněno, věnují různé vědecké oblasti, jako například psychologie, biologie, v současnosti dokonce informatika, a z nevědeckých primárně pak filozofie, která se tomuto tématu věnuje již od počátku a pomocí nejrůznějších metod hledá co nejkonkrétnější definici vědomí. Na vědomí je nahlíženo jako na jev, který

[1] Harley, T., *The Science of Consciousness*, s. 4.
[2] Marvan, T., Polák, M., *Vědomí a jeho teorie,* s. 21–22.
[3] Nagel, T., What Is It Like To Be a Bat? s. 436–440.
[4] Harley, T., *The Science of Consciousness*, s. 12.

nastává po zpracování konkrétních informací. Vnímání, které se stává vědomým, zahrnuje příjem podnětů z vnějšího světa specializovanými smyslovými receptory. Tyto receptory převádějí podněty na elektrické signály, které putují nervovými drahami do centrální nervové soustavy. V mozku jsou příchozí signály složitě zpracovávány a integrovány a dostávají se až do vědomí. Neurofyziologické procesy zpracování informací nám umožňuje sledovat neurověda, na základě jejíchž výsledků si pak utváříme lepší obraz o tom, jak běh informací v mozku souvisí s vědomím jedince.[5]

Jedním z neurovědců zabývajících se vědomím a jeho podstatou je Adam Zeman. Ve své práci *Consciousness* rozděluje vědomí do několika úrovní. Tou úplně základní úrovní je stav, při kterém je člověk vzhůru při plné své pozornosti, je připraven interagovat, komunikovat a jednat. Na tento stav navazuje druhá úroveň, charakterizovaná navíc prožíváním skutečných okamžiků. V této úrovni je zachycena subjektivita našeho vědomí, která limituje náš náhled na realitu kolem. Vědomí by mělo mít kontinuum obsahů, které jsou krátké, ale jednotné a stabilní. Mělo by být možné pozorovat své současné vědomí, ale stejně tak díky paměti obnovit vědomé prožívání z minulosti. A v neposlední řadě je selektivní a kapacitně omezené v daném okamžiku. Takto je popisováno vědomí ve své druhé úrovni podle Adama Zemana. V té poslední, třetí, je pak spojováno s myslí v její vědomé i nevědomé podobě. Být vědomý v tomto případě znamená mít v mysli obsahy, jako je například víra, předpoklady, či obavy. Takové obsahy nejsou výhradně spojovány pouze s vědomím, jelikož mohou být součástí nevědomé části naší mysli.[6]

Vycházejíce z nastíněných vymezení, budeme definovat pojem vědomí pro tuto práci následovně. Vědomí je složitý subjektivní fenomén, který je nepostradatelným aspektem lidské existence. Zahrnuje jedinečnou schopnost jednotlivců vnímat a interpretovat svět kolem sebe, utvářet jejich charakteristické zkušenosti a zásadně přispívat k jejich pocitu sebe sama. Naše vymezení zdůrazňuje dichotomii mezi subjektivně prožívaným vědomím v první osobě a objektivní perspektivou třetí osoby, která je nezbytná pro jeho vědecká pozorování a analýzu.

[5] Harley, T., *The Science of Consciousness*, s. 3–6.
[6] Zeman, A., Consciousness, s. 1265–1266.

2. Teorie vědomí

Získali jsme určitou představu o tom, jak může být vědomí definováno, nicméně dalším důležitým krokem je podívat se na některé základní přístupy pracující s tímto pojmem. Zkoumání vědomí v rámci kognitivní vědy a neurovědy je mnohostranné úsilí, které zahrnuje pronikání do různých teoretických rámců s cílem odhalit složitosti vědomé zkušenosti. V souvislosti s určováním kritérií vědomí slouží teorie jako nepostradatelné nástroje pro pochopení základní povahy vědomí, jeho neurálních korelátů a jeho vztahu k mozku. Zkoumání teorií, jako je teorie globálního pracovního prostoru, integrované informační teorie, teorie rekurentního zpracování a teorie vyššího řádu, nabízí cenné poznatky o mechanismech, které jsou základem vědomého vnímání a poznávání. V současnosti jsou tyto čtyři teorie (teorie globálního pracovního prostoru, integrované informační teorie, teorie rekurentního zpracování a teorie vyššího řádu) vnímány jako hlavní kandidáti na onu „pravou teorii vědomí“, a proto jsem si k představení vybrala právě je.

2.1. Teorie globálního pracovního prostoru

Jedná se o kognitivní teorii rozvinutou Bernardem Baarsem kolem roku 1980. Autor svoji teorii velmi srozumitelně ukazuje na příkladu s divadelním prostředím. Stejně tak jako v našem vědomí, i v divadle je ve středu pozornosti většinou jen jedinec, či velice úzká skupina herců a zbytek se odehrává mimo naši vědomou pozornost. Vědomí nám analogicky umožňuje vnímat pouze málo informací, zatímco procesů, které se dějí okolo bez našeho přímého vědomí, je nespočet. Je důležité poznamenat, že právě tyto okolní procesy tvarují to, co se bude dít ve světle našeho vědomí. Nejedná se ale o bezvýznamné přirovnání, neboť celá teorie je podle autora stavěna na základech neurovědy a psychologie, přičemž na pozadí je kontext a v pomyslném světle reflektorů stojí vědomí a vše kolem představuje naši aktivní paměť. Globální pracovní prostor (global workspace) je prostor, ve kterém se v jedné chvíli pracuje pouze s malým počtem informací. Kontext nám poskytuje informace ze smyslových vjemů, ty se dostávají k vědomí, které je naopak spojeno s těmi podvědomými složkami mysli, jako jsou: naučené schopnosti, jazykové znalosti, dlouhodobá paměť, systémy na rozpoznávání známých objektů a vše, co jsme nasbírali během života. Informace se stane vědomou, je-li zpracována v tomto globálním pracovním prostoru, a následně je jí otevřena cesta ke zmíněným podvědomým systémům. Z popisu takového systému posléze vyplývá, že

vědomí je podle Baarse skutečně funkční jednotkou kognitivního systému hrající ústřední roli.[7]

2.2. Informačně-integrační teorie

Informačně-integrační teorie (IIT) vědomí, kterou navrhl neurolog Giulio Tononi, se snaží vysvětlit vznik vědomí ve fyzikálních systémech, zejména v lidském mozku. Tato teorie nabízí novou perspektivu a zdůrazňuje zásadní roli integrovaného zpracování informací při vytváření subjektivní zkušenosti. V IIT hrají klíčovou roli při stanovení základních principů, na nichž je teorie postavena, axiomy a postuláty. Tyto axiomy a postuláty slouží jako základní předpoklady, jimiž se řídí vývoj a výklad rámce IIT pro pochopení vědomí. Axiomy jsou samozřejmé pravdy nebo principy, které jsou přijímány bez důkazů a tvoří základ pro uvažování v rámci určitého systému. V kontextu IIT axiomy stanovují základní předpoklady o povaze vědomí a vlastnostech, které vykazuje. Například jedním z axiomů IIT je tvrzení, že vědomí existuje, což představuje základní východisko pro zkoumání subjektivní zkušenosti v rámci teorie. Postuláty jsou naproti tomu konkrétní tvrzení nebo propozice, o nichž se předpokládá, že jsou v určitém teoretickém rámci pravdivé.

Giulio Tononi, Melanie Boly, Marcello Massimini a Christof Koch axiomy charakterizují v článku *Integrated Information Theory: from Consciousness to Its Physical Substrate*. První axiom IIT tvrdí, že vědomí existuje *inherentně*, což znamená, že vědomí je reálný a empiricky pozorovatelný jev. Postulát odpovídající prvnímu axiomu předpokládá, že fyzikální substrát vědomí[8] musí rovněž existovat inherentně, což znamená, že má pro sebe z vlastní perspektivy příčinnou moc. Příkladem tohoto pojetí je funkčnost neuronů v mozku, kde mají vnitřní stavy, které mohou být ovlivňovány vstupy a mohou ovlivňovat jiné neurony, čímž prokazují vnitřní existenci. Druhý, axiom *kompozice*, říká, že zkušenost je strukturovaná a skládá se z několika fenomenálních distinkcí v jejím rámci. Například v rámci jedné zkušenosti lze rozeznat různé prvky, jako jsou objekty, barvy, prostorová umístění atd. Axiom *integrace* tvrdí, že zkušenost je jednotná, což znamená, že ji nelze rozdělit na nezávislé části, ale že se skládá z integrovaných fenomenálních

[7] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 113–115.
[8] Axiomy IIT říkají, že každá zkušenost existuje sama o sobě a je strukturovaná, specifická, jednotná a určitá. IIT dále postuluje, že pro každou podstatnou vlastnost zkušenosti musí existovat odpovídající fyzikální substrát vědomí (physical substrate of consciousness). Jinými slovy, struktura a vlastnosti vědomí se odrážejí v základních nervových procesech a mechanismech mozku. (viz Tononi, G., Boly, M., Massimini, M., Koch, C., Integrated Information Theory: from Consciousness to Its Physical Substrate.*)*

distinkcí. Například vizuální zkušenost scény nelze rozdělit na nezávislé zkušenosti různých prvků v rámci této scény. Odpovídající postulát tvrdí, že příčinná struktura specifikovaná fyzikálním substrátem vědomí musí být rovněž jednotná a neredukovatelná na vzájemně nezávislé subsystémy. Tato unitární povaha vyžaduje obousměrné interakce uvnitř systému, měřené jako integrovaná informace (Φ Phi). Axiom *exkluze* říká, že zkušenost je určitá svým obsahem a časoprostorovým zrnem, což znamená, že má určitý soubor prvků a definované trvání. Například aktuální prožitek člověka zahrnuje specifické smyslové vjemy a trvá určitou dobu, nikoliv že by byl více či méně podrobný nebo že by trval různou dobu.[9]

2.3. Teorie rekurentního zpracování

Teorie rekurentního zpracování (Recurrent processing theory/TRZ) Victora Lammeho nabízí pohled na mechanismy, které jsou základem zrakové percepce, a zpochybňuje tradiční modely zpracování zraku. Ve své podstatě zdůrazňuje opakující se povahu nervového zpracování a zdůrazňuje význam rekurentních spojení v mozku. Na rozdíl od konvenčních modelů, které navrhují jednosměrný tok smyslových informací hierarchickými oblastmi mozku, TRZ předpokládá, že zraková percepce vzniká na základě neustálých interakcí a zpětnovazebních smyček mezi smyslovými oblastmi a oblastmi mozku vyššího řádu. V Lammeho rámci existují čtyři fáze normálního vizuálního zpracování, z nichž každá je charakterizována odlišnými neurálními aktivitami. První stadium zahrnuje povrchové dopředné zpracování, kdy jsou vizuální signály zpracovávány lokálně v rámci zrakového systému. Toto počáteční zpracování probíhá v primární zrakové oblasti V1.[10] Druhá fáze, hluboké dopředné zpracování, nastává, když zrakové signály postupují dále v hierarchii zpracování, kde mohou ovlivnit činnost. Ve třetí fázi dochází k lokálnímu rekurentnímu zpracování, kdy se informace vracejí zpět do dřívějších zrakových oblastí, což vede k lokálnímu rekurentnímu zpracování. A konečně ve čtvrté fázi dochází k rozsáhlému rekurentnímu zpracování, kdy informace aktivují rozsáhlé oblasti v mozku.[11]

[9] Tononi, G., Boly, M., Massimini, M., Koch, C., Integrated Information Theory: from Consciousness to Its Physical Substrate, s. 450–452.
[10] Primární zraková oblast V1 se významnou měrou podílí na zpracování vnímání viděných objektů.
[11] Wu, W., The Neuroscience of Consciousness, nečíslováno.

Jedním z klíčových poznatků TRZ je vysvětlení jevů, jako je zpětné maskování[12], kdy je krátce prezentovaný podnět zneviditelněn následným maskujícím podnětem. Podle Lammeho teorie zpětné maskování narušuje rekurentní tok informací, narušuje rekurentní zpracování a zabraňuje vědomému vnímání původního podnětu. Důkazy o rozsáhlém nevědomém zpracování informací pocházejí například z neurozobrazovacích studií, které ukazují aktivaci různých oblastí mozku v reakci na nevědomé podněty, včetně vizuální oblasti slovních tvarů (visual word form area)[13] a precentrálního kortexu, zatímco vědomé podněty aktivují rozsáhlou neuronovou síť mozkových oblastí, například ve zrakové, či temenní a frontální kůře. Tyto aktivace naznačují, že nevědomé informace mohou ovlivňovat percepční, emocionální, motivační, motorické a rozhodovací procesy, přestože nevyvolávají vědomé vnímání.[14] V jedné studii Lau a Passingham (2007) použili funkční magnetickou rezonanci ke zkoumání, zda lze nevědomě spustit přípravu na zadání úkolu. Účastníci byli vyzváni, aby provedli buď fonologický, nebo sémantický úsudek o nadcházejícím slově, kterému předcházely vědomé, nebo nevědomé primární podněty spojené se stejnými nebo alternativními úkoly. Výsledky ukázaly, že nevědomé podněty spouštěly aktivitu v korových sítích spojených s instruovaným úkolem, včetně premotorické kůry, což naznačuje, že neurální sítě související s úkolem, zahrnující prefrontální kůru, mohou být modulovány nevědomě.[15]

Podobná zjištění naznačují, že několik prefrontálních kognitivních funkcí vysoké úrovně může být spuštěno nevědomě, což indikuje, že kognitivní procesy zprostředkované prefrontálními oblastmi jsou iniciovány i při absenci vědomého uvědomění. Navzdory aktivaci těchto kognitivních funkcí však vědomá vizuální zkušenost zůstává nepřítomna, z čehož můžeme soudit, že výměna informací mezi mozkovými oblastmi

[12] Při pokusech se zpětným maskováním je účastníkům prezentován cílový podnět, po němž rychle následuje jiný podnět, tzv. prime, který je prezentován velmi krátce a v rámci obrysu cílového podnětu. Viditelnost základního podnětu je silně snížena a za určitých podmínek je až nemožné jej vidět. Důležité je, že tentýž krátce prezentovaný podnět je dokonale viditelný, je-li prezentován izolovaně. Dokonce i zcela maskované, a tedy nevědomé podněty mohou stále ovlivňovat percepční a behaviorální procesy, což dokládají rychlejší reakční časy a menší počet chyb, když jsou prime a cíl kongruentní, tedy ukazují stejným směrem.

[13] VWFA je funkční oblast nacházející se v levém fusiformním gyru. Předpokládá se, že VWFA hraje klíčovou roli při identifikaci slov a písmen, zejména ze zrakových podnětů, dříve, než jsou spojeny s fonologickými nebo sémantickými informacemi. To naznačuje, že VWFA se podílí na raných fázích vizuálního zpracování souvisejícího s psaným jazykem.

[14] Van Gaal, S., Lamme Af, V., Unconscious High-level Information Processing: Implication for Neurobiological Theories of Consciousness, s. 289–290.

[15] Van Gaal, S., Lamme Af, V., Unconscious High-level Information Processing: Implication for Neurobiological Theories of Consciousness, s. 291.

zprostředkovaná rekurentními neurálními interakcemi může být pro generování vědomí klíčová.[16]

2.4. Teorie vyššího řádu

Teorie vyššího řádu je teorie vědomí, podle níž je ke vzniku vědomí zapotřebí více než jediná úroveň mentálních stavů. Tomáš Marvan a Michal Polák v knize *Vědomí a jeho teorie* tuto teorii charakterizují následovně. Základní myšlenkou je, že mentální stav se stává vědomým, je-li na něho zaměřen stav vyššího řádu. První rovina zahrnuje mentální stav s obsahem a může jím být pocit, myšlenka, emoce, nebo vjem (nazvěme jej „M"). Druhá rovina obsahuje mentální stav druhého řádu, jenž je zaměřen na ten první a tímto stavem může být myšlenka, či percepce s obsahem „Právě se nacházím ve stavu M". Toto zaměření činí původní stav, do té doby nevědomý, vědomým. V případě percepce vyššího řádu (PVŘ) by tato percepce měla za pomoci určitého percepčního mechanismu monitorovat řád první. Nicméně kromě toho, že doposud žádný monitorovací mechanismus, či smysl, nalezen nebyl, je teorii PVŘ dále vytýkáno, že se ve skutečnosti jedná o proces na jediné úrovni.[17]

Marvan a Polák dále charakterizují i druhou variantu teorie vyššího řádu, kterou je myšlenka vyššího řádu (MVŘ). V tomto případě je na první úrovni opět nevědomý mentální stav M. Vědomým se stane ve chvíli, kdy se na něho zaměří myšlenka vyššího řádu typu „Nyní se nacházím ve stavu M." Myšlenka vyššího řádu je však sama o sobě také nevědomá, a proto je ale zapotřebí řád ještě vyšší, tzv. introspektivní, který zvědomí tuto myšlenku, jež je zaměřena na stav M. V takovém případě bude forma tohoto ještě vyššího řádu následovná: „Jsem si vědom toho, že se nacházím ve stavu M."[18] Takové teorie ulehčují pochopení rozdílu mezi tím, proč některé stavy vědomé jsou, a jiné ne, nebo, proč stejné stavy někdy vědomé jsou, ale jindy nejsou. Van Gulick[19] zmiňuje, že problém teorií vyššího řádu spočívá v něčem, co je označováno jako problém generality, který vyvolává otázky ohledně dostatečnosti těchto teorií při vysvětlování vědomí. Problém generality poukazuje na významný problém, s nímž se názory vyššího řádu potýkají, konkrétně při úvahách o tom, proč by přítomnost myšlenky nebo vnímání

[16] Van Gaal, S., Lamme Af, V., Unconscious High-level Information Processing: Implication for Neurobiological Theories of Consciousness, s. 295.
[17] Marvan, T., Polák, M., *Vědomí a jeho teorie,* s. 67.
[18] Marvan, T., Polák, M., *Vědomí a jeho teorie,* s. 67–68.
[19] Van Gulick, R., Consciousness, nečíslováno.

určitého duševního stavu (např. touhy nebo vzpomínky) měla z tohoto stavu činit vědomý duševní stav. Argumentace je následující. Pouhá myšlenka nebo vnímání vnějšího objektu neznamená, že tento objekt má vědomí. Například myšlení na vnější objekt X nebo jeho vnímání neznamená, že X sám o sobě je vědomý. Na rozdíl od toho ale teorie vyššího řádu předpokládají, že myšlenka nebo vnímání duševního stavu (např. touhy nebo vzpomínky) činí tento duševní stav vědomým. Úkolem je tedy podle Van Gulicka vysvětlit, proč existuje rozdíl v použití termínu *vědomý* pro vnější objekty a pro duševní stavy.[20]

Dle mého názoru je zkoumání teorií vědomí v kontextu kritérií vědomí u lidí i umělé inteligence nezbytné z několika důvodů. Zaprvé, pronikání do těchto teorií umožňuje hlubší pochopení naší vlastní mysli. Tyto teorie nabízejí rámce a perspektivy, které se snaží odhalit složitosti vědomého prožívání a vrhají světlo na to, co to je vědomí, jaký je vztah mezi vědomím a mozkem, či jaké vlastnosti a charakteristiky vědomí patří. Bez těchto základních poznatků o povaze našeho vlastního vědomí je jen těžké uvažovat o existenci vědomí u jiných organismů, či dokonce umělé inteligence. Bez dobrého pochopení vlastního vědomí se vypracování kritérií pro hodnocení vědomí jiných entit stává náročným úkolem. Teorie vědomí slouží jako základ, na němž lze kritéria stavět. Zkoumáním teorie globálního pracovního prostoru, integrované informační teorie, teorie rekurentního zpracování a teorie vyššího řádu se zapojujeme do kritického diskurzu, který pomáhá formulovat a zpřesňovat kritéria nezbytná pro identifikaci vědomí.

Dále se domnívám, že v oblasti umělé inteligence je pochopení teorií vědomí velmi důležité. Když se snažíme vytvořit inteligentní systémy, které napodobují nebo simulují aspekty lidského vědomí, řídí se náš přístup teoretickými rámci. Teorie poskytují vodítko pro to, jaké aspekty vědomí je nezbytné v systémech umělé inteligence replikovat nebo napodobovat, a pomáhají nám tak identifikovat kritéria, podle nichž hodnotíme přítomnost, či nepřítomnost vědomí u umělých entit. Diskuse o teoriích vědomí v úvaze o kritériích vědomí u lidí a umělé inteligence je v podstatě klíčová pro položení základů porozumění, stanovení kritérií a orientaci v hlubokých důsledcích uznání vědomí u biologických i umělých entit. Zajišťuje, že naše zkoumání je založeno na bohatém filozofickém a teoretickém rámci, což zvyšuje hloubku a komplexnost našeho zkoumání kritérií vědomí.

[20] Van Gulick, R., Consciousness, nečíslováno.

3. Kritéria vědomí u člověka

3.1. Neurální kritéria

3.1.1. Neurofyziologie mozku

Pochopení neurofyziologie mozku má zásadní význam pro vědecké zkoumání vědomí, protože umožňuje nahlédnout do mechanismů, které jsou základem našich subjektivních prožitků a kognitivních procesů. Přestože primárním cílem neurovědců není odhalit samotnou podstatu vědomí, ale spíše proniknout hlouběji do vztahu mezi chováním, prožíváním a neurofyziologickou strukturou mozku, tento výzkum významně přispívá i ke zkoumání vědomí. Výhodou neurovědy je tedy její možnost nahlédnutí do hmatatelného biologického substrátu a pozorování probíhajících procesů. Na rozdíl od filozofických výzkumů, které sahají od zájmu o evoluční počátky vědomí až po oblast umělé inteligence, neurověda podrobně zkoumá vědomí také prostřednictvím změněných stavů, ať už vyvolaných duševní nemocí nebo traumatem. Specializovaná oblast neurovědy se věnuje identifikaci měřitelných korelátů spojených s vědomím.[21]

Zkoumání neurofyziologie mozku má zásadní význam při diskusi o kritériích vědomí u lidí i umělé inteligence. Toto zkoumání poskytuje základ pro stanovení kritérií tím, že identifikuje klíčové faktory a vzorce spojené se stavy vědomí. Neurofyziologická měření slouží jako empirický důkaz pro ověření navržených kritérií a umožňují srovnávací analýzu mezi biologickými a umělými systémy. Stručný popis neurofyziologie mozku uvádí Susan Blackmore a Emily Troscianko v knize *Consciousness: An Introduction.* Mozek je sám o sobě velmi složitá struktura složená ze zhruba 86 miliard neuronů, bilionů synapsí, které je spojují a miliard dalších podpůrných gliových buněk. Základem centrální nervové soustavy (CNS) je mícha. Mícha slouží jako oblast, do které přichází informace díky motorickým a senzorickým nervovým buňkám a jsou skrze míchu posílány dále do mozkového kmene. Mozkový kmen je životně důležitou jednotkou zajišťující základní funkce lidského těla. V úrovni krční míchy začíná neuronová síť, jež dále prochází mozkovým kmenem. Je nazývaná jako retikulární formace a u obratlovců zajišťuje aktivaci regionů při přechodu organismu ze spánku do bdění nebo z uvolněného bdění do bdělé pozornosti.[22] V rámci fungování vědomí hraje důležitou roli, nikoli však zcela

[21] Havlík, M., *Vědomí a úrovně vědomí*, s. 195–196.
[22] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 79.

dostačující. Další důležitou součástí CNS je mozeček, řídící složité motorické funkce, například tělesnou rovnováhu při pohybu. Vedle této struktury je oblast zvaná thalamus, do něhož se dostávají veškeré informace ze smyslových senzorů, a právě zásluhou thalamu jsou tyto informace dále posílány do různých center mozkové kůry a tvoří tak důležité thalamo-kortikální smyčky. V rámci výzkumu vědomí bylo zjištěno, že právě tyto smyčky se na funkci vědomí také podílejí. A v neposlední řadě musíme v rámci tématu zmínit i mozkovou kůru a její roli. Pod ní schovaný limbický systém je složen z hned několika útvarů podílejících se na vědomí. S tvorbou dlouhodobých vzpomínek v paměti nám pomáhá hipokampus. Dále pak důležitou roli hrají amygdala, zodpovědná za emoční paměť, a hypothalamus s řídícím centrem naší autonomní nervové soustavy. U lidského mozku se pak nejvíce rozvinula část zvaná neokortex. Rozvrásnění mozku je v podstatě adaptivní vlastnost, která umožňuje přizpůsobit složitost struktury a funkce mozku v rámci lebeční dutiny. Kortex je obecně rozdělován podle role, kterou zastávají jednotlivé oblasti. Najdeme zde lalok čelní, spánkový, temenní, či týlní.[23]

3.1.2. Navrhovaná neurální kritéria

Autoři Anil Seth, Bernard Baars a David Edelman ve své práci *Criteria for Consciousness in Humans and Other Mammals,* věnované kritériím vědomí stanovují tři základní charakteristiky, které vědomí obsahuje. První se týká nepravidelné aktivity s nízkou amplitudou při elektroencefalografickém měření, druhá interakce mezi mozkovou kůrou a thalamem a třetí charakteristika mluví o rozsáhlé mozkové aktivitě. Tato tři základní pravidla by měla být dle autorů jasná a v rámci výzkumu jednoduše testovatelná, což z nich dělá ideální kandidáty na základní neurální kritéria.[24]

Berger spojil v roce 1929 vědomí se snímky z elektroencefalografie (EEG), které vykazovaly mozkovou aktivitu mezi 20-70 Hz, zatímco různé stavy bezvědomí, včetně hlubokého spánku, vegetativních stavů po poškození mozku, anestezie a epileptických absenčních záchvatů, byly charakterizovány pomalými, vysokoamplitudovými a pravidelnějšími vlnami o frekvenci nižší než 4 Hz.[25] Studie jednotlivých jednotek (single-

[23] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 80–81.
[24] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 120.
[25] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 122–123.

unit studies)[26] odhalily[27], že během spánku s pomalými vlnami (tzv. hluboký spánek beze snů) vykazuje velký počet korových neuronů výbuchy aktivity na vrcholu pomalé vlny a synchronně se zastavují během poklesu. Zatímco frekvence výbuchů jednotlivých neuronů v kůře během pomalého spánku se výrazně neliší od frekvencí pozorovaných během bdělosti, charakteristickým rysem spánku jsou synchronizované neuronální pauzy při frekvencích nižších než 4 Hz. Tyto synchronizované pomalé pauzy během pomalovlnného spánku mohou narušovat rychlé interakce mezi korovými oblastmi, které jsou nezbytné pro bdělé funkce, jako je vnímání, bezprostřední paměť, interakce mezi vzdálenými korovými oblastmi, vnitřní řeč a plánování činnosti. Kromě toho přítomnost pomalých vln v jiných stavech bezvědomí, ať už v důsledku poškození mozku, anestezie nebo epileptických záchvatů absence, naznačuje společný mechanismus přerušení vědomých funkcí.[28]

Druhé kritérium je dle autorů spojováno s thalamokortikálním systémem. Nicméně výzkum naznačuje, že různé oblasti mozku hrají ve vědomých prožitcích odlišnou roli. Integrativní chápání vědomí proto musí brát v úvahu zapojení mnoha oblastí mozku a jejich interakce. To zahrnuje zapojení takových oblastí, jako je hipokampus pro vědomé ukládání a vybavování epizod a mozeček pro vědomou zpětnovazební kontrolu jemné motoriky.[29] K poškození oblastí, jako je hipokampální systém a mozeček, může dojít, aniž by nutně došlo ke ztrátě vědomí. Avšak v oblasti mozkového kmene, kde se nachází thalamus, může při poškození dojít k částečné, nebo úplné ztrátě vědomí. Pokud se nejedná o úplnou ztrátu vědomí, může dojít ke ztrátě konkrétních vědomých funkcí, jako je barvocit nebo vědomé prožitky vizuálních objektů včetně obličejů. Mozkový kmen prokázal funkci udržování základních funkcí lidského těla včetně vědomí a dochází-li k interakci mezi kůrou a thalamem, pak je mozek schopný udržet si vědomé obsahy. Tímto je tedy dle autorů charakterizováno druhé testovatelné kritérium pro identifikaci vědomí.[30]

---

[26] Studie jednotlivých jednotek neuronů obvykle označuje typ neurovědního výzkumu, který se zaměřuje na aktivitu jednotlivých neuronů v mozku nebo nervovém systému. Při těchto studiích vědci obvykle zaznamenávají elektrickou aktivitu jednoho neuronu pomocí specializovaných technik, jako jsou mikroelektrodové záznamy.

[27] viz Steriade, M., Mccormick, D. A., Thalamocortical Oscillations in the Sleeping and Aroused Brain.

[28] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 122.

[29] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 124.

[30] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 123.

Třetí zmíněnou vlastností je rozsáhlá mozková kortikální aktivita. Velké množství experimentů prokázalo, že vědomé smyslové vjemy podněcují v mozku aktivitu, která se rozšíří do temenní, čelní a spánkové oblasti kortexu, zatímco nevědomé smyslové vstupy aktivují smyslové oblasti převážně lokálně bez zapojení rozsáhlých korových oblastí. Úkoly vyžadující vědomou pozornost zapojují širší síť korových oblastí ve srovnání s úkoly, které se postupem času stávají rutinními, automatickými a nevědomými. Jak úkoly přecházejí z nových a vědomě zpracovávaných na známé a automaticky prováděné, jejich zastoupení v kůře se omezuje, což naznačuje posun v neurálním zpracování od rozsáhlé aktivace k cílenějším oblastem.[31]

Takto jsou shrnuta nezbytná kritéria pro určení vědomí dle autorů, kteří závěr formulují následně: „*Tyto první tři vlastnosti společně naznačují, že vědomí zahrnuje rozsáhlé, relativně rychlé, nízkoamplitudové interakce v thalamokortikálním jádru mozku, řízené aktuálními úkoly a podmínkami. Stavy nevědomí jsou výrazně odlišné a mnohem méně reagují na senzorické vstupy nebo endogenní aktivitu. Tyto vlastnosti jsou přímo testovatelné a představují nezbytná kritéria pro vědomí u lidí.*“[32]

3.1.3. Neurální koreláty vědomí

Studium vědomí zahrnuje jak zkoumání neurálních kritérií potřebných pro existenci vědomí, tak identifikaci neurálních korelátů vědomí. Neurální koreláty vědomí (NKV) slouží jako empirické markery, které poukazují na pozorovatelné neurální procesy spojené s vědomím. Tyto koreláty jsou základem pro zkoumání a pochopení neurálních kritérií vědomí, protože nabízejí měřitelné a pozorovatelné jevy, které se shodují se změnami vědomých stavů. Vztah mezi neurálními koreláty a neurálními kritérii je symbiotický: empirické důkazy ze studií neurálních korelátů informují o stanovení a ověření koherence vědomí a fyzických stavů, zatímco kritéria, jakmile jsou identifikována, následně přispívají k upřesnění teoretických rámců a pochopení povahy vědomí. Toto vzájemné působení mezi neurálními koreláty a kritérii je zásadní pro lepší pochopení toho, jak neurální aktivita vede ke vzniku vědomých prožitků.

Neurální koreláty, jak popisují Blackmore a Troscianko, jsou běžně sledovány například na základě vizuálního vnímání, a to zejména u případu tzv. binokulární rivality, kdy jeden

[31] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 123–124.
[32] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 124.

vizuální objekt může vyvolávat rozdílné percepce. K tomu dochází, jelikož jsou oběma očím předkládány různé objekty. Jednomu oku se například může zobrazit obrázek mořského pobřeží, zatímco druhé oko vidí obličej. Namísto toho, aby došlo ke splynutí pobřeží a obličeje v jeden obraz, má mozek tendenci přepínat vnímání z jednoho na druhý. Takové podmínky jsou velmi vhodné pro výzkum vztahu mezi objektivními podněty a jejich subjektivním vnímáním. Při výzkumu právě takových situací na primátech byla nalezena neurální oblast v dolní spánkové kůře, která reagovala pokaždé, když se subjektivní percepce tvaru změnila. [33] Studie binokulárního soupeření umožnily nahlédnout do dynamiky vědomých vizuálních zážitků a přispěly k diskusím o hierarchických modelech zpracování a povaze vědomí. Tyto studie však stanovují pouze korelace mezi mozkovou aktivitou a vědomým prožitkem a ponechávají nezodpovězené otázky týkající se kauzálních mechanismů, které jsou základem vědomí.[34]

Jak již bylo řečeno, vizuální vnímání je jednou z nejběžnějších oblastí, ve které se provádí výzkum korelací vědomí. Vizuální percepce je dobře prozkoumaná, a výzkumníci tak mají dobrou kontrolu nad vizuálními objekty, které využívají. Francis Crick a Christof Koch patří mezi významné badatele hledající koreláty mezi vizuální percepcí světa kolem nás a neurálními oblastmi s odpovídající aktivitou. Při zkoumání vědomí navrhli model, v němž čelní mozková kůra představuje jakéhosi „nevědomého pozorovatele“, zatímco různé oblasti mozku se zapojují do přechodných koalic neuronů reprezentujících myšlenky, obrazy a vjemy. Tyto koalice mezi sebou soutěží, přičemž pozornost ovlivňuje jejich soupeření. V kontextu vidění se neurální aktivita rychle přesouvá do frontální kůry, kde poskytuje vědomý přehled o scéně, a pak se postupně vrací, aby poskytla podrobné informace. Vědomé vidění přirovnávají k sérii snímků s „namalovaným“ pohybem (motion ‘painted’ on)[35]. S tímto rámcem se pustili do identifikace neurálních korelátů vědomí a při jejich hledání zvažovali Crick a Koch různé možnosti, včetně toho, zda vědomí vzniká na základě konkrétních skupin pálejících neuronů, specifických vzorců pálení těchto neuronů, kombinací těchto dvou forem, či na základě něčeho zcela jiného. Žádná z těchto možností se však nezdá být zcela uspokojivá, když uvažujeme o neuchopitelné povaze vědomí. Výzva spočívá v pochopení toho, co způsobuje, že některé neurální aktivity jsou vědomé, zatímco jiné zůstávají nevědomé. Často se stává, že vědci přisuzují vlastnosti *těžkého problému* vědomí jen některým částem mozku, často popisují

[33] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 81–82.
[34] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 91.
[35] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 84.

určité oblasti mozku identifikované v neurozobrazovacích studiích jako spojené s vizuálním vědomím, ačkoli konkrétní vztah mezi těmito oblastmi a vědomím zůstává nejasný.[36] Crick a Koch navrhují, že se subjektivními zážitky (kválii) spíše korelují s přechodnými výsledky výpočtů v mozku, zatímco výpočty samé jsou nevědomé. Podobně i další vědci, jako Ray Jackendoff a Jesse Prinz, zkoumali tok informací ve vizuální hierarchii, aby určili, kde vědomí vzniká, přičemž Prinz navrhl, že vědomé stavy jsou reprezentace na střední úrovni modulované pozorností. Tyto perspektivy přispívají k pokračujícímu úsilí o pochopení nervového základu vědomí a jeho vztahu k mozkovým procesům.[37]

Do debaty o neurálních korelátech vědomí se zapojují autoři Stuart Hameroff a Roger Penrose, kteří sledují vliv anestetik na vědomé stavy. Jejich teorie se zaměřuje na nejnižší úroveň výkladu neurálních korelátů vědomí (neboť neurální koreláty vědomí mohou být hledány na různých úrovních, tj. molekulární, buněčné, kvantové apod.). Předpokládají, že vědomí vzniká z kvantových procesů probíhajících v mozku, konkrétně v mikrotubulech neuronů. Tyto mikrotubuly jsou válcovité bílkovinné struktury, které se nacházejí v cytoplazmě neuronů a tvoří nedílnou součást strukturálního rámce buňky. V oblasti kvantové mechaniky mohou částice existovat ve více stavech současně, což je jev známý jako superpozice. Kromě toho se částice mohou vzájemně proplétat, což znamená, že stav jedné částice je korelován se stavem jiné částice, i když jsou od sebe vzdáleny velkou vzdáleností. Penrose zavedl do kvantové teorie pojem *objektivní redukce*, který předpokládá, že superpozice se samovolně zhroutí, když je dosaženo určitého prahu, což vede k určitému stavu nebo výsledku. Penroseova a Hameroffova teorie kombinuje tyto koncepty a navrhuje, že kvantové superpozice v mikrotubulech dosahují prahu pro objektivní redukci. Podle Penrose a Hameroffa jsou události objektivní redukce v mikrotubulech považovány za základní mechanismus, který je základem vzniku vědomí.[38]

Hans-Joachim P. Flohr se také snaží o výklad neurálních korelátů vědomí, tentokrát ale na úrovni molekulární. Při výzkumu také využívá účinky anestetik měnících stavy vědomí. Flohrova práce se soustředila na pochopení molekulárních mechanismů, které jsou základem vědomí, zejména na úlohu NMDA (N-methyl-D-asparagová kyselina)

[36] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 85.
[37] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 85.
[38] Marvan, T., Polák, M., *Vědomí a jeho teorie,* s. 99.

receptorů při zprostředkování neurální aktivity a kognitivních procesů spojených s vědomím. NMDA receptory jsou typem glutamátového receptoru, který se podílí na synaptickém přenosu a plasticitě. Flohr navrhl, že NMDA receptory hrají klíčovou roli při integraci nervových informací a vytváření koherentních vědomých vjemů. Při normálním vědomí je aktivační práh kanálů NMDA receptorů[39] úzce spjat s depolarizací postsynaptické buněčné membrány. Flohr poznamenává, že působení ketaminu blokuje excitační účinky GABA (kyselina gama-aminomáselná), což vede ke změnám v aktivitě NMDA receptorů a v konečném důsledku ke ztrátě vědomí. Tvrdí, že standardní fungování NMDA receptorů je pro existenci vědomí nejen nezbytné, ale i dostačující, což naznačuje, že jiné neporušené fyziologické nebo výpočetní mozkové procesy jsou pro vznik vědomí nedostatečné. Flohrova hypotéza navíc dává úlohu NMDA receptorů do kontextu širšího rámce vzniku vědomí. Navrhuje, že kortikální NMDA synaptické procesy usnadňují vytváření velkých neurálních shluků v mozku, které slouží jako fyziologické vyjádření duševních stavů. Tyto shluky, které se řídí hebbovým[40] pravidlem neurální excitace, vytvářejí svou společnou aktivitou základ vědomé zkušenosti. Flohrova teorie tak zdůrazňuje význam aktivity NMDA receptorů při utváření vědomí na buněčné a molekulární úrovni.[41]

Podle mého názoru je studium neurálních korelátů vědomí v neurovědách zásadní činností, jejímž cílem je objasnit vztah mezi mozkovou aktivitou a vědomým prožíváním. Identifikací specifických neurálních procesů a mechanismů spojených s vědomím se vědci snaží odhalit tajemství toho, jak subjektivní prožitky vznikají z fyzikálních procesů v mozku. Tato snaha je významným příslibem pro zlepšení našeho chápání vědomí a jeho důsledků pro základní vědu i klinické aplikace. Jeden z hlavních přínosů studia NKV spočívá v jeho potenciálu prohloubit naše chápání samotného vědomí. Objasněním neurálních mechanismů, které jsou základem vědomí, mohou vědci získat náhled na základní otázky týkající se povahy vnímání, poznávání a subjektivního prožívání. Kromě

[39] NMDA (N-methyl-D-aspartátový) receptor je typ glutamátového receptoru, který se nachází v nervových buňkách. Aktivační práh NMDA (N-methyl-D-aspartátových) receptorů označuje úroveň depolarizace (změnu elektrického potenciálu) membrány potřebnou k otevření těchto receptorů a umožnění přítoku iontů, zejména vápenatých (Ca2+), do postsynaptického neuronu.
[40] Hebbovo pravidlo je základním principem neurovědy, který popisuje mechanismus, jímž se synaptická spojení mezi neurony posilují na základě jejich korelované aktivity. Základní myšlenka Hebbova pravidla spočívá v tom, že pokud jsou dva neurony opakovaně aktivovány ve stejnou dobu, spojení mezi nimi se posiluje. K tomuto posílení dochází prostřednictvím procesu známého jako synaptická plasticita, který zahrnuje změny v síle nebo účinnosti synaptických spojení. (viz Cooper, S.J., Donald O. Hebb's synapse and learning rule: a history and commentary, s. 868)
[41] Marvan, T., Polák, M., *Vědomí a jeho teorie,* s. 100–103.

zmíněného nabízí studium NKV odlišný přístup k identifikaci vědomí ve srovnání s jinými metodami, zejména například těmi, které jsou založeny na behaviorálních reakcích nebo introspekci. Jednou z klíčových výhod výzkumu NKV je jeho základ v biologickém substrátu skutečných fyzických mozků. Díky zkoumání neurálních procesů a mechanismů spojených s vědomím poskytují studie NKV korelace, které mají vědeckou platnost a mohou být studovány relativně nezávisle na jiných metodách.

Studium NKV však představuje také několik výzev a omezení. Jednou z významných výzev je inherentní subjektivita samotného vědomí. O této výzvě mluví i Blackmore a Troscianko. Vědomí je subjektivní fenomén, který se vzpírá přímému pozorování a měření, takže je obtížné zachytit jeho plnou bohatost a komplexnost pomocí objektivních měřítek. V důsledku toho se výzkum NKV často opírá o nepřímá měřítka vědomí, jako jsou vzorce neurální aktivity, které nemusí plně vystihovat subjektivní povahu vědomého prožívání. Další z problémů korelací, zejména v kontextu výzkumu vědomí, se vztahuje k problému stanovení kauzálního vztahu mezi pozorovanými neurálními koreláty a skutečnými mechanismy nebo procesy, které jsou základem vědomé zkušenosti. Výzkumníci sice mohou identifikovat vzorce neurální aktivity, které korelují s různými stavy vědomí, ale to nutně neznamená, že tyto neurální aktivity jsou kauzálně odpovědné za vědomí. Korelace neznamená kauzalitu a určení kauzality je v neurovědách a filozofii mysli složitý problém.[42]

3.2. Behaviorální kritéria

Při snaze o rozpoznání kritérií vědomí v rámci behaviorálního spektra je nutné si uvědomit omezení a složitosti, které jsou tomuto snažení vlastní. Na první pohled mohou behaviorální kritéria zdánlivě nabízet hmatatelný prostředek k objektivnímu hodnocení vědomí u lidí i jiných entit. Při bližším zkoumání je však zřejmé, že spoléhání se pouze na behaviorální reakce jako na měřítko vědomí má značné nedostatky. I když jsou behaviorální reakce snadno pozorovatelné, představují pouze vnější projevy spletité souhry vnitřních procesů. Tyto vnitřní procesy, počínající již na úrovni neurofyziologie, postupně přechází v subjektivní, introspektivně pozorovatelné prožitky a až ve své poslední fázi se tyto procesy projeví navenek v podobě tělesné reakce. Tvrzení, že behaviorální kritéria mohou izolovaně sloužit jako objektivní a pozorovatelné ukazatele vědomí, by proto dle mého názoru představovalo mylnou domněnku. Ve skutečnosti totiž

[42] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 82.

nelze behaviorální projevy dobře oddělit od neurofyziologických procesů a od našeho subjektivního vnitřního světa a bude zde tedy vždy docházet ke vzájemnému ovlivňování těchto tří rovin. Pokud k nim ale budeme přistupovat s obezřetností a budeme mít na paměti jejich komplexnost, mohou naše poznání obohatit o nové poznatky v rámci poznávání vědomí a jeho vzniku.

3.2.1. Odpovídající reakce

Mezi jeden z hlavních indikátorů vědomí u člověka patří *accurate report* neboli odpovídající reakce, jež je nyní pravidelně používána v klinickém i vědeckém prostředí a je aplikována nejen u zdravých jedinců, nýbrž i při různých poškozeních mozku. Ve své nejběžnější formě se jedná o schopnost jedince podávat zprávy o svých vědomých zkušenostech ať již verbálně, či neverbálně. Toto kritérium se často používá ve filozofických diskusích a psychologickém výzkumu jako způsob, jak posoudit, zda si jedinec uvědomuje určitý podnět nebo zkušenost. Vyvinutý jazykový aparát je při této metodě značnou výhodou, neboť jedinec může jasně a přímo sdělovat vědomou zkušenost.[43]

První známky této odpovídající reakce nacházíme již při těch nejjednodušších aktivitách od útlého věku. Malé děti ukazují na to, co právě vidí, a mají-li již vyvinuté základní jazykové schopnosti, pak svůj okamžitý vědomý prožitek i doprovází slovy o tom, co vidí. Při měření přítomnosti vědomého prožitku ve vědeckém prostředí většinou jedinec odpovídá na otázky o svém vědomí ve formě mluvené odpovědi, psané odpovědi, či v jiné formě, například na základě měření tělesných reakcí nebo stlačování tlačítek s odpověďmi. Pokud odpovídáme s pomocí přirozeného jazyka a své vědomí popisujeme, setkáme se s několika problémy.[44] O nevýhodách měření odpovídajících reakci mluví David Rosenthal v článku *Consciousness and Confidence.* Jedním z největších problémů u měření vědomí na základě verbální zpětné vazby je naše omezení sdělit vše formou slov a za pomoci omezeného jazykového aparátu. Jsme-li objektem měření, není zkrátka možné nahlásit každý jeden obsah našeho aktuálně probíhajícího vědomí, neboť takových obsahů je příliš velké množství najednou a mění se velmi rychle. Přesto, že určování přítomnosti vědomí na základě behaviorálních kritérií má svá omezení, má ve svém

[43] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 130.
[44] Gamez, D., The Measurement of Consciousness: a Framework for the Scientific Study of Consciousness, s. 2.

základě také velkou výhodu. Sledovaný subjekt je schopný jasně a ihned rozhodnout, zda se cítí být ve stavu M, či ne a reportovat svoji odpověď slovně (či jinou behaviorální formou). Jinými slovy reportuje, zdali si je vědom svého stavu M, ve kterém se nachází. Díky tomuto jednoduchému sebepozorování jsme schopni rozlišit například konkrétní psychologické stavy vědomé od těch nevědomých a je tím naznačena souvislost mezi vědomím a jeho reportovatelností.[45]

Zdůvodnění této souvislosti je ilustrováno podmínkami, jako je maskovaný priming, kdy subjekty vykazují přesné reakce na podněty, ale upřímně popírají, že by je vnímaly. V takových případech rozpor mezi výkonem a subjektivním vědomím naznačuje nevědomou znalost. Psychologický stav je tedy považován za vědomý pouze tehdy, pokud si jedinec uvědomuje, že se v tomto stavu nachází, a zároveň vykazuje odpovídající reakce. Podmínky takovýchto experimentů musí být dobře nastaveny a striktně dodržovány, aby výsledky byly spolehlivé, neboť i zde nastává prolínání behaviorální metody s introspekcí a toto sebepozorování s sebou přináší rizika. Existují metody, při kterých se pozorovaný jedinec napřed musí naučit, jak správně slovně popisovat své vědomé stavy. Přesto, subjekt si často nemusí být vlastními odpověďmi jistý, neboť ani introspektivní pozorování není vždy spolehlivé. A ačkoli za účelem vylepšení této metody vznikla tzv. konfidenční škála, o které bude zmínka v kapitole o subjektivních kritériích, vyvolává spolehlivost takových experimentů otázky.[46] Spoléhat se na pouhou verbální zpětnou vazbu nicméně není výhodné a kvůli hledání vědomí u jiných živočišných druhů byly navrženy i neverbální varianty experimentů. Například alternativní strategie získávání spolehlivých neverbálních zpráv o vědomých zkušenostech od zvířat, konkrétně od opic, pomocí metody komentářového klíče. Tato metoda spočívá v tom, že opice jsou trénovány, aby komentovaly chování, nebo prováděly diskriminaci druhého řádu na základě předchozí percepční diskriminace.[47]

Odpovídající zpětná vazba je považována za kritérium vědomí, protože odráží schopnost jedince behaviorálně reportovat známky vlastního vědomí. Je ale důležité brát v potaz, co již bylo zmíněno, tedy že takový jedinec musí napřed introspektivně přistupovat ke svým vnitřním duševním stavům, vjemům, myšlenkám a pocitům a až poté je schopen je vyjadřovat na základě behaviorální reakce. Je tedy otázkou, zda behaviorální kritéria

[45] Rosenthal, D., Consciousness and Confidence, s. 256–257.
[46] Rosenthal, D., Consciousness and Confidence, s. 256–257.
[47] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 130

můžeme považovat za právoplatná samostatná kritéria, když jsou nepopiratelně ovlivňována subjektivními i neurofyziologickými jevy.

3.2.2. Schopnost učení

Schopnost učení může být dle Setha, Baarse a Edelmana označena za jedno z dalších kritérií vědomí. Výzkumy poukazují na fakt, že učení, zvláště to ukládající se posléze do dlouhodobé paměti, vyžaduje přítomnost vědomí. Takové učení, ať již kognitivní, či behaviorální, nemusí být záměrné, aby k němu docházelo, nicméně musíme jej vědomě vnímat. K nezáměrnému učení dochází nejčastěji například v oblasti vizuální paměti. Ovšem v běžném životě bývá vědomé učení také záměrné. Takto probíhá například osvojování si schopnosti čtení, zavazování tkaniček, či vaření. Zpočátku nám úkon trvá delší dobu a musíme se soustředit, abychom postupovali správně. Postupem času se pro nás z tohoto úkonu stane rutina a zvládneme jej splnit i bez vědomého vnímání. Jak bylo nastíněno v kapitole o neurálních kritériích, celý thalamokortikální systém se prokazuje jako součást podstaty vědomí, a stejně tak tento systém prokazuje svoji roli při učení něčeho nového. Na rozdíl od této komplexní sítě se můžeme podívat na bazální ganglia, která se podílejí na projevech již naučeného, rutinního chování. Anatomie bazálních ganglií, charakterizovaná dlouhými paralelně probíhajícími polysynaptickými řetězci, je považována za vhodnou pro minimalizaci interferencí a efektivní vyjádření naučeného chování. Předpokládá se, že dynamika bazálních ganglií postrádá složitost a informativnost spojenou s thalamokortikálními sítěmi souvisejícími s vědomím.[48] Je tedy pravděpodobné, že vědomé učení zahrnuje záměrné zapojení, pozornost na podněty a komplexní nervovou dynamiku, která odpovídá kognitivním procesům vyššího řádu spojeným s vědomím. A proto by mohlo být vědomé učení považováno za kritérium vědomí.

V rámci experimentů posuzujících vědomí na základě behaviorálních reakcí se využívají různé strategie k zachování co největší přesnosti měření. Například strategické řízení zahrnuje určení vědomé znalosti na základě schopnosti osoby vědomě použít nebo nepoužít informace podle instrukcí. Tzv. Jacobyho procesní disociační procedura (process-dissociation procedure/ PDP) [49] rozlišuje mezi vědomými a nevědomými

[48] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 132.

[49] PDP je metoda kognitivní psychologie vyvinutá k rozlišení vědomých a nevědomých kognitivních procesů, zejména v paměti. Účastníci se nejprve naučí seznam položek a poté plní úkoly explicitní i

znalostmi na základě toho, zda se jedinci snaží vyhnout, nebo zajistit použití informací.[50] Procesní disociační proceduru využívají též Arnaud Destrebecqz a Axel Cleeremans při sledování vztahu mezi vědomím a učením a posléze tvrdí, že učení může skutečně probíhat nevědomě, kdy získané znalosti ovlivňují chování, ale zůstávají nedostupné vědomé kontrole. Jejich postup je aplikován na studium implicitního učení, konkrétně se zaměřením na úlohy sekvenčního učení. Účastníci se zabývají úkolem zvaným úloha sériové reakční doby (SRT). V úloze SRT je účastníkům předložena sekvence vizuálních podnětů, které se objevují na různých místech počítačové obrazovky, a účastníci musí na každý podnět reagovat co nejrychlejším a nejpřesnějším stisknutím odpovídající klávesy. Důležité je, že sekvence podnětů se opakuje podle určitého vzoru, o kterém účastníci nejsou výslovně informováni. Přestože subjekty v průběhu cvičení vykazují lepší reakční časy, často nedokážou verbalizovat své znalosti základního sekvenčního vzorce, což vede výzkumníky k tomu, že učení v tomto kontextu považují za implicitní, probíhající bez vědomí.[51]

3.3. Subjektivní kritéria

3.3.1. Introspekce jako metoda posuzování vědomých stavů

Jednou z nejvyužívanějších metod při posuzování existence vědomí je metoda introspekce. Jedinec vnitřně pozoruje a hodnotí své stavy, aby je pak mohl reportovat. Rosenthal v článku *Consciousness and Confidence* ale upozorňuje, že s touto metodou je neoddělitelně spjata subjektivita jedincových reportů. Jsou zde vážná omezení subjektivních reportů při hodnocení vědomí, zejména v situacích, kdy je podnět zhoršený, nebo blízko spodní hranice počitkového prahu. V takových případech si jedinci mohou být nejistí, zda něco vnímali, což vede k nespolehlivým subjektivním zprávám o vědomí. Různé faktory mohou zkreslovat subjektivní zprávy, což dále komplikuje jejich spolehlivost. V některých experimentech se od pozorovaných osob vyžaduje, aby provedly motorické reakce (např. stisknutí tlačítka), které naznačují jejich vnímání nebo

implicitní paměti. Explicitní úkol vyžaduje vědomé vybavování nebo rozpoznávání studovaných položek, zatímco implicitní úkol nepřímo měří paměť bez vědomého vybavování. Tím, že PDP zohledňuje strategickou kontrolu účastníků a jejich výkon v obou úlohách, odhaduje příspěvek vědomých a nevědomých procesů odděleně, což nabízí vhled do složitosti lidské paměti a poznávání. (PDP viz Jacoby, L.L., A Prcess Dissociation Framework: Separating Automatic from Intentional Uses of Memory.*)*

[50] Seth, A.K., Dienes, Z., Cleeremans, A., Overgaard, M., Pessoa, L., Measuring Consciousness: Relating Behavioural and Neurophysiological Approaches, s. 317.

[51] Destrebecqz, A., Cleeremans, A., Can Sequence Learning Be Implicit? New Evidence with the Process Dissociation Procedure. s. 343–344.

reakci na podnět. Výlohy spojené s provedením těchto motorických reakcí, jako je potřebné úsilí, nebo důsledky nesprávné reakce, mohou ovlivnit subjektivní zprávy. Subjekty mohou upravit své zprávy na základě vnímané obtížnosti, nebo důsledků motorické reakce, i když jejich skutečné vnímání zůstává nezměněno. Dalším jevem, který ovlivňuje výsledky subjektivních zpráv, by byla pozornost. Pozornost hraje při vnímání zásadní roli. Když se jedinci soustředí na určitý úkol nebo podnět, může jejich zaměření pozornosti zkreslit jejich subjektivní zprávy. Pokud například někdo věnuje větší pozornost jednomu aspektu podnětu, může nadhodnocovat, nebo podhodnocovat své vnímání ostatních aspektů, což vede ke zkresleným zprávám. To vnáší do subjektivních zpráv zkreslení, které komplikuje jejich spolehlivost jako měřítek vědomého vnímání. Subjektivní zprávy navíc nemusí vždy přesně odrážet subjektivní vědomí v důsledku nedokonalého vyjadřování.[52]

Konfidenční škála je další metodou posuzování přítomnosti vědomí, o které Rosenthal mluví. Tato škála slouží k tomu, aby subjekt behaviorálně zaznamenal sílu své jistoty při tvrzení, že se zrovna nachází ve stavu M. Nicméně samotná metodika posuzování konfidence je založená na introspekci jedince a s behaviorální reakcí tedy souvisí jen do té míry, že tělesná reakce slouží jako výsledek našeho introspektivního bádání. Všimněme si, že zde tedy dochází k přechodu od oblasti behaviorální k oblasti subjektivní a jsme nuceni se spoléhat na přesnost introspekce. Rosenthal upozorňuje, že ačkoli se běžně předpokládá, že vědomé vnímání něčeho s sebou nese pocit jistoty, není tomu tak vždy. Jedním z příkladů je vědomé periferní vidění, kdy se předměty mohou v periferní části zorného pole jevit jako jasné a zřetelné, avšak při pokusu o jejich identifikaci bez přesunutí pohledu se důvěra v jejich povahu výrazně snižuje. To poukazuje na rozpor mezi subjektivním prožitkem vnímání a úrovní důvěry s ním spojenou. Existují další faktory, jako například rozptýlení, které mohou vést k nedostatečné jistotě při vědomém vnímání. Tyto příklady argumentují proti názoru, že úroveň důvěry ve vědomé vnímání nutně odráží jasnost vnímaných podnětů, a zdůrazňuje potřebu rozlišovat mezi vědomým vnímáním a úrovní konfidence. Stanovení přímé souvislosti mezi vědomím a konfidencí představuje výzvu. Introspekce, která se běžně používá k měření subjektivních prožitků, nemusí vzhledem ke své subjektivní povaze spolehlivě určit korelaci mezi vědomím a konfidencí. Introspekce se navíc vztahuje pouze na vědomé vnímání, což omezuje její užitečnost pro pochopení nevědomých případů. I kdyby byla introspekce spolehlivá, bylo

[52] Rosenthal, D., Consciousness and Confidence, s. 258–259.

by použití introspektivní konfidence jako ukazatele vědomí nadbytečné. Je tomu tak proto, že introspektivní úsudky o vědomí vjemů již zahrnují vědomí jedince o jeho vědomých zkušenostech, takže další spoléhání na introspektivní důvěru je pro posouzení vědomí zbytečné.[53]

Další metodou kromě konfidenční škály, která je využívána při posuzování vědomých stavů v rámci introspekce, je hádání s nucenou volbou (forced-choice). Využívá se zejména proto, aby se zjistilo, zda dochází k vědomému vnímání navzdory upřímnému popírání subjektů. Při této technice subjekty popírají, že by si podnět uvědomovaly, ale přesto mohou přesně odhadnout jeho povahu nad úroveň náhody. K tomuto přesnému hádání, označovanému jako čisté hádání (pure guessing), dochází, když subjekty nemají o podnětu žádnou jistotu/konfidenci. Experimenty navíc naznačují, že k hádání dochází nejen v případech naprosté absence jistoty, ale také v situacích, kdy si subjekt podnět do určité míry uvědomuje, ale není si jist povahou svého uvědomění. Toto vyvolává otázku, zda stupně naší jistoty ohledně uvědomování mohou odrážet stupně našeho subjektivního uvědomování. Přesné hádání však nemusí nutně ukazovat na vědomí, protože neexistuje přímá souvislost mezi nedostatkem důvěry a nedostatkem vědomí. Absence vědomí je primárně určena subjektivními zprávami o neuvědomování si. Navíc, i když úroveň jistoty při hádání může odrážet stupně uvědomění, nemusí spolehlivě rozlišovat mezi vědomými a nevědomými stavy. Ačkoli by hodnocení důvěry mohlo poskytnout dodatečný vhled do stupňů subjektivního vědomí, mělo by se pro přesné posouzení vědomí používat ve spojení se subjektivními zprávami, a nikdy ne samostatně.[54]

Studium subjektivní zkušenosti, zejména v oblasti výzkumu vědomí, představuje pro vědce jedinečnou výzvu. Subjektivní zprávy sice poskytují cenný vhled do vnímání, myšlenek a pocitů, ale často se na ně pohlíží skepticky kvůli obavám o přesnost a spolehlivost. Kristian Sandberg a Morten Overgaard ve společné práci *Using the Perceptual Awareness Scale* uvádí, že subjektivní popisy postrádají objektivní ověření „z vně“, takže je obtížné zjistit jejich přesnost. V důsledku toho v kognitivní vědě převládá názor, že je třeba upřednostňovat objektivní měřítka, jako je správnost identifikace nebo reakční doba, a minimalizovat spoléhání se na subjektivní zprávy. Používání takových objektivních měřítek ve výzkumu vědomí však přináší řadu vlastních problémů. Určit, které objektivní měřítko adekvátně vystihuje subjektivní zkušenost, není jednoduché.

[53] Rosenthal, D., Consciousness and Confidence, s. 257.
[54] Rosenthal, D., Consciousness and Confidence, s. 257–258.

Například takové objektivní měřítko jako správná identifikace nemusí nutně přímo odpovídat zkoumanému subjektivnímu prožitku. Platnost objektivních měřítek často závisí na korelaci s introspektivními pozorováními a zprávami, které ze své podstaty postrádají přesnost. Kromě toho může být proces získávání subjektivních zpráv v experimentech problematický. Požádat účastníky, aby podali zprávu o svých vědomých prožitcích, vyžaduje stanovit kritéria pro určení přítomnosti, či nepřítomnosti těchto prožitků. To přináší subjektivitu a nejednoznačnost, protože neexistuje žádné vnější kritérium, kterým by se účastníci mohli řídit při definování svých subjektivních zážitků. V důsledku toho vědci čelí obtížím při zajišťování toho, aby subjektivní zprávy byly vyčerpávající a exkluzivní.[55]

V reakci na tyto problémy byla vyvinuta škála vnímání vědomí (perceptual-awareness scale / PAS) jako nástroj, který má řešit omezení tradičních metod subjektivního podávání zpráv. PAS byla zkonstruována na základě spolupráce s účastníky a jejím cílem bylo zjistit přímou shodu mezi hlášeními a vnitřními stavy.[56] PAS, kterou původně vyvinuli účastníci studie Ramsøye a Overgaarda (2004), prokázala svou užitečnost ve srovnání s jinými škálami při hodnocení vědomí. Účastníci studie měli za úkol vypracovat škálu během vizuálního identifikačního úkolu zahrnujícího jednoduché geometrické tvary různých barev a poloh. Byli požádáni, aby sdělili, nebo odhadli vlastnosti podnětů a ohodnotili zřetelnost každé vlastnosti pomocí škály, kterou sami vytvořili. Výsledná PAS se skládala ze čtyř kategorií: „žádná zkušenost/zážitek“ (no experience), „krátký/slabý záblesk“ (brief glimpse), „téměř jasný obraz/zážitek“ (almost clear image) a „naprosto jasný obraz/zážitek“ (absolutely clear image). Argumenty podporující používání PAS spadají do dvou kategorií: její korelace s objektivním výkonem a snadnost použití. Zatímco první předpokládá, že dobrá korelace s objektivními měřítky ukazuje na citlivost PAS, druhý vychází z předpokladu, že PAS účinně zachycuje introspektivní rozdíly, které účastníci prožívají. Tyto argumenty naznačují, že PAS může být nejvhodnějším dosud dostupným měřítkem vědomí.[57]

3.3.2. Subjektivita a kvália jako kritérium vědomí

Zkoumání vědomí již dlouho vzbuzuje zvědavost filozofů, vědců i myslitelů. Jedním z ústředních bodů tohoto zkoumání je i hledání základních kritérií, která by mohla nastavit

[55] Sandberg, K., Overgaard, M., Using the Perceptual Awareness Scale (PAS), s. 181.
[56] Sandberg, K., Overgaard, M., Using the Perceptual Awareness Scale (PAS), s. 181.
[57] Sandberg, K., Overgaard, M., Using the Perceptual Awareness Scale (PAS), s.182–184.

hranici mezi procesy, které jsou pro vznik vědomí podstatné, a těmi, které nehrají při vzniku vědomí podstatnou roli. Mezi těmito diskutovanými kritérii vědomí stojí subjektivita jako základní kámen, který nabízí optiku, skrze niž vnímáme a interpretujeme svět kolem nás. Subjektivita ve své podstatě ztělesňuje bohatou konstrukci individuálních perspektiv, emocí a zkušeností, které podbarvují naši vědomou existenci. Je to subjektivní čočka, kterou každý z nás vnímá realitu a která utváří naše jedinečné chápání a interpretaci světa. Na rozdíl od objektivity, která usiluje o nestrannost a odstup od osobních předsudků, nás subjektivita pevně ukotvuje v naší vlastní vědomé realitě a propůjčuje našim zkušenostem osobní význam a smysl. Tento subjektivní aspekt vědomí, zakořeněný v každodenní zkušenosti, představuje výzvu pro komplexní vědecký popis vědomí. Pokusy o objektivní, vědecký popis těchto subjektivních scén vědomí zůstávají nejasné a situace často vede k tomu, co je známo jako explanační mezera mezi vědeckou teorií a subjektivní zkušeností, označovaná Davidem Chalmersem jako těžký problém vědomí.[58]

Descartes, který je mnohdy považován za iniciátora problému mysli a těla a předchůdce současného zkoumání vědomí, zavedl dualistický rámec, podle něhož mysl a hmota existují jako oddělené entity a vzájemně se ovlivňují v epifýze. Jak ale podotýkají autoři Adrien Doerig, Aaron Schurger a Michael Herzog[59], dualismus se potýká s problémy, které je třeba sladit se základními fyzikálními zákony, jako je zachování energie a impulsu. V důsledku toho se moderní filozofické a vědecké teorie vědomí přiklánějí k fyzikalismu a tvrdí, že existuje pouze hmota a duševní události jsou buď totožné s fyzikálními procesy, nebo jsou jim nadřazené. I přesto ale mnoho filozofů vystupuje proti myšlence, že je možné redukovat mysl na hmotu. Kvália, jako subjektivní fenomenální kvality zkušenosti, jsou hlavním argumentem pro tato tvrzení proti redukci duševních stavů na hmotu.[60] Kvália se vztahují k nevýslovným, soukromým a subjektivním aspektům prožitků, jako je vůně kávy, pocit větru na tvářích nebo pohled na oblohu při západu slunce. Existence a povaha kválií jsou předmětem mnoha filozofických debat.

Existence kválií je například podpořena „argumentem z poznání“. Argument z poznání, známý také jako Mariin pokoj, je filozofický myšlenkový experiment, který navrhl Frank Jackson v roce 1982. Scénář se točí kolem Marie, geniální vědkyně, která strávila celý

[58] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 131.
[59] Doerig, A., Schurger, A., Herzog, M. H., Hard Criteria for Empirical Theories of Consciousness, s. 41.
[60] Doerig, A., Schurger, A., Herzog, M. H., Hard Criteria for Empirical Theories of Consciousness, s. 41.

svůj život v černobílé místnosti bez oken a naučila se vše, co se týká barevného vidění. Navzdory svým rozsáhlým znalostem Marie nikdy nezažila barvy na vlastní kůži. Argument Mariina pokoje tudíž předpokládá, že v Mariiných znalostech stále něco chybí: přímá zkušenost skutečného vidění barev. Když Marie konečně opustí svůj monochromatický pokoj a poprvé zažije barvy, získá novou zkušenost[61], která jí nebyla přístupna pouze na základě předchozího studia. Tato nově nabytá zkušenost s barvami se často interpretuje jako kvália, subjektivní, kvalitativní aspekty vědomé zkušenosti, které nelze plně zachytit nebo pochopit prostřednictvím fyzikalismu.[62] Argument z poznání podporuje kritérium subjektivity tím, že zdůrazňuje subjektivní povahu vědomé zkušenosti, zejména ve vztahu ke kváliím. Pokud bychom se přiklonili k názoru, že kvália jsou skutečně neredukovatelná na fyzické procesy, plynuly by z toho dva důsledky. Zaprvé, měli bychom přítomnost kválií považovat za kritérium vědomí. Zadruhé, vystoupili bychom proti fyzikalismu a poukázali na to, že fyzikalistické explanace jsou nedostačující.

Zatímco mnozí teoretici uznávají realitu kválií na základě našich bezprostředních zkušeností, jiní, jako například Daniel Dennett, jejich existenci popírají. Dennett zpochybňuje tradiční pojetí tohoto pojmu jako nevyjádřitelných, vnitřních a soukromých „syrových pocitů“ a tvrdí, že tyto vlastnosti nejsou pro vědomou zkušenost podstatné. Zpochybňuje, zda naše prožitky mají skutečně neredukovatelné, nevyjádřitelné kvality, nebo zda je naše chápání těchto prožitků ovlivněno kognitivními procesy a interpretacemi. Navzdory výzvám a neshodám, které se týkají kválií, zůstávají i nadále ústředním bodem diskusí o vědomí. Nejednoznačnost a různorodé interpretace tohoto pojmu zdůrazňují složitost chápání subjektivní zkušenosti a jejího vztahu k fyzické realitě. Pečlivé zvážení toho, jak je termín „kvália“ používán a definován, je proto ve filosofických diskusích o vědomí a jeho kritériích klíčové.[63]

Seth, Baars a Edelman mluví o subjektivitě jako o jevu, který může vyžadovat interakce mezi zadními oblastmi kůry, jako jsou parietální egocentrické mapy a orbitofrontální kůra. Předpokládá se, že tyto interakce přispívají k utváření vlastního já, ať už základního senzomotorického, nebo narativního vyššího řádu.[64] Tím se opět ukazuje, že zkoumat

[61] Detailní analýza a důsledky jsou podrobně rozebrány ve stati Polák, M., Co ví Marie o barvách
[62] Polák, M., Co ví Marie o barvách, s.161.
[63] Blackmore, S., Troscianko, E., *Consciousness: An Introduction*, s. 35.
[64] Seth, A. K., Baars, B. J., Edelman, D. B., Criteria for Consciousness in Humans and Other Mammals, s. 131.

kritéria vědomí odděleně není jednoduché, neboť všechny tři doposud zmíněné oblasti (tj. neurální, behaviorální, subjektivní) jsou vzájemně provázány a dochází u nich k ovlivňování.

4. Kritéria a jejich využití v případě umělé inteligence

Stejně tak, jako tomu je s definováním pojmu vědomí, tak i při pokusech o definici umělé inteligence narážíme na odlišné názory a přístupy. Jedná se o pojem, ke kterému různé vědecké oblasti různě přistupují. Jedna z definic by mohla být následovná. Umělá inteligence (UI) je mnohostranný obor, jehož cílem je vytvářet systémy schopné inteligentního chování. Stuart Russel a Peter Norvig[65] mluví o tom, že spektrum cílů UI je rozdělené do dvou kategorií. Zda je kladen důraz na vyrovnání se lidskému chování, či zda je kladen důraz na vyrovnání se lidskému myšlení.

Turingův test, o kterém budu mluvit také později v kapitole o behaviorálních kritériích, byl zaměřen právě na tu první oblast, v níž se soustředíme na tvorbu systémů, které budou vykazovat inteligentní lidské chování. Test spočívá v tom, že lidský tazatel komunikuje s počítačem i s člověkem prostřednictvím písemných otázek. Pokud tazatel nedokáže spolehlivě určit, které odpovědi pocházejí od počítače a které od člověka, má se za to, že počítač testem prošel. V druhém případě hovoří Russel a Norvig o tvorbě systémů se stejnou formou myšlení, přičemž inspirace vychází z prvků kognitivních procesů u člověka. Příkladem takového systému by byl General problém solver (GPS) vyvinutý Allenem Newellem a Herbertem Simonem.[66]

4.1. Neurální kritéria a umělá inteligence

Hlavní problém při pokusu o srovnání biologických mechanismů, které jsou základem vědomí u lidí, s umělými mechanismy, používanými v umělých agentech, vyplývá ze zásadních rozdílů mezi těmito dvěma typy systémů. Tento rozdíl charakterizují autoři Raúl Arrabales, Agapito Ledezma a Araceli Sanchis.[67] Zatímco lidé se při regulaci chování a projevování vědomí spoléhají na složité biologické systémy, jako je endokrinní a nervový systém, umělí agenti fungují pomocí umělé inteligence a výpočetních systémů. Tyto umělé systémy mohou být inspirovány biologickými principy, ale jsou navrženy spíše na základě inženýrských principů k dosažení specifických výpočetních úkolů než k přesné replikaci biologických mechanismů. Biologické mechanismy se vyvíjely miliony let, aby podporovaly vědomí savců, což vedlo ke složitým neuronovým sítím, systémům

[65] Norvig P., Russell S., *Artificial Intelligence*, s. 1–2.
[66] Norvig P., Russell S., *Artificial Intelligence*, s. 2–3.
[67] Arrabales, R., Ledezma Espino, A. I., Sanchis De Miguel, M. A., Criteria for Consciousness in Artificial Intelligent Agents, nečíslováno.

neurotransmiterů a anatomickým strukturám v mozku. Tyto struktury spolu dynamicky interagují, aby daly vzniknout vědomí, které vykazuje složité chování a funkce. Naproti tomu umělé systémy, včetně umělých neuronových sítí a výpočetních algoritmů, postrádají biologickou složitost a propracovanost, kterou lze nalézt u živých organismů.[68]

Na rozdíl mezi biologickou a umělou strukturou je možné se podívat například při porovnávání perceptronu a neuronu. O rozdílech mluví Ameet Joshi v knize *Machine learning and artificial intelligence*. Biologický neuron obsahuje tři důležité části, z nichž první je tělo neuronu obsahující jádro neuronu. Z těla vedou dendrity zodpovědné za příjem vstupních informací přicházejících od jiných neuronů. Poslední významnou částí je axon, který naopak vede přicházející signály od těla neuronu, aby je mohl přes synapse poslat dalším nervovým buňkám. Neurony komunikují prostřednictvím elektrochemických procesů, při nichž dochází k neustálé výměně aktivačních signálů mezi propojenými neurony. Neuron přijme aktivační signál od jiné nervové buňky, signál je zpracován a případně poslán dál k dalším nervovým buňkám. Tento složitý proces tvoří základ neuronové komunikace a umožňuje přenos informací v celém nervovém systému.[69] V roce 1957 byl vymodelován perceptron, umělý neuron, jehož stavba byla tou biologickou inspirována a jehož využití bylo zejména při řešení lineárních problémů.[70] Podobně jako u biologického neuronu, informace přicházejí do umělého neuronu skrze vstupy, v těle neuronu jsou poté sčítány, zpracovávány a nakonec předány prostřednictvím výstupů.[71] Z důvodu omezení jednotlivých perceptronů byla zavedena vícevrstevná architektura, která dala základ vývoji umělých neuronových sítí.[72]

Neuronové sítě v biologických systémech se skládají ze vzájemně propojených neuronů, které zpracovávají a přenášejí informace v rámci celé sítě. Podobně, spojením dvou a více umělých neuronů, vzniká umělá neuronová síť (artificial neural network/ANN). V roce 2013 vznikla komparativní studie [73] zaměřená na shody a rozdíly mezi ANN a biologickými sítěmi. Mezi nalezené podobnosti systémů patří podle studie následovné: Biologické neuronové sítě i umělé neuronové sítě zpracovávají informace paralelně. Oba

[68] Arrabales, R., Ledezma Espino, A. I., Sanchis De Miguel, M. A., Criteria for Consciousness in Artificial Intelligent Agents, nečíslováno.
[69] Joshi, A. V., Machine Learning and Artificial Intelligence, s. 57.
[70] V kontextu perceptronů se lineární problémy týkají klasifikačních úloh, kde jsou třídy lineárně oddělitelné. To znamená, že existuje přímka, která může dokonale oddělit datové body různých tříd.
[71] Suzuki, K., *Artificial Neural Networks – Methodological Advances and Biomedical Applications*, s. 5.
[72] Joshi, A. V., Machine Learning and Artificial Intelligence, s. 58.
[73] Eluyode, O. S., Akomolafe, D. T., Comparative Study of Biological and Artificial Neural Networks, s. 36–46.

typy sítí se učí z minulých zkušeností, aby zlepšily svůj výkon. Biologické neuronové sítě upravují synaptická spojení, zatímco umělé neuronové sítě upravují váhy na základě vstupních dat a požadovaných výstupů. Přenos informací v obou typech sítí zahrnuje používání elektrických signálů. Jak biologické, tak umělé neuronové sítě ukládají informace. V biologických sítích jsou informace uloženy v synapsích, zatímco v umělých sítích jsou informace uloženy v matici vah.[74]

Jako rozdíly byly ve studii uvedeny následující: Biologické neuronové sítě obecně zpracovávají informace pomalu vzhledem k době, kterou neurony potřebují k reakci na podněty. Zatímco umělé neuronové sítě vykazují velmi vysokou rychlost zpracování a často dosahují přepínacích časů několika nanosekund. Mozek dokáže efektivně řešit problémy, se kterými mohou mít digitální počítače problémy. Biologické neuronové sítě jsou konstruovány trojrozměrně z mikroskopických komponent s téměř neomezenými vzájemnými vazbami. Naproti tomu umělé neuronové sítě jsou jednodušší shluky primitivních umělých neuronů, obvykle uspořádané do vrstev s různou architekturou. Biologické neuronové sítě se skládají z obrovského počtu neuronů, řádově 10^11, zatímco umělé neuronové sítě obvykle propojují menší počet neuronů, od 10^2 do 10^4. Biologické neuronové sítě vykazují větší počet spojení mezi neurony s náhodnou spojitostí, zatímco umělé neuronové sítě mají méně hran[75] a přesněji specifikovanou spojitost, často podle předem daného plánu. Biologické neuronové sítě mají tendenci při částečném poškození elegantně degradovat, zatímco umělé neuronové sítě mají potenciál odolnosti vůči poruchám a zachovávají si robustní výkon i při částečném poškození sítě.[76]

4.2. Behaviorální kritéria a umělá inteligence

V kontextu vědomí strojů se simulace lidského chování zaměřuje na replikaci vědomého lidského chování a nebere ohled na to, zda stroj sám zažívá fenomenální stavy podobné lidskému vědomí. Výzkum v této oblasti může využívat různé přístupy, jako jsou kognitivní modely nebo architektury napodobující vědomí (např. počítačový model emocí nebo představivosti), ale může využívat i jednodušší metody, jako jsou

[74] Eluyode, O. S., Akomolafe, D. T., Comparative Study of Biological and Artificial Neural Networks, s. 44–45.
[75] V umělých neuronových sítích se „hranami“ obvykle označují spojení mezi neurony. Tato spojení přenášejí vážené signály z jednoho neuronu na druhý a umožňují tok informací v celé síti.
[76] Eluyode, O. S., Akomolafe, D. T., Comparative Study of Biological and Artificial Neural Networks, s. 45.

vyhledávací tabulky nebo logická pravidla pro generování chování. Důraz spočívá spíše v reprodukci chování než ve zkoumání fenomenálních stavů.[77]

Aplikace behaviorálních přístupů na umělé agenty představuje výzvu. Předpoklady a kritéria vyvinutá pro hodnocení vědomí u biologických organismů se nemusí přímo vztahovat na umělé agenty. Lidské chování, jak podotýkají Arrabales, Ledezma a Sanchis, se často řídí kulturními normami a pravidly, která jsou formována ontogenezí a kulturním prostředím. Naproti tomu vývoj umělých agentů probíhá po jiné trajektorii, což vede k vzorcům chování, které se nemusí shodovat se vzorci chování lidí. Proto by se posuzování chování umělých agentů nemělo opírat o stejná kritéria, jaká se používají u lidí. Další rozdíl mezi chováním lidí a umělých agentů spočívá dle autorů v absenci precizních verbálních komunikačních dovedností u umělých systémů. Přesné verbální hlášení (accurate verbal report/AVR) hraje zásadní roli při pochopení vnitřních prožitků lidských subjektů. Vzhledem k tomu, že umělí agenti nemají schopnost poskytovat AVR stejně precizním způsobem jako lidé, nelze na umělé agenty přímo aplikovat tradiční metody hodnocení vědomí, které se opírají o verbální komunikaci. Toto omezení představuje výzvu při posuzování vědomí nebo vnitřního života umělých systémů, protože absence precizního verbálního hlášení nám brání v nahlédnutí do jejich subjektivních prožitků. Proto jsou ke zkoumání a pochopení vnitřního fungování umělých agentů zapotřebí alternativní přístupy.[78]

Stejně jako tomu je u behaviorálních kritérií aplikovaných na člověka, behaviorální kritéria nejsou dostatečná sama o sobě k tomu, abychom na jejich základě mohli posuzovat, či dokonce replikovat vědomí v oblasti umělé inteligence. Přesto, že se nám podaří vyvinout umělou inteligenci vykazující stejné chování, jako vykazují lidé, nemusí se jednat o podmínku dostačující k tomu, aby u nich vzniklo i samotné vědomí. Při diskuzi na téma chování spojeného s vědomím se dříve či později opět dostaneme k myšlenkovým experimentům Johna Searla a Alana Turinga.

4.2.1. Argument čínského pokoje

Čínský pokoj je myšlenkový experiment Johna Searla, jehož cílem je zpochybnit představu, že počítačový program může na základě manipulace se symboly podle pravidel

[77] Gamez, D., Progress in Machine Consciousness, s. 888–889.
[78] Arrabales, R., Ledezma Espino, A. I., Sanchis De Miguel, M. A., Criteria for Consciousness In Artificial Intelligent Agents, nečíslováno.

skutečně porozumět. V tomto experimentu Searle žádá čtenáře, aby si představili, že se nacházejí v místnosti a podle anglicky psaných instrukcí manipulují s čínskými symboly, aniž by čínštině skutečně rozuměli. V místnosti přijímá jedinec čínské symboly zvenčí, zpracovává je podle souboru pravidel uvedených v angličtině a vytváří odpovědi v čínštině. Přestože je jedinec schopen vytvářet vhodné odpovědi na vstupní údaje, čínštině skutečně nerozumí. Tento scénář má ilustrovat rozdíl mezi pouhou manipulací se symboly a skutečným porozuměním.[79] Ze Searlovy práce lze vyvodit důležité myšlenky týkající se kritérií vědomí.

Searle ve své práci tvrdí, že rozhodujícím kritériem rozumění je sémantika, nikoli pouze syntax. Zatímco syntax se vztahuje k formálním pravidlům, jimiž se řídí manipulace se symboly, sémantika se týká významu nebo interpretace těchto symbolů. V argumentu čínského pokoje je jedinec uvnitř místnosti schopen řídit se syntaktickými pravidly pro manipulaci se symboly, aniž by chápal sémantický obsah těchto symbolů. Searle zmiňuje, že: „*Jde-li opravdu o počítač, pak jsou jeho operace definovány syntakticky, zatímco vědomí, myšlenky, pocity, city a všechno, co k tomu patří, zahrnují víc než syntax*.“[80] Tento důraz na sémantiku podtrhuje význam subjektivní zkušenosti a kválií ve vědomí. Searle tvrdí, že vědomí zahrnuje nejen schopnost syntaktického zpracování informací, ale také subjektivní zkušenost a kvalitativní aspekty spojené s pochopením významu těchto informací.[81] Tímto jeho myšlenkový experiment podporuje myšlenku, že subjektivita a kvália jsou základními kritérii vědomí, a zpochybňuje redukcionistické přístupy, které se zaměřují pouze na syntaktické procesy.

Zadruhé, experiment Johna Searla s čínským pokojem zpochybňuje základní principy a kritéria behaviorismu tím, že poukazuje na omezení spjaté s hodnocením přítomnosti vědomí pouze na základě pozorovatelného chování. Searlův experiment ukazuje, že schopnost produkovat reakce nebo chování nemusí nutně znamenat skutečné porozumění nebo vědomí. Ve scénáři čínského pokoje jedinec, který se řídí pokyny k manipulaci se symboly, prokazuje schopnost vytvářet příslušné reakce, aniž by ve skutečnosti chápal význam těchto symbolů. Tento rozpor mezi vnějším chováním a vnitřními duševními

[79] Searle, J. R., *Mysl, mozek a věda*, s. 34.
[80] Searle, J. R., *Mysl, mozek a věda*, s. 38.
[81] Searle, J. R., *Mysl, mozek a věda*, s. 34–35.

stavy podkopává behavioristickou představu, že vědomí lze odvodit na základě pozorovatelných činností.[82]

4.2.2. Turingův test

Strojový experiment Alana Turinga, známý také jako Turingův test, je zásadním pojmem v oblasti umělé inteligence a teorie výpočtů. Tento experiment, který Turing navrhl ve svém zásadním článku *Computing Machinery and Intelligence* v roce 1950, slouží jako měřítko schopnosti stroje vykazovat inteligentní chování nerozeznatelné od chování člověka. Test také naznačuje, že behaviorální kritéria jsou dostačující pro stanovení přítomnosti vědomí. Ve své práci ho popisuje Michael Wooldridge.[83] Podstata Turingova testu spočívá ve scénáři, kdy lidský hodnotitel komunikuje se dvěma entitami prostřednictvím textové komunikace bez přímého smyslového vnímání. Jednou entitou je člověk, zatímco druhou je stroj navržený tak, aby simuloval lidskou konverzaci. Úkolem hodnotitele je určit, která entita je člověk a která stroj, a to pouze na základě odpovědí, které obdrží prostřednictvím terminálu, například obrazovky od počítače. Turing navrhl, že pokud jsou odpovědi stroje nerozeznatelné od odpovědí člověka do té míry, že hodnotitel nedokáže spolehlivě rozlišit mezi oběma entitami, pak lze stroj považovat za subjekt, který prošel Turingovým testem a prokázal inteligenci srovnatelnou s inteligencí člověka. Tento koncept se často vyjadřuje jako stroj vykazující „silnou UI“ nebo schopný „myslet“. Pokud totiž tvrdíme o dvou entitách (v našem případě inteligenci lidí a inteligenci umělé), že jsou odlišné, ale na základě testů je nejsme schopni rozlišit, pak musíme přiznat, že jsou stejné.[84] Lze říci, že Turingův test poskytuje rámec pro hodnocení inteligence na základě pozorovatelného chování a podporuje behavioristické kritérium, které zdůrazňuje vnější chování jako hlavní ukazatel duševních stavů u lidí i umělé inteligence.

Kritika Turingova testu se zaměřuje především na jeho schopnost definitivně posoudit inteligenci a vědomí strojů. Kritici tvrdí, že úspěšné absolvování Turingova testu nemusí nutně znamenat skutečné porozumění nebo vědomí strojů, ale spíše jejich schopnost povrchně napodobovat lidské jazykové chování. Tato kritika zdůrazňuje omezení

[82] Searle, J. R., *Mysl, mozek a věda*, s. 39.
[83] Wooldridge, M., *The Road to Conscious Machines: The Story of AI.*
[84] Wooldridge, M., *The Road to Conscious Machines: The Story of AI*, s. 29–30.

Turingova testu při hodnocení hloubky inteligence a vědomí strojů přesahující jazykové schopnosti.[85]

Souhrnně řečeno, výzkum vnějšího chování v oblasti vědomí strojů zahrnuje i replikaci chování spojeného s vědomím, ale to nemusí nutně znamenat, že bude dosaženo prožívání fenomenálních stavů u umělých entit. Toto rozlišení je klíčové pro pochopení složitosti umělé inteligence a vědomí. Snahy v této oblasti, jako je absolvování Turingova testu a vývoj obecné umělé inteligence, se zaměřují na vytváření strojů schopných napodobovat vědomé lidské chování. Je však nezbytné si uvědomit, že replikovat chování se nerovná replikovat vědomí. Vědomí zahrnuje subjektivní prožitky, emoce a pocit sebeuvědomění, tedy prvky, které přesahují pouhé napodobování chování. Vzhledem k tomu, co bylo doposud řečeno, by mohlo být vyvozeno, že behaviorální kritéria sama o sobě nejsou spolehlivým ukazatelem přítomnosti vědomí. Hlavní nevýhodou je totiž právě fakt, že pouhé napodobování vědomého chování neznamená prokazování skutečného vědomí. Nicméně i přes to je chování tím, podle čeho lidé nejčastěji soudí přítomnost vědomí u druhých. V běžném životě interagujeme s lidmi a na základě jejich verbálního a neverbálního jednání soudíme, zda jsou vědomí (neklademe si při každé interakci otázku, zda u druhých nedochází k pouhému napodobování vědomého chování.). A tedy i přes značné nedostatky jsou behaviorální kritéria v běžném životě tou nejlépe pozorovatelnou složkou lidského vědomí.

4.3. Subjektivní kritéria a umělá inteligence

Koncept umělého vědomí je předmětem velkých diskusí, které zahrnují různé filozofické, vědecké a technologické perspektivy. O tématu vědomí u umělých agentů mluví Gunter Meissner[86] a definuje jej v rámci uvědomování si vlastní existence, včetně chápání vlastních činů a jejich příčin. Lze rozlišit různé stupně vědomí sebe sama, od primitivního vědomí o sobě samém, které zahrnuje základní vědomí vlastní existence, až po reflexivní vědomí, které zahrnuje schopnost analyzovat a uvažovat o vlastní existenci a existenci druhých. V zoologii a etologii se k posouzení přítomnosti sebeuvědomění u zvířat používá zrcadlový test. Několik druhů, včetně lidoopů, asijských slonů, delfínů skákavých, kosatek a strak, prokázalo známky vědomí vlastní existence tím, že

[85] Oppy, G., Dowe, D., The Turing Test, nečíslováno.
[86] Meissner, G., Artificial Intelligence: Consciousness and Conscience.

rozpoznávalo znaky na svém těle, které se odrážely v zrcadle. To naznačuje, že některá zvířata, která nejsou lidmi, mají určitou formu sebeuvědomění.[87]

Možnost umělého sebeuvědomění vyvolává otázky, zda jej lze naprogramovat do strojů. Někteří, například David Chalmers[88], podporují názor, že vědomí sebe sama může být dosažitelné prostřednictvím výpočetních procesů, jiní, např. Roger Penrose[89], se přiklání k myšlence, že vědomí, zejména pak (kromě vědomí sebe sama) fenomenální vědomí zahrnující subjektivní zkušenosti, jako jsou pocity a vjemy, může být jedinečné pro vnímající bytosti s biologickými neurony a smysly. Historicky se v mechanistickém pojetí navrhovalo, že mysl funguje jako složitý stroj, což podporovalo myšlenku umělého vědomí. Současné perspektivy, jako je například komputacionalismus, který nahlíží na mozek jako na počítač, jsou však mezi výzkumníky umělé inteligence široce diskutovány. Dosažení pokročilých stupňů vědomí, jako je reflexivní vědomí zahrnující metakognici a vůli, představuje pro umělé systémy značnou výzvu.[90]

[87] Meissner, G., Artificial Intelligence: Consciousness and Conscience, s. 230.
[88] Chalmers, D., *The Conscious Mind: In Search of a Fundamental Theory.*
[89] Penrose, R., The Emperor's New Mind.
[90] Meissner, G., Artificial Intelligence: Consciousness and Conscience, s. 230–231.

Závěr

Analýza kritérií vědomí, ať již u člověka, či u umělých agentů, představuje nelehký úkol. Tato práce zdaleka nepostihla všechny navrhované možnosti kritérií, neboť se jedná o rozsáhlou oblast plnou stále probíhajících debat. Nicméně byla vybrána a porovnána základní navrhovaná kritéria z oblasti neurální, behaviorální a subjektivní.

V první části práce byla analyzována navrhovaná kritéria vědomí u člověka. Dle mého názoru v oblasti každodenního života závisí určování vědomí převážně na pozorovatelných behaviorálních kritériích. I přes to, že vhodnějším kandidátem na objektivní posouzení přítomnosti vědomí by mohla být právě kritéria neurální (neboť mají možnost nahlížet do fyziologického substrátu mozku a jeho procesů), lidé ve všedních situacích nemají šanci se jimi řídit. Jak bylo naznačeno v předešlých kapitolách, stanovení neurálních kritérií vyžaduje například EEG a jiné metody umožňující sledovat neurofyziologické aktivity v mozku. Nicméně to je oblast, ve které se pohybují pouze specialisté. Laici ve svých každodenních životech soudí vědomí ostatních nejčastěji podle toho, jak jednají behaviorálně. Navzdory své přístupnosti jsou behaviorální kritéria zatížena subjektivitou a omezeními. I když chování může napodobovat projevy vědomí, nemusí prokazovat skutečnou existenci prvků vědomí, což představuje značnou nevýhodu. Nicméně díky své praktičnosti a bezprostřednosti přetrvávají behaviorální kritéria jako preferovaná metrika pro hodnocení vědomí v každodenních situacích.

V druhé části jsem se jednotlivé oblasti kritérií v analogickém pořadí (neurální, behaviorální, subjektivní) pokusila aplikovat na oblast umělé inteligence. V rámci aplikace kritérií vědomí člověka na oblast umělé inteligence se setkáváme hned s několika problémy. V případě neurálních kritérií spočívá hlavní překážka v rozdílu biologického a umělého substrátu. Lidská těla se evolučně vyvíjela výrazně déle než umělé systémy. Umělá inteligence tedy není schopna dosáhnout takové strukturní komplexnosti jako člověk, ačkoli jsou umělé systémy tvořeny tak, aby strukturně a funkčně napodobovaly biologické systémy (neurony, neuronové sítě). Posuzovat přítomnost vědomí pouze na základě behaviorálních kritérií se ukázalo být stejně problematické jak u umělých, tak u biologických systémů. U člověka v rámci behaviorálních projevů dochází k velkému ovlivňování subjektivními vnitřními procesy a také neurofyziologickou stavbou. U umělých systémů se pak při replikaci lidského chování ani nebere ohled na to, zda systém zažívá vědomé stavy. Oblast subjektivních

kritérií zůstává otázkou. Subjektivní prožívání ze své podstaty nemůže být nahlíženo nikým jiným, a o to více bude problematické, pokud jej budeme chtít přenést do umělých systémů.

Naše chápání vědomí vychází z principů lidského vědomí. Proto se při hodnocení vědomí u entit mimo člověka náš přístup opírá pouze o pozorování lidského vědomí. Tato metodika však představuje potenciální problém, protože zkresluje naše perspektivy a může způsobit, že nebudeme brát v úvahu alternativní projevy vědomí v umělé inteligenci. Je zcela pravděpodobné, že vědomí v UI se od lidského vědomí výrazně liší. Podobnou myšlenku formulují i Arrabales, Ledezama a Sanchis[91] a zmiňují, že vědomí produkované v UI by mělo zřejmě jinou povahu než to lidské. Měli bychom jej tedy pravděpodobně i terminologicky odlišit, např. „umělé vědomí“, a zvážit, zda nebude potřeba nastavit pro umělé vědomí kritéria vlastní, nezávislá na těch používaných pro lidi.[92]

Domnívám se, že s hledáním kritérií vědomí ale souvisí také problematika etiky. Pokud by umělá inteligence vykazovala známky vědomí, vyvstávají otázky týkající se jejího morálního postavení a odpovědnosti, kterou mají lidé vůči těmto vědomým entitám. Měli bychom se k vědomým systémům UI chovat jako k morálním subjektům s podobnými právy a povinnostmi, jako mají lidé? Určení etického přístupu k vědomé UI by bylo klíčové. Pokud by UI byly uznány za vědomé bytosti, mohly by mít nárok na určitá práva a ochranu. Zajištění etického zacházení s vědomými UI a jejich blahobytu by vyžadovalo vytvoření právních rámců, které by chránily jejich práva a zabránily jakékoli formě zneužívání nebo diskriminace.

Souhrnně řečeno, uvažování nad vědomím u UI s sebou nese mnoho dalších problémů a otázek. Ve chvíli, kdy se nám podaří stanovit kritéria pro existenci vědomí umělých agentů, budeme muset začít řešit otázky etické, sociální, právní atd. Umělá inteligence a její rychlý pokrok ovšem otevírá možnosti právě těmto hlubokým otázkám. Hledat a nastavovat kritéria vědomí UI je nelehký úkol (o to více, jestliže ani u kritérií vědomí člověka nepanuje konsenzus), ale je to úkol důležitý a nevyhnutelný, vzhledem k vývoji UI a jejímu zařazování do společnosti.

[91] Arrabales, R.; Ledezma E. A. I.; Sanchis De Miguel, M. A., Criteria for Consciousness in Artificial Intelligent Agents, nečíslováno.
[92] Arrabales, R.; Ledezma E. A. I.; Sanchis De Miguel, M. A., Criteria for Consciousness in Artificial Intelligent Agents, nečíslováno.

Zdroje:

ARRABALES, Raúl; LEDEZMA ESPINO, Agapito Ismael a SANCHIS DE MIGUEL, María Araceli. Criteria for Consciousness in Artificial Intelligent Agents. In: *Workshop Adaptive Learning Agents and Multi-Agent Systems*, 2008, s. 57–64.

COOPER, Steven J. Donald O. Hebb's Synapse and Learning Rule: a History and Commentary. *Neuroscience & Biobehavioral Reviews*, vol. 28 (2005), no. 8, s. 851–874.

BLACKMORE, Susan a T. TROSCIANKO, Emily. *Consciousness: An Introduction*. United Kingdom: Routledge, 2018.

DESTREBECQZ, Arnaud a CLEEREMANS, Axel. Can Sequence Learning Be Implicit? New evidence with the process dissociation procedure. *Psychonomic bulletin & review*, vol.8 (2001), s. 343–350.

DOERIG, Adrien; SCHURGER, Aaron a HERZOG, Michael H. Hard Criteria for Empirical Theories of Consciousness. *Cognitive neuroscience*, vol. 12 (2021), no. 2, s. 41–62.

ELUYODE, O. S. a AKOMOLAFE, Dipo Theophilus. Comparative Study of Biological and Artificial Neural Networks. Euro*pean Journal of Applied Engineering and Scientific Research,* vol. 2 (2013), no. 1, s. 36–46.

GAMEZ, David. Progress in Machine Consciousness. *Consciousness and cognition*, vol. 17 (2008), no. 3, s. 887–910.

GAMEZ, David. The Measurement of Consciousness: a Framework for the Scientific Study of Consciousness. *Frontiers in Psychology,* vol. 5 (2014), s. 1–15.

HARLEY, Trevor. A. *The Science of Consciousness: Waking, Sleeping and Dreaming.* Cambridge: Cambridge University Press, 2021.

HAVLÍK, Marek. Vědomí a úrovně vědomí. Dva rozdílné teoretické přístupy. *Akta Fakulty filozofické Západočeské univerzity v Plzni*, vol. 4 (2012), s. 186–208.

CHALMERS, David J. *The Conscious Mind: In Search of a Fundamental Theory*. Santa Cruz: University of California, 1997.

JOSHI, Ameet V. Perceptron and Neural Networks. In: *Machine learning and artificial intelligence*, Cham: Springer International Publishing, 2022, s. 57–72.

MARVAN, Tomáš a POLÁK, Michal. *Vědomí a jeho teorie*. Praha: v nakladatelství Vyšehrad vydala Tiskárna Bílý slon s.r.o., 2015.

MEISSNER, Gunter. Artificial Intelligence: Consciousness and Conscience. *AI & SOCIETY*, vol. 35 (2020), s. 225–235.

NAGEL, Thomas. *What Is It Like to Be a Bat?* Online. The Philosophical Review, vol. 83 (1974), no. 4, s. 435–50. Dostupné z: JSTOR, https://doi.org/10.2307/2183914. [citováno 2024-02-25].

OPPY, Graham a DOWE, David. *The Turing Test*. Online. The Stanford Encyclopedia of Philosophy, 2021. Dostupné z: https://plato.stanford.edu/archives/win2021/entries/turing-test/. [citováno 2024-03-12]

PENROSE, Roger. The Emperor's New Mind. *RSA Journal*, vol. 139 (1991), no. 5420, s. 506–14.

POLÁK, Michal. Co ví Marie o barvách. In: *Kognice 2009*. Hradec Králové: GAUDEAMUS, 2009, s. 160–175.

ROSENTHAL, David. Consciousness and Confidence. *Neuropsychologia*, vol. 128 (2019), s. 255–265.

RUSSELL, Stuart J. a NORVIG, Peter. *Artificial Intelligence: a Modern Approach*. New Jersey: Pearson, 2010.

SANDBERG, Kristian a OVERGAARD, Morten. Using the perceptual awareness scale (PAS). *Behavioral Methods in Consciousness Research,* (2015), s. 181–196.

SEARLE, John R. *Mysl, mozek a věda*. 1. vyd. Praha: Mladá fronta, 1994.

SETH, Anil K.; BAARS, Bernard J. a EDELMAN, David B. Criteria for Consciousness in Humans and Other Mammals. *Consciousness and cognition,* vol. 14 (2005), no. 1, s. 119–139.

SETH, Anil K.; DIENES, Zoltán; CLEEREMANS, Axel; OVERGAARD, Morten a PESSOA, Luiz. Measuring Consciousness: Relating Behavioural and Neurophysiological Approaches. *Trends in cognitive sciences*, vol. 12 (2008), no. 8, s. 314–321.

STERIADE, Mircea; MCCORMICK, David A. a SEJNOWSKI, Terrence J. Thalamocortical Oscillations in the Sleeping and Aroused Brain. *Science*, vol. 262 (1993), no. 5134, s. 679–685.

SUZUKI, Kenji (ed.). *Artificial Neural Networks: Methodological Advances and Biomedical Applications*. Croatia: InTech, 2011.

TONONI, Giulio; BOLY, Melanie; MASSIMINI, Marcello a KOCH, Christof. Integrated Information Theory: from Consciousness to Its Physical Substrate. *Nature Reviews Neuroscience*, vol. 17 (2016), no. 7, s. 450–461.

VAN GAAL, Simon a LAMME, Victor A.F. Unconscious High-level Information Processing: Implication For Neurobiological Theories of Consciousness. *The neuroscientist*, vol. 18 (2012), no. 3, s. 287–301.

VAN GULICK, Robert. *Consciousness*. Online. The Stanford Encyclopedia of Philosophy, 2018. Dostupné z: https://plato.stanford.edu/archives/win2022/entries/consciousness/. [citováno 2024-03-10]

WOOLDRIDGE, Michael. *The road to conscious machines: The Story of AI*. Penguin: United Kingdom, 2020.

WU, Wayne. *The Neuroscience of Consciousness*. Online. The Stanford Encyclopedia of Philosophy, 2018. Dostupné z: https://plato.stanford.edu/archives/win2018/entries/consciousness-neuroscience/. [citováno 2024-02-25]

ZEMAN, Adam. *Consciousness*. Online. Brain: A Journal of Neurology, vol. 124 (2001), no.7, s. 1263–1289. Dostupné z: https://doi.org/10.1093/brain/124.7.1263. [citováno 2024-02-25]

Resumé

The topic of this bachelor thesis is the analysis of consciousness criteria in humans and artificial intelligence. The first part of the thesis deals with the characteristics of four main theories of consciousness: global workspace theory, integrated information theory, recurrent processing theory and higher order theory.

In the second part I move on to an analysis of the criteria used to judge the presence of consciousness in humans. The criteria are divided into three areas: neural criteria, behavioral criteria, and subjective criteria.

The last section addresses the criteria of consciousness in artificial intelligence. Analogously, the work browses each criteria area (neural, behavioral, subjective) and analyse them.

The text concludes that in humans, consciousness is often judged based on observable behavioral criteria, despite the potential suitability of neural criteria, which observe changes in physical substrate of brain. However, behavioral criteria are subjective and may not truly indicate consciousness. When applying the criteria to AI challenges arise. Neural criteria face obstacles due to differences in biological and artificial substrates, while behavioral criteria may not account for the internal processes necessary for consciousness. Subjective criteria, inherent to human consciousness, present further complications in AI. The text proposes distinguishing "artificial consciousness" and developing criteria specific to AI rather than relying solely on those for humans.

The most used sources for this thesis include: *Consciousness and Its Theory* by Tomas Marvan and Michal Polák, *Criteria for Consciousness in Humans and Other Mammals* by Anil Seth, Bernard Baars and David Edelman, *Criteria for Consciousness in Artificial Intelligent Agents* by Raúl Arrabales, Agapito Ledezma and Araceli Sanchis.