FAKULTA FILOZOFICKÁ ZÁPADOČESKÉ UNIVERZITY V PLZNI **Katedra filozofie**	**PROTOKOL O HODNOCENÍ PRÁCE**

Práce (co se nehodí, škrtněte): bakalářská

Posudek (co se nehodí, škrtněte): oponentky

Práci hodnotil(a) (u externích hodnotitelů uveďte též adresu a funkci ve firmě): Mgr. Stefanie Dach, Ph.D.

Práci předložil(a): Denisa Turečková

Název práce: Kritéria vědomí u člověka a umělé inteligence

---

1. CÍL PRÁCE (uveďte, do jaké míry byl naplněn):

Cílem práce je podle autorky „představení a analýza kritérií, která vymezují vědomí u lidí, a možnost jejich aplikace v oblasti umělé inteligence.“ Cíl práce považuji vcelku za splněný. První část cíle týkající se kritérií vědomí u člověka je ovšem mnohem přesvědčivěji splněná než druhá část, která se týká kritérií vědomí u umělé inteligence.

2. OBSAHOVÉ ZPRACOVÁNÍ (náročnost, tvůrčí přístup, proporcionalita teoretické a vlastní práce, vhodnost příloh apod.):

Autorka zpracovala složité, aktuální a pro bakalářskou práci ambiciózní téma, které vyžaduje porozumění komplexní filosofické i empirické literatuře. Jak podotýká autorka v závěru, otázka, jestli mohou být uměle inteligentní systémy vědomé, je důležitá z hlediska našich etických povinností vůči takovým systémům. S narůstající komplexností a rozšířením těchto systémů bude tato otázka čím dál relevantnější a její konceptuální reflexe čím dál důležitější.

První dvě třetiny práce jsou na bakalářskou úroveň velmi zdařilé. Zde autorka pojednává o teoriích vědomí u člověka a o kritériích přítomnosti lidského vědomí (neurálních, behaviorálních, subjektivních). Tyto otázky autorka zpracovává obsáhle, s důrazem na empirické poznatky. Poslední třetina práce, kde se autorka snaží aplikovat získaná kritéria na umělé systémy, je zčásti méně zdařilá. Alespoň u subjektivních a neurálních kritérií chybí skutečná komparace s lidským vědomím, kterou autorka avizovala v úvodu. Autorčiny úvahy zde často zůstávají v zárodku a není jasná jejich relevance pro řešenou otázku.

Práce vykazuje některé konceptuální problémy. Předpokládám ale, že by je autorka zvládla odstranit v budoucnosti s rostoucími zkušenostmi se zpracováním takto obsáhlých témat. Z textu je patrné, že autorka má předpoklady pro intenzivnější odbornou práci, například v návaznosti na tuto závěrečnou práci. Právě z toho důvodu stojí za to zmíněným problémům věnovat pozornost. Dovoluji si u nich proto zastavit déle (na obhajobě není potřeba následující odstavce přečíst podrobně, jedná se spíš o zpětnou vazbu pro autorku).

Autorka se před vlastním zkoumáním kritérií vědomí snaží o vymezení pojmu vědomí. To je v mých očích správný postup. Její vymezení nakonec ale zůstává pro teoretické účely příliš vágní. Navíc se autorka v následujícím textu na svou definici už neodvolává. Ve výkladu dále kolísává mezi různými pojetími vědomí (fenomenální vědomí x sebe-uvědomění x bdělost). Důvod pro to může být, že literatura, kterou autorka zpracovala, pojednává o vědomí v různém smyslu. Tím více by bylo žádoucí, aby práce explicitněji reflektovala tuto nejednoznačnost pojmu vědomí.

Dále není jasné, jak druhá kapitola práce, kde autorka představuje různé současné teorie vědomí, souvisí s hlavní otázkou práce. Autorka sice tvrdí, že nám poskytují základ, na kterém lze postavit kritéria pro vědomí (str. 10), ale v práci pak dále není specifikováno, jak kritéria pro vědomí plynou nebo jinak souvisejí s těmito teoriemi. Text ve mně vzbudil očekávání, že autorka bude chtít těžit z těchto teorií například pro nějaký druh funkcionalistického uchopení vědomí, které by se pak mohlo aplikovat na umělé systémy. Autorka však teorie vědomí dále nijak nevyužije. Následující kapitola o neurálních kritériích už nenavazuje zřetelně na tuto druhou kapitolu.

Autorka zpracovává přesvědčující množství relevantní filosofické a empirické literatury. Na bakalářskou úrovni se ji daří úctyhodně uchopit velmi komplexní otázky a experimentální výsledky. Na některých místech by ale mohlo být víc reflektováno, do jaké míry je daná literatura relevantní pro otázku práce. Například kapitola 3.2.1. pojednává o verbálních zpětných vazbách jako o kritériu vědomí. Z autorčina výkladu relevantní literatury ale vyplývá, že tyto verbální zpětné vazby jsou relevantní pro posouzení přítomnosti určitých *obsahů* vědomí nebo jejich měření, ne pro přítomnost vědomí jako takového. Chybí vysvětlení, proč by mohlo být důležité přesně měřit přítomnost konkrétních obsahů vědomí, když nám jde o „pouhou“ otázku přítomnosti vědomí. Podobně zůstává v poslední části práce neujasněné, proč jsou Turingův test a Searlův myšlenkový experiment Čínského pokoje relevantní pro hlavní otázku práce. Oba myšlenkového experimenty se primárně soustředí na otázku přítomnosti jiných jevů než vědomí (sémantický obsah, myšlení, inteligence). Mohly by být přesto relevantní, ale v práci není vysvětleno jak.

V poslední třetině práce se autorka snaží kritéria vědomí aplikovat na umělé systémy. Nejlépe zpracovaná jsou zde behaviorální kritéria, kde autorka přichází se zajímavou myšlenkou, že tato kritéria jsou nejméně vypovídající o přítomnosti vědomí, když je aplikujeme na umělé systémy, ale naopak jsou pro naše každodenní posouzení přítomnosti vědomí u lidí nejvíce relevantní. Práce sice obsahuje krátkou kapitolu o subjektivních kritériích vědomí u umělých systémů, autorka tam ale o subjektivních kritériích v podstatě nepojednává.

Hodnocení aplikovatelnosti neurálních kritérií na umělé neuronové sítě se omezuje na popis fungování „umělého neuronu“ a konstatování, že tyto jednotky se standardně slučují do komplexnějších sítí. Pro hodnocení aplikovatelnosti neurálních kritérií z první části práce na neuronové sítě to ale nestačí. Neurální kritéria u člověka také netěží pouze z informací o stavbě jednotlivých neuronů a konstatování, že vytvářejí síť. Skutečně aplikovat neurální kritéria na umělé systémy bychom začali v momentě, kdy se například ptáme, jestli v umělých systémech najdeme něco podobného jako neurální koreláty vědomí v lidském mozku. Ale to se musíme už ptát na vzorce chování větších skupin „neuronů.“

Navíc v této části vychází autorka v klíčových místech z literatury o umělé inteligenci, která je 10 let stará, například když píše o velikosti umělých neuronových sítích. V oboru, který se vyvíjí takto rychle, to není optimální. Dále píše, že umělé systémy nejsou schopny podat stejně přesné verbální zprávy o „své zkušenosti“ jako člověk. Vzhledem k vývoji současných velkých jazykových modelů se mi zdá ale riskantní tvrdit, že umělé systémy nejsou schopny produkovat podobně přesné jazykové výstupy jako člověk.

3. FORMÁLNÍ ÚPRAVA (jazykový projev, správnost citace a odkazů na literaturu, grafická úprava, přehlednost členění kapitol, kvalita tabulek, grafů a příloh apod.):

Grafická úprava je pečlivá. Jazykový projev práce je většinou na adekvátní úrovni, občas se vyskytují neobratnosti (např. „tři základní charakteristiky, které vědomí obsahuje“ str. 12, „lidé ve všedních situacích nemají šanci se jimi řídit“ str. 37). Autorka používá některé termíny, které nejsou všeobecně známé, bez vysvětlení (např. těžký problém vědomí, str. 15, paralelní zpracování informací, str. 30). Pro anglické „belief“ bych v kontextu tématu práce spíše použila české „přesvědčení“ než „víra“ (str. 4). Práce je členěná logicky, citační aparát a bibliografie jsou v pořádku.

4. STRUČNÝ KOMENTÁŘ HODNOTITELE (celkový dojem z práce, silné a slabé stránky, originalita myšlenek apod.):

Celkově jde na bakalářské úrovni o zdařilou práci, která vykazuje některé konceptuální problémy. Vzhledem k cíli práce je škoda, že poslední třetina práce zaostává za prvními dvěma. Na druhé straně jsou tyto problémy vyváženy kvalitním zpracováním prvních kapitol práce. Také se mi tyto problémy zdají na bakalářské úrovni akceptovatelné.

5. OTÁZKY A PŘIPOMÍNKY DOPORUČENÉ K BLIŽŠÍMU VYSVĚTLENÍ PŘI OBHAJOBĚ (jedna až tři):

1. Autorka pojednává obecně o kritériích vědomí. Možná má ale smysl taková kritéria rozlišit na nutná a dostačující. Nutné kritérium by bylo něco, co musí být přítomné v každém (biologickém či umělém) systému, který je vědomý, ale samo nezaručuje, že systém je vědomý. Dostačující kritérium by bylo něco, co zaručuje, že systém je vědomý, zároveň ale nemusí být nutně přítomné (mohly by existovat jiná dostačující kritéria). Jsou některá z neurálních, behaviorálních a subjektivních kritérií, o kterých pojednáte, spíše nutná nebo spíše dostačující kritéria vědomí?

2. Vysvětlete prosím, jak se Searlův myšlenkový experiment Čínského pokoje a Turingův test vztahují k otázce, jaká jsou kritéria vědomí v umělých systémech.

3. Reflektujte prosím svou diskusi kritérií vědomí v umělých systémech ve světle rychlého současného rozvoje umělé inteligence (např. velkých jazykových modelů).

6. NAVRHOVANÁ ZNÁMKA (výborně, velmi dobře, dobře, nevyhověl):

při přesvědčivé obhajobě **výborně**

Datum: Podpis: