UNIVERZITA PARDUBICE
FAKULTA FILOZOFICKÁ

Katedra filosofie a religionistiky

Posudek oponenta bakalářské práce

Jméno studenta: David Walter

Název práce: Existenční hrozby obecné umělé inteligence řídící se vlastními cíli

Jméno vedoucího práce: Mgr. Ondřej Krása, Ph.D.

Jméno oponenta: doc. Filip Grygar, Ph.D.

I. Formální kritéria	**ano**	**zčásti**	**ne**
1. **Naplnění celkového záměru:** - práce odpovídá svému zadání a obsahu - vymezení tématu je přiměřené a není příliš široké	X		
2. **Metodické a systematické zvládnutí práce** **-** text je strukturně a logicky uspořádán **-** práce je adekvátně zpracována	X	X	
3. **Úvod a závěr** - úvod řádně seznamuje s cílem a postupem práce, s dosavadním bádáním na dané téma a stěžejní literaturou - závěr náležitě shrnuje a hodnotí celou problematiku, formuluje přínos k řešení problému a vlastní stanoviska	X		
4. Autor náležitě vypracoval všechny **formální požadavky kladené na bakalářskou práci** - zadání, prohlášení, obsah, úvod, pojednání, závěr, resumé v češtině a cizím jazyce - odpovídající počet stran	X		

Slovní ohodnocení

- Bakalářská práce Davida Waltera představuje na naší katedře první originální pokus o filosofické zpracování problematiky umělé inteligenci. Práce odpovídá zadanému tématu jak po obsahové, tak strukturální stránce. Téma bylo dobře vymezeno a autor se mu věnuje se zaujetím i důsledností. Práce má jasně strukturovaný úvod, přehlednou osnovu a smysluplný závěr, který shrnuje hlavní přínosy. Formální náležitosti (rozsah, jazykové mutace, citační aparát apod.) byly splněny.

- K některým problémům metodického a systematického charakteru se vyjadřuji níže v rámci podrobnějšího hodnocení.

II. Obsahová kritéria	**ano**	**zčásti**	**ne**
1. Práce prokazuje **porozumění** autora v dané problematice.	X		
2. **Myšlenková soudržnost a souvislost textu:** - je vyhovující - srozumitelnost argumentace a podloženost argumentů či myšlenek je adekvátní	X	X	
3. Práce se opírá o **relevantní prameny a literaturu**.	X		
4. **Práce s literaturou**: - v práci je zřetelně předvedeno, že student pracoval s odbornou literaturou a s myšlenkami jednotlivých autorů - autory uváděl (nejen odkazem na poznámku), srovnával, parafrázoval a citoval - zřetelně rozlišoval své myšlenky od myšlenek jednotlivých autorů	X	X	
5. Bakalářská práce je **filosofickým textem:** - autor prokázal fundovanou a analytickou práci s pojmy a myšlenkami - autorova práce nespočívala jen v převyprávění daného téma nebo pouhém poskládání hesel, parafrází či myšlenek - práce nebyla zatížena nepřiměřenou mírou historických, životopisných nebo jiných nefilosofických údajů	X	X	

Slovní ohodnocení

- Po studentovi bakalářské práce požadujeme zvládnutí základního filosofického řemesla. Tím rozumíme schopnost interpretační práce s odbornými texty – tedy aktivní rozvíjení diskuse mezi autory použité literatury, srovnávání jejich postojů, formulaci vlastních stanovisek a jejich kritické obhajování (opět v návaznosti na argumenty jiných autorů). Očekáváme, že student bude schopen vytvářet – na úrovni jednotlivých stran – argumentačně propojené pasáže ve stylu: *„Autor A tvrdí X, autor B namítá Y, autor C pak preferuje Z, přičemž já se kloním k…* "Takový způsob práce vede k filosoficky hutnému a spekulativnímu textu. Práce by se naopak měla vyhnout pouhému a postupnému převyprávění názorů jednotlivých autorů (např. ve formě parafrází s dlouhou posloupností odkazů „tamtéž"), stejně jako nekontrolované prezentaci vlastních nápadů, jež nejsou podloženy odbornou nebo filosofickou diskusí.
- V tomto ohledu vykazuje práce rezervy. Na mnoha místech není jasné, odkud autor své názory čerpá. Čtenář se tak nedokáže zorientovat, zda jde o parafráze určitého autora, nebo o osobní spekulaci.
- Tento nedostatek se ukazuje obzvlášť v pasážích věnovaných problematice morálních hodnot, kde zcela chybí odkazy na literaturu, z níž by autor čerpal podklad pro svá tvrzení.
-

III. Jazyková a grafická kritéria	**ano**	**zčásti**	**ne**
1. Práce je bez **gramatických a stylistických chyb** nebo jiných nedostatků (překlepy, nedokončené věty atd.)	X		
2. Práce je psána kultivovaným a **odborným jazykem**	X		
3. **Terminologické zpracování textu** - fundované porozumění terminologie v oboru a daném tématu - vhodné používání terminologie a vysvětlování pojmů	X	X	
4. **Odkazy, citace a poznámkový aparát** - vyhovující práce s odkazy - dodržení a jednotnost citačních norem - náležitý poznámkový aparát	X	X	
5. **Citování** - míra citování je vyvážená - vhodná volba a výstižnost citací	X	X	
6. **Grafická kritéria** - grafická úprava (písmo, obsah, názvy kapitol, odstavce, řádkování atd.) je vyhovující	X		

Slovní ohodnocení

- Důrazně doporučuji vyhnout se mnohačetným rétorickým obratům a otázkám, které jsou vhodné pro přednášky, rozhovory nebo novinové články, avšak nikoli pro akademický, respektive filosofický text. V takovém textu by měla být každá položená otázka důkladně propracovaná.
- V práci chybí důslednější vykazování zdrojů v poznámkovém aparátu, zejména v pasážích, kde autor rozvíjí či parafrázuje složitější argumenty, vyjasňuje pojmy apod.

IV. Použitá literatura	**ano**	**zčásti**	**ne**
1. kritický a vhodný výběr prvotní literatury a pramenů		X	
2. kritický a vhodný výběr druhotné literatury		X	
3. práce s cizojazyčnou literaturou	X		

Slovní ohodnocení

- Nedokáži s jistotou posoudit, zda je výběr použité literatury k tématu zcela adekvátní, nicméně práce působí dojmem, že čerpá z relevantních českých i cizojazyčných pramenů.
- Seznam použité literatury ve skutečnosti odpovídá seznamu 32 odkazů uvedených v poznámkovém aparátu. Při správném sestavení bibliografického seznamu by se ovšem ve výsledku objevilo pouze 6 autorů, respektive titulů, což je poměrně nízký počet vzhledem k rozsahu a povaze práce.
- V textu ani seznamu literatury nejsou uvedeny konkrétní odkazy na využití nástrojů umělé inteligence v procesu psaní práce. Není tedy jasně specifikováno, v jaké podobě byla či nebyla AI použita (např. jaká verze, rozsah použití apod.).

Celkové hodnocení:

Práci doporučuji k obhajobě: **ANO**

Návrh hodnocení: **D**

Náměty k rozpravě:

- Bude jednou možné přenést lidské morální hodnoty do algoritmického rámce AGI?
- Můžeme chápat *mesa-optimalizér* jakožto analogii například k nevědomým cílům u lidí?
- Jak byste do češtiny přeložil *deceptive alignment*?

V Pardubicích dne: 19. 5. 2025 oponent: Filip Grygar