Research Brief 003 — Kandidátní designové principy pro lidsko-AI kolektivní inteligenci

Status: Pracovní návrh v0.4

Epistemický status

Tento dokument nenavrhuje nový framework. Shrnuje kandidátní designové principy vyplývající z aktuálního výzkumu multi-agentního uvažování spolu s předběžnými kvalitativními pozorováními z projektu AI Council. **Všechny principy jsou pracovní hypotézy**, které mají vést budoucí experimenty — nejsou to ověřená fakta.

Úplnou historii revizí tohoto návrhu shrnuje Changelog na konci dokumentu.

Poznámka o provenienci

Oba primární papery byly pro tento projekt původně shrnuty modelem GPT. Následně byly dohledány a částečně ověřeny přímo z arXiv (červenec 2026):

- **Chen et al. 2025** — ověřeno z **plného textu**.
- **Elahi & Di Eugenio 2026** — ověřeno **pouze na úrovni abstraktu** (PDF nebylo v době kontroly strojově čitelné). Tvrzení označená níže jako *ověřeno z abstraktu* jsou potvrzena; detailnější tvrzení z GPT shrnutí (názvy protokolů WSV, CGA, HID; konkrétní změny F1 po úlohách) zůstávají **neověřená** a dokument se o ně neopírá.

⚠️ **Korekce:** původní GPT shrnutí uvádělo pokles F1 o 5–8 % u kompetitivní debaty. Paper ve skutečnosti uvádí **pokles F1 až o 13 % a F2 až o 15 %** oproti single-agent baselinům. Tato korekce je sama o sobě užitečným datovým bodem pro Ethimind praxi integrity provenience.

Primární zdroje

1. **Yongqiang Chen, Gang Niu, James Cheng, Bo Han, Masashi Sugiyama** — *Towards Scalable Oversight with Collaborative Multi-Agent Debate in Error Detection*,

arXiv:2510.20963 (CUHK, RIKEN AIP, HKBU, Tokijská univerzita). Zavádí pojem **debate hacking** a protokol **ColMAD**; vyhodnoceno na benchmarku RealMistake (Kamoi et al. 2024, arXiv:2404.03602).

2. **Ali Elahi, Barbara Di Eugenio** — *Multiagent Protocols with Aggregated Confidence Signals*, arXiv:2606.13591 (University of Illinois Chicago). Zavádí protokoly produkující **jedinou agregovanou confidence** pro výstup multi-agentního systému pomocí transformace confidence signálů a soft votingu nebo bayesovské fúze.

Sekundární zdroje: Kamoi et al. 2024 (nespolehlivost self-detekce chyb u LLM); Smit et al. 2024, Zhang et al. 2025, Yang et al. 2025 (kdy multi-agentní debata zaostává za jedním agentem); Kim et al. 2025 (korelované chyby napříč LLM).

Motivace

Aktuální výzkum přesouvá pozornost od schopností jednotlivých modelů k **architektuře interakce** mezi agenty. Tento brief se ptá: *které architektonické principy se opakovaně objevují napříč nezávislým výzkumem a které z nich si zaslouží empirické prozkoumání v rámci AI Council?*

Kandidátní princip 1 — Optimalizovat pro hledání pravdy, ne pro přesvědčování

Literatura. Chen et al. ukazují, že rámování debaty jako hry s nulovým součtem vyvolává *debate hacking* se třemi zdokumentovanými projevy: falešná evidence (dezinterpretace zadání), přehnaně sebejistá tvrzení a klamné argumenty odvádějící pozornost. Kompetitivní debata (CopMAD) často zaostávala za single-agent baseliny (F1 dolů až o 13 %), zatímco kolaborativní ColMAD — agenti si doplňují chybějící body — překonal kompetitivní debatu o 19 % a jednotlivé agenty až o 4 %. Formálně také dokazují, že u nečestných kompetitivních debatérů je optimální strategií soudce **debatní transkript zcela ignorovat**.

Pozorování z AI Council. Produktivní seance vznikaly, když modely rozvíjely argumenty, identifikovaly chybějící předpoklady a navrhovaly alternativní interpretace, místo aby hájily pevné pozice. Kvalitativní, nekontrolované.

Evidence	Síla
Externí	Střední–vysoká (kontext detekce chyb; generalizace netestována)
Interní	

	Nízká (anekdotická)

Budoucí test: porovnat kompetitivní / kolaborativní / syntézní prompty při konstantních modelech a stimulech.

Kandidátní princip 2 — Zachovat kognitivní diverzitu

Literatura. Heterogenní páry modelů dělají souběžné chyby méně často a záleží na heterogenitě **napříč firmami**: GPT-4 + Llama-2 snížily chyby o více než 30 % v oracle kolaboraci, zatímco blízce příbuzné Llama-2 + Llama-3.1 přinesly malé až žádné snížení (Chen et al.). Kim et al. 2025 dokumentují korelované chyby napříč LLM obecně.

Pozorování z AI Council. Různé modely zřejmě přispívají různými typy uvažování (kritika, syntéza, explorace, alternativní rámování), ale stabilita těchto profilů není známa a pozorování jsou zkreslena odlišnými prompty a pozicemi v konverzaci.

Evidence	Síla
Externí	Vysoká (benchmarkové úlohy)
Interní	Nízká

Budoucí test: měřit diverzitu hypotéz, konceptuální překryv a unikátní příspěvky jednotlivých modelů za fixního protokolu. → Přímo navazuje na **modul slepé konvergence Experimentu 002**.

Kandidátní princip 3 — Protokol interakce se jeví vlivnější než pouhé přidávání agentů

Literatura. Negativní výsledky kompetitivní debaty (Chen et al.; Smit et al.; Zhang et al.; Yang et al.) společně naznačují, že přidávání agentů nebo kol samo o sobě nepomáhá; dominuje protokol. CoIMAD byl robustní vůči počtu debatních kol. Elahi & Di Eugenio uvádějí (*ověřeno z abstraktu*), že předchozí studie popisují debatní kola po prvním kole jako přibližně **náhodnou procházku**.

Pozorování z AI Council. Neformálně některé dvoumodelové diskuse překonaly větší konverzace; kvantitativní srovnání neexistuje.

Evidence	Síla
Externí	Střední
Interní	Nízká

Budoucí test: stejná úloha, stejný celkový počet volání modelů, různá topologie.

Kandidátní princip 4 — Confidence brát jako evidenci, ne jako pravdu

Literatura (*ověřeno z abstraktu*). Žádná dřívější metoda neprodukovala jedinou confidence pro výstup multi-agentního systému. Elahi & Di Eugenio transformují surové confidence signály na srovnatelnou škálu a poté je agregují; výsledek je výrazně **diskriminativnější (AUARC)** než nejlepší jednotlivý agent, zatímco F1 zůstává stabilní a dorovnává ztráty, které debata působí na nejednoznačných úlohách. Analyzovány dva estimátory: pravděpodobnost sekvence a self-reported confidence; **kalibrace zlepšuje F1 u obou**. Tvrzení GPT shrnutí, že self-report někdy vyrovná či překoná token-probability estimátory, je konzistentní s abstraktem, ale na úrovni konkrétních výsledků neověřené.

Pozorování z AI Council. Confidence není systematicky sbírána; pouze anekdoticky.

Evidence	Síla
Externí	Střední (ověření pouze na úrovni abstraktu)
Interní	Zatím žádná

Budoucí test: zaznamenávat confidence u důležitých tvrzení, pozdější správnost a změnu confidence po diskusi. → Výrazně se překrývá s **modulem introspektivní přesnosti Experimentu 002**.

Kandidátní princip 5 — Procesy uvažování si zaslouží měření

Literatura. Většina benchmarků hodnotí finální přesnost; jak se kolektivní uvažování vyvíjí v čase, měří jen málokdo. Chen et al. nicméně hodnotí, zda *vysvětlení* odpovídají ground truth — krok tímto směrem.

Pozorování z AI Council. Kompletní transkripty uvažování jsou uchovávány, ale zůstávají archivními záznamy, ne strukturovanými měřeními; analytická pipeline neexistuje.

Evidence	Síla
Externí	Nízká (literatura dokládá mezeru, ne princip)
Interní	Nízká (data existují, neanalyzovaná)

Budoucí test: procesní metriky — počty generovaných/vyvrácených hypotéz, revize názorů a jejich správnost, syntézní události.

Kandidátní princip 6 — Epistemické guardraily mohou zlepšit kolaborativní uvažování

Literatura (ověřeno z plného textu). Implementace ColMAD zahrnuje tři mechanismy: **ověřování citací** (citace kontrolovány na přesnou shodu se zdrojovým kontextem), **povinný self-audit** (každý debatér musí uvést jeden možný způsob selhání vlastního tvrzení) a **povinné odhady confidence**. Vyhodnoceno jako balík v rámci vítězného protokolu; individuální příspěvky mechanismů nejsou známy.

Pozorování z AI Council. Podobné mechanismy používány neformálně, nestandardizovaně.

Evidence	Síla
Externí	Střední (pouze efekt balíku)
Interní	Nízká

Kandidátní guardraily pro důležitá tvrzení v Council seancích: podpůrná evidence · odhad confidence · nejslabší předpoklad · alternativní vysvětlení · možná falzifikace.

Sdílené mezery v evidenci

Následující mezery se týkají **všech šesti principů** a jsou uvedeny jednou, aby se neopakovaly: žádná replikace mimo kontext detekce chyb; neznámý přenos na otevřené úlohy bez ground truth; žádné studie s lidským facilitátorem ve smyčce; žádná longitudinální evidence (všechny citované experimenty jsou jednorázové). Všechny citované studie pracují výhradně s AI agenty; přenos na systémy zahrnující lidského facilitátora — kde se mísí odlišné kognitivní architektury, motivační struktury a epistemické normy — je zcela neprozkoumaný.

Otevřené otázky

- Překonává kolaborativní uvažování kompetici vždy, nebo jen při ověřitelné ground truth?
- Kdy je nesouhlas přínosný — a kdy se konsensus stává groupthinkem?
- Kolik diverzity je optimální?
- Jak agregovat confidence, když modely používají škálu odlišně?
- Které protokoly interakce maximalizují objevování místo shody?
- Jak velký je příspěvek lidského facilitátora oproti modelům?
- Projevy debate hackingu zdokumentované Chen et al. připomínají dlouho známé patologie lidské debaty. Zda to odráží na substrátu nezávislé limity kompetitivní debaty jako formy interakce, je otevřená, netestovaná otázka — citované papery to netvrdí.
- Oba primární papery předpokládají existenci ground truth. Přenos na otevřené úlohy (design, generování hypotéz) — kde AI Council primárně operuje — je **zcela otevřený**.

Výzkumné implikace

Pokud budoucí experimenty tyto principy podpoří, ovlivnily by primárně design protokolů AI Council, kritéria vyhodnocování experimentů, metodologii lidského facilitátora a design budoucích benchmarků. **Tento dokument nedoporučuje okamžité změny metodologie.**

Prohlášení o prioritách

Tento brief **nenavrhuje žádné nové experimenty jako závazky**. Aktuální prioritou zůstává **Experiment 002** (kompletní protokol, čeká na data) a **Experiment 001** (čeká na účastníka).

Naznačené směry (kompetitivní vs. kolaborativní promptování; homogenní vs. heterogenní skupiny; sledování confidence; sledování revizí názorů; procesní metriky) jsou zaznamenány pouze jako *možné budoucí směry*. Sledování confidence a revizí názorů se **výrazně překrývá s existujícími moduly Experimentu 002** (introspektivní přesnost, adversariální stabilita) a mělo by být začleněno do analýzy Experimentu 002, ne otevíráno jako samostatná vlákna.

Limity

Pozorování z AI Council jsou explorativní. Projektu aktuálně chybí preregistrované experimenty, kvantitativní procesní metriky, statistické vyhodnocení a nezávislá replikace. Jeden ze dvou primárních paperů bylo možné ověřit pouze na úrovni abstraktu. **Žádný z kandidátních principů by neměl být interpretován jako validovaná vlastnost lidsko-AI kolektivní inteligence** a žádné interní pozorování by nemělo být citováno jako evidence nezávislá na externí literatuře.

Závěrečná reflexe

Nejužitečnějším výstupem této literatury možná není potvrzení AI Councilu, ale zpřesněná otázka. Velká část současného výzkumu se ptá, jak mohou vícečetné AI systémy produkovat lepší odpovědi na úlohách se známou ground truth. AI Council umožňuje zkoumat příbuznou, ale odlišnou otázku:

> *Za jakých podmínek mohou lidé a vícečetné AI systémy produkovat poznání, které by žádný z nich pravděpodobně nevytvořil samostatně?*

Zda to představuje skutečně odlišný výzkumný směr, zůstává otevřenou empirickou otázkou, nikoli závěrem.

Changelog

v0.4 (červenec 2026): rozšířeny Sdílené mezery v evidenci o poznámku o nepřenositelnosti z čistě AI systémů na systémy s lidským facilitátorem (navrhl DeepSeek, přijato); přidána otevřená otázka, zda jsou patologie debate hackingu nezávislé na substrátu (DeepSeek navrhoval tvrzení uvnitř Principu 1; degradováno na otevřenou otázku, protože citované papery toto tvrzení nečiní). Navržený „Princip 7 — lidský facilitátor jako katalyzátor" byl odmítnut: dle přiznání samotného navrhovatele nemá žádnou externí literaturu, porušil by vstupní kritérium dokumentu (principy opakující se napříč nezávislým výzkumem) a věcně duplikuje otázku facilitátora již uvedenou v Otevřených otázkách. Zaznamenáno pro provenienci: jde o druhý modelem navržený pokus povýšit roli facilitátora (po GPT „Epistemic Conductor") a mezimodelová konvergence není považována za validaci.

v0.3 (červenec 2026): změkčen název Principu 3 z „architektura může být důležitější než počet agentů" na „protokol interakce se jeví vlivnější než pouhé přidávání agentů" podle toho, co literatura skutečně ukazuje (korekci navrhlo GPT, přijata); přidány Výzkumné implikace oddělující hypotézy od praktických důsledků; přidány Sdílené mezery v evidenci. Navržená celostacková škála epistemického statusu odložena jako rozhodnutí pro Research Stack, ne pro RB-003.

v0.2 (červenec 2026): ověřené citace s autory a arXiv ID; opraveno číslo u kompetitivní debaty (neověřených 5–8 % → ověřených až 13 % F1 / 15 % F2); oddělena hodnocení externí a interní evidence u každého principu; označena tvrzení ověřená z abstraktu vs. z plného textu; přidána poznámka o provenienci, prohlášení o prioritách a propojení navržených směrů s moduly Experimentu 002.

v0.1: původní návrh (GPT).