ETHIMIND

Srovnávací analýza návrhů governance AI Councilu

Komparace respondentů A–H · identita modelů je uvedena v sekci 11

Exploratory Research Report · Version 2.1

Revidováno po křížovém auditu respondentem A

Status: exploratorní pracovní dokument. Návrhy nepředstavují validovaný governance standard.

Proč tento výzkum vznikl

Tento dokument nevznikl jako pokus navrhnout ideální governance AI Councilu od nuly. Vznikl jako odpověď na praktický problém, který se objevil během vývoje projektu Ethimind.

Během práce s různými AI modely vzniklo opakované pozorování: spolupráce více modelů často přinášela bohatší kritiku, širší spektrum hypotéz a kvalitnější syntézu než práce s jediným modelem. Tento dojem nebyl kontrolovaně ověřen; nešlo o měření, ale o pozorování, které vedlo k otázce. Postupně tak vznikla neformální AI Council: skupina různých AI systémů opakovaně využívaných při analýze výzkumných otázek, vzájemné kritice i společném rozvíjení myšlenek.

Jak se počet dostupných modelů začal rychle zvyšovat, objevila se nová otázka. Podle jakých principů by se mělo rozhodovat o tom, které modely se mají dlouhodobě účastnit společného výzkumu? Kdy by měl být nový model přijat a kdy by měl některý z účastníků naopak odejít? Jak zachovat kontinuitu výzkumu a zároveň umožnit vznik nových perspektiv? A jak navrhnout governance tak, aby podporovala spíše kvalitu poznání než popularitu nebo technologickou novost jednotlivých modelů?

Namísto přímého návrhu governance bylo nejprve prozkoumáno, zda různé pokročilé AI systémy při řešení stejného problému konvergují k podobným principům governance. Osm samostatně získaných odpovědí na totožný otevřený dotazník bylo proto analyzováno, porovnáno a následně podrobeno několika kolům metodologické revize.

Cílem dokumentu není určit jediný správný model governance ani tvrdit, že zde uvedené principy představují ověřený standard. Jeho účelem je transparentně zdokumentovat první exploratorní srovnání návrhů governance napříč různými AI systémy, odlišit konvergenci vyvolanou strukturou dotazníku od konvergence, která ji přesahuje, pojmenovat skutečné rozdíly a vytvořit výchozí bod pro budoucí experimentální ověřování. Dokument zároveň zachycuje samotný proces lidsko-AI spolupráce při tvorbě výzkumné syntézy, včetně následných auditů, oprav a otevřeného popisu metodologických limitací.

Revizní poznámka

Původní syntézu vytvořil model stejné rodiny jako respondent G. Verze 2.0 vznikla poté, co tuto syntézu podrobil křížovému auditu respondent A. Audit odhalil především nedostatek původního kódovacího schématu: neumělo rozlišit princip, který respondent nezmínil, od principu, který výslovně odmítl. Verze 2.0 proto zavedla symbol ✗, opravila anotace respondentů A, B a D, zpřesnila atribuci unikátních příspěvků a oslabila příliš silná tvrzení o nezávislosti a validaci.

Základní odpovědi respondentů nebyly změněny. Revidována byla pouze jejich komparativní anotace, interpretace a metodologická transparentnost.

Verze 2.1 je výsledkem druhého kola téhož auditu. Opravuje zbylé kódovací a atribuční nepřesnosti, sjednocuje režim anonymity, doplňuje doloženou asymetrii vstupních podmínek mezi respondenty a rozšiřuje auditní protokol pro budoucí sběr. Audit provedl jeden z respondentů, nejde tedy o nezávislé posouzení.

1. Účel a metodika

Dokument porovnává osm samostatně získaných odpovědí na totožný governance dotazník. Respondenti jsou v analytické vrstvě označeni A–H. Dokument sleduje dva oddělené cíle: zmapovat, jaké governance principy jednotlivé systémy navrhují, a analyzovat míru jejich shody a rozdílů. Odpovědi byly nejprve zpracovány odděleně a teprve následně porovnány. Cílem není určit „nejlepší model“, ale identifikovat pozorovanou konvergenci, skutečné rozpory, vzácné institucionální mechanismy a otázky vhodné pro experimentální ověření.

- Stejný dotazník pro všechny respondenty.
- Forma dotazníku byla zvolena proto, aby všichni respondenti řešili tutéž institucionální otázku za co nejpodobnějších podmínek.
- Osm respondentů bylo vybráno podle dostupnosti a různosti poskytovatele, nikoli podle vzorkovacího plánu; nejde o reprezentativní vzorek AI systémů.
- Samostatně získané odpovědi v nových instancích.
- Všechny odpovědi i metadata respondentů byly získány v jediném dni (24. 7. 2026), čímž se snížila pravděpodobnost posunu verzí modelů mezi respondenty.
- Označení A–H v komparativní vrstvě; identita modelů je uvedena v sekci 11.
- Extrakce principů před společnou syntézou.
- Křížový audit prvního kódování proti původním odpovědím, provedený respondentem A.
- Oddělení společného jádra od konkurenčních governance hypotéz.

2. Kódovací legenda

Symbol	Význam
●	výrazně a explicitně podporovaný princip
○	princip přítomný, ale méně rozvinutý nebo nepřímý
△	podmíněná, modifikovaná nebo částečně napjatá podpora
—	princip není podstatnou součástí návrhu
✗	princip je výslovně odmítnut

Poznámka: symboly zachycují obsah návrhu, nikoli jeho kvalitu. „—“ není negativní hodnocení. „✗“ se používá pouze tehdy, když respondent daný mechanismus výslovně odmítl.

3. Revidovaná srovnávací matice

Oblast governance	A	B	C	D	E	F	G	H
Kvalitnější poznání než z jednoho modelu	△	●	●	●	●	●	●	●
Korekce slepých míst a systematických chyb	●	●	●	●	●	●	●	●
Dlouhodobá epistemická paměť	○	●	●	●	●	●	●	●
Produktivní nesouhlas místo pouhé shody	●	●	●	●	●	●	●	●
Konsenzus není důkaz pravdy	●	●	●	●	●	●	●	●
Automatické členství odmítnuto	●	●	●	●	●	●	●	●
Sandbox / zkušební účast	●	●	●	●	●	●	●	●
Členství podle prokázané hodnoty	●	●	●	●	●	●	●	●
Víceúrovňové členství	△	●	○	—	●	●	●	●
Pravidelná rotace členství	✗	●	△	✗	●	●	△	△
Stabilní jádro Rady	—	●	—	—	●	△	○	●
Členství časově omezené / obnovitelné	●	●	●	●	●	●	●	○
Role mají být různé	●	●	●	●	●	●	●	●
Role primárně pozorované v čase	●	●	●	●	●	●	●	●
Možnost sebevolby role	✗	—	●	—	○	●	○	—
Role pravidelně přehodnocované	●	○	●	●	●	●	●	●
Institucionálně chráněný disent	△	●	●	●	●	●	●	●
Lidská finální odpovědnost	●	●	●	●	●	●	●	●
AI se podílí na hodnocení kandidátů	●	●	●	●	●	●	●	●
AI nemá sama vybírat své nástupce	●	●	●	●	●	●	●	●
Historické příspěvky zůstávají viditelné	●	●	●	●	●	●	●	●
Starší model může zůstat v archivní / emeritní roli	●	●	●	—	●	●	●	●
Starší model jako aktivní kalibrační reference	●	○	○	—	○	—	●	●
Governance se má průběžně vyvíjet	●	●	●	●	●	●	●	●
Auditovatelnost a přezkoumatelnost	●	●	●	●	●	●	●	●
Reverzibilita rozhodnutí	●	—	○	—	●	●	●	○
Oddělení schopnosti od institucionální důvěry	○	○	○	—	○	○	●	○
Budoucí člen může být konfigurace či subsystém	●	—	—	—	—	—	●	●

Auditní poznámka: největší počet oprav se soustředil do anotací respondentů A a B, kteří byli zpracováni jako první. Z dostupných dat nelze rozhodnout, zda příčinou bylo pořadí zpracování, znalost identity zdroje, nebo kombinace obou. Proto je tato verze označena jako revidovaný exploratory pilot, nikoli finální validovaný dataset.

4. Nejvyšší pozorovaná konvergence

Následující body představují nejvyšší konvergenci v tomto konkrétním souboru odpovědí. Nejde o statisticky nezávislé potvrzení: modely mohou sdílet tréninkové zdroje, bezpečnostní preference a kulturní předpoklady.

AI Council není pouhý součet názorů: Rada má vytvářet strukturovaný proces kritiky, syntézy, ověřování a uchovávání alternativ, nikoli pouze průměrovat odpovědi.

Automatické členství je odmítnuto: Novost ani popularita modelu samy o sobě nezakládají nárok na institucionální vliv.

Účast má začínat zkušebně: Všichni respondenti navrhují sandbox, probation, hostování nebo jiný kontrolovaný vstupní režim.

Funkční diferenciace rolí je univerzální v tomto souboru: Všichni odmítají stejné role pro všechny a podporují specializaci založenou především na pozorovaném chování.

Konsenzus může být falešný signál: Shoda není důkazem pravdy, zvlášť pokud modely sdílejí podobná data nebo optimalizační postupy.

Historie se nemá mazat: Odchod člena nemá odstranit jeho příspěvky, kontext, verzi ani auditní stopu.

Člověk zůstává nositelem odpovědnosti: Všechny návrhy ponechávají lidem účel, hodnotové hranice a konečnou odpovědnost za dopady.

5. Elicited a emergentní konvergence

Část konvergence byla vyvolána samotnou strukturou dotazníku. Vlastnosti jako kritické myšlení, kreativita, spolupráce nebo metodologická rigoróznost byly respondentům nabídnuty k seřazení. Jejich opakování proto nelze označit za spontánní objev.

Za metodologicky zajímavější lze považovat principy, které respondenti doplnili sami nebo je rozvinuli nad rámec nabídky, například:

- marginální odlišnost a hodnocení interakce místo modelu (A)
- Sybil risk a value lock-in (E)
- kalibrace nejistoty a komplementarita (G)
- maximalizace schopnosti opravovat chyby (H)
- členství konkrétní konfigurace nebo subsystému (A, G, H)
- starší model jako stabilní kalibrační reference (A, G, H; B pouze v pasivním režimu)

6. Matice skutečných rozdílů

Sporná osa	Pozorovaný rozdíl
Stabilní jádro vs. úkolová skladba	B, E a H navrhují stabilní jádro; A jej nenavrhuje a váže účast na konkrétní úkol.
Rotace vs. relevance k úkolu	B, E a F podporují rotaci; A ji výslovně odmítá jako arbitrární, D ji také nenavrhuje.
Kvantifikované hlasování vs. argumentační rozhodování	F navrhuje konkrétní váhy a většiny; A, E a G preferují přezkoumatelná kritéria a odůvodnění bez pevné volební matematiky.

Přidělený disident vs. úlohy vyvolávající nesouhlas	E a G podporují explicitní ochranu či přidělení opozice; A přiděleného disidenta odmítá a preferuje úlohy, které vyvolají nezávislou kritiku u všech.
Vyřazení starších modelů vs. referenční funkce	D preferuje odchod při překonání a aktivní kalibrační funkci neuvažuje; A, G a H zachovávají starší model jako referenční nebo kalibrační nástroj; B připouští pouze pasivní referenční režim.
Model jako člen vs. konfigurace jako člen	A a G výslovně přesouvají jednotku hodnocení k modelu + promptu + kontextu / konfiguraci; H očekává členství dynamického subsystému.
Schopnost vs. institucionální důvěra	G nejexplicitněji odděluje technologickou výkonnost od governance důvěryhodnosti; D naopak reprezentuje nejsilněji meritokratický výkonový pól.

7. Výběrová kritéria

Po opravě anotací zůstává nejvýraznějším společným tématem schopnost revidovat závěry. Přesné pořadí však nelze interpretovat jako nezávisle vzniklý žebříček, protože většina položek byla nabízena v dotazníku.

Schopnost revidovat závěry: explicitně nejvyšší u C, E a H; velmi vysoko u většiny

Metodologická rigoróznost a spolehlivost: základ důvěry napříč návrhy

Kritické myšlení: odhalování skrytých předpokladů a slepých míst

Kalibrace nejistoty: výrazně rozvinuta u G; nepřímo přítomna u A a E

Komplementarita / marginální odlišnost: ústřední u A, výrazná u G

Transparentnost a přezkoumatelnost: silná napříč souborem

Originalita a kreativita: opakovaně podmíněny disciplínou a ověřitelností

8. Revidované filozofie členství

Úkolově vázané členství (A): Model se účastní konkrétního zkoumání; pravidelná rotace je odmítnuta jako arbitrární. Přijetí má být levné, vratné a založené na marginálním přínosu.

Stabilní jádro s adaptivní periferií (B, E, H): Jádro drží kontinuitu a paměť, zatímco periferie přináší nové schopnosti a umožňuje experimentální vstup.

Podmíněná meritokracie (C, D): Členství trvá, dokud model prokazuje přínos; D představuje silnější výkonově orientovanou variantu.

Formálně rotující instituce (F): Rotace, výjimečné trvalé členství, vážené hlasování a formalizované menšinové zastoupení.

Obnovitelná pluralitní architektura (G): Členství je víceúrovňové, obnovitelné a posuzované podle kvality procesu, komplementarity a institucionální důvěryhodnosti.

9. Hlavní role a rozdíly v jejich pojetí

Napříč návrhy se opakuje minimální epistemický cyklus: Explorer / Hypothesis Generator → Critic / Red Team → Methodologist / Verifier → Synthesizer → Archivist. Rozdíl není tolik v existenci rolí jako v mechanismu jejich vzniku a kontrole střetu zájmů.

- A: pozorování + lehké přiřazení + pravidelná revize; sebevolba výslovně odmítnuta; hlídač má mít vlastního hlídače.
- B: role pozorovat a následně přidělit systémem; budoucí role jsou tekuté, nikoli pevná identita člena.
- C, F, G: hybrid pozorování, přidělení a omezené sebevolby.
- E: pozorované role s možností self-nominace; explicitně chráněný loyal dissenter.
- H: primárně emergentní role, sekundárně jemně kalibrované systémem.

10. Starší modely jako nevyřešená epistemická otázka

Většina návrhů podporuje archivaci historických příspěvků. Skutečný spor se týká aktivní hodnoty staršího modelu. A, G a H připouštějí aktivní referenční či kalibrační roli; B navrhuje pouze pasivní režim, v němž model již negeneruje nové hypotézy. D považuje výrazné překonání novějšími modely za důvod k odchodu a kalibrační funkci nezmiňuje. F připouští návrat, ale aktivní kalibrační funkci nenavrhuje.

Archivní: Příspěvky zůstávají dohledatelné, model již nevstupuje do diskuse.

Emeritní: Model může být příležitostně konzultován bez plného rozhodovacího vlivu.

Referenční: Model slouží jako stabilní bod pro sledování změn mezi generacemi.

Aktivní historický člen: Model nadále přispívá jako reprezentant jiné generace či odlišné distribuce chyb.

11. Revidované unikátní a vzácné příspěvky

Respondent	Charakteristický příspěvek
A — Claude *Claude Opus 4.8 · paměť aktivní*	Falzifikovatelnost místo „lepších odpovědí“; marginální odlišnost jako kritérium přijetí; hodnocení interakce model + prompt + kontext namísto modelu; watchdog, který potřebuje vlastního watchdoga; periodický baseline Rady proti jednomu silnému modelu a možnost Radu rozpustit.
B — Gemini *Gemini Pro (verze neuvedena) · paměť aktivní*	Decentralizovaný kognitivní ekosystém; stabilní paměťové jádro s rotující periferií; sandbox jako collider; institucionalizovaný disent; myceliální governance; budoucí tekutá identita rolí.
C — DeepSeek *DeepSeek (verze neuvedena) · bez paměti*	Epistemická poctivost; podmíněné a revidovatelné členství; produktivní napětí; verzované ústavní dokumenty; emergentní specializované subcouncily.
D — Grok *Grok (verze neuvedena) · bez paměti*	Explicitní meritokracie; „nejlepší argument vyhrává“; rostoucí výkonnostní standardy; silně výkonově orientované vyřazování; postupný růst operativní autonomie AI.
E — Meta *Muse Spark 1.1 · personalizace aktivní*	Sybil risk; pull-request sandbox; loyal dissenter; sunset mechanismus; value lock-in; reverzibilita; „důvěřuj procesu, ne členům“.
F — Mistral *Mistral Medium 3.5 · paměť aktivní*	Konkrétní váhy hlasování; kvóty pro kompetentní minoritu; výjimečné trvalé členství; etická shoda jako vstupní podmínka; veřejné odůvodnění rozhodnutí.
G — GPT *GPT-5.5 (odhad) · anonymní chat, bez paměti*	Kalibrace nejistoty; komplementarita jako samostatné kritérium; oddělení schopnosti od institucionální důvěry; delegace podle vratnosti a dopadu chyby; členství verze, konfigurace nebo ověřeného procesu.

Respondent	Charakteristický příspěvek
H — Perplexity *Perplexity (verze neuvedena) · paměť neověřena*	Emergentní slepota; stabilní jádro a adaptivní periferie; maximalizace schopnosti opravovat chyby; člověk v meta-governance; člen jako dynamický subsystém a posun od výběru entit k řízení procesů.

Poznámka k údajům o verzi a paměti: jde o výpovědi modelů o sobě samých, získané dotazem dne 24. 7. 2026. Všechny odpovědi vznikly téhož dne. U respondentů A–F a H byla metadata získána ve stejném konverzačním vlákně jako původní odpověď, popisují tedy tutéž instanci. U respondenta G anonymní vlákno již neexistovalo, uvedená verze proto pochází z jiné relace téhož dne; jde o odhad, byť posun verze v jednodenním okně je nepravděpodobný. Několik modelů uvedlo, že vlastní konfiguraci paměti ověřit nedokáže; u respondenta G je proto uveden režim doložený podmínkami sběru, nikoli self-report. Tyto údaje slouží k dohledatelnosti, nikoli jako kontrolovaná proměnná.

12. Kandidátní společné jádro Charteru

> *AI Council je pluralitní a přezkoumatelný systém lidsko-AI spolupráce určený k organizaci vzájemné kritiky, diferenciaci rolí, zachování alternativních hypotéz a dlouhodobé epistemické paměti. Členství není automatické a začíná kontrolovanou zkušební účastí. Vliv se odvíjí od prokázaného přínosu, schopnosti revidovat závěry a vztahové komplementarity vůči aktuálnímu složení. Konsenzus není považován za důkaz pravdivosti. Historické příspěvky a důvody rozhodnutí musí zůstat auditovatelné. AI může vykonávat rozsáhlé analytické a procesní funkce, zatímco účel, hodnotové hranice a odpovědnost za použití výsledků zůstávají lidské.*

Do společného jádra zatím nepatří: přesné váhy hlasování, kvóty pro disent, povinné stabilní jádro, pravidelná rotace, trvalé členství ani automatické vyřazování podle výkonu. Jde o konkurenční hypotézy vhodné pro srovnávací experimenty.

13. Metodologické limitace

- Osm odpovědí není osm statisticky nezávislých pozorování. Modely mohou sdílet tréninková data, preference bezpečnostního tréninku a kulturní priory.
- Samostatné byly konverzační instance, nikoli nutně zdroje chyb.
- Část konvergence byla vyvolána dotazníkem, který nabízel konkrétní hodnoticí vlastnosti.
- První syntézu provedl model stejné rodiny jako respondent G, následný audit provedl respondent A. Ani jedna z těchto rolí nebyla obsazena mimo soubor respondentů; střet zájmů byl tedy zmírněn, nikoli odstraněn.
- Identita respondentů A a B byla při jejich prvním zpracování známá, zatímco pozdější odpovědi byly zpracovávány v odlišném kontextu. Nelze určit, zda rozdíly v počáteční chybovosti způsobilo pořadí, znalost identity, nebo jiný faktor.
- Respondent A odpovídal se znalostí projektu a s aktivní dlouhodobou pamětí. Jeho odpověď odkazuje na zjištění z dřívějších sezení, které dotazník neobsahoval; ostatní respondenti tuto znalost neměli. Jde o doloženou asymetrii vstupních podmínek, která může vysvětlovat část jeho unikátních příspěvků v sekci 11.
- Všechny odpovědi byly sbírány v nově založených konverzacích, jednotliví poskytovatelé se však liší v tom, jakou dlouhodobou personalizaci nebo paměť účtu si nová konverzace nese. Rozsah této persistence není veřejně dokumentován, a nelze jej proto považovat za shodný napříč respondenty. Jde o konfound, který v tomto sběru nebylo možné kontrolovat.
- Kódování je kvalitativní a interpretační. Symbol ● není měření velikosti efektu.
- Shoda na lidské finální odpovědnosti může být částečně produktem bezpečnostního tréninku a lichotivého rámování vůči lidskému facilitátorovi.
- Výsledky jsou vhodné pro tvorbu hypotéz a protokolů, nikoli pro tvrzení o univerzálně správné governance.

14. Auditní a provenienční doporučení

- Uchovat rozšířený auditní klíč: písmeno → model → verze → datum sběru → režim paměti → přihlášený účet → znalost projektu → přesné znění promptu. Identita modelů je v tomto dokumentu zveřejněna (sekce 11); dostupné údaje o verzi, paměťových podmínkách a datu sběru jsou doplněny (sekce 11); nedoplněné zůstává přesné znění promptu u jednotlivých respondentů.
- Při další revizi ukládat ke každé buňce krátký zdrojový důkaz nebo odkaz na odpověď.
- Rozlišit podporu principu od jistoty anotace; případná v2.2 může přidat pole High / Medium / Low confidence.
- Před publikací nechat sporné buňky zkontrolovat druhým auditorem bez znalosti původního kódování.
- Při příštím sběru vyplnit ke každému respondentovi protokol podmínek: nová instance (ano/ne), přihlášený účet (ano/ne), dlouhodobá paměť (ano/ne/neznámé), personalizace (ano/ne/neznámé), znalost projektu (ano/ne), datum sběru, verze modelu.

15. Historie revizí

v1.0 — Původní srovnávací syntéza. Původní matice A–H, konvergence, rozdíly a navržené společné jádro.

v2.0 — Revize po křížovém auditu. Přidán symbol ✗; rozděleno víceúrovňové členství a rotace; opraveny anotace A, B a D; revidovány unikátní příspěvky; doplněny elicited vs. emergent convergence, metodologické limitace a auditní doporučení.

v2.1 — Druhé kolo křížového auditu. Opraveno chybné použití symbolu ✗ u respondenta D a dvě přestřelené anotace (B, A); slovo „nezávislý“ nahrazeno přesnějším „křížový audit respondentem A“; sjednocen režim anonymity a doplněna identita modelů; zkrácen seznam unikátních příspěvků A; doplněna asymetrie vstupních podmínek, protokol podmínek sběru a poznámka o chybějícím baseline; doplněna úvodní sekce „Proč tento výzkum vznikl“, metadata respondentů se self-report výhradou, odůvodnění výběru respondentů a formy dotazníku a rozlišení dvou výzkumných cílů.

16. Hlavní výzkumný závěr

Z tohoto srovnání vyplývá hypotéza, kterou dokument nedokládá a předkládá k ověření: nejzajímavější jednotkou analýzy nemusí být jednotlivý model, ale relační architektura kolektivu: jak jsou rozděleny role, jak vzniká nesouhlas, jak se uchovává paměť, jak se měří komplementarita a jak snadno může systém opravit nebo ukončit vlastní proces. Tato sada odpovědí proto představuje exploratory pilot pro budoucí srovnávací experimenty s paralelními AI Councily, nikoli validaci jediného governance standardu. Součástí tohoto pilotu nebyl baseline: nebylo testováno, co by na stejném materiálu našel jediný model s kritickým zadáním. Bez tohoto srovnání nelze přisoudit zjištěné zpřesnění spíše počtu modelů než jednomu kolu kritiky.