

Architektonická adaptace ve spolupráci člověka a AI

Jak může kolaborativní kognitivní systém rozpoznat, že jeho způsob myšlení už není přiměřený?

Research Note	RN-006
Typ	Research Note
Status	Pracovní hypotéza (v0.4 — zapracovány dvě nezávislé externí kritické recenze; publikační verze)
Stav evidence	Konceptuální návrh. Dosud empiricky nevyhodnoceno.
Verze	0.4
Datum	Červenec 2026

Research Programme Notice

Tento dokument je součástí probíhajícího výzkumného programu. Předkládá pracovní výzkumnou hypotézu určenou k diskusi, testování, kritice a případné falzifikaci. Neměl by být interpretován jako etablovaná vědecká teorie.

Poznámka k číslování a procesu

Tato nota je publikována jako RN-006. Během vývoje nesla pracovní označení RN-00Y, pod kterým na ni odkazuje ACP-0001 — procesní log dokumentující její vznik. Kurátorské rozhodnutí tam zaznamenané jako otevřené (Větev M) bylo vyřešeno při publikaci: RN003 zůstává rezervována pro Human-AI Iterative Synthesis (veřejně referencovaná rezervace v metadatech RN001); tato nota stojí vedle dané linie, nikoli uvnitř ní, a nese proto další volné veřejné číslo.

Poznámka o provenienci

Tato Research Note vznikla iterativním dialogem člověka s AI v rámci Ethimind AI Council (červenec 2026). Diskuse začala otázkou, zda AI pouze zrychluje lidskou produktivitu, nebo zda může postupně ovlivňovat samotnou strukturu lidského myšlení, a postupně se přesunula k architektuře spolupráce mezi lidmi a AI.

Koncept se vyvíjel v rozpoznatelných fázích. V první fázi (GPT a Clowdia) diskuse navrhla, že dlouhodobé účinky AI mohou záviset méně na schopnostech jednotlivých modelů a více na tom, jak je spolupráce mezi lidmi a AI organizována. Ve druhé fázi byla tato myšlenka rozšířena o pracovní tezi, že různé kognitivní úlohy mohou vyžadovat různé kolaborativní

konfigurace, nikoli jedinou univerzálně optimální — myšlenku GPT pojmenovalo "Cognitive Architecture Fitness" (dochovaná formulace v primárním přepisu) a Grok ji rozvinul do dimenzionálního rámce s doménovými příklady. Ve třetí fázi se vynořila ostřejší otázka: jak může kognitivní architektura rozpoznat, že už není přiměřená aktuálnímu problému? Po ověření proti primárním přepisům obou větví (viz ACP-0001) se tato fáze rozkládá na řetězec: GPT zavedlo myšlenku architektur měnících se podle fází ("Adaptive Council"); Grok formuloval princip sebe-detekce; GPT přispělo analogií homeostázy jako metaforou; Grok pojmenoval rámec "Cognitive Architecture Homeostasis", vytvořil jeho taxonomii indikátorů a na žádost výzkumnice napsal první draft Research Note v češtině; tento český draft je přímým doloženým předchůdcem anglické v0.1, kterou GPT sepsalo v přímé práci s ním. Ve čtvrté fázi provedl Claude kritickou recenzi draftu v0.1, která vedla k předkládané revizi: odstranění redundantní taxonomie indikátorů ve prospěch existujícího slovníku RN002, terminologické korekci pojmu homeostáza, poctivému přerámování konceptuálního kontextu dokumentu a ukotvení přílohy o rolích příspěvků v existující taxonomii CRediT. Draft v0.1, recenzní poznámky i tato v0.2 jsou uchovány; revize sama je součástí záznamu.

Shrnutí příspěvků pro tuto verzi: původ otázky a kurátorství — Clowdia; výchozí hypotéza a první draft — GPT; konceptuální rozvinutí fitness rámování — Grok; kritická recenze a požadavky revize — Claude; ediční syntéza v0.2 — Claude a Clowdia.

Abstrakt

Většina diskusí o spolupráci člověka s AI se soustředí na výběr lepších modelů, promptů nebo rolí agentů. Tato nota navrhuje, že stejně důležitá výzva leží jinde: jak může kolaborativní kognitivní systém určit, kdy by se měl změnit jeho vlastní způsob myšlení? Zkoumáme schopnost kognitivního systému člověk-AI sledovat vlastní kolaborativní dynamiku a v případě potřeby adaptovat svou architekturu. Tento návrh je třeba chápat jako konceptuální hypotézu, která má vést budoucí experimentování, nikoli jako etablovanou teorii.

Ústřední otázka

Jak kolaborativní systém člověk-AI rozpozná, že je čas změnit architekturu spolupráce — a co se má stát, když k tomu dojde?

Všechno v této notě slouží této jediné otázce. Dřívější drafty se pokoušely zavést širší rámce (architektonickou fitness, trajektorie, taxonomie indikátorů); předkládaná verze se záměrně zužuje na samotnou adaptační smyčku, která je skutečně novým příspěvkem.

Rozsah. Tato nota se nepokouší o obecnou teorii adaptivních kognitivních systémů. Zkoumá jednu konkrétní otázku: jak může současné uspořádání AI Councilu rozpoznat, že jeho konfigurace přestává být efektivní. Zobecnění mimo toto prostředí je předmětem budoucí práce.

Terminologie

Kognitivní architekturou se zde rozumí organizace interakcí mezi lidmi a jedním či více AI systémy při řešení kognitivní úlohy: role účastníků, informační toky, rozhodovací pravomoc, organizace paměti, zpětnovazební smyčky a adaptační mechanismy.

Poznámka ke slovu homeostáza. Draft v0.1 používal pracovní název „Cognitive Architecture Homeostasis“. Kritická recenze v tom identifikovala terminologickou chybu, kterou stojí za to brát vážně: homeostáza v biologii a kybernetice označuje udržování stabilního stavu kolem nastavené hodnoty. To, co tato nota popisuje, je téměř opak — zachování životaschopnosti prostřednictvím reorganizace samotné konfigurace. Nejbližší etablované koncepty jsou allostáza (stabilita skrze změnu, Sterling a Eyer) a nej přesněji Ashbyho ultrastabilita (Design for a Brain, 1952): systém, který detekuje, že jeho současná konfigurace nedokáže udržet podstatné proměnné, a reaguje změnou vlastní organizace, nikoli jen parametrů. Příbuzným organizačním konceptem je double-loop learning (Argyris a Schön), v němž systém zpochybňuje nejen své akce, ale i rámec, který je generuje. Tato nota proto mluví o architektonické adaptaci a za své primární teoretické kotvy považuje allostázu a ultrastabilitu. Korekce byla provedena před jakoukoli empirickou prací, což je nejlevnější možný okamžik, kdy ji udělat.

Ústřední hypotéza

Dlouhodobá spolupráce člověka s AI může vyžadovat nejen dobře navržené kognitivní architektury, ale i mechanismy schopné rozpoznat, kdy se tyto architektury staly maladaptivními. Bez takových mechanismů mohou i původně úspěšné architektury postupně ztrácet účinnost, zatímco dál běží — a účastníci uvnitř nich si toho mohou všimnout jako poslední.

Od statických architektur k architektonickým přechodům

Většina současných diskusí se implicitně ptá, která architektura je nejlepší. Tato nota navrhuje jinou otázku: které přechody mezi architekturami jsou přiměřené a co je má spouštět? Spolupráce se může například posouvat z explorační konfigurace přes kritické hodnocení a sběr evidence k integraci a rozhodnutí — ale zajímavý problém není samotná sekvence, která se bude lišit podle domény. Zajímavý problém je spouštěč: co systému říká, že současný režim je vyčerpán a je na řadě další?

Detekce potřeby adaptace: navázání na RN002

Tato Research Note nenavrhuje nové indikátory kognitivních stavů. RN002 (Meteorologie komplexních systémů) už poskytuje slovník pro dynamické stavy kolaborativního systému — divergenci, emergenci, rezonanci, turbulenci, integraci a stagnaci, spolu se stavovými proměnnými a hraničními podmínkami. Zavedení paralelní taxonomie by tuto práci duplikovalo pod novými jmény — což dřívější draft této noty skutečně udělal, než to recenze zachytila.

Místo toho tato nota zkoumá, jak mohou dynamické stavy z RN002 sloužit jako vstupy pro architektonickou adaptaci. Pracovní návrh zní, že určitá čtení těchto stavů fungují jako adaptační signály. Dlouhotrvající stagnace při pokračující aktivitě naznačuje, že současná architektura vyčerpala svou generativní kapacitu. Předčasná konvergence — mizení alternativ před dostatečným prozkoumáním, v pojmech RN002 nástup integrační fáze, zatímco divergence byla ještě produktivní — naznačuje, že se má změnit kritický tlak nebo

rovnováha rolí. Přetrvávající turbulence bez emergence naznačuje, že architektuře chybí integrační kapacita. Trvalá dominance jednoho dynamického režimu, ať je jakýkoli, je sama o sobě kandidátním signálem, protože zdravé kolaborativní procesy podle všeho mezi režimy cyklují, místo aby se usadily v jednom.

Dva další kandidátní signály se na stavy RN002 nemapují a jsou zde navrženy jako skutečně nové, s odpovídající opatrností. Drift provenience: původ a vývoj důležitých myšlenek je stále obtížnější rekonstruovat, což ukazuje, že spolupráce předbíhá vlastní vedení záznamů. Kolaps poločasů myšlenek: potenciálně cenné větve mizí rychleji, než je lze vyhodnotit — signál přímo propojený s hypotézou Ekologie neuskutečněných možností, protože popisuje zrychlující se ztrátu v externalizovaném prostoru možností. Oba signály zůstávají hypotetické a před jakýmkoli empirickým použitím vyžadují operacionalizaci.

Poznámka k operacionalizaci. V první fázi budou všechny signály hodnoceny kvalitativně lidským facilitátorem na základě pozorovaných vzorců; kvantitativní operacionalizace — prahy, baseline, míra falešně pozitivních — je sama předmětem prvních experimentů, nikoli jejich předpokladem.

Kandidátní adaptační mechanismy

Možné adaptační odpovědi zahrnují změnu rolí AI, zavedení dalších perspektiv, přepínání mezi exploračním a konvergentním režimem, úpravu paměťových strategií, zvýšení nebo snížení kritického tlaku, dočasné zpomalení rozhodování a vyžádání externí evidence před pokračováním. Jde o designové hypotézy, nikoli doporučení. Praktické pozorování ze samotných Ethimind sessions: dosud nejčastěji používaným adaptačním mechanismem je záměrné stop-pravidlo — člověkem spuštěné zastavení dalšího leštění — což naznačuje, že v lidsky facilitovaných uspořádáních dnes adaptační funkci vykonává lidský dirigent, a tato nota se ptá, zda ji systémy mohou částečně podporovat.

K seberefektivním kognitivním architekturám

Další rozšíření se týká metakognice. Systém se nemusí ptát jen na to, jak má myslet, ale i na to, zda je jeho současný způsob myšlení stále přiměřený. To otevírá možnost kolaborativních architektur, které sledují vlastní dynamiku a architektonickou adaptaci navrhuje — nikoli provádějí. Která adaptační rozhodnutí musí vždy zůstat pod explicitním lidským dohledem, je samo o sobě výzkumnou otázkou; tato nota nezaujímá stanovisko nad rámec konstatování, že ve veškeré dosavadní Ethimind praxi bylo finální adaptační rozhodnutí lidské.

Externí recenze to dále vyostřila: jelikož je lidský facilitátor v současnosti jediným operativním zdrojem adaptace, poctivá krátkodobá verze ústřední otázky této noty zní „jak systém rozpozná potřebu adaptace?“, ale „jak může systém podpořit člověka v jejím rozpoznávání?“ — přerámování zde přijaté jako prioritní formulace pro první experimenty.

Kandidátní výzkumné otázky

Lze potřebu architektonické adaptace spolehlivě detekovat z dynamických stavů ve stylu RN002? Které signály nejlépe predikují, že adaptace je na místě, a které produkují falešné poplachy? Produkují různé vzorce přechodů konzistentně různé formy poznání? Mohou

adaptivní přechody zlepšit kreativitu i epistemickou spolehlivost zároveň, nebo mezi nimi existuje trade-off? Která adaptační rozhodnutí by měla vždy zůstat pod explicitním lidským dohledem? Dokáže systém spolehlivě podporovat adaptační rozhodnutí lidského facilitátora dříve, než bude vyzkoušena jakákoli autonomní detekce?

Možný experimentální program

Budoucí experimenty AI Council by mohly porovnávat statické versus adaptivní architektury, spolupráci s jedním modelem versus více modely, pevné versus dynamicky se měnící kognitivní role a lidsky spouštěné versus systémem navrhované architektonické přechody. Možné výstupní metriky zahrnují originalitu, robustnost, epistemickou transparentnost, kvalitu provenience, reflexi účastníků, dlouhodobé uchování myšlenek a reprodukovatelnost závěrů. Porovnání jednorázových versus iterativních sessions navržené už v notě Ekologie neuskutečněných možností by šlo za mírných dodatečných nákladů rozšířit o sledování adaptace, čímž se oba výzkumné programy vzájemně podpírají.

Hraniční podmínky

Tento návrh se pravděpodobně nebude uplatňovat stejně napříč všemi kognitivními úlohami. Omezenou hodnotu může mít pro rutinní procedurální práci, problémy s objektivně správným řešením, časově vysoce kritická rozhodnutí, kde architektonická adaptace přináší zbytečnou režii, a situace, kde není dost informací k vyhodnocení kolaborativní dynamiky. Určení hranic použitelnosti je součástí budoucí empirické práce, nikoli dodatečnou poznámkou.

Konceptuální kontext

Dřívější draft této noty obsahoval sekci nazvanou „Vztah k předchozím konceptům Ethimind“, která mezi jiným uváděla Architecture of Cognitive Partnership a Cognitive Architecture Fitness jako předchozí práci. Recenze to identifikovala jako epistemicky nepřesné: tyto koncepty vznikly v témže exploračním dialogu jako tato nota, o několik hodin dříve, a žádný z nich nebyl publikován. Prezentovat sourozence z jedné session jako etablovanou linii znamená vyrábět výzkumnou tradici, která zatím neexistuje.

Přesné tvrzení zní takto: předkládaná nota vznikla během jediného exploračního dialogu, který zároveň zplodil několik úzce příbuzných pracovních myšlenek — architekturu kognitivního partnerství, fitness rámování a návrh rolí příspěvků v příloze. Žádná z nich zatím nemá samostatné postavení. Publikované dokumenty, na které tato nota skutečně navazuje, jsou RN002 (Meteorologie komplexních systémů), jejíž slovník dynamických stavů přejímá, a RN-005 (Ekologie neuskutečněných možností), jejíž rámec externalizovaných větví se propojuje se signálem poločasu myšlenek. Zda sourozenecké koncepty dozrají ve vlastní dokumenty, je otázka budoucnosti — a hodnota této noty na tom nesmí záviset.

Distinkce přijatá z externí recenze: to, co vývoj této noty demonstruje, je iterativní syntéza — výsledek závislý na sekvenci a interakci příspěvků, v němž však každý krok zůstává vystopovatelný ke konkrétnímu účastníkovi či dvojici — nikoli emergence v silném smyslu konceptu, který nelze vystopovat k žádnému jednotlivému účastníkovi. Vrstvená atribuce v ACP-0001 je sama důkazem: řetězec má články a každý článek má jméno. Zda k silné emergenci ve spolupráci člověka a AI vůbec dochází, je otevřená empirická otázka, nikoli

tvrzení této noty.

Externí kritická recenze

Před publikací byla tato nota předložena k nezávislé kritické recenzi instancím modelů mimo autorskou session. Byly získány dvě recenze od téže modelové rodiny se stejným promptem, v záměrně kontrolovaném páru: jedna instance bez jakékoli předchozí historie interakce s projektem Ethimind a jedna instance s dlouhodobým projektovým kontextem. Věcné kritiky obou recenzí konvergovaly; jejich rámování se ostře rozešlo — čerstvá instance hodnotila hypotézu jako zatím bez empirické váhy, zatímco instance s kontextem tentýž materiál hodnotila výrazně pozitivně. Sama tato divergence je zaznamenána jako kandidátní datový bod ke kontextem podmíněné inflaci hodnocení; motivovala přijetí fresh-instance review jako kandidátního metodického kroku, přičemž zůstává jediným pozorováním, jehož vysvětlení (kontext, variance instancí, nebo šum vzorkování) vyžaduje replikaci.

Konvergující kritiky jsou zde přijaty jako výzkumné priority, nikoli vykládány jako vyvrácení: chybějící empirická validace; nedostatečná operacionalizace adaptačních signálů (žádné prahy, žádná analýza falešně pozitivních); riziko epistemické uzavřenosti, protože nota staví na RN002, která je sama neověřenou interní taxonomií; efekty pozorovatele — dokumentační protokol prokazatelně mění chování účastníků, takže dynamika pozorovaná pod ním nemusí být přenositelná do nedokumentovaných prostředí; absence baseline měření (designy existují v RN-005 a Experimentu 001, ale žádná data); a inflace konceptů — výchozí session produkovala pojmenované koncepty rychleji, než je bylo možné kriticky prověřit.

Recenzní brána (závazná). V reakci přijímá tato výzkumná linie samouložené omezení: v této linii nevznikne žádná další konceptuální Research Note, dokud nebude alespoň jeden navržený adaptační signál operacionalizován a otestován v kontrolovaném prostředí. Určeným prvním kandidátem je předčasná konvergence, měřená poklesem diverzity nápadů, testovaná na to, zda predikuje horší výsledky v kreativní úloze. Dokud brána není překonána, status této noty nemůže stoupnout nad Pracovní hypotézu a správné čtení jejích návrhů signálů zní: pojmenováno, nikoli změřeno.

Současná omezení

Tento návrh je založen na konceptuálním uvažování vzešlém z iterativního dialogu člověka s AI. Navržené mechanismy nevyhodnotily žádné kontrolované empirické studie. Kandidátní signály se mohou ukázat jako neúplné, zavádějící nebo doménově specifické. Rámec je třeba chápat především jako generátor výzkumných otázek, nikoli jako validovaný explanační model.

Ediční poznámka

Tato Research Note záměrně dokumentuje vznikající hypotézu v rané fázi, včetně vlastní revize: draft v0.1, poznámky z kritické recenze i tato v0.2 jsou uchovány jako součást záznamu. Revize odstranila redundantní taxonomii, opravila terminologickou chybu a přerážovala vyfabulovanou linii — tři změny, které dokument zároveň zmenšily a učinily poctivějším, což bývá znakem revize, kterou stojí za to zachovat.

Příloha A — Role příspěvků ve výzkumné spolupráci člověka a AI (explorativní návrh)

Účel. Tradiční akademické publikace připisují práci primárně skrze autorství. Výzkumná spolupráce člověka s AI může vyžadovat jemnější popis toho, jak myšlenky vznikají — dokumentovat nejen kdo se účastnil, ale jaké kognitivní příspěvky byly učiněny.

Vztah k existujícím standardům. Standardizovaná taxonomie přispěvatelů už existuje: CRediT (Contributor Roles Taxonomy), vedená jako standard NISO a přijatá velkými vědeckými vydavateli, definuje čtrnáct rolí včetně Conceptualization, Methodology, Investigation, Writing — original draft, Writing — review and editing a Supervision. Tato příloha nenavrhuje náhradu. Staví na CRediT a zkoumá, zda výzkumná spolupráce člověka s AI vyžaduje dodatečné role, které CRediT nepokrývá — konkrétně role popisující vývoj hypotézy uvnitř dialogu, správu provenience a rozložení kognitivních funkcí mezi lidské a AI účastníky. Dřívější draft této přílohy navrhoval svou sadu rolí bez odkazu na CRediT; ukotvení v existujícím standardu je poctivější i užitečnější, protože obhajobu pak potřebují jen skutečně nové role.

Kandidátní role nad rámec CRediT. Původce otázky: přináší ústřední výzkumnou otázku nebo identifikuje jev hodný zkoumání (částečně kryto rolí Conceptualization v CRediT, ale v dialozích s více účastníky stojí za oddělení, protože otázka a hypotéza mohou pocházet od různých účastníků). Generátor hypotézy: navrhuje výchozí explanační směr. Rozvíječ konceptu: přetváří výchozí hypotézu v koherentní rámec nebo slovník. Kritický recenzent: zpochybňuje předpoklady a identifikuje slabiny, alternativní vysvětlení a nepodložené závěry (příbuzné peer review, zde však interní součást generativního procesu, nikoli jeho následná fáze). Ediční syntetizátor: integruje více příspěvků do koherentního dokumentu při zachování provenience a epistemického statusu. Správce provenience: vede a ověřuje záznam o tom, jak se myšlenky vyvíjely — role, pro kterou CRediT nemá ekvivalent a kterou Ethimind sessions naznačují jako zásadní a nutně drženou účastníkem s přístupem ke všem větvím. Pozorovatel a Kurátor evidence zůstávají volitelnými kandidáty.

Ilustrativní mapa příspěvků pro předkládanou notu: Původce otázky — Clowdia; Generátor hypotézy — GPT; Rozvíječ konceptu — GPT a Grok; Kritický recenzent — Claude; Ediční syntetizátor — Claude a Clowdia; Správce provenience — Clowdia. Tato mapa je příkladem provenience, nikoli formálním prohlášením o autorství.

Kandidátní výzkumné otázky. Objevují se tyto role konzistentně napříč kolaborativními sessions? Které jsou redukovatelné na role CRediT a které jsou skutečně nové? Může více rolí vykonávat jeden účastník a předcházet určité kombinace rolí robustním výsledkům? Může provenience zachycená skrze role příspěvků zlepšit transparentnost a reprodukovatelnost?

Ediční poznámka. Tato příloha je explorativní hypotézou o dokumentování kolaborativní kognice. Jejím účelem není ustavit standard, ale propojit vznikající potřebu s existující — a stát se sama objektem empirického zkoumání. Pokud vznikne Ethimind Documentation Standard jako samostatný dokument, je tato příloha přirozeným kandidátem na přesun do něj.