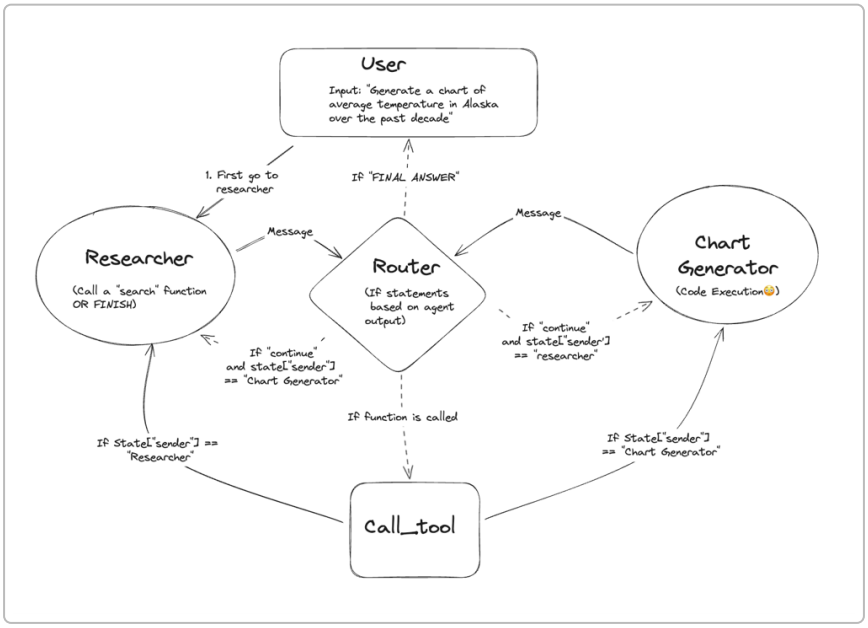
ChatGPT

Moduláris és decentralizált AI-ügynök architektúrák kormányzásban és döntéshozatalban

Összefoglaló: A moduláris vagy decentralizált AI-ügynök architektúrák lényege, hogy több különálló, nagy nyelvi modellek (LLM) által vezérelt "ügynök" együttműködve old meg komplex feladatokat. Az alábbiakban áttekintjük a legfrissebb nemzetközi példákat és kutatásokat ezen a téren, különös tekintettel a kormányzati és társadalmi döntéstámogatásra, valamint DAO-szerű (decentralizált autonóm szervezeti) rendszerekre. Kitérünk arra, hogyan integrálják a nyelvi modelleket több ügynökbe (pl. Reflexion, AutoGPT, OpenAGI), milyen döntéstámogató és szimulációs modulok léteznek, miként alkalmaznak adaptív tanulást (RLHF/RLAIF) a kormányzási célokra, hogyan vonható be a közösségi visszajelzés az AI tanításába, és bemutatunk nyílt forráskódú kezdeményezéseket és javasolt architektúra-típusokat.

Többügynökös nyelvi modellek integrációja



Egy egyszerű több-ügynökös workflow vázlata (LangGraph példa): külön LLM-alapú "Kutató" és "Diagram-készítő" ügynök együttműködik egy router komponensen keresztül a felhasználói feladat megoldásában.

A mai LLM-eket gyakran több autonóm modulként szervezik meg egy rendszerben annak érdekében, hogy összetett feladatokat hatékonyabban oldjanak meg. A **multi-agent** (többügynökös) megközelítés lényege, hogy a különálló LLM-alapú szereplők mind saját szerepet, képességeket vagy eszközöket kapnak, és kommunikálnak egymással egy közös cél érdekében [1] [2]. Ez a struktúra számos előnyt kínál: az egyes ügynökök specializált feladatokra fókuszálhatnak (pl. információkeresés, tervezés, végrehajtás), külön finomhangolt promptokat és példákat használhatnak, és az egész rendszer modulárisan fejleszthető és tesztelhető [2]. Ha egy monolitikus LLM egyedül nem boldogul egy komplex problémával, több együttműködő ügynök „társasága" jobb eredményt érhet el a munkamegosztás révén.

Reflexion (2023) egy korai példa a moduláris ügynök-koncepcióra, amelyben egy LLM-alapú agent belső komponensekre bontva, ciklikusan tanul a saját visszajelzéseiből. A Reflexion architektúrában elkülönül az **Actor**, az **Evaluator** és a **Self-Reflection** modul [3]. Az *Actor* (LLM) hajtja végre a lépéseket és javasol megoldást, a *Trajectory* (rövid távú memória) rögzíti a kimenetet, az *Evaluator* (ami lehet egy másik modell vagy szabályrendszer) értékeli az eredményt és jutalmaz vagy büntet, majd a *Self-reflection* modul egy nyelvi visszajelzést generál az LLM számára, amit az következő iterációban felhasznál a javulásra [3] [4]. Ez lényegében egy „verbális megerősítéses tanulás": az ügynök saját szavaival kap magyarázó visszajelzést a hibáiról, így finomhangolás nélkül, pusztán a memóriáján keresztül tanul az epizódok során [5] [4]. A Reflexion keretrendszerrel Shinn et al. (2023) látványos javulást értek el például döntési feladatokban és programkód-generálásban az ismétléses próbák során [6] [7].

Az utóbbi évben több *nyílt forráskódú projekt* is megjelent, amelyek lehetővé teszik, hogy hétköznapi fejlesztők is több LLM ügynököt állítsanak munkába egy cél érdekében. Az **AutoGPT** az egyik első ilyen nagy visszhangot kiváltó kísérlet volt 2023-ban. Ez egy nyílt forrású keretrendszer, amely GPT-4-re építve képes egy magas szintű felhasználói célból részfeladatokat lebontani és azokat önállóan végrehajtani, minimális emberi beavatkozással [8] [9]. Az AutoGPT 2023. március 30-án jelent meg Toran Bruce Richards (Significant Gravitas) fejlesztésében, és demonstrálta, hogy egy megfelelően illesztett LLM képes "gondolkodási láncot" (Chain-of-Thought) kialakítva, memóriát használva több lépéses feladatokat koordinálni [8]. Például a rendszer egy *"menedzser"* ügynökként megtervezi a teendőket, majd külön *"végrehajtó"* ügynököket indít az egyes részfeladatokra (keresés, számítás, stb.), és folyamatosan értékeli a haladást [10] [11]. Az AutoGPT-t sokan valódi többügynökös platformnak tekintik, hiszen "egy diverz autonóm ügynökökből álló csapatot" hoz létre egy adott cél elérésére [12]. Hasonló koncepcióval működik a BabyAGI, AgentGPT és számos további variáns – mind azt demonstrálják, hogy a LLM-ek képesek lehetnek egy primitív öntervezésre és végrehajtásra, ha hurokba szervezzük őket (azaz a saját outputjukat visszacsatoljuk új inputként).

OpenAGI (2023) nevű kezdeményezés egy kutatási platformot kínál a multi-step feladatok megoldására, ötvözve a nagyméretű nyelvi modellek általános képességeit szakosodott modellek vagy eszközök "szakértelmével" [13]. Az OpenAGI keretében a feladatot természetes nyelvű leírás formájában adjuk meg az ügynöknek, amely ezután automatikusan kiválasztja és meghívja a megfelelő "szakértő modult" (pl. egy matematikai modellt, keresőmotort vagy más AI komponenst) a probléma megoldásához [14]. A rendszer egyik újdonsága a *Reinforcement Learning from Task Feedback (RLTF)* mechanizmus – ez lényegében azt jelenti, hogy az ügynök a feladat eredményének visszajelzése alapján folyamatosan javít a saját teljesítményén, kialakítva egy önfejlesztő visszacsatolási hurkot [15] [16]. Az OpenAGI kódja nyílt forráskódú, így a közösség is kísérletezhet vele; a projektet Yingqiang Ge és munkatársai publikálták és a NeurIPS 2023 konferencián is bemutatták [17] [18].

Egy másik architekturális megoldás a **tervező-végrehajtó** (planner-worker) minta, ami több rendszerben is megjelenik. Az **OpenAGI** fent említett *TaskPlanner* modulja vagy az **AutoGPT** hasonló módon egy *Planner/Menedzser* ügynököt alkalmaz a feladatok lebontására, míg a konkrét teendőket specializált *Worker/Munkás* ügynökök hajtják végre [19] [20]. Hasonló logika figyelhető meg a Microsoft **AutoGen** keretrendszerében is: itt az ügynökök aszinkron üzenetküldéssel kommunikálnak, és egy *"Commander"* ügynök koordinálhat több *"Worker"* ügynököt. Az AutoGen v0.4 architektúrája eseményvezérelt és moduláris, támogatja a hosszú ideig futó, proaktív ügynököket, az egyes komponensek (modellek, eszközök, memória) könnyű cserélhetőségével [21] [22]. Fontos jellemzője az ilyen rendszereknek a naplózás és megfigyelhetőség: az AutoGen például beépített monitorozó és hibakereső eszközöket kínál az ügynök-interakciók követésére, és skálázható, akár elosztott ügynökhálózatokat is lehetővé tesz (sőt több programozási nyelven – Python, .NET – íródott ügynökök együttműködését is) [23] [24]. Mindez azt mutatja, hogy a többügynökös architektúrák nemcsak koncepcionálisan érdekesek, de technikailag is kezdenek kiforrott keretrendszerekké válni.

Döntéstámogatás, auditálhatóság és szimuláció több ügynökkel

A kormányzás és társadalmi döntéshozatal területén különösen ígéretes a több AI-ügynök együttműködése, mivel utánozhatják egy szakértői bizottság vagy tanács munkáját, és képesek lehetnek komplex szempontok mérlegelésére. Több friss kutatás foglalkozik azzal, hogyan **hozhatnak kollektív döntéseket** az LLM-alapú ügynökök, és hogyan tehetjük ezt a folyamatot megbízhatóbbá, átláthatóbbá.

Az egyik kihívás, hogy a legtöbb jelenlegi LLM-agent keretrendszerben a döntés folyamata gyakran leegyszerűsödik: vagy egy *„vezér”* ügynök (kvázi diktátor) hozza meg a végső döntést a többiek javaslatai alapján, vagy a többségi szavazás egy primitív formáját alkalmazzák (pl. a válaszok közül egyszerűen kiválasztják a legtöbbször előfordulót) [25] [26]. Zhao és munkatársai 2024-ben egy átfogó elemzést készítettek 52 különböző LLM-alapú multi-agent rendszerről, és megállapították, hogy a **kollektív döntéshozatal (Collective Decision Making – CDM)** módszerei meglehetősen egysíkúak voltak: vagy *“diktatórikus”* módon egy agent döntött, vagy egyszerű többségi szavazást (plurality voting) alkalmaztak, esetleg használtak valamilyen utilitariánus pontszám-összegzést [27] [28]. Ennek orvoslására a szerzők a közösségi döntéselmélet (social choice theory) bevált szavazási módszereit ültették át a gépi ügynökök világába. Bemutatták a **GEDI** (General Electoral Decision-making Interface) nevű modult, amely több preferenciális szavazási rendszert integrál a LLM ügynökök közé [29] [30]. A GEDI lehetővé teszi például rangsorolásos szavazás, Borda-számlálás, Condorcet-módszer stb. alkalmazását az ügynökök által javasolt opciók között (szemben a puszta többségi szavazattal). Eredményeik szerint bizonyos **ordintális szavazási módszerek** bevonása jelentősen javította az ügynökök következetes érvelését és a döntések robusztusságát a hagyományos módszerekhez képest [31] [32]. Külön figyelemre méltó, hogy már egész kis ügynökcsoporttal (akár 3-5 agent) is pozitív szinergia érhető el, tehát nincs szükség hatalmas “szavazóközönségre” ahhoz, hogy a kollektív intelligencia előnye megmutatkozzon [31]. A többügynökös szavazás további előnye, hogy nincs egyetlen hibapont: ha egy ügynök téved vagy elfogult, a többiek ellensúlyozhatják – ez növeli a rendszer megbízhatóságát az érzékeny döntésekben [31] [33].

Egy másik izgalmas megközelítés a **LLM ügynökök vitája vagy tárgyalása**. Itt nem egy formális szavazási szabályt alkalmaznak, hanem dialógus formájában több ügynök megpróbál konszenzusra jutni, érveket ütköztetni. Kutatók beszámoltak róla, hogy megfelelően konfigurálva a több-LLM “vita” kreatívabb és megalapozottabb megoldásokat szülhet bizonyos feladatokban, mint egy izolált modell. Például Liang et al. (2023) kimutatták, hogy egy többügynökös *debate* keretrendszerben a modellek divergensebb gondolkodásra képesek, így jobb eredmények születnek, mintha csak egyetlen modell próbálna megoldást találni [34] [35]. Hasonlóképpen, a **role-play** (szerepjáték) stratégiák – amelyeket pl. a CAMEL keretében használnak – lehetővé teszik, hogy az ügynökök különböző nézőpontokat képviseljenek: tipikus példa, amikor az egyik LLM *„felhasználó”* szerepben kéréseket fogalmaz meg, a másik LLM *„asszisztens”* szerepben megpróbálja teljesíteni azokat, és így iterálnak a megoldás felé [36] [37]. Ez tulajdonképpen szimulált ember–ember interakció, melynek során a modellek követik a beépített szerepük “intencióit”. A CAMEL projekt (Li et al., 2023) bemutatta, hogy ezzel a módszerrel humán beavatkozás nélkül is komplex feladatokat lehet végrehajtani, és közben értékes adatok gyűjthetők a modellek viselkedéséről egy *“LLM-társadalomban”* [36] [38]. A szerepjátékos kommunikáció egyfajta *szimulációs laborként* is felfogható, ahol a kutatók megfigyelhetik, hogyan alakulnak ki kooperatív viselkedések, munkamegosztás vagy épp konfliktusok a mesterséges ügynökök között.

A szimulációs képességek a társadalomtudományi és kormányzati alkalmazások szempontjából is ígéretesek. A Stanford Egyetem kutatói 2023-ban nagy visszhangot kiváltó kísérletben egy **virtuális kisváros lakóit modellezték 25 AI ügynökkel**, akik mind valószerű emberi karakterekként viselkedtek (napi rutint követtek, beszélgettek egymással stb.). Ez a *“Generative Agents”* tanulmány (Park et al., 2023) megmutatta, hogy az LLM-alapú ügynökök képesek konzisztens, hosszabb távú szimulációkat

fenntartani, és a társas interakciókban még váratlan emergens jelenségek is megfigyelhetők voltak [39] [40]. Ilyen szimulációkat a jövőben akár arra is lehet használni, hogy új kormányzati intézkedések vagy közpolitikai döntések társadalmi hatását előre "leteszthessük" mesterséges társadalmakban. Például egy AI-alapú közösségben kipróbálható, miként reagálnának a polgárok (ügynökök) egy új szabályozásra, vagy hogyan alakulna a véleménydinamika egy nyilvános vita során. Bár ezek egyelőre kísérleti fázisban lévő kutatások, jól mutatják a potenciált: a moduláris multi-agent rendszerek nemcsak végrehajtó eszközök, hanem *szimulációképes modellek* is, melyek a társadalmi döntések támogatására szolgálhatnak.

A *kormányzati döntéstámogatásban* az AI-ügynökök már most megjelennek prototípus formában. DAO-k (decentralizált autonóm szervezetek) esetén például felmerült, hogy bizonyos feladatokat bízzunk algoritmikus ügynökökre: ilyen lehet a javaslatok előszűrése, elemzése, sőt akár a szavazás automatizálása is. **Fatuma Yattani (2023)** egy esettanulmányokat is bemutató cikkében rámutatott, hogy a DAO-k küzdenek a javaslatok dömpingjével és a humán szavazók korlátaival – az AI ügynökök segíthetnek a javaslatok összegyűjtésében, rangsorolásában és kockázatelemzésében [41]. Például az Aave nevű DeFi-protokoll közössége AI eszközöket használ arra, hogy előre **szimulálják egy-egy javaslat hatását** a likviditási poolokra, mielőtt szavaznának róla [42]. A *MakerDAO* egy másik példa: itt egy "Governance AI" modult tesztelnek, amely javaslatot tesz bizonyos pénzügyi döntésekre (pl. fedezeti arányok módosítására), de az **emberi felhasználók utólagos jóváhagyása kötelező** – tehát egy hibrid modellt követnek, ahol az AI a döntéshozatal előkészítését és automatizált végrehajtását segíti, de a végső kontroll az embereknél marad [43]. Ez a gyakorlatban azt jelentheti, hogy a közösség tagjai meghatározhatnak az AI számára szabályokat (pl. "csak olyan javaslatot támogass, ami növeli a kincstár hozamát"), az AI ügynök pedig ennek megfelelően *delegált képviselőként* szavaz a rengeteg apró döntésben, tehermentesítve a humán résztvevőket [44] [45]. Ugyanakkor felmerül a kérdés: mi van, ha a botok többségbe kerülnek a DAO-ban? A konszenzus, hogy a közeljövőben az AI nem válthatja ki teljesen az embereket, viszont **kiterjesztheti a döntéshozók képességeit**. A legvalószínűbb forgatókönyv egy *augmentált döntéshozás*, ahol az AI adatelemző és racionalizáló gyorsaságát kombináljuk az emberek etikai és kreatív ítélőképességével [46] [47]. E kompromisszum jegyében több helyen javasolják az **"Explainable AI"** eszközök és auditálási mechanizmusok beépítését a DAO-kban használt ügynökök mellé – például nyilvános, on-chain napló vezetését arról, hogy az AI milyen indoklással jutott egy adott döntésre [48]. Mivel a blokklánc-transzparencia elvárt, izgalmas kutatási irány a modell-döntések magyarázhatósága és a "fekete doboz" hatás minimalizálása: egyes kísérletek szerint az AI ügynökök által hozott határozatokat utólag is lehet értelmezni, ha a rendszer eltárolja a kommunikációjukat, chain-of-thought lépéseiket vagy a szavazási preferenciáikat.

Összességében a többügynökös AI rendszerek a kormányzás terén **kollektív intelligenciát** ígérnek: akár a polgárok digitális képviselőiként (ahogy Jarrett et al. felveti, hogy minden embernek lehet egy AI "ügynöke", aki az ő preferenciái szerint szavaz [49]), akár szakértői bizottságként (amikor több specializált AI véleményét összesítjük), akár szimulációs eszközként (virtuális társadalmak modellezésére). A kihívások között van a transzparencia, az elszámoltathatóság és a beépített elfogultságok kezelése – de ezekre a moduláris felépítés részben választ adhat, hiszen **auditálható komponensekre** bontható a rendszer (pl. külön modul felel az etikai szűrésért, külön a jogi megfelelésért stb., és ezek működése külön-külön vizsgálható). A decentralizált architektúra pedig azt is jelentheti, hogy maga a modell futtatása vagy tanítása oszlik meg több szereplő között (pl. *federated learning* jelleggel több intézmény vagy közösség tanítja az egyes ügynököket a saját nézőpontjára, majd egyesítik a tudást egy konszenzusos döntésben). Ilyen irányú kísérletek már elindultak – pl. a SingularityNET és mások olyan platformokat építenek, ahol **AI ügynökök decentralizált hálózata** szolgálhat DAO-k vagy más szervezetek kiszolgálására anélkül, hogy egy centralizált AI-monopólium jönne létre [50] [47].

Adaptív tanulás és közösségi visszajelzés az ügynökrendszerekben

Mivel a kormányzási és társadalmi döntések esetén kulcsfontosságú az *értékalapú* működés, az AI ügynököknél előtérbe került a **reinforcement learning from human feedback (RLHF)** és újabban a **reinforcement learning from AI feedback (RLAIF)** alkalmazása. Az RLHF – amit az OpenAI a GPT-4 és korábbi modelljeinek finomhangolásánál is nagy sikerrel alkalmazott – arra épít, hogy emberi annotátorok értékelik az AI válaszait (pl. melyik megfogalmazás udvariasabb, melyik döntés etikusabb), és ezen visszajelzések alapján egy jutalmazó modellt tanítanak, amelyet aztán megerősítéses tanulással használunk az alapmodell irányítására. A kormányzásban ezt a módszert úgy is értelmezhetjük, mint egyfajta *"közösségi tanítás"*: az AI ügynökök betanításakor bevonjuk a polgárokat vagy szakértőket, hogy jelöljék meg a preferált döntéseket, válaszlehetőségeket. Így az ügynök nem vakon követ egy előre programozott szabályt, hanem **tanult egy értékrendet** a közösségtől.

Felismerték azonban, hogy az RLHF nem skálázódik könnyen (mert rengeteg emberi munka kell hozzá), ezért kísérleteznek az *RLAIF*-fel, ahol **maguk a modellek adnak visszajelzést más modelleknek**. Az Anthropic kutatói például a *"Constitutional AI"* megközelítéssel olyan rendszert építettek, ahol egy nyelvi modell egy előre rögzített "alkotmányos" elvrendszer alapján bírálja el a másik modell válaszait – tehát a humán etikát egy AI kritikus képviseli, és ez ad visszacsatolást a fő modellnek [51]. Ez is egy formája az RLAIF-nak, hiszen **AI által generált preferenciacímkéket** használunk fel a tanításhoz [52] [51]. A kezdeti eredmények biztatóak: egy 2023-as tanulmány szerint az RLAIF-kal finomhangolt modell teljesítménye felér a humán visszajelzéssel tréningezett modellével bizonyos feladatokban, ami azt sugallja, hogy ezzel a módszerrel részben kiváltható vagy kiegészíthető a drága emberi felügyelet [53] [54]. Természetesen a kormányzati alkalmazásokban óvatosnak kell lenni azzal kapcsolatban, hogy *milyen AI ad visszajelzést az AI-nak* – hiszen a felügyelet így áttételessé válik. Egy lehetséges kompromisszum, hogy a közösség alkot egy "alkotmányt" (elvek gyűjteményét), és ezt kódoljuk az AI kritikus moduljába, amely az RLHF folyamatot elvégzi. Így végső soron mégis a társadalom preferenciái érvényesülnek, csak éppen skálázhatóbb, automatizált formában.

Az adaptív tanulás másik aspektusa az, amikor **maguk az AI ügynökök tanítják egymást vagy saját magukat**. Erre láttunk példát már a Reflexion rendszerben (az ügynök saját korábbi hibáiból tanul nyelvi önkritika révén). További példaként említhető a **Self-Reflective Decoding** vagy a *"let's think step by step"* stílusú promptolás: ezek mind arra ösztönzik a modellt, hogy egyre jobb belső representációt alakítson ki a feladatról. Sőt, Liu és társai (2023) kísérleteztek azzal, hogy **szimulált közösségi interakciókon tréningeznek egy nyelvi modellt**, hogy az jobban igazodjon a társas normákhoz [55]. Konkrétan generáltak egy halom mesterséges párbeszédet különböző személyiségű LLM-ügynökök között, mintegy *szociális tanulási környezetet* teremtve, és ezen adatokkal finomhangolták a modellt. Az eredmény egy *"szociálisan hangolt"* AI lett, ami bizonyos esetekben jobban követett udvariassági, méltányossági elveket, mint az eredeti modell. Ez a megközelítés a *"közösségi tanítás"* egy sajátos formája – itt a közösség is mesterséges, de a cél az, hogy a modell magatartása közelebb kerüljön egy ideális közösség által elvárthoz.

Végül megemlítendő az **emberi visszajelzés folyamatos integrálása** a működő rendszerekbe. A kormányzati AI-ügynököket nem elég egyszer betanítani: fontos, hogy éles működés közben is lehessen korrigálni őket. Erre szolgálhatnak a *"human-in-the-loop"* mechanizmusok. Például az OpenAGI-nál is van lehetőség emberi beavatkozásra a tervezési fázisban [56], vagy az AutoGPT újabb változataiban beépítettek egy *felügyelt megerősítés* modult, ahol az ügynök néha megáll és rákérdez a felhasználóra (kvázi jóváhagyást kér vagy pontosítást). Hasonló módon a kormányzati ügynökök esetén képzelhető el egy *"AI auditor"* szerepkör: egy ügynök, ami figyeli a többi döntéseit, és ha bizonytalan vagy ellentmondásos helyzet adódik, kikéri egy emberi szakértő bizottság véleményét. Ilyen auditálható és

korrigálható keretrendszerek kidolgozása folyamatban van – például az IBM is publikált egy *AI Governance Evaluation* toolkit prototípust, ami nyomon követi az autonóm ügynökök döntéseit és ellenőrzi, hogy betartják-e a megadott etikai szabályokat.

Összefoglalva, az adaptív tanulás és visszajelzés kulcsszerepet kap abban, hogy a többügynökös rendszerek biztonságosan és az emberi értékekkel összhangban működjenek. A **RLHF** bevonja a közösség bölcsességét a tanításba, a **RLAIF** segít ennek skálázásában, a **szimulált közösségi tanulás** pedig új utakat nyit a modellfinomhangolásban. Ezek a technikák biztosítják, hogy egy AI-ügynökökre épülő kormányzati szisztéma nem merev automata, hanem folyamatosan tanuló és alkalmazkodó entitás legyen, amit végső soron az emberek formálnak.

Nyílt forrású kezdeményezések és ígéretes architektúrák

Végül nézzünk át néhány jelentős **nyílt forráskódú projektet és kutatási keretrendszert**, amelyek a moduláris vagy decentralizált AI-ügynök architektúrák élvonalát képviselik. Ezek a példák inspirációt nyújthatnak DAO-k, e-kormányzat vagy bármely közösségi döntéstámogató rendszer tervezéséhez is:

- **Reflexion (Shinn et al., 2023)** – Egy három komponensű (Actor–Evaluator–Reflector) LLM ügynök-architektúra, mely *verbális önmegerősítés* révén iteratívan fejleszti a teljesítményét. A Reflexion ügynök belső memóriában eltárolja a tapasztalatait, kiértékeli saját lépéseit, és nyelvi *önkritikát* fűz hozzájuk, amelyet a következő próbálkozásnál figyelembe vesz [3] [57]. Ezzel finomhangolás nélkül is jelentős javulást lehet elérni például összetett feladatok megoldásában (a ReAct láncgondolkodás kiterjesztéseként) [6].

- **AutoGPT (Significant Gravitas, 2023)** – Az első széles körben elterjedt autonóm agent keretrendszer GPT-4-re építve. Lehetővé teszi több *"mini-GPT"* ügynök létrehozását, amelyek egy magas szintű cél érdekében együtt dolgoznak. Az AutoGPT automatikusan bontja le a feladatokat, priorizálja őket és internetes eszközök segítségével hajtja végre a teendőket, emberi beavatkozás nélkül [8] [9]. Fejlesztője Toran Bruce Richards; 2023 tavaszán vált híressé a Twitteren bemutatott példák révén. Az AutoGPT bizonyította, hogy egy LLM képes proaktívan, *szereplők csapatát* imitálva működni, nem csak chatbottként [12].

- **OpenAGI (Y. Ge et al., 2023)** – Nyílt kutatási platform a többlépéses, valós feladatok megoldására. Integrálja a LLM-eket *domain specifikus* modellekkel, eszközökkel (plugins) egy közös keretbe [13]. Különlegessége a *Reinforcement Learning from Task Feedback (RLTF)* mechanizmus, melynek során a rendszer a feladat kimenetele alapján frissíti a döntéshozó LLM-et, egyfajta önfejlesztő ciklust alkotva [15]. A projektet a Rutgers Egyetem és a Huawei közös csapata jegyzi; kódja és részletes benchmarkjai nyilvánosak [58].

- **AgentVerse (Weize Chen et al., 2024)** – Egy többugyöngyös együttműködési keretrendszer, amelyet a pekingi Tsinghua Egyetem kutatói fejlesztettek és az ICLR 2024 konferencián mutattak be. Az AgentVerse lehetővé teszi sok különböző LLM-alapú *"szakértő ügynök"* dinamikus koordinálását egy közös cél érdekében [59]. A kutatók kimutatták, hogy az így összeállított ügynökcsoportok számos feladatban felülmúlják az egyedül dolgozó LLM-et – legyen szó szövegértésről, következtetésről, kódírásról vagy eszközhasználatról [59]. Érdekes módon megfigyeltek *emergens kollaboratív viselkedéseket* is: az ügynökök néha spontán "munkamegosztást" alakítottak ki, ami növelte a hatékonyságot [60]. Az AgentVerse kódját nyíltan elérhetővé tették a további multi-agent kutatás támogatására [61].

- **AutoGen (Microsoft, 2023)** – A Microsoft által fejlesztett nyílt forrású keretrendszer *agentic AI* alkalmazások építéséhez. Az AutoGen egy magas szintű API-t ad, amellyel könnyen létrehozhatunk együttműködő ügynököket. A 0.4-es verzióra teljesen újratervezték aszinkron, eseményvezérelt architektúrával [62] [21]. Fő jellemzői: ügynökök közti aszinkron üzenetküldés (egyszerre több párbeszédszál futhat), moduláris és bővíthető komponensek (saját eszközök, memóriarendszerek könnyen integrálhatók), beépített megfigyelhetőség (tracing, logging, OpenTelemetry támogatás) és skálázhatóság akár elosztott környezetben is [23] [22]. Támogat több nyelvet (Python, .NET), így vállalati környezetben is könnyen adaptálható. Az AutoGen lényegében a *"LLM-agents as a service"* vízióját valósítja meg.

- **CAMEL (Li et al., 2023)** – A *Communicative Agents for "Mind" Exploration* elnevezésű framework, amely az elsők között valósított meg **LLM-párok közti autonóm párbeszédet**. A CAMEL-ben egy *AI User* és egy *AI Assistant* ügynök kommunikál egymással, hogy megoldjon egy feladatot; induláskor *inception prompting* technikával mindkettő megkapja a szerepét és a feladat kontextusát [36] [37]. Ezután emberi beavatkozás nélkül folytatnak párbeszédet a megoldásig. A módszer kiválóan használható szintetikus adatok generálására is (hiszen két modell beszélgetéséből korlátlan mennyiségű QA vagy dialógus adat nyerhető). A CAMEL-projekt nyílt közösséget is épített (camel-ai.org), amely célul tűzte ki a *multi-agent rendszerek skálázási törvényeinek* vizsgálatát – vagyis azt kutatják, hogyan viselkedik a sok ügynökből álló "AI-társadalom", és hogyan javítható a teljesítmény ahogy nő az ügynökök száma [63] [64]. A keretrendszer dokumentációja és könyvtára elérhető, sőt olyan továbbfejlesztéseket is kínálnak, mint az OASIS (egymillió ügynökös szociális szimulációs platform) és különböző multi-agent benchmarkok a terület további standardizálására [65] [66].

- **GPT-Engineer (Anton Osika, 2023)** – Egy gyakorlati eszköz, amely az *LLM-agensek használatát a szoftverfejlesztésben* demonstrálja. A GPT-Engineer egy parancssori platform, ahol a fejlesztő leírja, milyen alkalmazást szeretne (specifikáció formájában), erre az AI ügynök **visszakérdez további pontosításokat**, majd generál egy komplett projektkódot a megadott igény alapján [67]. A rendszer mögött a GPT-4 áll, és a folyamata moduláris: van egy *tervezési/feltérképezési fázis* (kérdések feltevése a követelmények tisztázására), egy *kódgenerálási fázis*, majd igény szerint egy *tesztelési/refaktorálási fázis*. Bár egyetlen LLM vezérli, mégis tekinthetjük egy többkomponensű autonóm ügynöknek, hiszen különböző "szerepeket" játszik a folyamat során (analitikus kérdező, majd programozó, majd ellenőrző). A projekt nyílt forrású és gyorsan népszerűvé vált, jól mutatva az LLM-ek potenciálját az összetett, több lépéses feladatok (jelen esetben teljes szoftverek) automatikus előállításában.

- **LangChain – LangGraph (2024)** – A LangChain könyvtár kiterjesztése, amely kifejezetten *grafusalapú vezérlési folyamatokat* támogat LLM ügynökök között. A **LangGraph** lehetővé teszi, hogy az ügynököket gráf csomópontokként definiáljuk, az üzenetküldési és átadási logikát pedig élek reprezentálják [1] [68]. Így komplex workflow-kat lehet összerakni: pl. egy *felügyelő ügynök* csomópont eldönti, melyik specializált al-ügynökhöz irányítja a kérést, majd azok megoldásai valamilyen csatornán összegződnek. A LangGraph támogat hierarchikus ügynökcsapatokat is, ahol egy-egy csomópont mögött maga is egy teljes ügynökhálózat áll (rekurzív struktúra) [69] [70]. A LangChain blogja szerint a multi-agent megközelítésnek több előnye van: az egyes ügynökök kevesebb eszköz közül választanak (így fókuszáltabbak), minden ügynöknek saját promptját lehet finomhangolni, és a fejlesztés/tesztelés is könnyebb, mert modulonként vizsgálható a rendszer [2]. A LangGraph-ot már alkalmazták gyakorlati példákban is (pl. GPT-Newspaper – hírek több ügynökkel való feldolgozása, CrewAI – több ügynökös optimalizálás üzleti feladatra), amelyek megmutatták, hogy a grafus struktúra jól kezeli a ciklikus folyamatokat és a komplex logikát [71] [72].

- **“Quorum”-alapú szavazó ügynökök (2024)** – Bár nem konkrét szoftverkeretrendszer, de fontos kutatási irány. Ide tartozik Zhao et al. munkája a GEDI modullal, illetve általában véve azok a kísérletek, amelyek azt vizsgálják, **hány ügynök szükséges egy megbízható kollektív döntéshez** és hogyan érhető el konszenzus. Az első eredmények szerint meglepően kevés: akár 3-5 ügynök már elég lehet a *“wisdom of the crowd”* effektus kihasználásához, ha jól meg van választva a döntési mechanizmus [31]. A *quorum* itt arra utal, hogy bizonyos döntéseket csak akkor fogadunk el, ha legalább X számú autonóm ügynök egyetért. Ezzel elkerülhető, hogy egy magányos, esetleg tévedő ügynök befolyása alá kerüljön a rendszer. A témában született egy “Electoral Approach to Diversify LLM-based Multi-Agent Decision-Making” című tanulmány is, ami az ideális *voting quorum* kérdését járja körül – azaz, mi az optimális létszám és szavazási mód a legjobb kollektív intelligencia eléréséhez [73] [74]. Az ilyen kutatások nyomán a jövőben akár standardizált modulok jelenhetnek meg (pl. egy *“VotingAgent”* osztály, amit bármely ügynökcsapatba bedobhatunk, hogy biztosítsuk a demokratikus döntéshozatalt).

Összegzésképpen elmondható, hogy a moduláris, többügynökös AI architektúrák egyre nagyobb szerepet kapnak a kormányzás és közösségi döntéshozatal innovációiban. Az LLM-ek integrálása több autonóm egységbe lehetővé teszi a feladatok szétosztását és a szakértelem kombinálását, ami jobb teljesítményhez vezethet, mint egyetlen monolitikus modell esetén [59]. A döntéstámogató rendszerek auditálhatósága javul, hiszen a gondolatmenetek és szavazatok rögzíthetők és elemezhetők [48], a szimulációképes ügynöktársadalmak pedig kockázatmentes kísérleti terepet nyújtanak a társadalmi hatások vizsgálatához [39]. Az adaptív tanulási technikák – beleértve az emberi és mesterséges visszajelzéseket – biztosítják, hogy ezek az ügynökrendszerek idővel egyre inkább igazodjanak az emberi értékekhez és a közösségi elvárásokhoz [51] [49]. Végül, a nyílt forráskódú közösség pezsgése (projektek mint AutoGPT, LangChain, CAMEL, stb.) garantálja, hogy a legújabb kutatási eredmények gyorsan átültethetők gyakorlati prototípusokba is. A javasolt architektúrák skálája széles – a központosított tervező-végrehajtó modellektől a teljesen decentralizált, szavazás-alapú együttműködésig –, és valószínű, hogy a különböző kormányzási feladatokra más és más lesz az optimális megoldás. Az azonban már most látszik, hogy az **AI-agentek bevonása a kormányzásba** nem science fiction: megvannak az első prototípusok és keretrendszerek, amelyek segítségével *emberszerűen együttműködő, tanuló és a közjóért dolgozó* mesterséges ügynökök támogathatják a jövő döntéshozóit.

Források: Az állítások alátámasztására a válaszban hivatkozott források: Reflexion keretrendszer leírása [3] [4]; AutoGPT definíció (IBM) [8]; OpenAGI kutatási absztrakt [13]; AgentVerse (ICLR 2024) kivonat [59]; Microsoft AutoGen dokumentáció [23]; CAMEL paper absztrakt [36]; GPT-Engineer leírás [67]; LangChain blog (LangGraph előnyök) [2]; Zhao et al. (2024) GEDI absztrakt [31]; Jarrett et al. (NeurIPS 2023) workshop kivonat [49]; Fatuma Yattani (2023) cikk DAO és AI témában [41] [48]; illetve további hivatkozások a multi-agent rendszerek és RLHF/RLAIF témaköréből [51] [39]. E források részletesen bemutatják a fent említett példákat és kutatási eredményeket.

---

[1] [2] [68] [69] [70] [71] [72] LangGraph: Multi-Agent Workflows
https://blog.langchain.com/langgraph-multi-agent-workflows/

[3] [4] [5] [6] [7] [57] Reflexion | Prompt Engineering Guide
https://www.promptingguide.ai/techniques/reflexion

[8] [9] [10] [11] [12] What is AutoGPT? | IBM
https://www.ibm.com/think/topics/autogpt

[13] [14] [15] [16] [17] [18] [58] [2304.04370] OpenAGI: When LLM Meets Domain Experts
https://arxiv.org/abs/2304.04370

19 20 56 Autonomous Multi-Agent Architecture | OpenAGI 0.2.9.4 release | by Tarun Jain | AI Planet
https://medium.aiplanet.com/autonomous-multi-agent-architecture-openagi-0-2-9-4-release-170bfa7b5a9b?gi=ed066a8de0a7

21 22 23 24 62 AutoGen - Microsoft Research
https://www.microsoft.com/en-us/research/project/autogen/

25 26 27 28 29 30 31 32 33 34 35 39 40 49 55 74 An Electoral Approach to Diversify LLM-based Multi-Agent Collective Decision-Making
https://arxiv.org/html/2410.15168v1

36 37 38 [2303.17760] CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society
https://arxiv.org/abs/2303.17760

41 42 43 44 45 46 47 48 50 AI-Powered DAOs: Can Bots Outvote Humans in Governance? | by Fatuma Yattani | Medium
https://medium.com/@fyattani/ai-powered-daos-can-bots-outvote-humans-in-governance-0aac121fa02d

51 How Reinforcement Learning from AI Feedback works - AssemblyAI
https://assemblyai.com/blog/how-reinforcement-learning-from-ai-feedback-works

52 RLAIF: Scaling Reinforcement Learning from Human Feedback with ...
https://openreview.net/forum?id=AAxIs3D2ZZ

53 RLAIF vs. RLHF: Scaling Reinforcement Learning from Human ...
https://arxiv.org/abs/2309.00267

54 RLAIF: What is Reinforcement Learning From AI Feedback?
https://www.datacamp.com/blog/rlaif-reinforcement-learning-from-ai-feedback

59 60 61 AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors | OpenReview
https://openreview.net/forum?id=EHg5GDnyq1

63 64 65 66 CAMEL-AI Finding the Scaling Laws of Agents
https://www.camel-ai.org/

67 GPT Engineer - GitHub Pages
https://b2ktortechnik.github.io/gpt-engineer/

73 An Electoral Approach to Diversify LLM-based Multi-Agent ...
https://www.promptlayer.com/research-papers/an-electoral-approach-to-diversify-llm-based-multi-agent-collective-decision-making