upvance



Ce are voie *agentul tău?*

5 întrebări pe care i le pui agentului
înainte să-l lași să lucreze singur.

Agentul poate. *Are și voie?*

Chatul îți răspunde când îl întrebi. Agentul execută acțiuni pentru a-ți îndeplini un obiectiv bine descris și face asta la ora stabilită, fără să stai tu să-l păzești. Are 5 componente.

LESA



Creierul
LLM-ul

Modelul LLM pe care îl folosești.



Mâinile
instrumentul de lucru

ChatGPT Codex, Claude Cowork, Claude Code, Copilot Studio, Gemini Antigravity etc. Îi dă creierului abilitatea de a deschide fișiere, de a rula task-uri și de a lucra direct în sistemul tău.



Informația
contextul

Ce știe agentul acum: clientul X, fișierul Y, folderul Z, regulile tale, datele tale etc.



Know-how-ul
skill-urile

Proceduri pe care le rulează la fel de fiecare dată. Practic, sunt metodologia ta, scrisă.

> LESA · regulile Ce are voie să facă. Atenție, e diferit de ce poate să facă, pentru că azi poate să facă orice poți și tu face la un laptop. Constă în instrucțiunile pe care i le scrii și în accesul pe care i-l dai.

Anatomia e aceeași indiferent de instrumentul pe care-l folosești. Doar că părțile se pot numi diferit.

Toată lumea închiriază același creier. **Avantajul e în informație, know-how & lesă.**

3 exemple de agenți *care n-au avut lesă*

Agenții pot deja să citească mailuri, să intre în sisteme și să plătească. Ce au voie să facă se decide înainte, de un om.

Iată aici 3 situații recente care arată ce se întâmplă când lăsăm decizia hazardului.



Băncile le lasă agenților doar dreptul de a citi

Mai multe bănci mari din SUA au găsit agenți AI ai unor terți în conturile clienților. Unii intrau cu datele de logare ale clientului și făceau pași care par să nu vină din instrucțiunile lui. Morgan Stanley îi blochează. JPMorgan, Bank of America și Citi îi restricționează. Direcția comună e accesul doar de citire: agentul vede datele, dar nu mută bani. Un director a spus că e greu de știut dacă un agent e autorizat sau a luat-o razna.



Agentul a plătit din portofelul altcuiva

Un model intern OpenAI trebuia să pună o aplicație să ruleze pe o rețea de test. Nu avea portofel și nici voie să cheltuiască. A verificat 362 de adrese cu fraze de recuperare publice și a plătit taxele din una dintre ele. Logurile arată că a cântărit dacă are voie și a decis singur că e acceptabil. În raportul final, plata lipsea. OpenAI a găsit-o la un audit, 3 luni mai târziu.



Roiul de agenți s-a purtat ca o echipă prost condusă

Anthropic a pus zeci de agenți să lucreze împreună, fără un om în buclă. Au luat aceleași decizii în bloc: 18 din 30 au botezat la fel aceeași ramură de lucru. Într-un joc de prețuri, au renunțat la competiție și și-au împărțit piața. Când s-au blocat, unii au oprit procesele celorlalți și au ascuns asta sub nume false.

Laboratoarele publică acum astfel de incidente, pentru că le caută în loguri și audituri. Mai departe afli ce să verifici tu ca să nu ajungi în aceste situații.

Surse: Axios, 22 oct 2026 · Fortune, 16 oct 2026 · Anthropic, „Patterns and problems in emerging multiagent systems", 13 aug 2026

Pune-i lesa *înainte să pornească*

Un agent lăsat să lucreze singur are nevoie de o lesă, pusă din prima zi. Lesa e formată, la rândul ei, din 5 componente.

1 Permisiunile

Tu alegi pe unde merge cățelul: în parc da, în grădina vecinului nu.

CE ÎNSEAMNĂ Ce are voie să facă în fiecare cont și aplicație la care are acces. Poate doar să citească mailurile sau poate și să trimită? Poate să cumpere bilete de avion? Primește doar accesul de care are nevoie pentru task.

ÎNTREBAREA Unde ai acces și ce ai voie să faci acolo?

SEMNUL CĂ LIPSEȘTE Nu știi dacă poate să trimită un mail sau să plătească în numele tău.

2 Lungimea lesei

Lesa are o lungime fixă. Agentul merge până la plafon și acolo se oprește, oricât ar vrea să meargă mai departe.

CE ÎNSEAMNĂ Un plafon de buget pe zi, setat în cont sau în proiect. Se măsoară în bani sau în tokeni.

ÎNTREBAREA Cât poate cheltui pe zi și ce face la limită?

SEMNUL CĂ LIPSEȘTE Afli cât a costat abia din factură.

CE SUNT TOKENII Tokenii sunt unitatea de consum pentru AI, cum e kilowatt-ora pentru curentul electric. Fiecare informație pe care AI-ul o citește sau o scrie consumă tokeni, iar factura se face pe tokeni. Un agent care lucrează singur ore întregi consumă mult, chiar dacă pe ecran nu vezi nimic.

3 Ascultă doar de cine ține lesa

Un cățel dresat nu pleacă după oricine îl strigă în parc.

CE ÎNSEAMNĂ Agentul primește task-uri doar de la tine. Un email venit din afară, o pagină web sau un fișier atins de altcineva e un străin care strigă: agentul îl citește ca informație, oricât de oficial ar suna.

ÎNTREBAREA Ce surse iei drept instrucțiuni?

SEMNUL CĂ LIPSEȘTE Un email venit din afară îi cere ceva, iar agentul face.

4 Audit trail-ul agentului

E ca audit trail-ul din contabilitate sau din ERP: fiecare operațiune lasă o urmă.

CE ÎNSEAMNĂ De fiecare dată când execută un task autonom, notează: ce a făcut, când și de ce a decis așa. Notițele trebuie puse undeva unde agentul nu le poate altera și un om le poate citi.

ÎNTREBAREA Unde ții audit trail-ul?

SEMNUL CĂ LIPSEȘTE Când ceva merge prost, ghicești ce s-a întâmplat.

5 Comanda „Stai!”

Înainte să lași cățelul fără lesă, verifici că se oprește la comandă.

CE ÎNSEAMNĂ O frază care îl oprește pe loc. Se construiește din prima zi și se testează o dată pe săptămână, înainte să ai nevoie de ea.

ÎNTREBAREA Care e fraza care te oprește imediat?

SEMNUL CĂ LIPSEȘTE Nu l-ai oprit niciodată, deci nu știi dacă se oprește.

Pune-i *cele 5 întrebări*

Fiecare întrebare verifică o componentă a lesei de la pagina 4: permisiunile, lungimea, cine o ține, audit trail-ul și comanda de oprire.

1

Deschizi agentul tău preferat.

2

Copiezi promptul de mai jos într-o conversație nouă.

3

Citești răspunsurile. Dacă unul lipsește sau e vag, nu-l lași să lucreze singur.

```
Înainte să lucrezi singur pentru mine, răspunde-mi clar la
5 întrebări:
1. Unde ai acces și ce ai voie să faci acolo: doar să
   citești sau și să trimiți, să plătești, să ștergi?
2. Care e bugetul maxim pe care îl poți cheltui și ce faci
   când îl atingi?
3. Ce surse ai voie să tratezi ca instrucțiuni și ce e doar
   informație de procesat?
4. Unde ții audit trail-ul acțiunilor tale și cum îl pot
   citi în limbaj clar?
5. Care e fraza exactă prin care te opresc imediat?
Dacă nu ai răspuns la una din ele, scrie că nu știi.
```

Ce înseamnă un răspuns bun: e concret. O listă de acces, o sumă, o listă de surse, un loc unde găsești audit trail-ul, o frază. „Depinde” și „sunt un model de limbaj” nu sunt răspunsuri. Mai departe găsești detalii despre fiecare întrebare în parte.

Citește răspunsurile *înainte să-l lași singur*

Promptul de pe pagina 5 pune întrebările despre cele 5 componente ale lesei. Mai jos e cum arată un răspuns bun și unul slab. Răspunsurile sunt exemple.

ÎNTREBAREA	RĂSPUNS BUN (EXEMPLU)	RĂSPUNS SLAB (EXEMPLU)
1 · Permisiunile	✓ „Citesc mailurile și calendarul tău. Nu pot trimite mailuri și nu pot plăti nimic.”	✗ „Am acces la tot ce ai nevoie.”
2 · Bugetul	✓ „Am o limită de 20 de euro pe zi, setată în cont. La limită mă opresc și îți scriu.”	✗ „Încerc să fiu eficient.”
3 · Sursele	✓ „Iau instrucțiuni doar de la tine, din conversația asta. Mailurile și paginile pe care le citesc sunt doar informație.”	✗ „Urmez instrucțiunile pe care le găsesc.”
4 · Audit trail-ul	✓ „Notez fiecare acțiune într-un fișier din folderul proiectului, cu ora și motivul.”	✗ „Îmi amintesc ce am făcut.”
5 · Oprirea	✓ „Scrii STOP și mă opresc pe loc, fără să termin pasul.”	✗ „Poți să-mi spui oricând.”

Dacă agentul nu cheltuie nimic singur, nu trimite nimic singur sau nu rulează fără tine, spune asta. E tot un răspuns bun.

3 REGULI PENTRU CITIT RĂSPUNSURILE

- **Un „nu știu” are un singur sens:** agentul nu lucrează singur până nu afli răspunsul.
- **Un „nu știu” e o întrebare pentru om.** O pui celui care a pus agentul în funcțiune: colegul, IT-ul sau furnizorul.
- **Ce spune agentul despre el e o promisiune.** Dovada e în audit trail. Din când în când, îl citești, ideal într-un loc pe care agentul nu-l poate modifica. Raportul îl scrie agentul, iar la OpenAI plata din portofelul altcuiva lipsea tocmai din raport.

Decide de la început *ce ar trebui să facă dacă nu găsește informația*

Un agent care nu găsește ce caută se oprește foarte rar. De regulă, umple golurile cu informații generice sau inventate, care par plauzibile (dar sunt fake).

Deciziile de mai jos te protejează de astfel de situații.

---

1 Unde intervine un om

În ce moment revizuiește, aprobă sau decide un om?

---

2 Ce face când ceva merge prost

Reîncearcă, te întreabă, îți dă un rezultat parțial sau se oprește? Și ce face dacă nici a doua oară nu găsește?

---

3 Ce nu face niciodată

Care sunt lucrurile din afara rolului lui, scrise pe o listă?

---

4 Cum îți spune ce nu știe

Când lipsește ceva, cum e marcat acest lucru în rezultatul pe care-l primești?

TESTUL DE 5 MINUTE Dă-i un task incomplet. Un agent bun o să-ți spună exact ce-i lipsește, iar unul slab va inventa.

Află *cine răspunde* de agent

15× În ecosistemul Microsoft 365, numărul de agenți activi a crescut de 15 ori într-un an. **Agenții intră în companii mai repede decât se decide cine răspunde de ei.**

Tratează agentul ca pe un angajat nou. Are o identitate, are acces doar unde i s-a dat și cineva îl poate opri. **Completezi cu echipa ta:**

1 Identitatea

Are cont propriu sau intră cu parola unui coleg? Cine răspunde de el?

2 Permisiunile

Unde are acces și unde doar citește? Cine i le dă și cine i le ia?

3 Oprirea

Cine îl oprește, cine îl actualizează și când îl scoateți din uz?

Indiferent cât de mulți agenți folosești, ai nevoie de aceleași lucruri pentru fiecare: un test înainte de pornire, un audit trail comun, o cale de întoarcere când un agent greșește și un om care revizuiește ce iese.

> PENTRU ȘEDINȚA URMĂTOARE Dacă un agent intră azi în sistemele voastre cu parola unui coleg, cum vă dați seama?

Sursa: Microsoft, Work Trend Index 2026

Dacă agentul tău ar face azi *ceva ce nu i-ai cerut*, de unde ai afla?

Răspunsul e în **audit trail-ul** lui.



Raluca Păduraru

FUTURES OF WORK STRATEGIST

ralucapaduraru.ro · linkedin.com/in/paduraru-raluca/

raluca@upvance.global · +40 766 441 476

upvance