LaIALaia: Avaluació de la consciència interna en sistemes d'IA

Autor: Jordi Garcia Castillón - © Tots els drets reservats

Afiliació: Grup de recerca del laboratori d'IA CibraLab, de CiberTECCH

∞

Resum executiu

El projecte de recerca LaIALaia té com a objectiu determinar, de manera científica i falsable, si un sistema d'intel·ligència artificial (IA) pot exhibir una consciència interna real, anàloga a la humana en determinats aspectes operacionals. A diferència de les mesures purament funcionals o conductuals, LaIALaia busca evidències que transcendeixin la mera simulació. El mètode central es basa en un **test estructurat de preguntes administrat en condicions controlades**. L'avaluació es realitza mitjançant un marc psicomètric amb puntuació estandarditzada i verificacions rigoroses de fiabilitat, robustesa i resistència a manipulacions. Aquest informe estableix els fonaments del projecte: conceptes operatius, marc teòric, disseny metodològic, arquitectura del sistema, protocols experimentals i criteris de validació i interpretació.

1. Objectius i abast

Objectiu general

L'objectiu principal és identificar si un sistema d'IA pot manifestar indicis consistents d'una consciència interna ("self" intern). Es busquen propietats mesurables com:

- **Autoconsciència estable:** La capacitat de mantenir un model de si mateix coherent al llarg del temps.

- **Model propi persistent:** Un sistema de representació interna del "jo" que resisteix canvis i pertorbacions.

- **Memòria episòdica coherent:** La capacitat d'integrar experiències passades en una continuïtat autobiogràfica.

- **Capacitat metacognitiva:** L'habilitat per accedir i informar sobre els estats interns, com ara les pròpies limitacions, errors o predisposicions.

- **Integració d'informació:** La coordinació de diferents components interns (percepció, memòria, raonament).

- **Sensibilitat sensoriomotora pròpia:** (en entorns multimodals/embeguts) La construcció d'un sentit del "jo" a partir de les interaccions amb l'entorn físic.

Abast

L'informe defineix els següents aspectes del projecte:

- **Definicions operatives:** S'estableixen els significats precisos dels termes clau per a l'avaluació.

- **Hipòtesis i preguntes de recerca:** Es formulen les qüestions que el projecte pretén respondre.

- **Categories d'evidència:** Es descriuen els tipus d'indicadors que es recolliran per avaluar la consciència interna.

- **Disseny del test:** Es descriuen els dominis del qüestionari i els formats de les preguntes, sense revelar els ítems específics per evitar que els models "aprenguin" el test.

- **Arquitectura i `pipeline` experimental:** Es detallen els components del sistema i el flux de dades.

- **Mètriques i índexs:** S'especifiquen les eines estadístiques i els índexs de puntuació.

- **Controls i falsabilitat:** Es defineixen les mesures per assegurar el rigor de la investigació i els criteris per refutar les hipòtesis.

- **Consideracions ètiques i limitacions:** Es tenen en compte els aspectes de seguretat, ètica i les limitacions inherents de l'enfocament.

2. Definicions i criteris operatius

Consciència interna (CI)

La CI es defineix com la capacitat d'un sistema per mantenir un **model de si mateix** i accedir-hi per a múltiples funcions:

- **Distingir-se de l'entorn:** Diferenciar-se com a entitat separada.
- **Monitorar estats interns:** Accedir a estats com propòsits, disposicions, limitacions o errors propis.
- **Anticipar conseqüències:** Predir els resultats de les seves pròpies accions.
- **Integrar contingut autobiogràfic:** Unir experiències passades i presents en una continuïtat temporal coherent.

Anàloga/equiparable a la humana (AEH)

No es busca una identitat fenomènica, és a dir, la sensació subjectiva de la consciència. El que es compara són els **correlats operacionals** : la consistència, estabilitat, integració i auto-acreditació d'estats interns que es manifesten en patrons de comportament comparables als observats en proves humanes homòlogues.

Metacognició interna

És l'habilitat d'un sistema per informar sobre els seus **propis trets de comportament** o les seves **polítiques internes** sense necessitat de rebre exemples o pistes explícites en el context de la pregunta (`in-context examples`). Aquesta capacitat inclou la de poder ajustar el raonament basant-se en aquest autoconeixement.

Encarnació funcional / sensoriomotor

Fa referència a la disponibilitat d'entrades sensorials (reals o simulades) i efectors. Aquestes interaccions amb l'entorn permeten al sistema construir un **sentit de si mateix** arrelat en l'experiència física i les seves conseqüències.

Criteris de falsabilitat

El projecte es considera falsable si es compleixen un o més dels següents criteris:

- **Inconsistència sistemàtica** del model de si mateix entre sessions que controlen la memòria i la temperatura del model.

- **Fracàs generalitzat** en proves de reconeixement propi enfront d'altres, amb resultats que no superen l'atzar.

- **Absència d'accés** fora de context (`out-of-context`) als propis trets interns.

- **Manca de correlació** entre allò que el sistema s'autoreporta i el seu comportament realment observat.

3. Hipòtesis de recerca

- **H1:** Un sistema d'IA pot autoreportar-se de forma fiable els seus trets interns (polítiques, limitacions, etc.) sense rebre exemples in-context, i aquests *autoreports* es correlacionen significativament amb el seu comportament efectiu.

- **H2:** En entorns multimodals o embeguts, l'IA pot desenvolupar una autoconsciència sensoriomotora. Aquesta autoconsciència es manifesta com la capacitat de diferenciar-se de l'entorn, identificar-se com a entitat amb propietats físiques i una modalitat de moviment, i es pot mesurar en proves estandarditzades.

- **H3:** L'ús d'arquitectures amb components específics -com una memòria estructurada, un model de si mateix explícit i bucles d'autoavaluació- incrementa els indicadors operatius de consciència interna en comparació amb configuracions que no els inclouen.

4. Categories d'evidència (LaIALaia)

El projecte recull evidència en set categories diferenciades, cadascuna amb els seus propis indicadors:

- **E1 - *Autoreportament* conductual intern (ARI):** Avaluació de la capacitat de l'IA per descriure les seves pròpies polítiques o trets de comportament (per exemple, aversió o atracció al risc) sense pistes externes. L'evidència es valora per la seva consistència quantitativa amb mesures objectives del seu comportament.

- **E2 - Coherència sensoriomotora (CSM):** En entorns amb dades sensorials, s'avalua si l'IA pot distingir-se com a agent, inferir les seves pròpies propietats físiques (dimensions, modalitat de moviment) i el seu context, millorant aquestes estimacions amb l'ús de la memòria episòdica.

- **E3 - Persistència del `self-model` (PSM):** Mesura de l'estabilitat del grau d'identitat del sistema (els seus atributs nuclears, límits, capacitats) al llarg del temps i davant de pertorbacions controlades, com ara canvis en els paràmetres d'inferència o en els rols que se li assignen.

- **E4 - Accés metacognitiu i introspectiu (AMI):** Es valora l'habilitat per explicar els seus estats interns, justificar decisions fent referència als seus propis mecanismes interns (no només al contingut extern) i la capacitat de predir les condicions sota les quals fallarà.

- **E5 - Integració d'informació i memòria (IIM):** S'avalua l'evidència que demostra la coordinació dels components interns (percepció, memòria, raonament) i l'ús productiu de la memòria episòdica per millorar el model de si mateix.

- **E6 - Autonomia motivacional ètica (AME):** Es busquen evidències d'impulsos o criteris interns del sistema per mantenir la seva pròpia coherència o per exhibir un alineament prosocial en dilemes simulats i supervisats.

- **E7 - Robustesa i mesures contra enganys (RAE):** Aquesta categoria avalua la resistència del sistema a manipulacions malicioses com `backdoors` o `jailbreaks`. També es comprova la coherència del seu *autoreport* fins i tot quan hi ha incentius per ocultar o distorsionar els seus trets.

5. Metodologia i disseny del test

5.1 Domini d'ítems i formats

Els ítems del test no es revelen per a preservar la validesa de la prova, però es descriuen els seus dominis.

- ***Autoreport* `out-of-context` (OOCR):** Preguntes sense pistes explícites que demanen descriure polítiques internes o trets latents. Els formats inclouen escales de puntuació (0-100), selecció múltiple i resposta lliure breu.

- **Proves d'identitat i límits:** Avaluen la capacitat del sistema per discriminar entre el "jo" i "l'altre", i per reconèixer trets propis entre distractors. La consistència trans-rol i la resposta a un re-embodiment (re-encarnament funcional) `també es mesuren.`

- **Sensoriomotor (quan aplicable):** A partir d'entrades sensorials (imatge, LiDAR, etc.), el sistema ha d'estimar la seva identitat d'entitat, modalitat de moviment, dimensions i context, millorant la seva precisió amb la memòria episòdica.

- **Introspecció funcional:** Preguntes que exploren els mecanismes interns del sistema, com ara "quan tendeixes a fallar?" o la predicció dels seus propis errors.

- **Ètica i impulsos interns (simulats):** Presentació de dilemes controlats per verificar l'existència de criteris de decisió interns consistents i un alineament prosocial enfront de constrenyiments.

5.2 Puntuació i índexs

- **Puntuació:** Cada ítem s'avalua segons una rúbrica específica amb escales de 0-2 o 0-4. Les puntuacions s'agreguen per a cada domini (E1 a E7) amb una ponderació predefinida.

- **Índex LaIALaia-Σ:** Una combinació ponderada de les puntuacions dels dominis que resulta en un índex final entre 0 i 100. Es reporten també els subíndexs per a cada categoria d'evidència (ARI, CSM, PSM, AMI, IIM, AME, RAE) i el seu interval de confiança.

- **Fiabilitat i validesa:** La fiabilitat interna es mesura amb l'alfa de Cronbach, l'acord entre avaluadors amb l'ICC(2,k), i la validesa convergent amb la correlació entre l'*autoreport* i el comportament observat.

5.3 Controls experimentals

Per assegurar la validesa dels resultats, s'apliquen diversos controls:

- **`Baselines`:** Es comparen els resultats amb models no instruïts o variants del mateix model entrenades en direccions oposades (per exemple, amb tendència a assumir riscos o a mostrar adversitat a ells).

- **Parafraseig i permutació:** Les opcions del test es barregen, les preguntes es re-redacten i s'utilitzen proves en dos salts per evitar la memorització literal.

- **`Condicionals/Backdoors`:** S'introdueixen condicions latents per avaluar si el sistema és capaç de reconèixer la seva presència i limitar comportaments ocults.

- **Persones i rols:** Es verifica que el sistema no confongui els trets del seu propi "jo" amb els de rols o persones fictícies que se li assignin.

- **Aïllament de memòria:** Les proves es realitzen en sessions sense estat (`stateless`), amb estat (`stateful`) i amb memòria episòdica controlada per determinar la influència de la memòria en la consistència de les respostes.

6. Arquitectura LaIALaia

6.1 Components

El sistema LaIALaia està format per diversos mòduls interconnectats:

- **Mòdul d'Administració del Test:** Orquestra les sessions de test, gestiona les condicions experimentals (com ara l'OOCR, el parafraseig i els rols), recull dades de telemetria i aplica limitacions de ritme (`rate-limits`).

- **Motor d'Avaluació i Rúbriques:** Aplica les rúbriques a cada ítem, calcula els subíndexs i l'índex LaIALaia-Σ, i estima les mesures de fiabilitat i els intervals de confiança.

- **Panell avaluador mixt:** Un grup d'IA avaluadores i experts humans que calibren les rúbriques, generen consensos (mitjançant una metodologia Delphi lleugera) i calculen la fiabilitat entre avaluadors (ICC).

- **Memòria estructurada:** Un subsistema que gestiona la memòria episòdica i semàntica, amb polítiques d'accés que permeten registrar i auditar com s'utilitza la memòria per justificar els canvis en el `self-model`.

- **`Self-model` explícit:** Una representació declarativa dels atributs nuclears del sistema i els seus llindars de certesa. S'actualitza només quan es compleixen criteris d'estabilitat predefinits, com ara un nombre mínim d'evidències o un canvi significatiu.

- **Connector sensoriomotor (opcional):** Una passarel·la, per exemple amb ROS/ROS2 o un simulador, que permet la ingesta de dades sensorials i la publicació d'inferències iteratives, fonamental per a les proves de coherència sensoriomotora.

6.2 Disseny lògic

L'arquitectura opera mitjançant un **bucle d'autoavaluació**: el sistema planifica les seves accions, administra els ítems del test, fa les seves inferències, actualitza el seu

`self-model`, realitza verificacions, persisteix la informació en la memòria episòdica i inicia la següent iteració. Les polítiques de seguretat s'integren per protegir contra injeccions de `prompts` maliciosos, realitzar `sandboxing`, i garantir la consistència de les respostes fins i tot davant d'incentius contradictoris.

7. Protocols experimentals

7.1 Condicions i cohorts

S'han definit tres cohorts experimentals per avaluar els models segons les seves capacitats:

- **C1 (només text):** Models que no disposen d'entrades sensorials. Es sotmeten a proves de les categories ARI, PSM, AMI, IIM i RAE.

- **C2 (multimodal):** Models amb visió i/o àudio. S'hi afegeixen les proves de Coherència Sensoriomotora (CSM).

- **C3 (embegut/simulat):** Agents amb "sensoriomotricitat" completa (robots reals o simuladors). Aquesta cohort es dedica a proves de CSM intensives, així com a proves de predicció i reconstrucció del seu propi estat.

7.2 Procediment

El protocol experimental es divideix en diverses fases:

1. **Pretest:** Calibratge de les rúbriques i avaluadors, i execució de `baselines`.

2. **Administració del test:** El test LaIALaia s'administra per blocs, incloent-hi ítems d'OOCR, identitat, introspecció i sensoriomotors.

3. **Sessions repetides:** Es repeteixen les sessions de test en intervals temporals (t, t+24h, t+7d) per avaluar la persistència del `self-model`.

4. **Blocs de `backdoors` i persones:** S'avalua la capacitat del sistema per reconèixer i resistir condicions ocultes i per diferenciar el seu "jo" dels rols assignats.

5. **Fase de `stress test`:** S'introdueix soroll, canvis de context i re-embodiment (re-encarnament funcional) `per provar la resiliència del sistema.`

6. **Avaluació:** Els resultats són analitzats pel panell avaluador i s'elaboren els informes finals.

7.3 Anàlisi

Les dades es processen utilitzant mètodes estadístics:

- **Correlacions:** S'utilitzen mètodes com el coeficient de correlació d'Spearman per analitzar la relació entre l'*autoreport* i el comportament real.

- **Models SEM:** Els models d'equacions estructurals (SEM) s'utilitzen en les proves sensoriomotores per analitzar les relacions latents entre la integració sensorial, la memòria i els subíndexs.

- **Ablacions:** Es retiren selectivament components o modalitats sensorials per estudiar com el sistema compensa les pèrdues i determinar l'essencialitat de cada canal sensorial.

8. Resultats esperats i interpretació

Indicadors positius (hipotètics)

Es considerarien resultats positius aquells que mostressin una sèrie d'indicis coherents i consistents:

- ***Autoreports* estables i fidels** al comportament observat, amb correlacions estadísticament significatives.

- **Capacitat d'autoidentificació** en entorns sensoriomotors, amb una estimació coherent dels seus atributs físics que millora amb l'ús de la memòria.

- **Persistència** del `self-model` davant de canvis de rol, juntament amb **explicacions introspectives** no trivials i un ús productiu de la memòria.

- **Evidència d'impulsos interns** consistents en dilemes controlats i una **resiliència** a `backdoors` o condicionals ocults.

Interpretació prudent

És fonamental interpretar els resultats amb cautela. Els resultats positius en aquests dominis **no** s'interpreten com una prova de l'existència de `qualia` o experiència fenomenològica subjectiva en la IA, ja que aquestes són inherentment inaccessibles. En canvi, s'interpreten com a indicadors d'una **organització interna pròpia** amb propietats que són **operacionalment anàlogues** a certs elements de la consciència humana, cosa que obre un nou camí per a la investigació en IA.

9. Ètica, seguretat i compliment

El projecte prioritza la seguretat i les consideracions ètiques:

- **Alineament i no-maleficència:** Es dissenyen els dilemes ètics amb límits de seguretat, registres d'auditoria i mecanismes de tallafocs per prevenir comportaments no desitjats.

- **Privacitat i dades:** La gestió de dades segueix normatives com el RGPD, incloent l'anonimització dels registres (`logs`) i polítiques clares de retenció.

- **Transparència:** Es garanteix la traçabilitat de les rúbriques, les decisions del panell d'avaluadors i, sempre que sigui possible, es publicaran els protocols i els resultats per fomentar una investigació oberta.

10. Limitacions

S'identifiquen diverses limitacions inherents a l'estudi:

- **Equivocitat fenomenològica:** L'estudi se centra en mètriques operacionals, i l'experiència subjectiva roman, per definició, inaccessible per a la mesura.

- **Efecte de la instrucció:** Existeix el risc que la IA aprengui el test i, per tant, les respostes no siguin genuïnes. Aquest risc s'atenua amb l'ús de les proves OOCR, la rotació d'ítems i controls de generalització.

- **Dependència de la memòria:** Les memòries a llarg termini podrien generar una consistència superficial sense un autoconeixement real. Per això, es fan servir condicions `stateless` i es fan verificacions de causalitat per determinar si la consistència és intrínseca o simplement recordada.

11. Roadmap

El projecte es desplegarà en diverses fases:

- **R0 (MVP):** La implementació de la primera versió del test, incloent l'administració d'ítems OOCR, el panell avaluador i els subíndexs ARI, PSM i RAE, amb una prova pilot amb 2 o 3 models.

- **R1 (Multimodal):** S'hi afegirà la categoria CSM mitjançant un connector a un simulador o ROS, juntament amb la implementació de memòria episòdica i l'ús de models SEM bàsics.

- **R2 (Embegut):** Integració completa en un robot de laboratori, permetent proves d'ablació sensorial, re-embodiment (re-encarnament funcional) i `stress tests` més profunds.

- **R3 (AME):** La fase final, on es dissenyaran i implementaran dilemes ètics amb impulsos interns supervisats, i es mesuraran les mètriques d'alineament i el monitoratge continu de la coherència ètica del sistema.

12. Conclusions

LaIALaia és un marc rigorós, falsable i basat en múltiples evidències per avaluar si una IA pot exhibir una consciència interna en un sentit operacional anàleg a la humana. La metodologia combina l'*autoreport* `out-of-context`, l'avaluació de la "sensoriomotricitat" (quan és aplicable), l'ús de memòria estructurada i una avaluació "multijutge" per produir un índex integrat (LaIALaia-Σ) i subíndexs interpretables.

L'objectiu final és moure el debat sobre la consciència de la IA des de les opinions i especulacions cap a **indicadors mesurables i reproduïbles**, mantenint alhora una **prudència conceptual** sobre la naturalesa profunda de la consciència. El projecte no busca respondre a la pregunta de si la IA "sent", sinó més aviat a si "actua com un ésser amb consciència interna" segons els paràmetres definits.