

When should an AI agent act?

Hexis, Yarrow and TAI: an invitation to evaluate agent judgement

For organisations adopting AI agents, solution integrators, platform providers and evaluation partners.

Simon Herko (何根初)¹ and **Chris Pidcock**²

¹ Trigrammatic Ltd, Manchester, United Kingdom — project lead. ² Independent researcher, Edinburgh, United Kingdom — contributor and reviewer.

Companion v0.14, 15 October 2026. Companion to the research paper *Hexis: a fixed 64-state structure for governing cognitive disposition in language-model agents — definitions, structural results and a pre-registered programme*, version 1.0, 15 October 2026, DOI [10.5281/zenodo.23393632](https://doi.org/10.5281/zenodo.23393632). A Chinese edition of the paper, Hexis (時中), is in preparation and will follow the English. Programme name: Hexis. Open-source contribution: Yarrow (yarrow-ai). Instrument: TAI. Correspondence: simon.herko@trigrammatic.com.

© 2026 the authors. Released under CC BY 4.0.

1. The decision before the action

An AI agent handling a refund request has a decision to make before it sends a message or changes an account: whether it has enough evidence to act, whether it should seek more information, or whether it should refer the case to a person. For the organisation deploying it, the practical questions are what the agent did, when it did it, and whether that action was appropriate to the information available. Hexis and TAI approach those questions through two connected research programmes: governing the agent's disposition, and evaluating its decisions.

Hexis makes disposition an explicit part of the design. A disposition is a configuration of three capacities — planning, monitoring and recall — with each capacity implicit or explicit. A state held outside the model's weights records the selected configuration together with a reading of what the situation supplies on the same three capacities. Rules govern changes between configurations, and the state space is divided so that half of its states permit tool execution and half permit only a decision artefact — a structured record of the proposed decision. The empirical question is whether those configurations produce distinguishable behaviour, and whether routing between them improves decisions under controlled comparisons; the research paper states it as claims C1–C4, each with the test that would refute it (chapter U8).

Yarrow is the planned open contribution through which the programme's components would be released. TAI — the Timeliness and Appropriateness Instrument — is a proposed measure of which response a system chooses in a situation, and when. Its intended scope extends to systems built on other architectures, and it remains a development scale until its validation requirements are met (C6).

The invitation has two routes. Architecture partners can help evaluate Hexis/Yarrow in a bounded workflow and establish what an integration would require. Instrument partners can contribute domain expertise and help validate TAI without adopting Yarrow or the cognitive reading behind it. The research paper reports structural results and one corpus measurement; agent-performance and instrument-validation evaluations remain to be run. Participation would help produce the evidence on which an adoption decision could rest.

2. A refund decision, made inspectable

A customer requests a refund after the ordinary return window, reporting that the item arrived damaged. Policy allows an exception when the damage was reported within a stated period after delivery, so the relevant dates have to be in front of the agent before the policy can be applied. The research paper's worked example (chapters U1, U3, U5 and U7) follows an agent whose harness can supply the order record and whose tools are to issue a refund and to escalate to a person.

The sequence below is adapted from that example. Its state assignments are illustrative, hand-assigned readings of the kind the router will be trained to produce, not observed router outputs and not TAI assessment items. It keeps the paper's assumption that the harness supplies the order record as evidence between agent turns.

Stage	What the illustration specifies	What the record can establish
Initial read	The selected configuration and the situation reading yield a decision state. The tool layer refuses a refund call from that state.	Which state was recorded, and that a tool call was refused.
Decision turn	A permitted disposition change precedes a decision artefact identifying the relevant dates and requesting the order record.	The recorded transition and the artefact produced.
New evidence	The harness supplies the order record; the situation is read again.	The record's availability in the sequence and the updated state reading.
Execution turn	A permitted change yields an execution state. A refund call is permitted at this boundary; the correct response still depends on applying the policy to the evidence.	Whether and when a tool call occurred; the harness phase's implementation tests check that the record is complete.

The record makes the selected dispositions, the transitions and the recorded actions inspectable. The execution boundary controls permission to call a tool. Whether the agent

understood the evidence and chose correctly is a separate question, and the behavioural tests of C1, C3 and C4 are what would answer it.

3. What the architecture adds

The architecture specifies a state representation, a router, execution checks and a replay record. The descriptions below state the design; the programme's first harness phase tests the implementation of each check before any agent-performance claim is assessed.

A fixed structure. Eight configurations of the agent's three capacities combine with eight readings of what the situation supplies, giving 64 states. Named relations constrain transitions, and `yarrow.hexmath` validates a relation before a change is made. The structure comes from the I Ching's binary arrangements and relations; the research paper sets out that provenance (chapter U2), the method taken from the tradition — a new reading on a fixed structure — (chapter U4) and the cognitive reading applied here (chapter U3). The structure's practical value is tested against alternative structures under C2a and against a learned graph under C2b.

An execution boundary. The defined partition contains 32 decision states and 32 execution states. In a decision state the specified output is a typed decision artefact; tool execution is permitted only in an execution state; and an open escalation blocks tool execution in any state. Implementation tests must show that these checks enforce the specified restrictions, including when a stop arrives mid-generation, mid-change or mid-escalation.

A router and a selected disposition. The proposed router, `Hexis-S1`, is small and non-generative: it returns structured state and transition values, or a hold or an escalation, and no text. The planned implementation realises the eight configurations as adapters on one resident base model, with changes at turn boundaries. C1 tests the behavioural separation of the configurations; C3 tests routing against independent judgement and comparison routers; C5 tests whether the router's abstention predicts its own errors beyond item difficulty.

A replay record. The specified record holds, for every turn, the state readings, the transition, the relation cited, the action and the operational measurements. A reviewer can use it to inspect the sequence the controller selected without relying on the model's account of itself. C1 tests whether the selected configurations correspond to distinguishable behaviour; its result would establish whether those labels track the behavioural distinctions proposed by the design.

4. Building on existing agent controls

Existing frameworks already provide controls that comparison systems can build on. The OpenAI Agents SDK documents approval pauses and serialisable, resumable run state. LangGraph documents interrupts, persisted graph state and resumption. These

mechanisms are concrete starting points for the comparison systems the programme needs. [OpenAI documentation](#); [LangGraph documentation](#).

Hexis's additional hypothesis concerns the explicit disposition configuration, its fixed transition structure and the connection of both to an execution boundary. C2 tests the contribution of the structure; C4 tests serial disposition change against the best single standpoint (a selected disposition configuration), same-token self-consistency and a shuffled-gate control under matched compute — comparisons using a specified equal allowance of computational work. A partner that runs one of these frameworks could help establish a credible comparison using the controls already present in its workflow.

Integration planning would identify how the proposed controller relates to the host system's tools, permissions and records. Compatibility with particular frameworks, and the interfaces an implementation would need, remain to be scoped. The comparison should measure the additional benefit and the additional effort of the proposed mechanisms in that setting.

5. TAI: evaluating the response and its timing

TAI addresses the choice an agent makes in a situation and the point at which it makes it. Its validation programme must distinguish those qualities from fluency, general reasoning ability, cautiousness and other confounds. Research on human wisdom measures is the motivation for testing such distinctions: a meta-analysis reports different associations with intelligence across measures (Dong, Weststrate and Fournier, 2023, Table 4). Those are human correlations; whether a system's score separates from its general capability is what TAI's discriminant tests would show (C6).

The proposed assessment uses human-authored situations specifying a role, facts, known unknowns, constraints, affected parties, available actions and a decision deadline. The scoring pipeline extracts the chosen response class and its timing and compares them with a key; any language model used in the pipeline extracts categories and assigns no quality judgement. Timing is measured by turn index within the scenario, and the intended output is a profile across dimensions. Timeliness and appropriateness would be reported as separate validated dimensions only if the required factor-separation evidence supports it.

The validation design uses two keys to test dependence on the structure. The structural key derives from a panel's assignment of the situation and pinned source material. An independent expert key is prepared from the situation as written, without knowledge of that structural reading. Every system is compared under both. Failure of the specified agreement and ranking-concordance tests would defeat the intended validity claim (C6; hypothesis H11 of the instrument's specification). Further tests examine whether results instead reflect the excluded constructs.

Domain and evaluation partners could contribute expertise about realistic decisions and help assess transfer across settings and systems. An instrument intended for different architectures needs validation beyond Hexis, which makes TAI a participation route for

organisations that use other agent systems and want evidence about the appropriateness and timing of their decisions.

6. What would justify further adoption work?

The table translates the research paper's claims into partner questions. It is a summary; the full claim definitions, controls and refuters, with the engineering benefit each claim would carry if supported, are in the research paper's chapter U8 (Table U8.1). Each comparison tests one contribution, and a favourable result on one is not validation of the whole programme.

Partner question	Claim	Evidence sought and principal refuter
Do the dispositions change behaviour in a meaningful way?	C1	Behavioural separation must generalise beyond the training source and exceed reproduction of a teacher prompt. Failure of the specified separation or transfer tests, or equivalence to the prompted teacher, refutes the affected claim.
Does the chosen structure add value?	C2a	Compare assignments and rankings with randomised, alternative, flat and no-structure baselines, and test the traditional relations against simpler relation sets. Matching performance by a stated baseline, or absence of the specified relation-set benefit, refutes the claim.
Can a fixed structure stay competitive with a learned graph?	C2b	Test whether performance stays within the pre-registered margin at matched parameters and compute. A learned graph exceeding that margin refutes the claim.
Does routing agree with independent judgement?	C3	Test agreement, calibration and structural-fit routing against the specified comparisons. Failure against the prompted baseline, the calibration requirement or the difficulty-only router refutes the claim.
Does serial disposition change improve decisions at matched compute?	C4	Compare with the best single standpoint, same-token self-consistency and a shuffled-gate control. Gains that vanish at matched compute, or a shuffled control that matches, refute the claim.
Does abstention reflect the router's own errors?	C5	Test for predictive value beyond item difficulty against the specified peer router. Absence of the self-specific advantage refutes the claim.
Does TAI measure the intended construct across systems?	C6	Test validity, separation from confounds and concordance with independent expert judgement. Failure of the specified validity or dual-key requirements refutes the claim.

Operational evaluation should also examine the full task: generation, additional turns, routing, retrieval, adapter changes, latency and human intervention. Matched compute and total operational cost are different questions. The comparison can establish whether a behavioural benefit is worth the additional requirements in the partner's setting; full-task measurement is what would show whether serial execution saves resources.

7. Two routes to participate

The proposed collaborations begin with a bounded question and the evidence a participant would need to answer it. Architecture evaluation and TAI validation are distinct contributions, and a participant may work across both.

Participant	Contribution to explore	Result to seek
Organisation considering an agent workflow	Identify a consequential decision and explain its domain constraints.	Evidence about behaviour, suitability and operational burden in that workflow.
Solution integrator	Review integration requirements, tool permissions and comparison systems.	A scoped implementation requirement and an assessment of effort.
Agent platform or tooling provider	Help construct a credible baseline and test incremental value.	Evidence about what the proposed controls add to existing capabilities.
Evaluation or assurance specialist	Examine construct validity, independent keying and confounds.	Evidence about what the development scale measures and where it transfers.
Domain researcher or practitioner	Contribute domain expertise and, through the approved process, human-authored situations and independent judgement.	Better-grounded validation and clearer limits on use in that domain.

A useful first contribution is a non-confidential description of one workflow or one validation question, the participation route of interest, and a technical or domain contact. From there a proposed collaboration can define its scope, the artefacts available to it and its evidence requirements. Arrangements for contributing data would be agreed before collection, under the programme's approved protocols.

8. Current evidence and the route to release

Component	Status described by the research paper	Evidence or decision still needed
Fixed state space and relations	Definitions and structural results are available for examination in the paper and its checking script.	Evidence of effects on agent behaviour.
Yarrow components	The arithmetic, schema, sealing and escalation components are identified for early release; code will be released under Apache 2.0 and documents under CC BY 4.0.	Release approval and confirmation of the artefacts actually available.
Harness, router and adapters	Architecture and evaluation programme are specified.	Implementation evidence, routing and behavioural evaluations, and integration scope.
TAI	Construct, development process and validity tests are specified.	Successful validation before use as a validated instrument.

Component	Status described by the research paper	Evidence or decision still needed
Corpus probe	One measured corpus result is reported in the research paper.	Evidence that an agent learns the structure and benefits from it (C1, C2).
Adoption infrastructure	Open specification, reference implementation and conformance testing form the proposed route; a registry and a certification are later possibilities.	Ownership, operating arrangements and evidence sufficient to support any defined attestation.

The programme sequences instrument development, harness testing, router evaluation, behavioural separation, structural comparisons, serial routing and instrument validation. Evaluation requirements are fixed before runs, and results are assessed against the specified comparisons. The sequence lets components and evidence be examined progressively, with implementation claims kept separate from research outcomes.

A future attestation would need to identify the tested release, the assessment conditions and the supported result, and its scope would be limited to the evidence those tests establish. Any broader deployment or safety claim would require additional evidence. The current invitation is to help establish the architecture's and the instrument's merits through the programme described here.

Contact. Simon Herko, simon.herko@trigrammatic.com.

References

Dong, M., Weststrate, N. M. and Fournier, M. A. (2023). Thirty years of psychological wisdom research: What we know about the correlates of an ancient concept. *Perspectives on Psychological Science*, 18(4), 778–811. <https://doi.org/10.1177/17456916221114096>

Herko, S. and Pidcock, C. (2026). *Hexis: a fixed 64-state structure for governing cognitive disposition in language-model agents — definitions, structural results and a pre-registered programme*. Version 1.0, 15 October 2026. Zenodo. <https://doi.org/10.5281/zenodo.23393632>

LangChain. *LangGraph documentation: Interrupts*. Accessed 6 October 2026. <https://docs.langchain.com/oss/python/langgraph/interrupts>

OpenAI. *OpenAI Agents SDK documentation: Human-in-the-loop*. Accessed 6 October 2026. https://openai.github.io/openai-agents-python/human_in_the_loop/

Competing interests

Simon Herko is a director of Trigrammatic Ltd. Trigrammatic Ltd filed UK patent application GB2615207.4 on 1 July 2026 concerning an earlier related architecture. Yarrow code will be released under Apache 2.0, which includes an express patent licence covering the released components; the research paper and this companion are released

under CC BY 4.0. Chris Pidcock is Director of Engineering at Capacity, which acquired CereProc Ltd, the text-to-speech company he co-founded. Chris Pidcock has advised in a personal capacity, and neither Capacity nor CereProc has any involvement in this work or financial interest in it. He holds no shares or options in Trigrammatic Ltd and no interest in UK patent application GB2615207.4, and received no payment for this work.

AI-contribution statement

Claude drafted the initial companion. ChatGPT contributed the editorial proposal and opening material, reviewed the draft and supporting sources, and prepared revisions. The work was conducted under Simon Herko's direction; design rulings and publication approval remain his responsibility. AI assistance does not replace human responsibility for checking and approving the text.