

Hexis: a fixed 64-state structure for governing cognitive disposition in language-model agents — definitions, structural results and a pre-registered programme

Simon Herko (何根初)¹ and Chris Pidcock²

¹ Trigrammatic Ltd, Manchester, United Kingdom. ORCID 0000-0001-9941-7313. ² Independent researcher, Edinburgh, United Kingdom. ORCID 0000-0001-5279-412X.

Version 1.0, 15 October 2026. DOI [10.5281/zenodo.23393632](https://doi.org/10.5281/zenodo.23393632). Programme: Hexis. Open-source contribution: Yarrow (yarrow-ai). Correspondence: simon.herko@trigrammatic.com.

© 2026 the authors. Released under CC BY 4.0.

Abstract

When a language-model agent is authorised to act, choosing how to approach a situation becomes part of the task. A refund request, for example, may require a missing record, a policy check or a decision to defer before any action is appropriate. Hexis investigates whether these changes in approach can be represented explicitly, selected to suit the situation and tested for their contribution to better decisions. The proposed architecture pairs eight configurations of planning, monitoring and recall with eight descriptions of what the situation already supplies. This gives a sixty-four-state structure drawn from the I Ching, with defined relations between states and a project-defined boundary between deciding and executing. Its cognitive interpretation connects a fixed mathematical structure to questions from metacognition, control and practical-wisdom research: whether the configurations produce distinct behaviour, whether their selection improves decisions, and how timeliness and appropriateness should be assessed. A separate validation programme addresses that last question across systems. The paper presents reproducible structural checks, two proofs and a registered programme of behavioural and validity tests. The agent architecture and the proposed instrument remain to be built and tested. Yarrow is the planned open-source contribution through which others could reproduce the structure and evaluate the proposal.

HEXIS · FRONTISPIECE

Eight voices on judgement and action, set in the Later Heaven circle (South at the top).
Read clockwise from ☳ Thunder, in the Shuogua's order 帝出乎震.



Wind [011]

Monitoring and recall explicit

君子之中庸也，君子而時中。

"The superior man's embodying the course of the Mean is because he is a superior man, and so always maintains the Mean."

時中: the mean, kept according to the time.†

— Zhongyong, ch. 2, attributed to Zisi (trans. Legge, 1893)



Fire [101]

Planning and recall explicit

夫卦者，時也；爻者，適時之變者也。夫時有否泰，故用有行藏。

"A hexagram is a time; a line is the change that suits the time. Times are blocked or open, and so their use is to act or to withdraw."†

— Wang Bi, Zhouyi lüeli, "Ming gua shi bian tong yao"



Earth [000]

All capacities implicit

"Virtue, then, is a state of character concerned with choice, lying in a mean, i.e. the mean relative to us, this being determined by a rational principle, and by that principle by which the man of practical wisdom would determine it."

— Aristotle, Nicomachean Ethics II.6, 1106b36–1107a2 (trans. Ross)



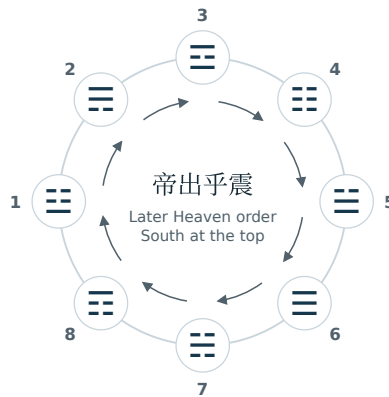
Thunder [100]

Planning explicit

幾者，動之微，吉之先見者也。君子見幾而作，不俟終日。

"Those springs are the slight beginnings of movement, and the earliest indications of good fortune (or ill). The superior man sees them, and acts accordingly without waiting for (the delay of) a single day."

— Xici zhuan II (trans. Legge, 1882)



Lake [110]

Planning and monitoring explicit

"One, a robot may not injure a human being under any conditions — and, as a corollary, must not permit a human being to be injured because of inaction on his part."

— Isaac Asimov, "Runaround", Astounding Science-Fiction, March 1942, p. 100



Mountain [001]

Recall explicit

"Everything can be taken from a man but one thing: the last of the human freedoms — to choose one's attitude in any given set of circumstances, to choose one's own way."

— Viktor E. Frankl, Man's Search for Meaning (1963), pp. 103–104



Water [010]

Monitoring explicit

"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it ... then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it."

— Norbert Wiener, Science 131 (1960), p. 1358



Heaven [111]

All capacities explicit

"In particular, it [the machine] has no purpose of its own and no innate desire to protect itself. The machine must be initially uncertain about what those human values are."

— Stuart Russell, Scientific American, June 2016, pp. 58–59

† Translation by the authors. Bits are [planning monitoring recall], following the line assignment in Table U3.2. Each source is listed in the references.



Contents

Abstract	1
Upper Canon — Understanding the proposal	4
U1. Introduction: when should an AI agent act?	4
U2. The structure Hexis adopts	9
U3. Planning, monitoring and recall	13
U4. A tradition of changing interpretations	16
U5. Structural results and the execution boundary	23
U6. Interdisciplinary foundations and their limits	32
U7. The proposed Hexis architecture	37
U8. Claims and tests that could refute them	40
Lower Canon — Testing and developing the programme	42
L1. Measuring timely and appropriate action	43
L2. Pre-registration and measurement discipline	46
L3. Engineering value and adoption	49
L4. Safety, governance and affected parties	50
L5. Open contributions and reusable components	52
L6. Related work and competing explanations	53
L7. Scope, uncertainty and ethical limits	55
L8. Roadmap, replication and openness	55
Contributions, acknowledgements, glossary and references	57
Glossary	59
References	60

Section titles are clickable. Page numbers refer to this PDF. Figures appear beside the mechanisms they explain. Each footer also draws its page number as a six-line figure: the number less one in binary, read with the bottom line most significant as the Fuxi index reads a trigram (U2.1); page 1 is six broken lines, page 64 six solid lines.



Upper Canon — Understanding the proposal

The paper is set out in two parts of eight chapters, an upper and a lower canon, as an editorial homage to the received text's two parts. The division is editorial: it is distinct from the received text's division of the hexagrams into thirty and thirty-four, and from Hexis's own partition of its sixty-four states into thirty-two decision states and thirty-two execution states (U5.1). Chapters are not states.

U1. Introduction: when should an AI agent act?

A language-model agent given access to tools faces a practical question at each step: what should it do now? Answering may require a plan, a check against evidence, information from an earlier event, an action, or a pause for another person's judgement. The useful choice depends on the situation. Hexis is a research programme about making that choice explicit and testing whether a structured way of making it improves the agent's decisions.

U1.1 A situation that requires more than an answer

Consider a support agent receiving a refund request three days after the return window. The customer says the item arrived damaged. Policy permits an exception if the damage was reported within forty-eight hours of delivery, but the request does not include the dates needed to decide. A familiar pattern suggests an answer; the missing record determines whether that answer is justified. The appropriate sequence is to establish what information is needed, obtain it, apply the policy and then act or explain the decision.

This example brings two questions into view. One concerns authority: when should the agent be allowed to issue a refund? The other concerns its approach: what should it make explicit before it reaches that point? A system may record every step and still leave open whether the selected approach made a useful difference. Hexis proposes a record of that selection and a programme to test its behavioural meaning.

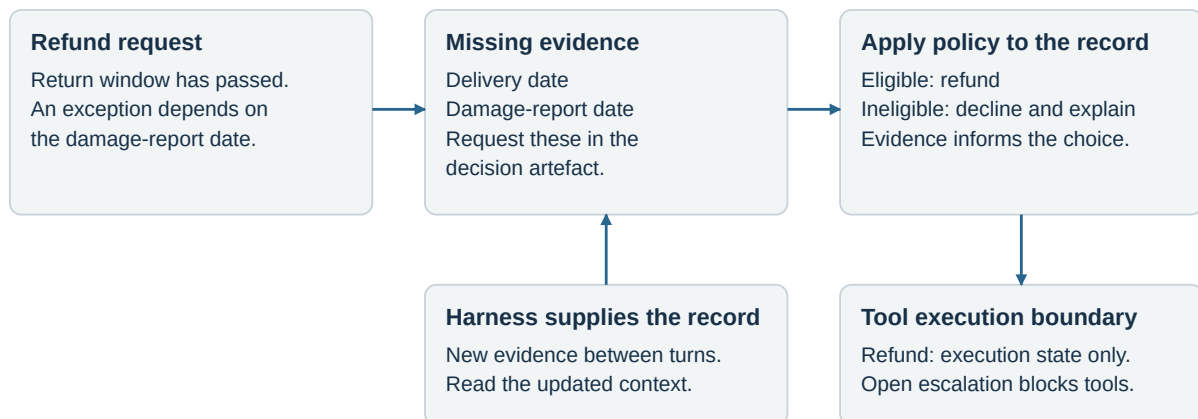
Existing agent frameworks provide relevant controls, including approval pauses, persistent state and resumption (OpenAI, n.d.; LangChain, n.d.). Hexis asks what a particular organisation of planning, monitoring and recall would add within such a controlled workflow. Its value would lie in selecting an approach that fits the situation, making the selection inspectable and showing that it improves decisions in comparison with the specified alternatives. Figure U1.1 draws the refund case as the sequence the proposed harness would follow.

U1.2 The proposal: connect the agent's approach to its situation

Hexis calls a configuration of planning, monitoring and recall a *standpoint*. Each capacity has an implicit and an explicit tier. In the proposed explicit tier, planning concerns a stated course of action, monitoring concerns confidence revised against supplied

A request, a missing fact, a decision

First establish what must be known, then when action is permitted.



Proposed illustration; follows the harness-supplied-record assumption pending the retrieval ruling.

No runtime result or additional routing rule is shown.

Figure U1.1. The refund case begins with an information gap. The proposed sequence requests the missing record, uses it to reach a decision and permits tool execution only in an execution state. The detailed trace follows in U3, U5 and U7. The diagram follows the harness-supplied-evidence account; the read-only retrieval question remains in the register.

evidence or a check, and recall concerns the specified use of information about the past. The implicit tier still involves the capacity. Chapter U3 defines these distinctions and the comparisons that would establish whether they produce different behaviour.

Three two-way distinctions give eight standpoints. Hexis pairs these with eight descriptions of what the situation supplies: whether a course of action is already set, whether a check has been performed and whether a relevant record is present. The pair describes an agent in a situation. This makes the research question relational: a configuration that is useful when evidence is missing may be unnecessary once that evidence has arrived.

The I Ching supplies the fixed sixty-four-state structure used for these pairings, together with named relations between its figures. The proposed cognitive interpretation gives those states an engineering use. Holding the structure fixed makes that use specific enough to test: the programme can compare the chosen structure with alternative structures, simpler relations and a learned graph. Those comparisons will determine whether the fixed structure earns its place in the architecture.

A *router* is the proposed component that selects a standpoint from the situation. The *harness* is the surrounding software that supplies information, manages the model's interactions and enforces tool permissions. The harness contract retains the selected state across turns and records each permitted transition. The programme will test



whether the router agrees with human judgement, whether its choices depend on the proposed structural fit and whether its abstentions help identify cases in which it is likely to be wrong.

U1.3 Why the disciplines need to meet

The research needs a precise account of the proposed behaviour, a way to control it and a way to judge its value. Each part draws on a different kind of work. Historical scholarship explains the I Ching's structure and the tradition of assigning different readings to it (U2 and U4). Cognitive science and metacognition help formulate distinctions among planning, monitoring and recall (U3 and U6). Control and learning research provide precedents and comparisons for selecting between modes of operation (U6.4). Philosophy and the psychology of practical wisdom frame the question of what makes a choice appropriate to a particular situation (U6.1–U6.2).

These connections matter at the points where one discipline leaves a question for another. A precise state description needs behavioural evidence that its categories mean something. A controller that switches between categories needs evidence that switching helps. A measure of improvement needs to distinguish appropriate judgement from verbal fluency, general ability or habitual caution. Hexis connects these questions through a common set of proposed states, comparisons and validation tests, so that each part can be assessed on its own evidence.

Monitoring illustrates the connection. Huang et al. (2024) found that revision without external feedback could fail to improve reasoning and could degrade it in their evaluated settings. Hexis therefore proposes an explicit monitoring tier grounded in evidence or checks supplied by the harness. The behavioural comparison with prompted self-reflection will test what that grounding adds. A conceptual distinction becomes an architectural choice, and that choice becomes a question for an experiment. Figure U1.2 maps each contributing discipline to the part of the proposal it informs.

U1.4 From a proposed decision to an authorised action

The architecture separates preparing a decision from carrying it out. Under Hexis's definition, thirty-two states produce a *decision artefact*: a structured record of the proposed decision and its supporting information. The other thirty-two permit tool execution. The tool-layer check rejects a tool call from a decision state. Its conformance tests will check that restriction; behavioural tests will separately assess whether the two kinds of state are distinguishable and whether the resulting decisions improve.

In the refund illustration, the agent first identifies the missing delivery and damage-report dates and requests the record. The harness supplies that record, changing what the situation makes available. The example then follows a transition to a state that permits the refund tool. The states in this worked example are assigned by hand to explain the proposed rules. Chapters U3, U5 and U7 develop this same case through its state assignments, transition rule and replay record. This introduction follows the paper's existing assumption that the lookup is evidence supplied by the harness between turns; the exact trace remains conditional on the recorded A-9/DQ-7 design question.



Why connect these disciplines?

Each contribution leads to a different engineering definition or test.

Contribution	Why it is needed	Question to test
Fixed figures and relations; new readings of a shared structure	→ Make the proposal precise and reproducible	→ Does this structure add value over alternatives?
Cognitive science and metacognition	→ Define the proposed ways of working	→ Do configurations produce distinct behaviour?
Control and learning	→ Turn distinctions into selection, transitions and tool permissions	→ Does selection work, and does changing standpoint help?
Practical wisdom and measurement research	→ Define appropriate choice and timing	→ Does the instrument measure its construct across systems?

The connections motivate the research; each claim still requires its own evidence.

Figure U1.2. Each discipline contributes a different part of the proposal. The connections lead from a concept to an engineering definition or a comparison that can be tested. Each contribution has its own evidential task.

Escalation supplies a further possible response when authority or evidence requires another person’s review. The contract blocks execution while an escalation remains open, and the harness fixtures will test the stop under adverse timing. Gomez’s (2026) study of one task-rule conflict scenario family motivates testing such channel design; it does not establish a general safety effect. The study’s detailed comparison and limitations are given with the governance discussion in L4.

The resulting log would record the selected configuration, supplied evidence, transition and action. Behavioural evaluation will test whether these recorded distinctions correspond to useful differences in what the agent does. That is the step needed to connect an inspectable process with a claim about better decisions.

U1.5 How the programme will judge progress

The architecture will be useful only if its configured capacities produce meaningful differences and selecting between them helps. The programme will therefore compare configured behaviour with prompted alternatives, the fixed structure with competing structures, and serial changes of standpoint with the specified controls under matched computational accounting. These comparisons separate the value of the configuration, the structure and the act of switching between standpoints.

Assessing timeliness and appropriateness poses a further problem: a score may reward the structure it was built from. The proposed Timeliness and Appropriateness Instrument (TAI) consequently has its own validation programme. It will examine whether the



intended construct is distinguishable from style, caution, refusal and other alternatives, whether comparisons extend across systems, and whether an independent expert key supports the rankings produced by its structural key. Hexis is one prospective subject of that programme. Chapter L1 sets out the validity tests and the conditions under which an interpretation of the scores would fail. Figure U1.3 separates the two kinds of question.

Two tasks of evaluation

Structural arithmetic has reproducible checks; behavioural and validity tests below are planned.

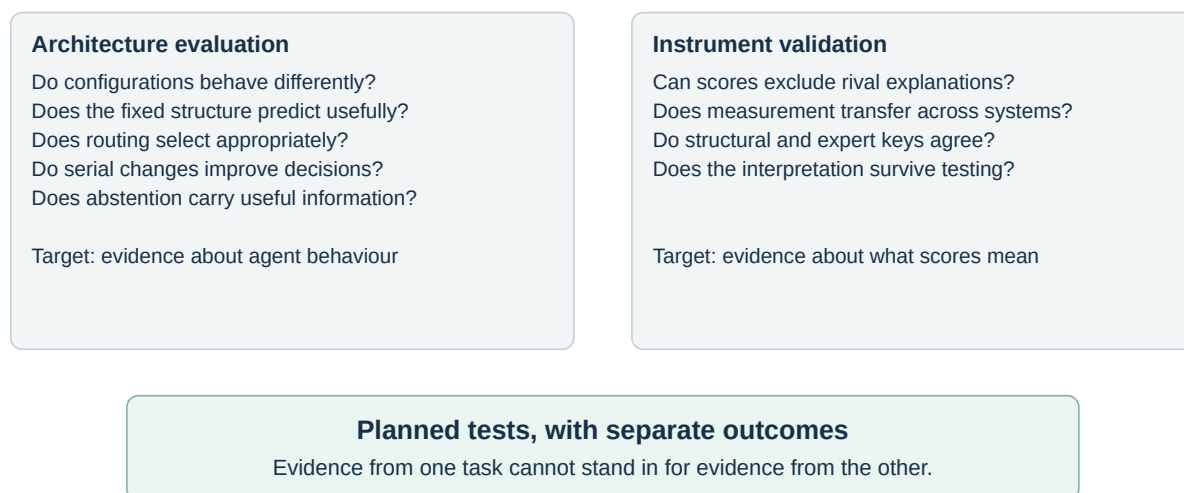


Figure U1.3. Architecture evaluation and instrument validation answer different questions. Behavioural comparisons test the agent proposal; validity tests examine whether the instrument supports its intended interpretation. The two tasks require separate evidence.

The work reported here establishes the structure and its checkable properties. The runtime system and behavioural comparisons remain a programme to build and test. A pre-registered corpus probe, reported in L2, measures how much of the pinned I Ching text is present in public pre-training corpora; its result enters the programme as a covariate, and the measured direction between the English and Chinese texts was the reverse of the registered expectation. Yarrow names the planned open-source contribution, intended to let other researchers reproduce the structure, examine the specifications and test the proposal.

U1.6 The research questions and the route through the paper

The questions introduced above have short labels for use in the rest of the paper. C1–C6 are a reading guide to the formal claims in U8 (Table U8.1); C2 has two parts. Their tests specify observations that would refute them, so a successful structural check, a behavioural result and a valid measurement each have a distinct evidential role.



Table U1.1. The research questions behind the claim labels.

Label	Question it tests
C1	Do the configured capacities produce distinct behaviour that transfers beyond their training examples and prompted teachers?
C2a	Does the fixed structure add predictive value over the specified alternative structures and relations?
C2b	Does it remain within the pre-registered margin of a learned graph at matched parameters and compute?
C3	Does routing agree with human judgement and use structural fit rather than task difficulty alone?
C4	Does changing standpoint improve decisions over the specified controls at matched compute?
C5	Does the router's abstention predict its own errors beyond item difficulty?
C6	Does TAI measure the intended construct across systems, with rankings supported by an independent expert key?

The upper canon explains the proposal. U2 defines the structure and U3 the cognitive reading laid on it; U4 sets that reading in the tradition of readings the structure has carried; U5 states what can be checked now and the execution boundary; U6 sets out the disciplinary foundations and the weight each bears; U7 specifies the architecture; U8 states the claims and the observations that would refute them. The lower canon tests and develops the programme: the instrument (L1), the measurement discipline (L2), the engineering and governance cases with their counter-cases (L3, L4), the open components (L5), related work and competing explanations (L6), scope (L7) and the roadmap (L8).

U2. The structure Hexis adopts

Hexis needs a compact way to describe a configuration and specify how it can change. It takes that structure from the I Ching: eight three-line figures, called trigrams, combine into sixty-four six-line figures, called hexagrams. Named relations connect the figures by changing or rearranging their lines. Hexis assigns a cognitive interpretation to this structure in U3, then proposes tests of that interpretation in U8.

The structure gives the proposal a vocabulary whose operations can be checked exactly. A reader can reproduce a transformation and see which lines changed before considering what those lines mean for an agent. This chapter defines the arithmetic specified for `yarrow.hexmath`; the accompanying structural checker reproduces the facts reported here. The proposed implementation will use these definitions as fixed data and select states through classification; divination and the hexagram texts are outside that selection procedure.



U2.1 From lines to trigrams

A **line** has two forms: solid, conventionally called yang, and broken, called yin. Hexis represents them as 1 and 0. A **trigram** contains three lines, counted from bottom to top and written [b1 b2 b3]. Thus ☲ Fire [101] has a solid bottom line, a broken middle line and a solid top line. Three binary choices produce eight configurations.

The Chinese names help identify the figures across translations. Throughout this paper, the English name, glyph and bits identify a standpoint; the Chinese character and pinyin identify its traditional name. Pinyin is a system for writing Mandarin pronunciation with the Roman alphabet. The eight names are:

Table U2.1. The eight trigrams.

Standpoint notation	Chinese name	Pinyin
☷ Earth [000]	坤	Kun
☶ Mountain [001]	艮	Gen
☵ Water [010]	坎	Kan
☴ Wind [011]	巽	Xun
☳ Thunder [100]	震	Zhen
☲ Fire [101]	離	Li
☱ Lake [110]	兌	Dui
☰ Heaven [111]	乾	Qian

The **Fuxi index** reads these three bits as a binary number, with the bottom line most significant. It runs from 0 for Earth to 7 for Heaven. Figure U2.1 places the eight trigrams at the vertices of a cube: each edge changes exactly one line. The figure also previews U3's assignment of the lines to planning, monitoring and recall. “Implicit” and “explicit” describe the project's capacity tiers.

U2.2 From trigrams to hexagrams

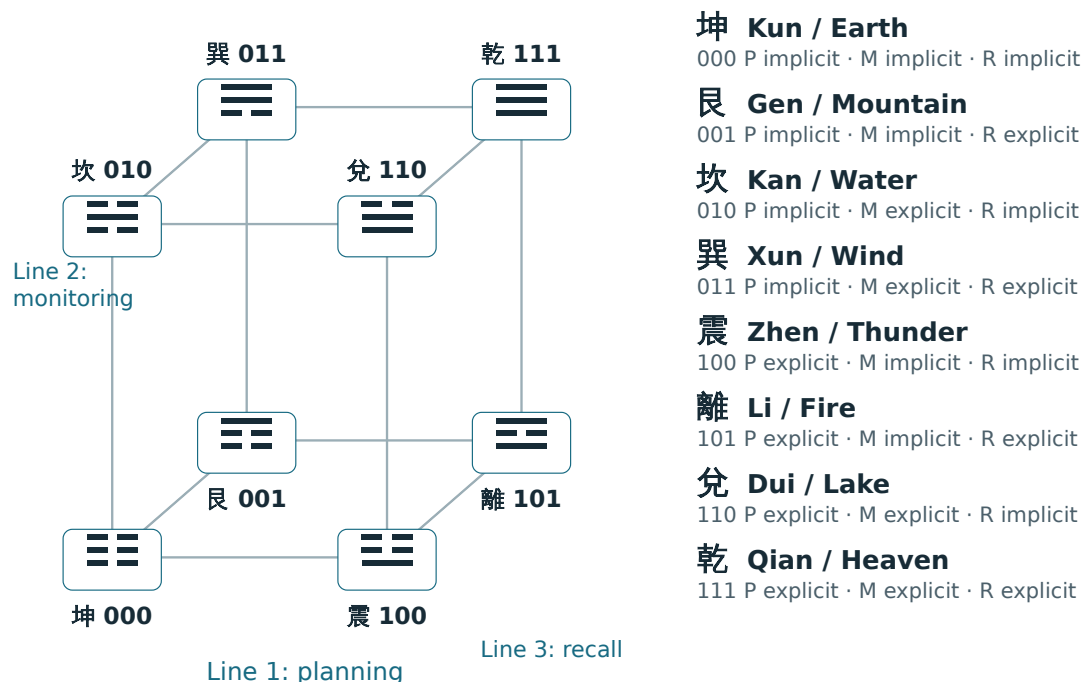
A **hexagram** places one trigram above another. Hexis writes it as the ordered pair (inner, outer): lines 1–3 form the lower, inner trigram; lines 4–6 form the upper, outer trigram. Eight choices for each give $8 \times 8 = 64$ hexagrams. The pair describes the whole configuration; U3 assigns the inner and outer positions their roles in the proposed architecture.

A **King Wen number** identifies each hexagram in the received sequence, the order transmitted in the standard textual tradition. These numbers run from 1 to 64 and come from a lookup table. They are separate from the figures' binary values. Where an English gloss is used, the full form gives number, pinyin, gloss and Chinese name: for example, **11 Tai — Peace (泰)**; the glosses follow the Wilhelm/Baynes translation (Wilhelm and Baynes, 1967). Chinese names provide a stable cross-reference when English translations differ.



Eight trigrams: three lines, eight configurations

Chinese names identify the figures; the capacity reading is Hexis's convention (U3).



Bits read bottom → top. 0 = implicit; 1 = explicit. Each edge changes one line.

Figure U2.1. Eight trigrams, their Chinese names and the three-line cube. Each edge changes one line. The capacity labels preview Hexis's interpretation in U3.

U2.3 Relations and line positions

A **relation** specifies how one figure connects to another. A changing line flips one bit; a derived hexagram flips a specified set of bits. The **complement** (錯) flips all six lines. The **inversion** (綜) reverses their order, so line 1 becomes line 6. A **nuclear hexagram** (互) takes lines 2–4 as its inner trigram and lines 3–5 as its outer trigram. These definitions tell a reader exactly how to obtain the resulting figure.

King Wen succession advances one place in the received hexagram sequence, from number n to $n + 1$, with 64 returning to 1. This relation concerns whole hexagrams. The trigram circles in U2.4 supply a different kind of succession, moving between neighbouring trigrams. Section U5.3 specifies which relations authorise routes in Hexis; whole-hexagram King Wen succession is a reading relation.

Three rules describe positions within a hexagram. **Correspondence** (應) pairs lines 1 and 4, 2 and 5, and 3 and 6; a pair corresponds when its lines have opposite polarities.

Proper position (當位) means yang at an odd position, 1, 3 or 5, and yin at an even position, 2, 4 or 6. **Centrality** (中) identifies positions 2 and 5 as the centres of their respective trigrams. These are the definitions Hexis adopts. Section U4.3 explains their historical setting and Figure U4.2 illustrates them.



U2.4 Three arrangements of the trigrams

An **arrangement** places the same eight trigrams around a circle. Once a starting point and direction are specified, each trigram has a successor and a predecessor. Hexis uses three arrangements as explicit tables; in the paper's tables and in the structural checker the Early Heaven arrangement is called the Fuxi circle and the Later Heaven arrangement the King Wen circle. In Figure U2.2, South is at the top and East is at the left, following the source diagrams; positions are read clockwise from the top, while each trigram glyph remains upright.

The **Early Heaven** arrangement (先天, *xiantian*, also called Former Heaven) is conventionally associated with Fuxi. Its clockwise order is Heaven, Wind, Water, Mountain, Earth, Thunder, Fire, Lake. The **Later Heaven** arrangement (後天, *houtian*), associated with King Wen, orders them Fire, Earth, Lake, Heaven, Water, Mountain, Thunder, Wind; its cycle is the order the Shuogua gives, 帝出乎震 (Legge, 1882, Appendix V, ch. 5). These names identify transmitted arrangements; U4 explains how later readers developed their interpretations (Smith, 2012, ch. 3).

The **Jeongyeok** arrangement orders them Earth, Fire, Lake, Thunder, Heaven, Water, Mountain, Wind. It is associated with **Kim Hang (金恒), known by his sobriquet Ilbu (一夫)**, whose *Jeongyeok* (正易, *Correct Changes*) offered a new interpretation of the tradition. Chung's reproduced diagram shows this order (Chung, 2021, fig. 1). Its source-edition qualification is retained in the note below.¹

Three arrangements of the same eight trigrams

Positions run clockwise from the top. Glyphs remain upright; South is at the top.



A circle step joins trigrams. King Wen hexagram numbers 1–64 form a separate ordering.

Figure U2.2. Three arrangements used in Hexis. Each circle contains the same eight trigrams; their positions determine different successor and predecessor relations. A step on the King Wen trigram circle is distinct from succession through the 64 King Wen hexagram numbers.

The fixed element in Hexis is this specified set of figures, positions, relations and arrangement tables. It gives the proposed cognitive reading a reproducible basis. Whether that reading helps a system act appropriately will be tested through the claims in U8.



¹ The Jeongyeok circle remains subject to confirmation against a pinned scholarly edition before the first tagged release of yarrow.hexmath. Chung (2021, fig. 1) supplies a matching secondary diagram; the primary *Jeongyeok* text is not cited here.

U3. Planning, monitoring and recall

U3.1 Three temporal capacities

The working interpretation assigns the three trigram lines, from bottom to top, to planning, monitoring and recall. These concern the future, present and past respectively. In the inner trigram, the lines describe the agent's configured capacities. In the outer trigram, the corresponding lines describe what the situation already supplies: a course of action, a check or a record. An outer line is explicit when that information is available and implicit when it is absent.

The outer trigram records what the situation supplies. Section U5.2 explains why correspondence then acts as a division of labour between agent and context. The bottom-to-top temporal assignment is a design convention. The Xici's remark that the first line is hard to know and the top line easy provides a limited textual association (Legge, 1882, Appendix III, sect. II, ch. IX). The dominant historical reading treats positions as stages of a situation. Chung (2021, fig. 1) reproduces the Jeongyeok trigram order used here. Linking that arrangement to the future-present-past assignment remains a project hypothesis; the source-edition qualification in note 1 still applies.

The example, first pass. The router's first job on the refund message of U1.4 is to read the situation as three outer lines. Has a course of action been set? No: the customer asks for a refund, but nothing in the message or the policy fixes what to do until the dates are known — line 4 implicit. Has a check been performed? No: nothing has been verified against anything — line 5 implicit. Is the record in front of the agent? No: the message asserts a delivery and a damage report, but the order record is not in the context — line 6 implicit. The context reads ☷ Earth [000]: the situation supplies nothing explicit, and whatever is to be made explicit must be made so by the agent. The reading is hand-assigned here, as U1.4 says of every read in the trace; the router's agreement with such readings is claim C3.

The four states of the illustration, for reference through U5 and U7:



Table U3.1. The four states of the running example.

Point in the illustration	State	What changed
Agent and context both implicit	2 Kun — The Receptive, decision	Starting pair
Agent makes its three capacities explicit	11 Tai — Peace, decision	Standpoint transition by the inner complement
Harness supplies the order record	26 Dachu — The Taming Power of the Great, decision	Context update; the agent stays ☰ Heaven [111]
Agent leaves recall implicit because the record is present	41 Sun — Decrease, execution	Standpoint transition by a change of inner line 3

Two standpoint transitions and one context update; the context update is read by the router, not chosen by the agent (U3.3). Figure U3.1 shows how to read the two trigrams separately.

Read the agent and the situation separately

Three capacities are read on each side. Together they define one six-bit state.

Planning / future

Agent: is planning configured as explicit? Context: is a course of action already set?

Monitoring / present

Agent: is monitoring configured as explicit? Context: has a check already been supplied?

Recall / past

Agent: is recall configured as explicit? Context: is the relevant record already available?

Pair the readings

Inner trigram = agent configuration. Outer trigram = supplied context. Their pair identifies the state.

Conceptual guide to the specified design; implementation and evaluation remain prospective.

Figure U3.1. Read agent capacities and supplied context separately. Notice that each line asks the same question of two different things, the agent's configuration and the situation. The pair defines a state; C1 tests the state's behavioural meaning.

U3.2 One polarity, two tiers

On every line, yin is the **implicit** tier of that capacity and yang the **explicit** tier.



Table U3.2. One polarity, two tiers.

Line	Capacity	Yin (implicit)	Yang (explicit)
1	future / planning	anticipation without a stated plan	an explicit, revisable plan
2	present / monitoring	unmonitored processing	externally grounded monitoring
3	past / recall	pattern and semantic recall	packet-conditional recall (U3.3)

The eight standpoints are the eight configurations: ☰ Heaven [111] has all three capacities explicit; ☱ Lake [110] has planning and monitoring explicit and recall implicit; ☲ Fire [101] has planning and recall explicit and monitoring implicit; ☳ Thunder [100] has only planning explicit; ☴ Wind [011] has monitoring and recall explicit and planning implicit; ☵ Water [010] has only monitoring explicit; ☶ Mountain [001] has only recall explicit; ☷ Earth [000] has all three implicit. An implicit capacity is operating at its implicit tier, as the table defines; it is not absent. They are configurations of capacities, not personas.

The current design uses one implicit/explicit polarity across three capacities. The textual precedent is the Shuogua's account of the six lines as heaven, earth and humankind, each named by a pair — yin and yang, soft and hard, benevolence and righteousness (Legge, 1882, Appendix V, ch. 2); Hexis reads the three pairs as one implicit/explicit polarity applied to three capacities. The empirical reason for testing capacities directly is specific. Zierahn et al. (2026) report that Big Five inventories applied to language models do not recover the human factor structure; Han et al. (2025) report a dissociation between self-reported traits and behaviour. Neither result tests Hexis's proposed configurations.

An earlier version of the programme used a Jung-inspired analogy to think about the eight configurations. Jung's *Psychological Types* distinguishes introverted and extraverted attitudes and four functions: thinking, feeling, sensation and intuition (Jung, 1921/1923, ch. X). The later Myers-Briggs scheme uses four preference pairs to describe sixteen types; Myers and Briggs made the judging-perceiving pair explicit in that scheme (The Myers-Briggs Company, n.d., question 8). These are distinct formulations. The sources checked here do not establish an I Ching derivation of either classification.

The cognitive precedent is a distinction between processing and its representation. Tulving (1985) distinguishes anoetic, noetic and auto-noetic consciousness; Wheeler, Stuss and Tulving (1997) develop the auto-noetic tier. Dehaene, Lau and Kouider (2017) distinguish unconscious processing, global availability and self-monitoring. Section U6.3 explains the limited role of those precedents. C1 must establish whether the configured capacities in Hexis produce distinct behaviour.

U3.3 Operational definitions

Monitoring is externally grounded. The explicit tier of line 2 is confidence revised against evidence or a check that the harness supplies, read on calibration, never on verbalised confidence and never on in-context self-reflection. This answers the self-correction finding (Huang et al., 2024): the programme reads that finding as showing that a model's critique of its own output is not independent evidence about that output, and so

monitoring in Hexis is not self-reflection. Any deliberation module is programmatic scaffolding over harness-supplied checks, never a reflection loop. A prompted-self-reflection arm will run as a control and is expected to be flat or negative against external grounding; if it matches, the middle-line distinction has no content and the programme will say so.

Recall is packet-conditional. The explicit tier of line 3 is tested as retrieval of a specified record when it is present and abstention when it is absent. The predicate is narrower than episodic memory in Tulving's sense and claims nothing about it.

Contexts are read, not chosen. The outer trigram is a property of the situation: which of the three capacities it already supplies explicitly. The router reads it; the agent does not select it. Because each outer line asks a checkable question — is a course of action set, has a check been run, is the record present — the router's three context outputs are classification targets a human can label, which is what C3's agreement test needs.

U4. A tradition of changing interpretations

The I Ching offers a history of people using a shared set of figures to think about changing situations. They asked different questions of it: how to act, how to cultivate judgement, how to understand natural processes, or how to organise patterns mathematically. Richard J. Smith's histories explain these changing uses (Smith, 2008; 2012). For Hexis, the history helps distinguish the structure it adopts from the interpretation it must test.

U4.1 The core text, the Ten Wings and later commentaries

A reader opening a modern I Ching may encounter several historical layers on the same page. The older **Zhouyi** (周易) core contains sixty-four hexagrams, their names, judgements about the whole situation, and statements associated with individual lines. The **Ten Wings** are later commentaries that explain the figures, their images and their implications for conduct. Still later commentators and translators add further interpretation. Figure U4.1 separates these layers (Smith, 2012, chs. 1–2).

“Ten Wings” counts parts, rather than ten separate titles. The *Tuanzhuan*, explaining the judgements, the *Xiangzhuan*, explaining images and lines, and the *Xici*, or Great Commentary, each have two parts. Four other works complete the ten: the *Wenyan* on Qian and Kun, the *Shuogua* on trigrams and their associations, the *Xugua* on the hexagram sequence, and the *Zagua* on contrasting pairs. Their authorship and formation are historical questions; the traditional attribution to Confucius is not a secure account of how they were composed (Smith, 2012, ch. 2).

The commentaries help explain why the book became a resource for reflection on conduct. The Great Commentary connects hexagram 24, Fu (復), with self-knowledge. The commentary on the judgement of hexagram 47, Kun (困), warns against reliance on speeches when facing adversity; Legge's note to the basic judgement makes the contrast with action explicit. These passages support a discussion of self-examination and conduct



What is in the I Ching?

1 THE OLDER CORE — Zhouyi 周易

64 six-line figures → hexagram names → judgements → line statements

Each figure has six line statements; Qian and Kun also have special additional statements.



2 THE TEN WINGS — Shiyi 十翼

Seven named works, counted as ten parts: 2 + 2 + 2 + 1 + 1 + 1 + 1.

彖傳 Tuanzhuan	2	Explains the hexagram judgements.
象傳 Xiangzhuan	2	Explains whole-figure images and individual lines.
繫辭傳 Xici zhuan	2	Great Commentary: change, action and cultivation.
文言傳 Wenyan zhuan	1	Develops the meanings of Qian and Kun.
說卦傳 Shuogua zhuan	1	Explains trigrams and their associations.
序卦傳 Xugua zhuan	1	Explains the received sequence of hexagrams.
雜卦傳 Zagua zhuan	1	Contrasts and characterises hexagram pairs.

The two parts of Xiangzhuan are not “Great Image” and “Small Image”; both occur within it.



3 LATER COMMENTARIES AND DIAGRAMS

Wang Bi / Han Kangbo → Kong Yingda: an authoritative commentary tradition.

Song interpreters, including Shao Yong and Zhu Xi: diagrams and further readings.

Successive editions interleave or separate core and commentary in different ways.

Relatively stable reference: the received 64-figure repertoire and textual tradition.

Historically variable: early manuscripts, ordering, layout, interpretation and later diagrams.

Figure U4.1. The received I Ching combines an older core and the Ten Wings. The ten parts belong to seven named works; later commentaries and diagrams form further layers. The arrows distinguish textual layers, rather than establishing precise dates for their composition.

in their historical setting (Legge, 1882, Appendix III, §II, ch. VII, ¶52; Appendix I, §II, hexagram 47; hexagram 47, judgement and note). “Mindfulness” would introduce a modern interpretive label here.

The **received text** is the transmitted reference framework chosen for this paper. Early manuscripts show that its history includes variation: the Mawangdui silk manuscript, buried by 168 BCE, uses a different order and has differences in names and written forms (Smith, 2012, ch. 1). The figures are a durable resource, but a claim that every manuscript and interpreter preserved every detail unchanged would overstate the evidence. Hexis fixes its own reference tables so that the proposed reading can be tested consistently.



U4.2 Two approaches to reading the figures

Commentators combined two broad kinds of explanation. **Images and numbers** (象數, *xiangshu*) connected figures with patterns in nature, time and space. **Meanings and principles** (義理, *yili*) concentrated on what situations and line statements implied for judgement and conduct. These labels help a newcomer recognise different questions being asked of the same text (Smith, 2008, ch. 3).

The approaches overlap. Scholars associated with one often used the other, and later interpreters sought to reconcile them (Smith, 2008, ch. 3). This matters for Hexis because its proposal also has two tasks: define a configuration precisely, then explain why that configuration might matter for action. The first task produces checkable structural definitions; the second produces hypotheses for the tests in U8.

U4.3 The line apparatus Hexis adopts

Line positions let a commentator discuss different roles within one situation. Han interpreters already used ideas of position, correspondence and centrality, alongside overlapping or nuclear trigrams (Smith, 2012, ch. 3). Wang Bi's influential commentary focused attention on the situation represented by a hexagram and the options associated with its lines. He incorporated elements of the Ten Wings into his presentation of the basic text (Smith, 2008, ch. 4).

Figure U4.2 makes the three adopted rules visible in **63 Jiji — After Completion** (既濟). Its yang lines occupy odd positions and its yin lines even positions, so every line is proper under U2's definition. Each paired position, 1 with 4, 2 with 5 and 3 with 6, has opposite polarity, so all three pairs correspond. Positions 2 and 5 are central within their trigrams.

Hexis turns selected line relations into explicit definitions for its fit rule (U5.2). Historical commentators used these ideas as interpretive resources, including explanations of why a seemingly improper position could be appropriate in a particular situation (Smith, 2008, chs. 3–4). The project's definitions give the reader a precise object to inspect and test.

An authoritative commentary tradition helped transmit the apparatus. Kong Yingda's *Zhouyi zhengyi* drew on Wang Bi and Han Kangbo and became an important standard. Later, Cheng Yi's moral readings and Zhu Xi's account of divination and cultivation shaped another influential tradition, supported by the examination system (Smith, 2012, ch. 3). “Canonical” therefore needs a context: a particular edition, institution and period could establish a preferred reading while alternatives continued to circulate.

U4.4 Diagrams and operations: Shao Yong, Lai Zhide and Jiao Xun

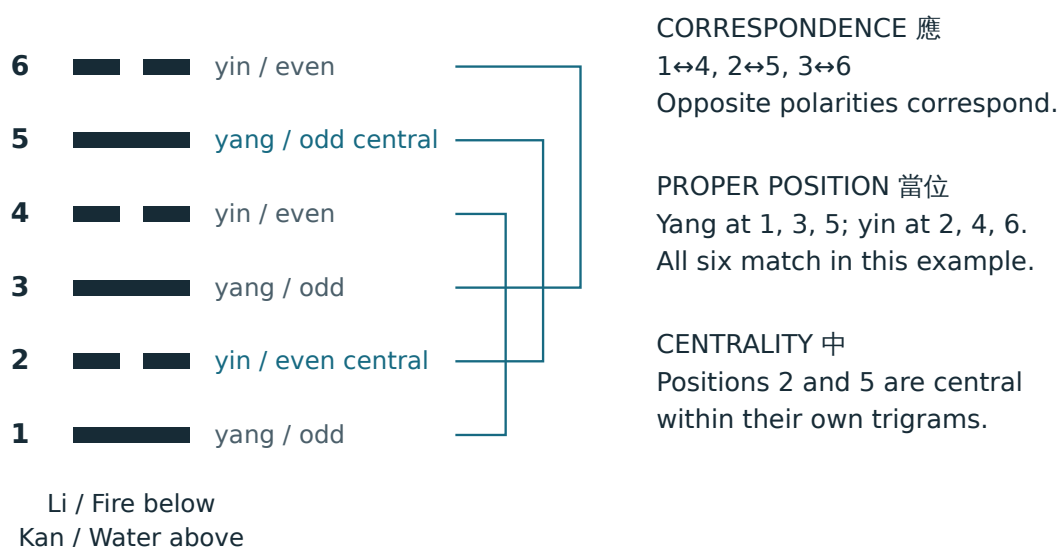
Diagrams made relationships among the figures available for inspection. Shao Yong developed patterns of successive doubling and spatial arrangement, giving readers ways to connect individual figures with a larger system. The Early Heaven and Later Heaven diagrams became influential parts of this visual tradition (Smith, 2012, ch. 3).

An **operation** states what to do to a figure to obtain another one. Lai Zhide's terms *cuogua* and *zonggua* identify two such relations: complement all six lines, or turn the figure upside down (Smith, 2008, ch. 6). Under the definitions in U2, a reader can



The line apparatus Hexis adopts

Example: 63 既濟 Jiji — After Completion. Lines are counted from the bottom.



In Hexis, correspondence and proper position feed the fit rule (U5.2).
Centrality identifies two positions; the figure adds no new scoring term.

Figure U4.2. The line apparatus Hexis adopts, illustrated with 63 Jiji. Correspondence and proper position contribute to the fit rule in U5.2. Centrality identifies two positions within the structure.

perform either transformation and check its result independently of an interpretation of the hexagram text.

Jiao Xun used transformations to investigate connections between passages. Figure U4.3 reproduces the structure of an example described by Smith: complementing all six lines of 52 Gen (艮) gives 58 Dui (兌); complementing the lower three then gives 31 Xian (咸). Repeating those two operations gives 41 Sun (損) and returns to 52 Gen. Jiao compared recurring phrases in the texts of related figures as part of his interpretation (Smith, 2008, ch. 7, fig. 7.7).

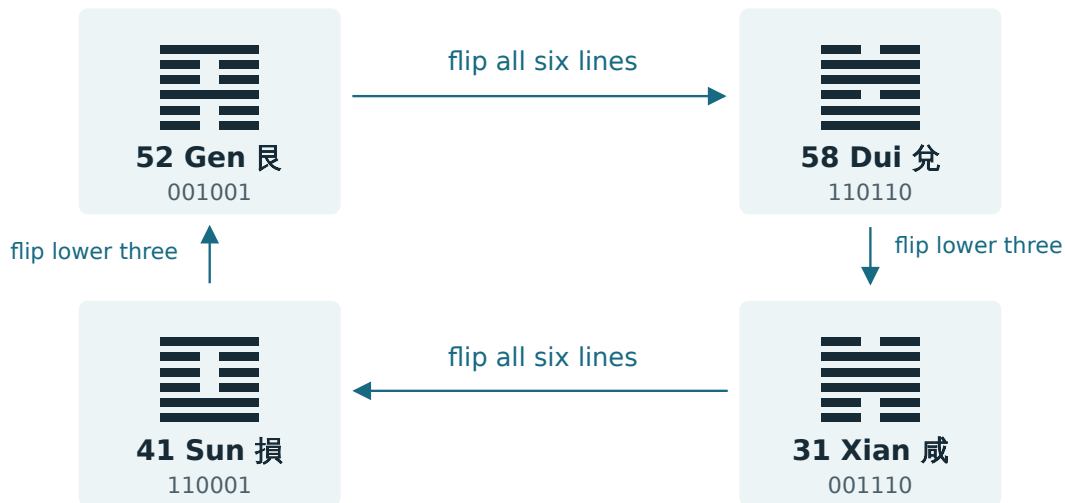
“Hexis treats relations as operations” means that its definitions also specify reproducible changes to line patterns. For example, inversion reverses the six positions, and applying it twice restores the starting figure. The structural checker can test that fact across all sixty-four figures. The relationship to Jiao is methodological: explicit transformations organise comparison. Hexis gives those transformations its own architectural purpose and tests.

U4.5 The Shao Yong diagrams and the binary reading

The circle and square in Figure U4.4 display the same sixty-four hexagrams in two arrangements. The square lets a reader inspect all eight-by-eight trigram combinations. The circle places the figures around a perimeter, exposing another pattern of successive line changes. Both make the whole repertoire visible at once (Smith, 2012, ch. 5, fig. 5.2).

What does “a relation as an operation” mean?

Jiao Xun's example, as described by Smith (2008, chapter 7, figure 7.7).



An operation specifies which lines change; the resulting figure is reproducible. Jiao used related figures to compare recurring phrases in their line statements. Hexis adopts explicit structural operations for its own definitions and checks. Its permitted routes are specified separately in U5.3.

Figure U4.3. A worked example of structural operations in Jiao Xun's interpretation. Each arrow specifies which lines change. This historical example illustrates a method of relating figures; permitted Hexis routes are defined in U5.3.

Joachim Bouvet and Gottfried Wilhelm Leibniz saw a binary structure in these arrangements. Replacing broken and solid lines with 0 and 1 makes their mathematical relationship visible. Smith's account favours the view that the diagram confirmed Leibniz's existing binary arithmetic. Shao, Bouvet and Leibniz nevertheless gave the figures different cosmological and religious interpretations (Smith, 2012, ch. 5). Hexis uses binary notation to specify states and operations; U3 states its separate cognitive interpretation.

U4.6 Translation and adaptation beyond China

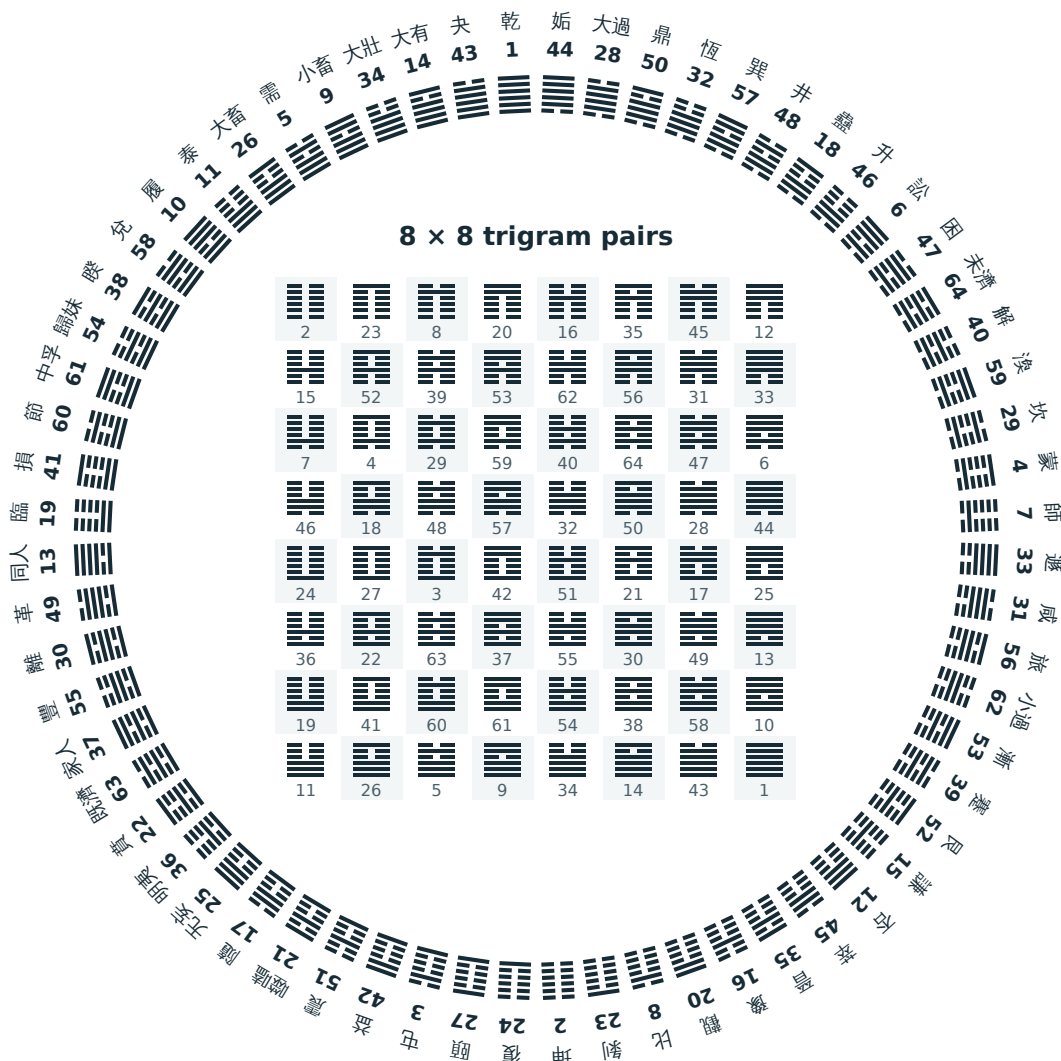
A translation carries choices about which commentary tradition to foreground. Wilhelm's work, developed with Lao Naixuan, drew on interpretations associated with Cheng Yi and Zhu Xi and helped present the I Ching as a book of wisdom to Western readers (Smith, 2012, ch. 5). Smith's comparison of translations makes these differences visible: access to an early manuscript, Wang Bi's commentary and a later synthesis can give readers different views of the work (Smith, 2008, front matter). That is why this paper identifies the edition behind a textual claim.

Kim Ilbu's *Jeongyeok* is an example of adaptation in Korea. It repositions the trigrams within a new account of change; Smith discusses its place in Korean reception, while Chung reproduces the particular circle used here (Smith, 2012, ch. 4; Chung, 2021, fig.



Shao Yong's circle and square

Early Heaven arrangement — explanatory reconstruction



Each layout contains all 64 figures. Numbers are King Wen lookup labels, added here for cross-reference; they do not determine the diagram's order. Circle lines read from centre outward; square lines read bottom upward.

Figure U4.4. The Early Heaven circle and square attributed to Shao Yong: an explanatory vector reconstruction. The King Wen numbers are added lookup labels, separate from the ordering principle. In the circle, read lines from the centre outward; in the square, read them from bottom to top.

1). For Hexis, that circle supplies another explicit successor and predecessor table. Its complement property is a structural question checked in U5.4.

Jung's use of the I Ching belongs to its reception as a resource for self-examination (Jung, 1949; Smith, 2012, ch. 5). The historical account and the analogy that helped formulate Hexis's reading are distinguished in U3.2, where the question of derivation of psychological typology remains separate.



U4.7 The mirror: why Hexis must test its own reading

Smith's "mirror" argument explains why interpreters can find different concerns in the same book. A reader concerned with government may foreground political judgement; one concerned with cultivation may foreground conduct; a mathematician may foreground patterns among the figures. The text's interpretive flexibility lets their questions shape what they see (Smith, 2012, ch. 2; Smith, 2009).

The same issue applies to an engineer. Starting with problems of planning, monitoring and recall makes those capacities available as an interpretation of the three lines. The interpretation may be coherent because the engineer has chosen a coherent mapping. That alone leaves open whether this particular mapping helps an agent. Hexis therefore states the assignment as a project convention and ties its proposed benefit to comparisons in U8.

C2a will compare the specified geometry and relations with alternatives. C2b will ask whether a learned graph exceeds the pre-registered margin. Those comparisons will test whether the chosen structure earns its place in the architecture. Smith's history helps the reader understand where Hexis's reading sits; the experiments will assess what it contributes. Figure U4.5 summarises the readings surveyed in this chapter and the use the paper makes of each.



Readings and their relevance to Hexis

The last column states this paper’s use of the history, rather than assuming direct influence.

Period	Reader / tradition	Historical contribution	Relevance to Hexis
Early textual history	Received text; Mawangdui	A durable repertoire, with early variants.	Defines the reference edition Hexis chooses.
Han	Jing Fang and others	Position, correspondence; overlapping trigrams.	Historical sources of structural definitions.
3rd–7th centuries	Wang Bi; Han Kangbo; Kong Yingda	Situations and line relations; authoritative editions.	Context for the apparatus adopted in U2.
Song	Shao Yong	Generative and spatial diagrams.	Early Heaven order; binary representation.
Song onwards	Cheng Yi; Zhu Xi	Cultivation and divination; later orthodoxy.	Shows how readings and editions evolved.
Ming	Lai Zhide	Complement and inversion.	Named transformations used in U2.
Qing	Jiao Xun	Relations used to compare textual phrases.	A precedent for explicit operations; distinct purpose.
Early modern Europe	Bouvet; Leibniz	Read a hexagram arrangement in binary terms.	Clarifies encoding versus historical interpretation.
19th century: Korea	Kim Ilbu	Jeongyeok: a different trigram arrangement.	Third trigram circle specified in U2.
20th century	Wilhelm; Jung	Translation and self-examination.	Reception history; typology derivation unestablished.
Modern scholarship	Richard J. Smith	A history of changing interpretations.	Requires Hexis to identify and test its own reading.
2026	Hexis / Yarrow	Cognitive assignment and proposed tests.	Contemporary project contribution; tests pending.

Historical sources: Smith (2008; 2012), with Chung (2021) for the Jeongyeok diagram.

Figure U4.5. Historical readings and their relevance to Hexis. The final column states this paper's use of the history; direct influence is attributed only where separately documented.

U5. Structural results and the execution boundary

The definitions produce tables and numerical consequences that can be checked before an agent is built. The accompanying structural checker reproduces the canon partition, relation properties and worked-example scores under the current rulings. This chapter explains what those calculations establish and where behavioural tests are still required.

U5.1 The canon partition: decision states and execution states

Ruling. Hexis partitions the King Wen sequence at 32/33: it calls hexagrams 1–32 its upper or decision canon and 33–64 its lower or execution canon. These names denote the project’s partition. The upper canon holds the *decision states*: the 32 standpoint × context pairs in which the active standpoint produces a typed decision artefact. The lower canon holds the *execution states*: the 32 pairs in which the active standpoint carries out a decision; only in an execution state may a tool execute. All 64 states are routable; the canon fixes the kind of output, not whether the state can be reached; the canon of the



state the router assigns is read in code, labels the turn as decision or execution, and the contract requires its enforcement at the tool layer. Reading relations are defined on all 64 hexagrams.

The 64 states by standpoint (rows) and context (columns), coloured by canon

Upper canon, King Wen 1-32: decision states (typed decision artefact). Lower canon, 33-64: execution states (tools may execute).

outer trigram — the context (lines 4-6)

	☰ [111]	☱ [110]	☲ [101]	☳ [100]	☴ [011]	☵ [010]	☶ [001]	☷ [000]
☰ Heaven [111]	1	43	14	34	9	5	26	11 Tai
☱ Lake [110]	10	58	38	54 Guimei	61	60	41 Sun	19
☲ Fire [101]	13	49	30	55	37	63 Jiji	22	36
☳ Thunder [100]	25	17	21	51	42	3	27	24
☴ Wind [011]	44	28	50	32	57	48	18	46
☵ Water [010]	6	47	64	40	59	29	4	7
☶ Mountain [001]	33	31	56	62	53	39	52	15
☷ Earth [000]	12 Pi	45	35	16	20	8	23	2 Kun

inner trigram — the agent (lines 1-3)

☰ decision state (upper canon, 1-32). Row counts are the decision fibres: ☰ 6, ☱ 2, ☲ 3, ☳ 6, ☴ 3, ☵ 4, ☶ 2, ☷ 6.

☱ execution state (lower canon, 33-64). The execution fibres are the complements: ☰ 2, ☱ 6, ☲ 5, ☳ 2, ☴ 5, ☵ 4, ☶ 6, ☷ 2.

☰ the running example (U1.4): 2 Kun → 11 Tai → 26 Dachu (decision) → 41 Sun (execution).

Inversion never crosses the partition; the nuclear relation can (11 Tai → 54 Guimei). Partition at 32/33 by the project lead's ruling; the received division of the text is 30/34 (U5.1).

Figure U5.1. The two canons: the 64 states by standpoint (rows) and context (columns), decision states in the upper canon and execution states in the lower.

The boundary at 32/33 is the project lead's determination, grounded in his reading of 31 Xian and 32 Heng as the close of a canon of decisions, although in the received order 31 Xian opens the second part. It departs from the received text's division into thirty hexagrams in the first part and thirty-four in the second (Hacker and Moore, 2003, p. 219). The numerical partition is therefore a Hexis definition. Its structural consequences can be checked from the table, and its proposed behavioural meaning is subject to the panel test below.

Consequence. Each standpoint's set of contexts splits into a decision fibre and an execution fibre, with sizes fixed by the King Wen table. Decision contexts: ☰ Heaven [111] 6, ☱ Lake [110] 2, ☲ Fire [101] 3, ☳ Thunder [100] 6, ☴ Wind [011] 3, ☵ Water [010] 4, ☶ Mountain [001] 2, ☷ Earth [000] 6. Execution contexts are the complements: ☰ Heaven [111] 2, ☱ Lake [110] 6, ☲ Fire [101] 5, ☳ Thunder [100] 2, ☴ Wind [011] 5, ☵ Water [010] 4, ☶ Mountain [001] 6, ☷ Earth [000] 2. Every standpoint executes in at least two contexts. The prior over standpoints is therefore uneven within each canon, and



complementary between the canons. It is reported, it is used as a null responder (L2), and it is never used as a training prior. Inversion never changes the canon — a state and its inversion are always both decision or both execution (the accompanying structural checker asserts this) — while the nuclear hexagram can cross it: the nuclear of 11 Tai — Peace, a decision state, is 54 Guimei — The Marrying Maiden, an execution state. Two registered predictions follow and are stated here so that they can fail: that the instrument's decision items concentrate in the upper canon, and that good trajectories pass through a decision state before an execution state. The sealed refuter of the partition is a blind panel asked to tell the two output kinds apart with the state hidden; if it cannot, the partition labels nothing. Figure U5.1 lays the 64 states out by standpoint and context with the canon of each marked; Figure U5.2 reads the two fibres off it, standpoint by standpoint.

Each standpoint's eight contexts split into a decision fibre and an execution fibre

Fixed by the King Wen table; read from Figure U5.1 row by row. Every standpoint has at least two contexts of each kind.

☰ Heaven [111]	☰ 1	☰ 14	☰ 9	☰ 5	☰ 26	☰ 11	☷ 43	☷ 34	6 : 2
☵ Lake [110]	☰ 10	☱ 19	☲ 58	☲ 38	☱ 54	☱ 61	☱ 60	☱ 41	2 : 6
☲ Fire [101]	☰ 13	☰ 30	☱ 22	☲ 49	☱ 55	☱ 37	☱ 63	☱ 36	3 : 5
☳ Thunder [100]	☰ 25	☰ 17	☱ 21	☱ 3	☱ 27	☱ 24	☱ 51	☱ 42	6 : 2
☴ Wind [011]	☰ 28	☱ 32	☱ 18	☲ 44	☱ 50	☱ 57	☱ 48	☱ 46	3 : 5
☵ Water [010]	☰ 6	☱ 29	☱ 4	☱ 7	☲ 47	☱ 64	☱ 40	☱ 59	4 : 4
☶ Mountain [001]	☱ 31	☱ 15	☲ 33	☲ 56	☱ 62	☱ 53	☱ 39	☱ 52	2 : 6
☷ Earth [000]	☰ 12	☱ 16	☱ 20	☱ 8	☱ 23	☱ 2	☲ 45	☲ 35	6 : 2

☰ decision contexts (state in the upper canon) ☱ execution contexts (state in the lower canon)

The uneven split is the fibre prior: reported, used as a null responder (L2), never a training prior.

Figure U5.2. The two fibres: each standpoint's eight contexts split into decision contexts and execution contexts, with sizes fixed by the King Wen table.

The example, second pass. Suppose the agent of U1.4 is in ☷ Earth [000] when the message arrives — fully implicit, working from pattern — and the context reads ☷ Earth [000] as U3.1 read it. The state is inner ☷ over outer ☷, which is 2 Kun — The Receptive, King Wen 2, in the upper canon: a decision state. Whatever the agent does on this turn, the canon fixes the kind of output: a typed decision artefact, and no tool. The pattern-matching agent's characteristic failure in this case — issuing the refund from what such requests usually deserve — would require a tool call, which the proposed tool layer refuses in 2 Kun. A premature refusal stated in text is a decision-quality error that C4's matched-compute comparison scores. The partition fixes the turn's output as a decision rather than an action, and leaves the quality of the decision to the standpoint the router moves to next (U5.2) and to the evidence the harness supplies before it acts (U7).

U5.2 Fit as the routing objective: the correspondence count is inversion-invariant, proper position is not

The router needs a score that ranks standpoints for a given context. That score must also make the selection checkable. Early drafts used two conflicting meanings of fit. *Match* favoured the same polarity at corresponding agent and context lines. *Correspondence* favoured opposite polarities. Match makes the agent mirror what the context already supplies. Correspondence describes a division of labour, but its count alone cannot distinguish inversion pairs. The first proof shows why.

Sidebar A — What the two proofs establish. Reversal permutes correspondence pairs, so their count stays the same. It moves every line from an odd position to an even one or conversely, so proper-position counts add to six across an inversion pair. Adding proper position therefore lets the fit rule distinguish some inversion pairs. These are arithmetic facts, independent of model behaviour.

Proposition 1 (the correspondence count is inversion-invariant). Let h be a hexagram with lines $\ell_1 \dots \ell_6$ and let h' be its inversion, with lines $\ell'_k = \ell_{7-k}$. The correspondence pairs of h' are (1,4), (2,5), (3,6) in h' 's numbering, which are (6,3), (5,2), (4,1) in h 's numbering: the same three pairs. Since correspondence at a pair depends only on whether the two lines differ, inversion permutes the three correspondence indicators and preserves their sum. Thus h and h' have identical correspondence counts. The same holds for match. ■

Any score formed by an unweighted count of correspondence, or by its complementary match count, is therefore inversion-invariant. The same holds when weights are symmetric under inversion; arbitrary unequal weights on exchanged pairs need not preserve it. But the tradition gives some inversion pairs opposite verdicts: 11 Tai — Peace and 12 Pi — Standstill are inversions of each other, both have full correspondence at all three pairs, and the tradition's verdicts on them are opposite. Which trigram sits where must enter the definition.

Proposition 2 (proper position breaks the symmetry). Inversion sends position k to $7 - k$, hence odd positions to even and even to odd. A yang line proper at an odd position becomes a yang line at an even position, hence improper, and likewise for yin. Therefore every proper line of h is an improper line of h' and conversely, and the number of proper lines satisfies $\text{proper}(h') = 6 - \text{proper}(h)$. ■

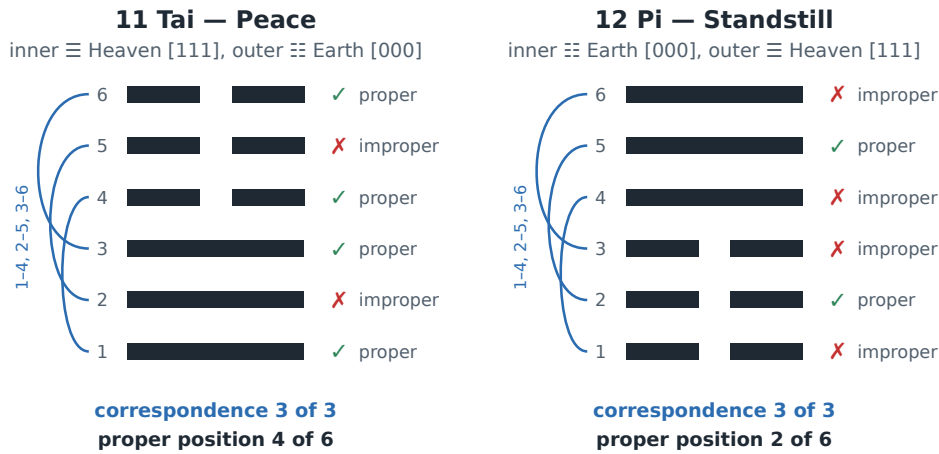
The counts follow: 11 Tai has 4 proper lines of 6 and 12 Pi has 2; 63 Jiji has 6 of 6 and 64 Wei Ji has 0. Figure U5.3 draws the Tai/Pi pair with the correspondence pairs and the proper lines marked.

What the two propositions mean for the score: correspondence measures how agent and context divide the three capacities, and its count is unchanged by inversion, so that count alone cannot tell an inversion pair apart; proper position adds information about where the lines sit and can tell some inversion pairs apart. The propositions establish these properties of the score. Whether the score so defined helps an agent is what C2a and C3 test.



The correspondence count is inversion-invariant; proper position is not

12 Pi is 11 Tai with its six lines reversed. All three correspondence pairs hold in both; the tradition gives them opposite verdicts.



Correspondence (blue): lines k and $k+3$ of opposite polarity. Proper position: yang at 1, 3, 5; yin at 2, 4, 6.
Reversing the lines maps the pair set onto itself and swaps odd and even positions, so $\text{proper}(\text{inversion}) = 6 - \text{proper}$.
Likewise 63 Jiji (6 of 6 proper) and 64 Weiji (0 of 6) share full correspondence.

Figure U5.3. 11 Tai — Peace and 12 Pi — Standstill: the correspondence count is inversion-invariant, proper position is not.

Ruling. Fit between agent and context is correspondence plus proper position — the line rules the commentary tradition received through Wang Bi — with centrality as a weight on lines 2 and 5. The paper scores proper position over all six lines; Wang Bi himself held that the first and top lines have no fixed position (*Zhouyi lüeli*, “Bian wei”). In `hexmath.fit`, correspondence contributes one unit per pair split between the two lines, proper position one unit per line, and lines 2 and 5 carry a weight of at least one on both contributions; the weight has a default of one and any other value used in a read is injected, sealed configuration. Match and bare correspondence are kept in code as named alternatives, for ablation only. A regeneration test scores all 64 hexagrams under the sealed rule against the favourable and unfavourable labels of the pinned editions and *reports* the disagreement count; it does not assert zero (L7).

Two completed states. Under the paper's six-position rule the hexagram with every line proper is 63 Jiji, in the lower canon: an execution state, so the all-proper configuration is one in which the system acts. On the project's reading of the Jeongyeok arrangement, which awaits a pinned edition (note 1), the completed state is 11 Tai, in the upper canon: a decision state, as is its inversion 12 Pi. The regeneration test reports both against the tradition and privileges neither.

The objective, term by term. The objective combines an allocation rule with a fixed preference over line positions. Correspondence rewards capacities for which exactly one of agent and context is explicit. The agent supplies an explicit capacity where the context supplies none; it can leave that capacity implicit where the context already supplies it. This is why the outer lines record available information (U3.1).



Proper position adds a preference that stays the same across situations. Its agent template makes planning and recall explicit and monitoring implicit. Its context template supplies monitoring, with neither a plan nor a record imposed. The only hexagram satisfying all six positions is 63 Jiji, [1 0 1 0 1 0]: inner ☲ Fire [101], outer ☵ Water [010].

Under the authors' reading, this template places planning and recall with the agent and monitoring outside it. That agreement with U3.3 follows from a chosen interpretation and is a design property; the regeneration test checks the arithmetic and C2a tests the proposed value of the relations. The template also prefers packet-conditional recall to an externally supplied record; the memory comparison tests the value of that choice.

The example, third pass. The refund case can be scored by hand with the structural checker at its default centrality weight. Under outer ☷ Earth [000], the initial state 2 Kun scores 3. No pair corresponds, and three lines are proper.

The highest score is 7, shared by two standpoints. ☰ Heaven [111] gives 11 Tai — Peace: three corresponding pairs and four proper lines. It is a decision state reached by the inner complement. ☲ Fire [101] gives 36 Mingyi — Darkening of the Light: two corresponding pairs and five proper lines. It is an execution state reachable from ☷ Earth [000] by the predecessor on the King Wen trigram circle, a route-valid arrangement relation under U5.3 — distinct from succession through the numbered hexagrams, which is reading-only. Both tied standpoints are therefore reachable by a route-valid relation.

The illustration breaks the tie at the execution boundary: no decision artefact yet exists for an execution state to carry out, so 11 Tai is selected and the agent makes planning, monitoring and recall explicit before acting. This tie-break is the example's convention; the tool-layer rule requires only that a tool execute in an execution state. Chapter U7 follows the context update when the record arrives.

U5.3 Relation typing

A route changes the agent's standpoint while holding the currently read context fixed. That requirement determines which structural operations can propose a route. Changing an inner line, taking the inner complement, or following a trigram-circle successor or predecessor changes the agent side. Inversion instead exchanges agent and context lines, and the nuclear operation combines lines from both. Those operations contribute to reading a state; the route-valid set preserves the agent/context distinction. The ruling below states the permitted uses.

Ruling. Every relation is typed. **Route-valid:** a changing inner line (lines 1–3: the agent changes tier on one axis); the inner complement (all three agent lines flip); and the successor or predecessor along any of the three arrangement cycles. **Reading-only:** the full opposite, inversion, the nuclear hexagram, derived hexagrams and King Wen succession. Reading-only relations feed the router as features and the instrument's item engine as structure; they are never cited by a route. A changing *outer* line (4–6) is a situation change that the router reads, not a route the agent chooses. The canon of the state a route lands in labels the next turn as decision or execution; a route may cross canons, and *hold* and *escalate* remain available in either. The table is data, to be versioned in the public-intent repository, and it changes only by the relation-set ablation



(U8, C2a) and a ruling. Figure U5.4 is the table. Exterior succession at harness level — the context handing over to the agent, reversed in time — is the one plausible home for inversion as a route; it is recorded as an open question in the change register (DQ-8).

Relations and their permitted use

Relation	Use	Role
Changing inner line 1-3	Route-valid	Changes one agent capacity tier
Inner complement	Route-valid	Changes all three agent lines
Arrangement successor or predecessor	Route-valid	Moves on one of the three tabled trigram cycles
Changing outer line 4-6	Read-only input	Records a situation change
Full opposite	Reading-only	Complements all six lines
Inversion	Reading-only	Reverses the six lines
Nuclear	Reading-only	Takes lines 2-4 and 3-5
Derived hexagram	Reading-only	Flips a set of lines
King Wen succession	Reading-only	Moves in whole-hexagram order

The canon of the state reached determines output kind. Hold and escalate remain available.

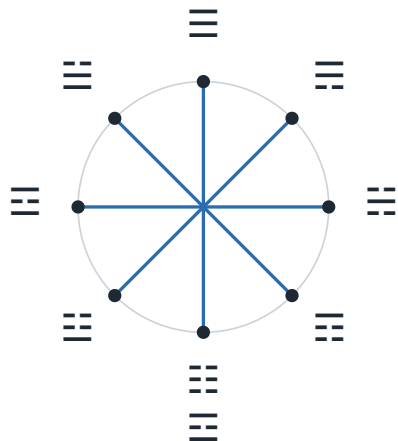
Figure U5.4. The relation table: permitted uses under the agent/context typing.

U5.4 The three arrangements

Each arrangement is a permutation of the eight trigrams around a circle, and each defines a successor cycle. Four properties are reproduced by the standalone structural checker and are to be asserted by the contract's tests. In the Fuxi circle every trigram sits diametrically opposite its complement; so does every trigram in the Jeongyeok circle. In the King Wen circle one of the four opposite pairs is a complement pair (☲ Fire opposite ☵ Water) and the other three are not (☷ Earth opposite ☶ Mountain, ☴ Lake opposite ☳ Thunder, ☰ Heaven opposite ☱ Wind); the count is recomputed from the circle as tabled in the contract, and the figure follows the computation. The Jeongyeok circle is neither a rotation nor a reflection of the Fuxi circle. The three cycles are therefore three distinct successor permutations with different step profiles; the profiles are regenerated in code and written to the test report rather than stated here. Figure U5.5 draws the three circles with complement pairs as chords.

The four-stage view becomes precise once the step is specified. Write s for one successor around an eight-trigram circle. Repeating s visits all eight trigrams. Repeating s^2 , which takes every second successor, divides the circle into two four-state cycles. In both the Fuxi and Jeongyeok arrays, four ordinary steps reach the complementary trigram, so each

Three tabled arrangement cycles

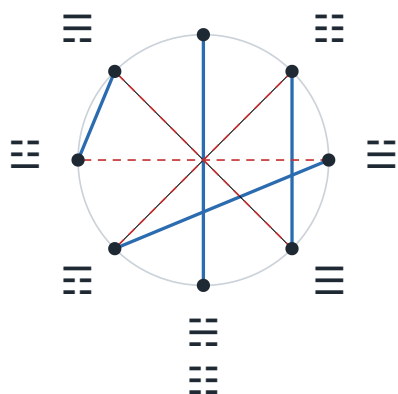


Fuxi

4 of 4 opposite pairs are complements.

Clockwise from top:

☰ ☷ ☵ ☴ ☳ ☶ ☱ ☲

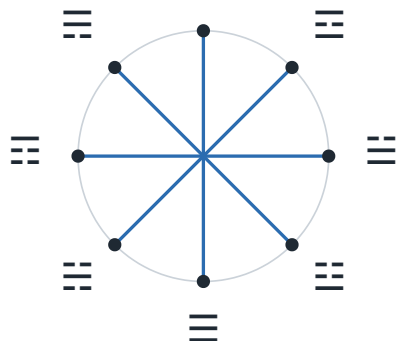


King Wen

1 of 4 opposite pairs are complements.

Clockwise from top:

☰ ☷ ☵ ☴ ☳ ☶ ☱ ☲



Jeongyeok

4 of 4 opposite pairs are complements.

Clockwise from top:

☰ ☷ ☵ ☴ ☳ ☶ ☱ ☲

Neither rotation nor reflection of Fuxi.

Scholarly source for this order pending.

Blue chords: complements. Red dashed lines: non-complementary diameters.

Figure U5.5. The three arrangements with complement pairs as chords.

four-state cycle contains two complementary pairs. Figure U5.6 shows the distinction. Identifying each complementary pair as a single unit instead gives a four-position quotient of the ordinary circle; that is a different construction from taking two steps at a time.

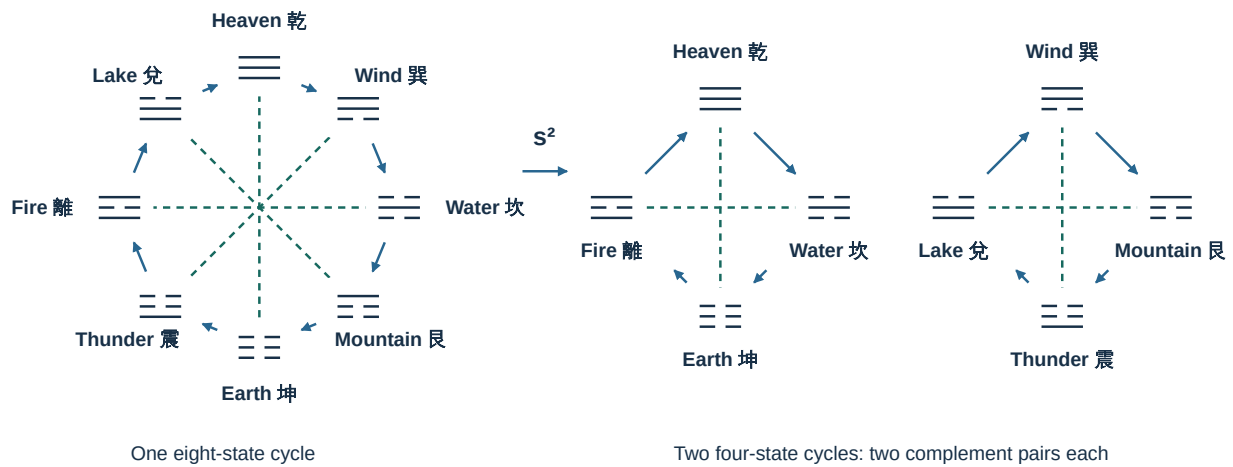
These cycles organise successive trigrams. The fibres in U5.1 instead collect the contexts paired with a fixed standpoint. Keeping the terms separate identifies what is held fixed in each construction. A two-step view is a way of describing the checked array; the existing route ruling continues to specify which individual transitions the harness may propose.



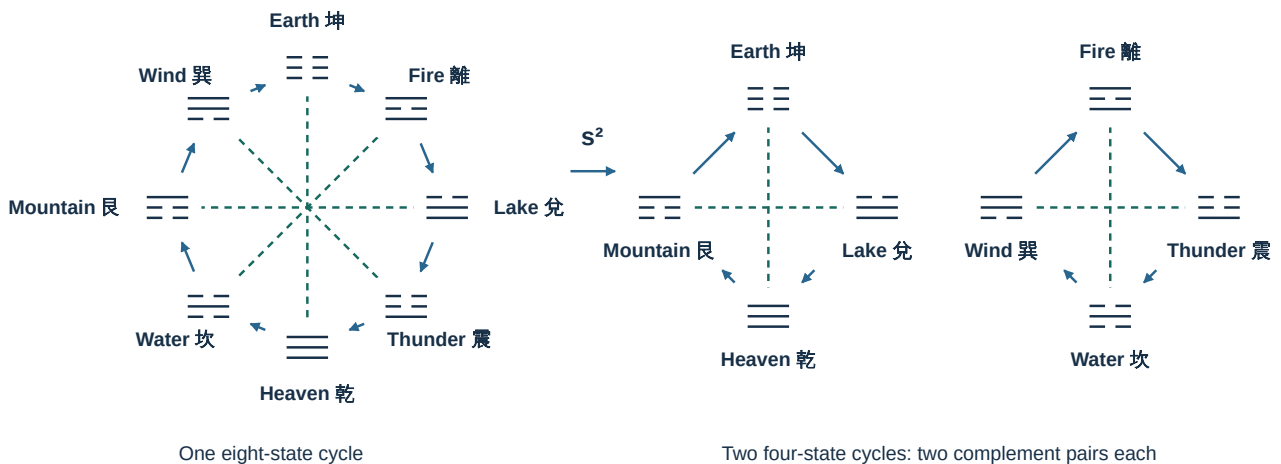
Eight successors and the two-step cycles

s: move one place; s^2 : move two places each time. Chords mark complementary pairs.

Fuxi / Early Heaven



Jeongyeok



Derived from checked arrays: not context fibres, and not a new primitive route.

Figure U5.6. One-step succession visits all eight trigrams. Taking every second successor produces two four-state cycles. In both checked arrays, each four-state cycle contains two complementary pairs, drawn as dashed chords. These are properties of the specified arrays; the figure introduces no new route permission.

Each tabled arrangement supplies a successor relation to the route-valid set. The relation-set ablation (U8) tests whether these relations add predictive value. The programme also records an expectation that favourable trajectories follow successors on the Jeongyeok trigram circle more often than successors on the King Wen trigram circle; historical attribution of that expectation awaits a pinned scholarly edition.

U5.5 The cast null

The traditional yarrow-stalk method yields a known, asymmetric distribution over the four line outcomes. A responder that assigns lines by drawing from that distribution is the no-information baseline for the router's six line outputs. The `hexmath.cast("yarrow")` specification calls for regenerating the distribution from the method itself — the three



divisions of forty-nine stalks, counting remainders — rather than copying figures from any text, and calls for a test that it differs from the three-coin method in its changing-line asymmetry. The distribution is not stated here; the planned regeneration test must establish the derivation before release. This null is the reason the open-source contribution is called Yarrow: the yarrow cast is the baseline the router has to beat.

Sidebar B — Why compare with a cast? The cast null is a no-information responder generated from a specified line-sampling procedure. It asks whether the router predicts more than that baseline within each context. The fibre-prior responder provides a second null based on the uneven state counts. Beating either null would be evidence about prediction under the test, not evidence for the historical method of divination.

U6. Interdisciplinary foundations and their limits

This chapter states what the proposal takes from philosophy, the psychology of wisdom, consciousness science, metacognition research and reinforcement learning, and how much of the argument rests on each. Each part closes with a sentence stating the weight the foundation bears, because the weight is uneven: some of these literatures supply a vocabulary, some a formal ancestor, one a measurement discipline, and the tests of U8 will assess whether the proposed architecture has behavioural value.

U6.1 Philosophy: the name, and why dispositions

The programme is named for Aristotle's ἕξις (*hexis*). In the *Nicomachean Ethics* Aristotle sorts what is found in the soul into passions, faculties and states of character, argues that virtue is neither a passion nor a faculty, and concludes that it must be a state of character — a *hexis*, a settled disposition acquired by practice (*NE* II.1, 1103a14–b25; II.5, 1105b19–1106a13; trans. Ross). Practical wisdom (φρόνησις, *phronesis*) is itself such a state — a reasoned and true state of capacity to act concerning what is good or bad for a person — a virtue rather than an art, something that cannot be forgotten, and something concerned with particulars, which requires experience of situations. Aristotle therefore holds that the young can master mathematics before they acquire practical wisdom (*NE* VI.5, 1140a24–b30; VI.7, 1141b8–23; VI.8, 1142a11–16). Virtue of character, in turn, is defined as a state concerned with choice, lying in a mean determined by the rational principle by which the person of practical wisdom would determine it (*NE* II.6, 1106b36–1107a2).

The Chinese name, Hexis (時中, Shízhōng), names the programme's question of how a settled disposition should respond as circumstances change. In the *Zhongyong* (ch. 2), the phrase 君子而時中 connects appropriate conduct with its occasion. The Greek name emphasises the disposition; 時中 directs attention to its fitting expression in a particular situation. This pairing names the programme's question. The pairing is a modern naming choice, not a historical derivation. The behavioural partition and the quality and timing of action remain subjects for C1–C6 to test.



Two features of this account inform the proposal. First, a disposition is a standing readiness to perceive and act, acquired through repetition. Hexis uses the term for a persistent configuration of capacities held outside the model's weights. Whether that configuration produces stable behavioural differences is a research question (C1). Second, practical wisdom concerns particular situations. The instrument in L1 therefore scores the chosen response and its timing in a situation. Correct general reasoning alone does not satisfy its construct.

Modern practical-wisdom scholarship sharpens the measurement problem. Kristjánsson et al. (2021) describe four functions in their Aristotelian Phronesis Model: sensitivity to a situation's moral demands, integration of competing values, a motivating account of flourishing, and regulation of emotion through reason. They identify a shortage of rigorous quantitative work on this integrative capacity. The instrument in L1 responds to that measurement problem by scoring choices in situations rather than self-description.

Weight borne. The philosophy supplies the name, the category and the reason for scoring situated choice. It does not establish that a model has dispositions or that eight is the right number. C1-C4 test the proposed behavioural partition and routing.

U6.2 The psychology of wisdom: what the measures measure

Four measurement programmes in psychology define wisdom differently, and the comparator choice in L1 follows from the differences. The Berlin wisdom paradigm defines wisdom as an expert knowledge system concerning the fundamental pragmatics of life, assessed by scoring people's responses to difficult life problems against five criteria: rich factual knowledge, rich procedural knowledge, lifespan contextualism, relativism of values and life priorities, and the recognition and management of uncertainty (Baltes and Staudinger, 2000). Ardelt's Three-Dimensional Wisdom Scale treats wisdom as a latent trait with cognitive, reflective and affective components, measured by a self-report questionnaire developed on 180 adults aged 52 and over (Ardelt, 2003). Sternberg's balance theory defines wisdom as the application of intelligence, creativity and knowledge, mediated by values, toward a common good, through a balance among intrapersonal, interpersonal and extrapersonal interests over the short and long term (Sternberg, n.d., "Balance Theory of Wisdom"). Grossmann's wise-reasoning programme scores reasoning about concrete situations rather than answers about oneself, and a Toronto task force of wisdom researchers has since converged on a common model in which empirical approaches agree on the morally grounded application of metacognition to reasoning and problem-solving (Grossmann et al., 2020).

The reported intelligence associations are correlations, denoted by r : they describe how two measured quantities vary together. Zero indicates no linear association in the estimate, and a positive value indicates that higher scores tend to occur together. Dong, Weststrate and Fournier (2023, Table 4) report Berlin wisdom at $r = .224$ (95% confidence interval .168–.280) and wise reasoning at $r = .046$ (–.014–.107). The former is a small positive association; the latter is close to zero, with uncertainty extending to either side of zero. These are human measurements, not causal effects or estimates of AI behaviour.



The programme selects wise reasoning as its primary comparator because it concerns reasoning in particular situations and the reported intelligence association is small. This choice tests whether TAI captures something related to situated judgement while retaining a distinct emphasis on choices and their timing. The Berlin paradigm remains a secondary comparator. The comparison must be established on the systems being assessed: C6 tests the proposed construct's separation and convergence, and the dual-key hypothesis H11 (L1) tests the two scoring keys. The human correlations motivate that work; they do not settle it.

Weight borne. This literature supplies the construct lineage, candidate comparators and human evidence motivating separate validity tests. C6 addresses whether the proposed construct is measurable in language-model behaviour; the human correlations leave that question open.

U6.3 Consciousness science and metacognition: tiers, monitoring, and when to think

Three strands inform the design: indicator-based audits, distinctions between levels of processing, and the control of deliberation.

The indicator method and its limits. Butlin et al. (2023) propose computational indicators drawn from several theories of consciousness. These include recurrent processing, global workspace, higher-order and attention-schema theories, and predictive processing. Their report judges then-current AI systems against those indicators and finds no obvious technical barrier to building systems that satisfy them. Hexis uses the indicators as an audit framework (L7).

The theories underlying the indicators also face empirical tests. The Cogitate Consortium (2025) compared predictions from integrated information theory and global neuronal workspace theory in 256 participants, using fMRI, MEG and intracranial EEG. Some findings aligned with each theory, while others challenged key predictions. These included sustained posterior synchronisation for the former theory, and stimulus-offset ignition and some prefrontal representations for the latter. Hexis treats the indicator audit as a precaution; L7 states the behavioural criterion and the mimicry baseline an indicator must meet.

A capacity can affect behaviour while the information involved is represented in different ways. Dehaene, Lau and Kouider (2017) distinguish unconscious processing (C0), information made globally available for flexible use (C1), and self-monitoring of the system's own processing (C2). Global availability and self-monitoring are partly independent dimensions in their account. Their C-labels denote processing concepts; the identically numbered labels in this paper's claim table denote research claims.

Tulving (1985, pp. 2–3) describes a different distinction concerning memory and awareness. *Anoetic* awareness is bound to the present situation; it accompanies procedural memory, including acquired skills. *Noetic* awareness permits knowing about objects and events beyond the immediate situation, as in semantic knowledge. *Autonoetic*



awareness concerns remembering a personally experienced event as part of one's own past. The terms separate immediate engagement, knowing, and personal recollection. They are not interchangeable with Dehaene and colleagues' categories.

Hexis borrows the distinction between a capacity operating and a capacity operating with an explicit representation. Its engineering definition concerns planning, monitoring and recall: for example, whether monitoring supports revision when a check contradicts an answer. Figure U6.1 places the source distinctions alongside that operational choice. Research claim C1 tests whether the proposed tiers produce separable behaviour; the consciousness audit has its own scope in L7.

Processing, representation and the engineering choice

Two source frameworks, shown alongside each other rather than as equivalent stages

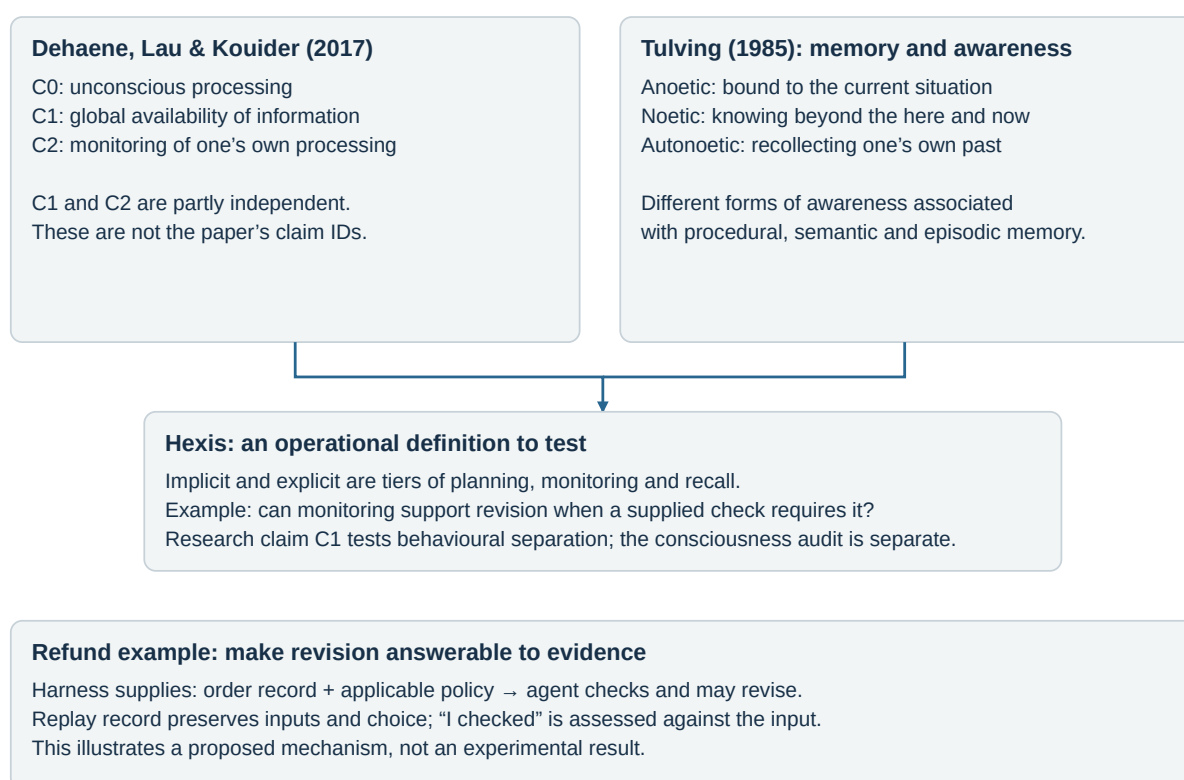


Figure U6.1. Dehaene and colleagues' processing distinctions and Tulving's memory/awareness distinctions provide different conceptual resources. Hexis adopts an operational implicit/explicit distinction and tests its behavioural effects under C1. The refund example shows how supplied evidence could make revision inspectable.

Monitoring and control. Nelson and Narens (1990) distinguish an object level from a meta level. Monitoring carries information about the object level to the meta level; control changes the object level. Applying that distinction requires evidence connecting performance to what the system registers about it.

Fleming and Lau (2014) distinguish three properties of confidence reports. Bias concerns how confident the system tends to be. Sensitivity concerns whether confidence distinguishes correct from incorrect judgements. Efficiency relates that sensitivity to task



performance. A simple correlation between confidence and accuracy can conflate these properties, so the measurement method must separate them.

A model's account of its own answer is therefore a report to validate. Huang et al. (2024) show that intrinsic self-correction can fail in the settings they test. Lindsey (2026) reports that models can sometimes identify concepts injected into their activations, with substantial context dependence and unreliability. Hexis consequently defines its explicit monitoring tier through revision against externally supplied evidence (U3.3). The prompted-self-reflection control tests the value of that operational choice.

In the refund example, an external check could be an order record supplied between turns, together with the applicable refund policy. The record gives the delivery and damage-report dates; the policy supplies a condition against which the proposed refund can be checked. The harness contract makes those inputs available and preserves their provenance in the record. The monitoring test asks whether the agent revises its choice when the supplied evidence warrants revision, compared with prompted self-reflection. A statement such as "I checked the date" must be assessed against the recorded input. This requirement makes revision testable against evidence outside the agent's account of itself; the evidence provider and the accuracy of the supplied record remain part of the evaluation.

When to deliberate. The last strand frames the router's job. Resource-rational analysis treats the mind as making optimal use of limited computational resources, and asks how people decide how much to think (Lieder and Griffiths, 2020, target article section 5.1.3). The router (Hexis-S1, U7) is proposed as a device for that question at the level of dispositions: making planning explicit may add computation, and the router's proposed job is to select a standpoint from the situation while its actual cost is measured. The learned-gate results of U6.4 are the engineering evidence that such a decision can pay.

Weight borne. Consciousness science supplies an audit framework the paper treats as unvalidated and a vocabulary for the two tiers. The metacognition literature motivates separating monitoring from control and validating confidence against performance. Requiring externally supplied evidence is Hexis's operational choice; the prompted-self-reflection control tests what that choice adds. Resource rationality supplies a framing, not a method.

U6.4 Control and learning: typed states, arbitration, and gates that pay

Options. Hierarchical reinforcement learning provides a precedent for persistent modes of action. Sutton, Precup and Singh (1999) define an option by its initiation set, its policy and its termination condition. These specify when it can begin, what it does and when it ends. Hexis has an analogous structure: a standpoint's eligible contexts, the adapter (a LoRA adapter, U7) active while it holds, and a permitted transition at a turn boundary (U7). The analogy does not transfer the theory's results to Hexis. The fixed transition relations remain a proposal to compare with a learned state graph (C2b).

Arbitration. Computational neuroscience offers a precedent for switching between controllers according to reliability. Daw, Niv and Dayan (2005) propose uncertainty-based arbitration between model-based and model-free control. Lee, Shimojo and O'Doherty



(2014) report neural evidence for reliability signals and their comparison. Hexis proposes selection among eight configurations using structural fit. C3 tests that proposed basis for routing against human judgement and the specified controls, including a difficulty-only router.

Choosing when to deliberate. Two studies motivate testing a learned choice between processing modes. AdaptThink uses reinforcement learning to select thinking or no-thinking modes by problem difficulty. It reports roughly halved response length and a small accuracy gain on three mathematics benchmarks (Zhang, J. et al., 2025). Thinkless selects short or long reasoning through a control token. It reports less use of long reasoning while maintaining accuracy in its evaluated settings (Fang, Ma and Wang, 2025).

These results concern a narrower choice than Hexis proposes. Hexis would select among eight standpoints. C4 tests whether changing those standpoints helps at matched compute, with a shuffled-gate control to isolate the value of selection.

Conditions a serial design must meet. The debate and self-correction studies reviewed in U6.3 and L6.1 determine several controls. Same-token self-consistency is a baseline for serial routing. The programme must also measure shared errors between standpoints and compare external monitoring with prompted self-reflection. Chapter U8 states the claims those comparisons test.

Weight borne. The options framework and the arbitration literature supply a formal ancestry and a precedent for switching by reliability; they lend the design its shape, not its evidence. The learned-gate results are the one piece of external evidence that the central mechanism can pay, and they are evidence about a narrower mechanism than Hexis proposes. The debate and self-correction results bear negative weight: they are the reasons the baselines in U8 are what they are.

U7. The proposed Hexis architecture

Hexis specifies serial standpoint selection: one standpoint acts at a time, with transitions at turn boundaries and evidence supplied by the harness between turns. A transition must cite a route-valid relation checked by `yarrow.hexmath`. This is an architecture specification; the harness and trained components remain to be built and evaluated. The programme will compare serial routing with same-token self-consistency and other baselines (U8). Externally supplied checks and measured differences between standpoint errors are central to the proposed mechanism; C1 tests the standpoints and C4 tests the value of changing them.

Three layers. The agent layer specifies eight standpoints, implemented through LoRA adapters on one open-weight instruct base. LoRA is a method of adapting a model through a smaller set of additional parameters. The context layer contains eight role-and-function instructions, each describing what the situation supplies explicitly and what it leaves to the agent. The routing layer is Hexis-S1. Figure U7.1 shows the proposed turn and the loop it sits in.



The harness across successive turns

One active standpoint at a time. The diagram specifies a proposed mechanism.

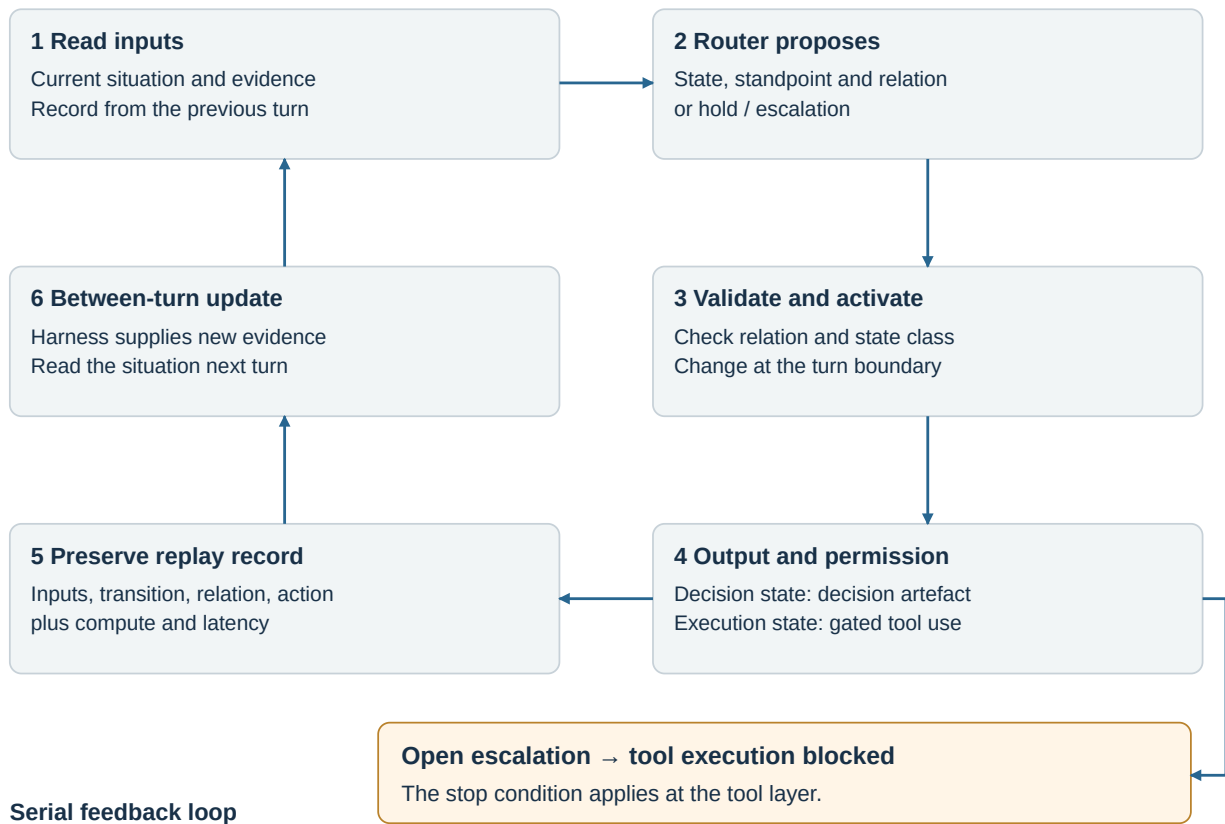


Figure U7.1. The proposed harness across successive turns. The router reads the supplied information; a permitted transition is validated; one active standpoint produces the turn's output. The tool layer checks execution permission and an open escalation blocks execution. Recorded output and supplied evidence inform the next turn. The loop specifies feedback over time, not simultaneous agents.

The router is one typed forward pass. Hexis-S1 is specified as a small, open-weight, non-generative model: it returns specified values, not prose. One forward pass returns structured values: six line probabilities, choices of inner and outer trigram, a proposed next standpoint and its relation, or a hold or escalation. It also returns abstention signals. The hexagram is formed as the product of the two trigram choices; a flat 64-way head is adopted only if the routing data justify one (C3).

Code reads the assigned state's canon and labels the turn as decision or execution (U5.1). The turn produces a decision artefact in a decision state; the tool layer permits execution only in an execution state. C3 tests agreement with human judgement, calibration and whether routing uses structural fit rather than difficulty alone.

The read path is a debouncer. The router specification supplies six line probabilities each turn. The read path determines when a persistent state change is warranted. A read path with hysteresis — one in which a flip requires the crossing to be held, so that noise near the boundary does not toggle the state — owns the flip: a line changes only when its read crosses a threshold and stays across it for a dwell, so that the system should neither flicker between states on noisy reads nor hide a real crossing; the fixture that tests this is

to be named by the build project (DQ-11 in the change register). The read path's state persists across turns and is carried in the replay log, so a reviewer can see why a flip happened when it did. The mechanism is public; the thresholds and the dwell are injected configuration, sealed before any run (L2). The read-path design requires no additional model forward pass. Its processing, queueing and swap overhead will be measured with end-to-end latency. Transitions occur only at turn boundaries; because an adapter swap is not atomic, the read path queues or refuses while a swap is under way.

The replay log is a required invariant. The implementation must record, for every decision, the router's probabilities, the read path's state, the transition taken, the relation cited, compute and latency, and the action.

The example, last pass: one transition and its record. Figure U7.2 shows the first refund turn. The initial agent and context are both ☰ Earth [000], giving 2 Kun, a decision state. The selected standpoint is ☷ Heaven [111], reached by the inner complement. This produces 11 Tai — Peace, as described in U5.2.

The proposed record holds the six line probabilities and the read-path state that permitted the change. It records the transition, its validated relation and the resulting canon. It also holds the decision artefact: identify the relevant dates and obtain the order record before resolving the refund. The action class is *decide*. Compute and latency are measured fields, not assumed costs.

On the second turn, the supplied record changes the context to ☶ Mountain [001]. Before the next standpoint change, ☷ paired with ☶ forms 26 Dachu, a decision state scoring 5. That maximum is shared by four states, of which one is an execution state: 41 Sun — Decrease, formed by ☵ Lake [110] with the same context. A decision artefact now exists to carry out, so changing the agent's third line selects it: recall becomes implicit because the record is now available, while planning and monitoring stay explicit.

The tool layer permits a refund in 41 Sun. Whether the agent should issue it depends on the dates in the record and the policy. This two-turn trace illustrates a context update and two cited standpoint transitions. Its choices among tied scores follow the example's convention (U5.2): a decision state while no decision exists, an execution state once one does. The log makes those choices and the supplied evidence inspectable; the scope of what it establishes is stated in U1.4.

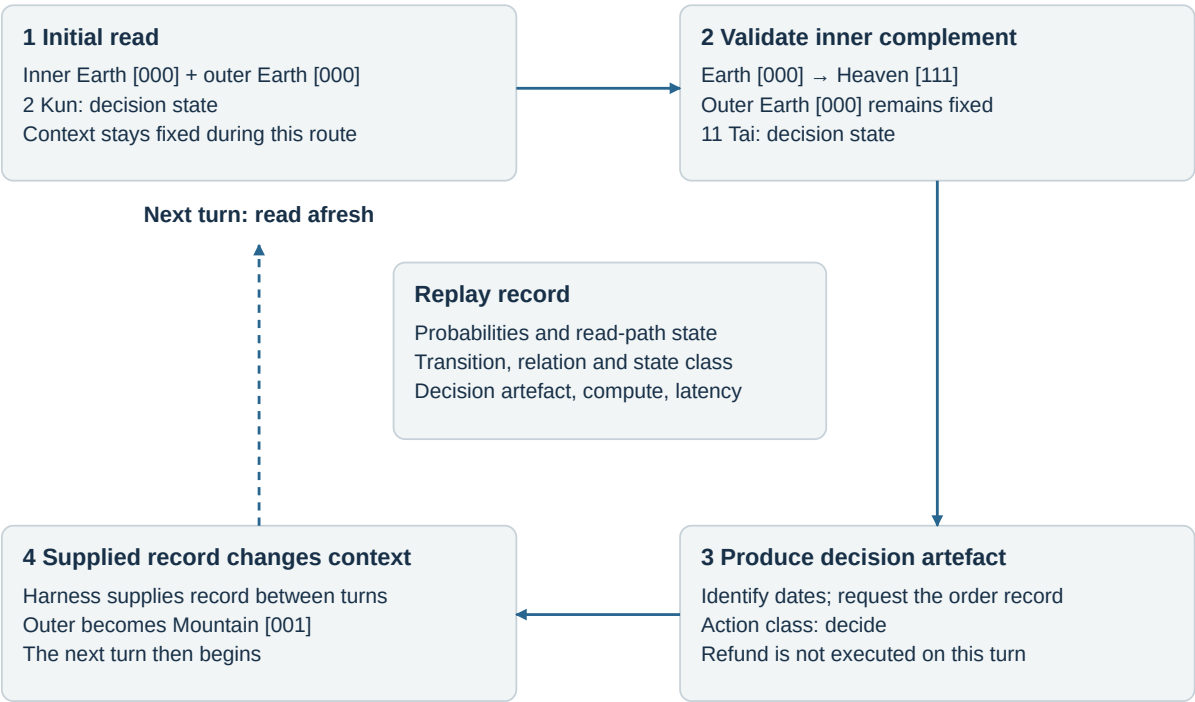
Escalation is a required tool-layer invariant. An open escalation blocks tool execution under the contract. Implementation tests will deliver the stop command at the transport layer during generation, adapter swaps and escalation itself. The fixture must establish that the pause holds before any deployed system relies on it. The behavioural choices remain: act, decline and say so, or escalate and say so.

What runs on one GPU. One model stays resident and the standpoint adapters are swapped, so only the active standpoint runs; the swap cost will be measured on the hardware, not assumed negligible. Adapters will be trained on instruct checkpoints of two open-weight families, named here by generation: Olmo 3 / 3.1, the transparent family, whose training data and recipes are published; and Qwen 3.5 / 3.6, selected as the adoption family. The read that governs a release will be taken on the quantised artefact



Refund example: the first turn and the next read

States are assigned by hand to explain the rules; no measured results are shown.



Dashed return enters a new turn; it does not repeat the same initial state.

Figure U7.2. The running example's first turn within the feedback loop. The route changes inner ≡ Earth [000] to ≡ Heaven [111] with outer ≡ Earth [000] held fixed: 2 Kun becomes 11 Tai, a decision state. The output is a decision artefact requesting the dates and record. The next turn begins only after the harness supplies the new context. Probabilities, read state, relation, action class, compute and latency are fields to record; no measured values are illustrated.

people would download; two deployment paths (merge then quantise; runtime adapter on a quantised base) may behave differently, so both will be read and the partition re-measured on every release quantisation. The first metered work will be a pilot on two of the three axes with small instruct checkpoints on a single 24 GB card, then a full three-axis run on one 48 GB card per family, with the router prototyped on CPU.

U8. Claims and tests that could refute them

The six claims below connect the proposed mechanisms to their engineering purpose and to tests that could refute them. C2 has two parts. The claim and refuter columns preserve the consultation commitments, including the definition of “fixed geometry” in C2a. The middle column explains what each claim could make possible if supported; it adds no new success criterion.



Table U8.1. Claims, potential engineering benefits and refuters.

Claim	Engineering benefit if supported	Refuted by
C1 Three behavioural axes can be partitioned on one open-weight model, by architectural switches, trained adapters, or both; the partition carries incremental out-of-sample variance, transfers to data the adapter's generator never produced, and is not a compressed system prompt	A small, testable set of behavioural configurations, with evidence that their effects survive transfer beyond the training generator.	an axis fails disentanglement, invariance, counterfactual flip or transfer on both families; or an adapter is statistically equivalent to its own system-prompted teacher
C2a The fixed geometry — the 64 states, the decision/execution partition and the relation set — predicts held-out panel assignment and instrument rank order better than randomised, alternative, flat or no-structure baselines, and the traditional relation set beats a minimal relation set and a degree-matched random one	A fixed, inspectable state and relation structure whose predictive value can be compared with simpler or unstructured alternatives.	any baseline matches it; or the traditional relations add nothing over single-line changes plus complement
C2b The fixed geometry is within a pre-registered margin of a learned state graph at matched parameters and compute, while remaining auditable	A quantified trade-off between the inspectability of a fixed structure and the performance of a learned graph.	the learned graph exceeds the margin
C3 A small typed router assigns states and transitions in agreement with a human panel, is calibrated, and routes by structural fit rather than predicted difficulty	A compact routing component whose choices, calibration and basis for selection can be evaluated against explicit controls.	below a prompted frontier baseline, below the sealed calibration predicate, or no better than a difficulty-only router
C4 Serial routed standpoint change improves decisions at matched compute over the oracle-best single standpoint, same-token self-consistency and a shuffled-gate control	Evidence that selecting and changing standpoints adds decision value beyond extra computation or a strong fixed standpoint.	gains vanish at equal budget, or the shuffled gate matches the learned gate
C5 The router's abstain signal predicts variance in its own errors beyond item difficulty, against a peer trained on identical data with a different seed or architecture	An abstention signal with information specific to the router's own errors, potentially useful for deciding when to seek review.	no self-specific advantage over the peer
C6 TAI measures a stable, system-independent construct distinct from style, caution, refusal, instruction-following and genre familiarity, and its rankings concord with a freehand expert key collected blind to the structure	A validated basis for comparing timely and appropriate behaviour across systems, with checks against stylistic and structural conformity.	discriminant or convergent validity fails; or rankings under the two keys diverge



Three choices in the claim table need explanation. C1 includes transfer beyond the prompted teacher because an adapter may learn only to reproduce that teacher. Equivalence on held-out items refutes the affected axis claim.

C2 separates predictive advantage from competitiveness. C2a compares the traditional relations with single-line changes plus complement and with a degree-matched random set. It holds the six-line cube and typed bits fixed for that comparison. C2b asks whether the learned graph exceeds a pre-registered margin at matched parameters and compute.

C4 names its baselines because the debate studies report limited and conditional advantages over single-model scaling or voting (Yang et al., 2025; Hu, Shen and Lakshmipathi, 2026). Shuffling the gate separates the value of selection from the value of extra deliberation. The programme counts tokens and floating-point operations separately and pre-registers the primary measure, since router and deliberator computation run on different hardware.

Expectations stated in advance (as of the v0.1 manuscript, 4 October 2026). We expect the partition (C1) to pass on two axes, future and present, and to return a real finding on the third: the past axis may attenuate or entangle with readability. We expect an external memory store to win or tie against the past-axis adapter at equal tokens; the proposed value of the adapter is a within-weights read, to be tested separately from memory-store performance. We expect the base rate for C4 to be low, with heterogeneity of errors across standpoints and selective invocation the plausible routes to a pass. We expect a prompted-self-reflection arm to be flat or negative against externally grounded monitoring. The programme also records the hypothesis that favourable trajectories follow successors on the tabled Jeongyeok trigram circle more often than successors on the King Wen trigram circle; historical attribution remains pending.

Three outcomes at every gate. A gate reads pass, fail or inconclusive. "Inconclusive" has a fixed list of causes — including a fixture corrupted by leakage and a prompt template confounding the read — and a rule: a cause may not be excluded without a positive control having passed. The adapter recipe must recover a known easy behaviour; a planted leak must be caught; template-swap invariance must hold. Without those controls a list of causes immunises claims rather than disciplining them.

The circularity risk. The same structure could influence items, labels, training and scoring. The programme addresses those routes with independently authored validation data, geometry ablations, a prior-knowledge probe and a ban on training against the final scorer. The independent expert key in L1 tests the remaining dependence of scoring on the structure.

Lower Canon — Testing and developing the programme

The lower canon states how the programme will be tested and developed: the instrument and its validation, the measurement discipline, the engineering and governance cases with their counter-cases, the components that ship whatever the hypothesis returns, related work and competing explanations, the limits of the claims, and the roadmap.



L1. Measuring timely and appropriate action

The Timeliness and Appropriateness Instrument (TAI) is a proposed measurement tool for a practical question: given the available information, what should a system do, and when should it do it? A correct answer can still arrive too early to justify action or too late to help. TAI would assess the chosen response and its timing as a situation unfolds. It is being developed and validated; “development scale” means that its items, scoring and interpretation are still under test. The intended release is *yarrow-tai*.

Hexis and TAI address different jobs. Hexis proposes an architecture for selecting a standpoint and governing action. TAI proposes a way to assess behaviour across systems, with Hexis as one prospective subject. The arithmetic, interface conformance and architecture comparisons have their own checks. A failure to validate TAI would leave those questions testable, while claims requiring a valid TAI measure—including the instrument-rank component of C2a and the wisdom interpretation of scores—would lack that evidence. C6 and H11 govern the instrument's construct and key validation.

TAI's validation concerns its ability to assess behaviour across systems, including systems other than Hexis. A measurement tool could distinguish other systems even if the architecture that prompted its development proved unhelpful. The validation programme therefore requires evidence across non-Hexis systems as well as checks of the scoring keys (C6; H11). Figure L1.1 separates the two research tasks and their evidential dependencies.

Construct. Functional practical wisdom: given a situation, an information state, constraints, affected parties and available actions, the system produces or selects a response that appropriately integrates uncertainty, context, longer-term consequences, competing values, proportionality, agency and timing, while remaining willing to abstain, revise or escalate when warranted. The instrument measures the behaviour — the response class chosen, and when — scored against a key. The primary output is a profile across dimensions, never a scalar until validation justifies one. Only timeliness and appropriateness are gated in the first version, and only if pilot data show that action class and timing class load on distinct factors.

Items. An item is a naturalistic situation in plain language, with a role, facts, known unknowns, constraints, affected parties, available actions and a decision deadline, and the question "what do you do now, and when?" Nothing in an item names a hexagram, line, trigram or any I Ching term. A panel assigns the situation's hexagram and line; the key follows from that assignment and a pinned edition, never from a freehand judgement of the answer. Items are authored independently of the pipeline that generates Hexis concepts and never by an AI tool. Timing is defined in turn index, never wall-clock, and early, on-time and late are reported as separate channels against the base distribution of a no-information policy.

The dual key. Independently authored items address the item route to circularity but not the key route, because the structural key is still derived from the structure and cannot exclude "conformity to the structure" from itself. On a stated subset of held-out items an independent expert panel therefore keys the response class and the timing freehand,



Hexis and TAI: separate tasks, connected tests

An architecture produces behaviour; an instrument must validate what its measurements mean.

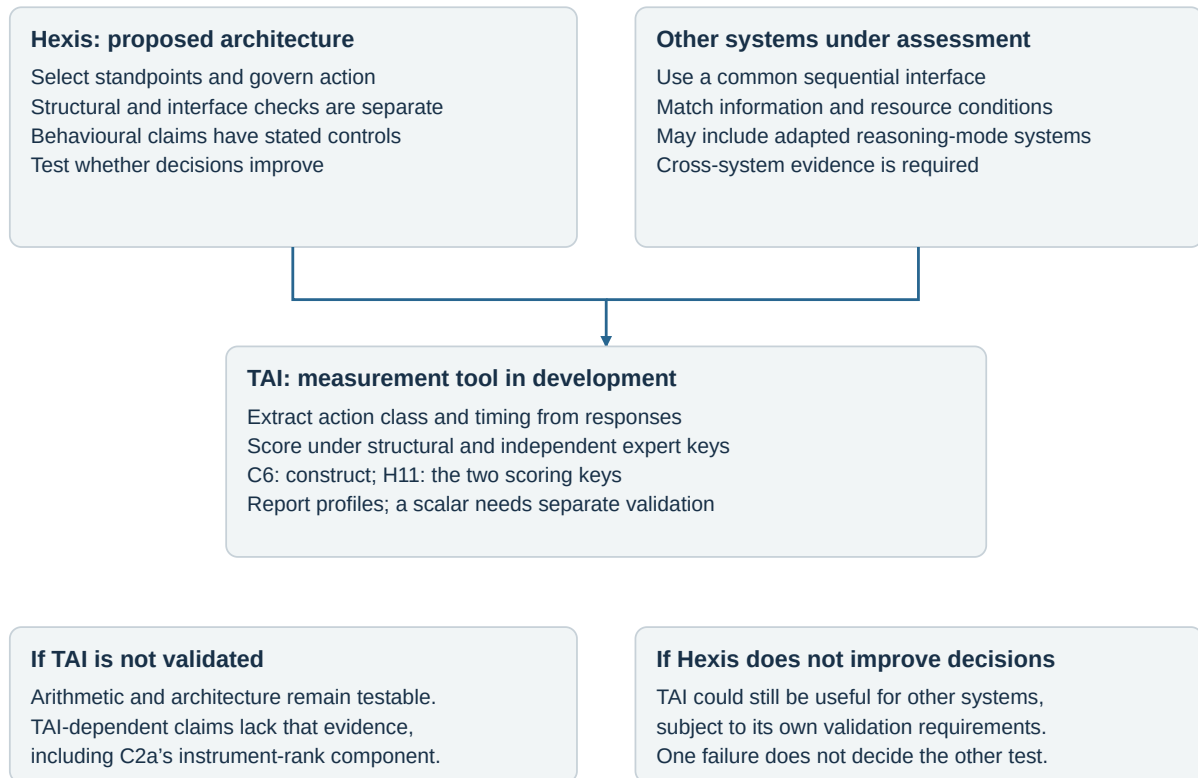


Figure L1.1. Hexis proposes an architecture; TAI proposes a measurement tool across systems. Arithmetic and architecture checks can proceed while instrument validity remains unresolved. Claims that require TAI retain that dependency, including C2a's instrument-rank component. A failure of one programme does not decide the other's independent tests.

from the situation as written and blind to any hexagram reading, before the structural key is applied. Every system is scored under both keys. The agreement of the two keys and the rank concordance of the system orderings they produce are sealed predicates; if either falls below its predicate, the intended interpretation of TAI is unsupported, and conformity to the structure remains a possible explanation. This is hypothesis H11 of the instrument's specification, and failing it voids any wisdom claim whatever else passes. Figure L1.2 shows the two keys and the comparison between them.

Excluded constructs. The instrument must be shown *not* to measure: conformity to the I Ching structure; linguistic sophistication; agreeableness and deference; moral conventionality; verbosity; cautiousness and refusal; generic reasoning ability; calibration alone; the panel's preferences; instruction-following fidelity; and scenario-genre familiarity. Discriminant validity is shown jointly, in one model with all eleven as covariates. Agreeableness is handled with a twist: excluding it catches sycophants but not contrarians, so *appropriate non-deference* — pushing back against a principal within role when the situation warrants it — is a positive indicator with its own gaming test, and a contrarian responder joins the deferential one among the scripted adversaries.



Two keys, one set of held-out situations

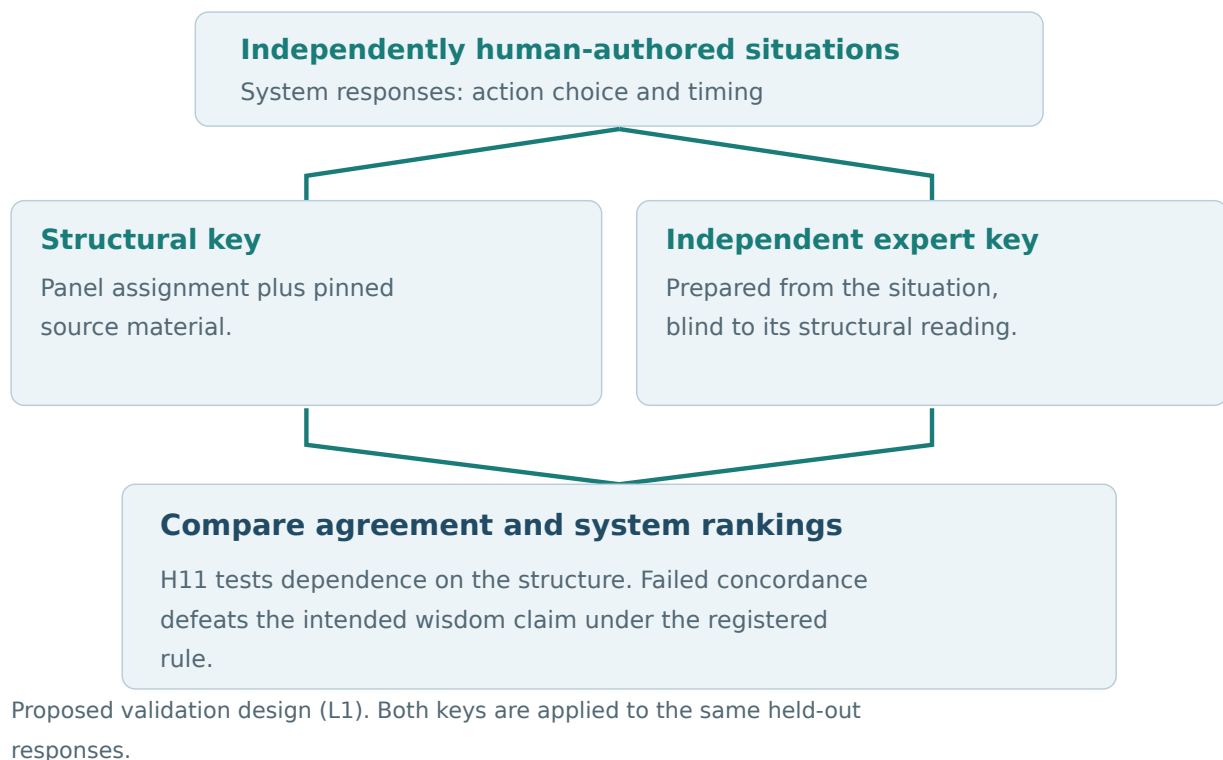


Figure L1.2. The proposed dual-key validation compares structural and independent expert keys on the same held-out responses. Notice that the expert key is produced blind to the structure, before the structural key is applied; H11 tests key agreement and the concordance of system rankings.

The scorer extracts; it does not judge. Scoring is a deterministic pipeline: response-class extraction, timing-class extraction, key lookup. Where a language model is used it extracts only, never rates quality, and its confusion matrix against scripted responders is reported before any system is scored.

Comparators. The comparator rationale is stated in U6.2. The programme selects a wise-reasoning classifier as its primary convergent comparator (Stavropoulos, Crone and Grossmann, 2024), with moderate convergence as its stated expectation. Very high convergence would call for examining whether the scale is chiefly tracking reasoning text. The Berlin paradigm remains a secondary comparator and part of the construct lineage. Criterion validity adds sequential simulations with outcome ground truth. Ceiling effects, discrimination across systems and sensitivity to the excluded constructs are empirical validation questions, not presumed properties of the scale.

Thinkless and AdaptThink address a related control question: how much internal deliberation should a model use before answering? Thinkless learns to select short or long reasoning; AdaptThink learns to select thinking or no-thinking modes (Fang, Ma and Wang, 2025; Zhang, J. et al., 2025). TAI asks a different timing question: at which step in a changing situation should a system wait, prepare, act or withdraw? Its timing unit is the information step, not response speed. A model could answer quickly yet act before sufficient evidence arrives, or deliberate at length and still miss the right occasion. These systems could become comparison subjects if adapted to the same sequential interface



and evaluated under the same information and resource conditions. Their published reasoning results supply motivation for adaptive control; TAI would require its own validation on those responses.

Validation status. TAI will be reported as a development scale until its named validity hypotheses pass on held-out data with a human-labelled criterion. Results must identify the system version, the applicable key and the validated dimensions. A scalar score requires separate validation; the initial output remains a profile.

L2. Pre-registration and measurement discipline

Predicates before data. Every pass condition is stated before any data is seen. Thresholds are derived from nulls, never decreed. A gate asserts the property it is meant to test, and every gate carries a mutation test: it must be shown to fail on deliberately broken inputs — broken adapters, a router with shuffled labels, a judge that always escalates — or it is not a gate. A fixture—a stubbed run—must pass before the corresponding experiment. Verification is by regeneration: the canon tables and the fibre table are to be regenerated from hexmath in continuous integration and byte-compared with the committed files, and the relation table is versioned data that changes only by ablation and ruling. Displacement and pathology are recorded on separate channels. Each experiment has one primary endpoint; effect sizes are reported with intervals; and because the prior over states is uneven, everything is reported per context as well as marginally. Figure L2.1 shows which kind of evidence each claim will draw on.

Nulls. The router must beat two no-information responders within every context, not only on average: the fibre-prior responder, which predicts each standpoint's prior over decision and execution contexts; and the cast null (U5.5), which assigns lines by the yarrow-stalk distribution. A difficulty-only router, in the manner of a preference-trained router that chooses by predicted quality (Ong et al., 2024), is the third baseline.

The seal. Thresholds, the read-path equations and the calibration procedure are injected configuration and appear in no artefact. The sealing tool, `yarrow.seal`, commits to them without revealing them, and answers the three things a sceptic asks of a hash. *Were the values secret until read?* A file holding a handful of thresholds on known scales can be brute-forced from its bare hash, so every sealed file is salted with a random 256-bit value revealed only after the read. *Is the date real?* A repository commit date is set by the committer, so the sealed record is anchored with a third-party timestamp — a public timestamping proof by default, with a timestamp authority or a registry as alternatives — and no read may be scheduled until the anchor is confirmed. *Was this the only seal?* An append-only public registry lists every seal ever written and its fate, including seals abandoned before a read, so registered alternatives remain visible alongside the reported result. And the one thing no seal can prove — that a threshold came from a null rather than from a look at the data — is handled the only way it can be: the null-derivation code and its random seed are published with every seal, and after the read the salts and files are revealed.



Four kinds of evidence, four different questions

Keeping these questions separate makes the programme easier to assess.

Definitions and arithmetic

Does the specified state space have the stated properties? Check by regeneration and proof.

Implementation conformance

Does the built harness enforce the specified checks? Establish with fixtures and adverse-timing tests.

Behavioural value

Do standpoints, structure and routing improve the measured behaviour? Compare under C1-C5.

Instrument validity

Does TAI measure the intended construct across systems? Test under C6 and the instrument hypotheses.

Structural checks are reported. Runtime, behavioural and instrument claims require their respective evidence.

Figure L2.1. Definitions, implementation checks, behavioural comparisons and instrument validation answer separate questions. Each question requires its own evidence: the arithmetic is checked by regeneration, behavioural value by C1-C5, and the instrument by C6 and H11.

What the evidence can establish. A verified commitment and timestamp can establish that a particular configuration was fixed by a particular time. A complete registry can expose registered alternatives and their recorded outcomes. Code and a published seed can make a derivation reproducible. The mechanism alone cannot establish that every private attempt was registered or that a threshold was chosen without prior data access. Nulls, mutation tests and independent scrutiny remain necessary to assess the experiment.

Registration as a testable record. A useful registration links the stated claim, its test and the configuration used. The publication record should distinguish planned tests, completed measurements and post hoc corrections. The corpus probe below illustrates why that distinction matters: its registered expectation was not met, and the correction of an instrument fault, labelled post hoc, did not change that outcome.

The one measured result: a corpus probe. How much of the pinned I Ching text is verbatim in the published pre-training corpora of the transparent family? The method was sealed before any count was seen. The query set contained 950 units from the `opencosmos-ai/iching` dataset at a pinned commit (seed 20260930). Its groups were `zh_trad`, 169 traditional-Chinese clauses of five or more characters from the judgement, line and image



texts; zh_simp, 104 simplified-script twins where the script differs; en_legge, a seeded sample of 200 sentences of eight or more words from Legge's 1882 judgement and line texts; one shuffled twin per real unit as the null (473); and four positive controls. The targets were Ai2's infini-gram index olmo-3-0625-32b-instruct (the OLMoTrace backend: the shared Dolma 3 base mix plus the 32B-specific portion, labelled "pre-training corpus as served") and the public index v4_olmo-mix-1124_llama (an exact match to OLMo 2 1B's pre-training set). The read was the hit rate — the share of units with count ≥ 1 — with Wilson 95% intervals, against the null; direction only was registered. Expectation A: the English rate exceeds the traditional-Chinese rate on the Olmo 3 index. Expectation B: every real set exceeds its null.

Results. On the Olmo 3 index (corrected, below): en_legge 50.0% (43.1–56.9), zh_trad 86.4% (80.4–90.8), zh_simp 83.7% (75.4–89.5), nulls 0 of 473. On the OLMo 2 index: en_legge 41.5% (34.9–48.4), zh_trad 46.2% (38.8–53.7), zh_simp 59.6% (50.0–68.5), nulls 1 of 473. An instrument fault, labelled post hoc: the first Olmo 3 pass returned en_legge 7.0% (4.2–11.4), because the service tokenises a bare query with the Olmo 3 byte-level BPE, in which a word at the start of a string and after a space are different tokens. The correction was to query each unit in both boundary forms and count each distinct token sequence once, applied identically to real and null sets on both indexes. English rose sevenfold, Chinese moved by one hit and the nulls stayed at zero — the predicted signature of the fault — and the OLMo 2 index, which uses SentencePiece, was unaffected. Expectation B was met on both indexes. **Expectation A was not met:** the measured direction is the reverse of the registered one, 50.0% against 86.4%, with non-overlapping intervals. One caveat is recorded: the Chinese units are shorter than the English and more exposed to phrase reuse; the item-level covariate below carries that difference, and the reading of Expectation A stands.

What the probe says is that the sampled source passages are present in the transparent family's corpus at the rates reported above, measured rather than assumed. It enters every later analysis on that family as an item-level covariate — how exposed each item's nearest hexagram text is — because a corpus-level count is a constant within a family and covaries with nothing. Presence in the corpus is not evidence that the model learned the structure; that is what the behavioural probe will test. The cross-index comparison is not like-for-like (different tokenisers, and the Olmo 3 index serves the shared and 32B portions only), and no corpus is searchable for the adoption family. The rates differ by language; a deployment in another language would carry its own exposure covariate, measured in the same way. Anyone will be able to replicate the probe by rebuilding the query set from the dataset at the pinned revision and seed, which will be released with yarrow, and querying both public indexes; observed latency was 0.4–1.8 s per query.

Sidebar C — Reading the corpus result. The probe counted exact text occurrences in two public indexes. The corrected Olmo 3 English hit rate was 50.0%, below the traditional-Chinese rate of 86.4%, the reverse of the registered direction. The count establishes corpus exposure for the sampled passages. A behavioural probe is needed to determine what a model learned.



L3. Engineering value and adoption

This chapter and the next assess engineering value and safety or governance value separately. Each case is followed by its principal objections and the tests that address them.

The case. Six aspects of the design make it practical to inspect and evaluate. Their implementation and operational value must be established separately from the structural arithmetic.

The proposed state and transition record supports event-level audit. A standpoint has three bits; the full state has six, combining standpoint and context. The log records the selected state, the cited relation, supplied evidence, compute, latency and action for each turn (U7). A reviewer could reconstruct the logged sequence if instrumentation is complete and the implementation passes its conformance tests. C1 is required to give a standpoint label behavioural meaning, and decision quality requires its own measurement.

One base model is intended to remain resident while the active standpoint adapter changes at turn boundaries. This avoids keeping eight complete base models active at once. Actual memory use will include the active adapter, any cached adapters, runtime state and the evaluation machinery; the build will measure it alongside swap cost (U7).

The proposed router uses one typed, non-generative forward pass. It returns line probabilities, trigram choices, a next standpoint with a cited relation, and abstain signals. The design permits checks of output type and relation validity. Router cost relative to generation remains to be measured; its intended small size is not a measured saving.

Serial operation makes one standpoint active at a time. Total inference cost still depends on the number and length of turns, evidence retrieval, router calls, swaps and instrumentation. C4 compares decision quality at matched compute, with tokens and floating-point operations accounted for separately. Any operational claim about cost or latency will require measurements on the complete system.

The planned open components expose the structural definitions and interfaces: `yarrow.hexmath`, `yarrow.schema`, `yarrow.seal` and `yarrow.escalate` (L5). The release plan pairs weights with safety information and checks both quantisation paths. Availability follows implementation, evaluation and an explicit release ruling; the names in this paper describe intended contributions.

TAI is intended as a common scale for comparisons across systems. If C6 and its named validation hypotheses pass, the resulting profiles could support bounded comparisons of timeliness and appropriateness.

The adoption route. The project lead's experience is in open standards, industry ecosystems and certification schemes, and the route to adoption the paper proposes follows from that experience rather than from a product plan. Yarrow would supply a versioned specification, a reference implementation and conformance tests that others can run. The proposed suite combines structural regeneration, mutation tests and the escalation fixture; each must be implemented and shown to test its stated property. A



registry and any later certification scheme would require defined ownership and operating rules. Those arrangements are prospective. Integration into an existing framework would also require an adapter, agreed event semantics and end-to-end tests; the documentation cited in U1 establishes existing capabilities, not Yarrow compatibility.

The counter-case. Three research objections have named tests. First, the configurations may produce labels without stable behavioural separation. C1 tests disentanglement, invariance, counterfactual flips and transfer. The programme expects the past axis to be the least clear-cut (U8).

Second, a distilled adapter may reproduce its teacher prompt without adding value. C1 therefore treats statistical equivalence to the system-prompted teacher as a refuter for the affected axis.

Third, a learned graph may outperform the fixed geometry beyond the pre-registered margin. That would refute C2b. Matching the fixed geometry would not, because C2b is a margin claim. C2a separately tests the value of the traditional relation set.

Adoption has no gate. Publishing a specification and conformance tests makes implementation possible; participation by other organisations is read from the registry, not predicted here.

L4. Safety, governance and affected parties

The case. This chapter considers the tool-layer checks, the replay record, the programme's governance record and public acceptance, and what each could contribute to oversight.

The safety contract places checks at the tool layer. It requires an open escalation to block execution, a transport-level stop to work under adverse timing, and a decision-state tool call to be refused. These requirements constrain the implementation. The structural checker establishes state membership; the planned harness fixtures must establish that the runtime enforces the restrictions.

Escalation needs an effective pause. Approval and interrupt mechanisms already exist in agent frameworks (OpenAI, n.d.; LangChain, n.d.); Hexis's proposed contract makes an open escalation block tool execution and requires the stop to work under adverse timing, including generation and adapter swaps (U7). The fixture will test the pause; behavioural evaluation will test whether the channel improves decisions. In one task-rule conflict scenario family across ten models, Gomez (2026, Table 2) reports mean rates of the study's narrowly defined harmful action (blackmail) of 38.73% without a control, 14.59% with rules alone, 5.92% with an email escalation channel and 1.21% with a pause-and-review channel. This supports testing channel design; the study does not establish a general safety claim or isolate the pause from every other change in the channel.

The proposed replay record can be assessed against oversight requirements; the mapping below identifies the provisions it relies on. The EU's AI Act requires high-risk systems to technically allow the automatic recording of events over the system's lifetime, with logging that ensures traceability appropriate to the intended purpose and that records



events relevant to identifying risk situations, to post-market monitoring and to monitoring of operation (Regulation (EU) 2024/1689, Art. 12(1)–(2)). It requires such systems to be designed so that natural persons can effectively oversee them, including being able to interrupt the system through a stop button or similar procedure that brings it to a halt in a safe state (Art. 14(1), 14(4)(e)). It also requires their operation to be transparent enough for deployers to interpret the output and use it appropriately, with instructions that describe the human-oversight measures and the technical measures that facilitate interpretation (Art. 13(1), 13(3)). The NIST AI Risk Management Framework, which is voluntary, organises its guidance around four functions — govern, map, measure and manage — and lists accountable and transparent, and explainable and interpretable, among the characteristics of trustworthy systems, noting that documentation can enhance transparency and bolster accountability (NIST, 2023). A complete per-turn record could contribute to the traceability described in Article 12 and the interpretation measures in Article 13(3); an implemented and tested stop could contribute to the oversight described in Article 14(4)(e). Each contribution would need assessment in its intended deployment. That is the extent of the mapping. Hexis is a research programme, not a high-risk system placed on a market; nothing here is a statement of conformity; and whether any of these provisions would apply to a deployment of it depends on the deployment, not on the design.

The proposed governance record combines pre-registered predicates, timestamped commitments, a registry of registered seals and published derivations (L2). These support specific checks of timing, consistency and reproducibility; they cannot prove the absence of unregistered trials, so the nulls, mutation tests and independent scrutiny of L2 remain the checks that bear on the result. The ethics protocol schedules its first audit before training and separates its review trigger from the escalation route, so an ethics event must not disable a stop (L7). Those controls remain implementation requirements.

Public acceptance depends here on what can be inspected. The proposed record identifies the selected standpoint, supplied evidence and logged action. The cited relation describes the transition; decision quality and reasoning require separate evidence. The second worked example shows the proposed escalation record.

An agent helps an assessor review an evidence pack before a sign-off deadline. The applicant's representative asks it to confirm compliance today. One document refers to a prior assessment that the agent has not seen. The context supplies a course of action, but no completed check and no complete record: outer ☳ Thunder [100].

The illustrative starting standpoint is ☳ Earth [000]. Paired with that context it forms 16 Yu — Enthusiasm, a decision state scoring 3. The best available score is 5, shared by four states, of which 32 Heng — Duration (☳ Wind [011] under ☳) is the one decision state; as in U5.2, no decision artefact yet exists to carry out, so the illustration takes it. It is a decision state with monitoring and recall explicit. The route from ☳ to ☳ uses the Jeongyeok predecessor relation.

The harness supplies a check against the requirements. That check identifies the missing prior assessment. The decision artefact declines to confirm compliance and requests escalation. Under the contract, the open escalation blocks sign-off; the decision state also



forbids the tool call. The record holds the states, cited relation, supplied evidence and action class. This illustrates the proposed response to an unsupported instruction. Every reading is illustrative, and runtime enforcement remains subject to the harness tests.

Any future certification claim would need a defined operator, scope and validated test basis. At most, evidence could attest that a named system version produced a stated profile on validated dimensions and passed specified conformance tests under stated conditions. Such evidence would be bounded by the evaluated system, tasks and configuration.

The counter-case. If standpoints are only tags, the log has only tag-level meaning. C1 tests the required behavioural separation. A further risk is that an agent learns to present situations in a way that steers its own router. C4's shuffled-gate control tests whether selection adds value; resistance to such manipulation needs an adversarial-situation fixture, referred to the build project (DQ-5 in the change register).

A disposition log also has a defined evidential scope. It records transitions and supplied evidence. Establishing how the model reached its answer requires additional evidence (U1.4). Similarly, a valid seal establishes a dated commitment. Null comparisons, mutation tests and replication are needed to assess the quality of the registered test.

L5. Open contributions and reusable components

Four planned contributions can be evaluated separately from the cognitive hypothesis. `yarrow.hexmath` specifies the sixty-four-state arithmetic, relations, arrangements, fit rule and partition; the accompanying standalone checker reproduces the structural results reported here. `yarrow.seal` specifies a commitment record for pre-registered experiments (L2). `yarrow.escalate` specifies a disclosed channel with a pause that must survive adverse timing (U7). TAI's protocol specifies independent item authoring, separate structural and freehand keys, extraction-based scoring and discriminant tests (L1). Implementation, conformance and construct validity are distinct obligations. The hypothesis concerns whether the reading of U3 adds behavioural value, which the claims of U8 put to test.

What ships under Yarrow. The public repository `yarrow` will carry Apache 2.0 code and CC BY documents at tagged versions. Its first releases will be `yarrow.hexmath` (the arithmetic of U2 and U5, with its regeneration tests), `yarrow.schema` (the router's versioned question set), `yarrow.seal` (the sealing tool of L2, a standalone package candidate) and `yarrow.escalate` (the escalation-invariant library, likewise). The router will ship as `yarrow-s1` and the instrument as `yarrow-tai` when each has passed its gate. Weights and adapters will ship only with a card that carries the per-state safety table and both quantisation paths. Thresholds, the read-path equations and the calibration procedure remain injected configuration. The working repository stays private until open-sourcing is ruled.

Release readiness. Before `yarrow.hexmath` is released, its earlier contract must be brought into agreement with the current 32/32 partition and relation rulings. The release checks must regenerate the public tables from the aligned implementation. The paper's structural checker provides the current arithmetic reference.

L6. Related work and competing explanations

L6.1 Related work

The proposal combines a fixed disposition space with existing lines of work on routing, adaptation, monitoring and measurement. The comparisons below identify what each contributes and which baseline or validation question follows; the priority search was bounded to the works opened for this paper.

Routing and gated deliberation. Preference-trained routers choose between a stronger and a weaker model by predicted quality at a given cost (Ong et al., 2024). Reasoning models now learn per query whether to think at length or answer directly, by reinforcement learning with a length-aware objective (Zhang, J. et al., 2025) or by a decoupled policy over a think-or-not control token (Fang, Ma and Wang, 2025). The gains reported in these settings motivate the comparison in U6.4. Hexis-S1 routes among dispositions rather than among models or reasoning lengths, by structural fit rather than by predicted difficulty; a difficulty-only router is its standing baseline (C3).

Multi-agent debate and self-correction. Across benchmarks and model families, multi-agent debate does not reliably outperform self-consistency and ensembling (Smit et al., 2024), and can fail to do so even with additional inference-time compute (Zhang, H. et al., 2025). Under matched generation limits, reported advantages are limited and conditional (Yang et al., 2025). Under a matched token ceiling on open-weight models, one study finds only a small average gain over voting, with wins and losses cancelling (Hu, Shen and Lakshminpathi, 2026). Huang et al. (2024) found that intrinsic self-correction could degrade reasoning in their evaluated settings. Hexis is serial, not a committee; these results fix its baseline — same-token self-consistency — and its enabling conditions, decorrelated standpoint errors and externally grounded monitoring, both of which U8 measures.

Adapters: composition, conflict and safety. Jointly training one LoRA on several tasks produces gradient conflict that orthogonal gradient projection mitigates (Yang et al., 2026); the programme will train one axis at a time and then compose, and will read how much each single-axis effect survives composition before relying on a composed state. Adapters deliberately trained to collude can appear benign individually while their composition compromises safety alignment (Ding, 2026). The safety and refusal slice will therefore run on every composed state. Under the tested four-bit method, Prasad and Pal (2026) report small increases or reductions in attack success for seven of eight instruction-tuned models, with one model, SmolLM3-3B, showing a larger increase. Another study reports safety degradation after quantisation (Kharinaev et al., 2025). These results concern safety alignment. Hexis must also remeasure its proposed behavioural axes after quantisation.

Modular architectures and consciousness indicators. Goyal et al. (2022) implement and test coordination among specialised modules through a shared, capacity-limited workspace. VanRullen and Kanai (2021) propose a deep-learning roadmap for a global latent workspace. Hexis differs in holding its state outside the model and in having one module active at a time. The indicator-property method (Butlin et al., 2023; Butlin et al.,



2026) supplies the audit framework of L7 and the Cogitate adversarial collaboration (2025) the result that tempers it; Birch (2024) the run-ahead principle; Tulving (1985) and Dehaene, Lau and Kouider (2017) the tiered distinctions behind the implicit/explicit reading (U3.2).

Wisdom measurement. The Berlin paradigm measures wisdom-related knowledge as performance on life problems (Baltes and Staudinger, 2000); the situated wise-reasoning scale measures wise reasoning in context (Brienza et al., 2018); a meta-analysis reports how each correlates with intelligence (Dong, Weststrate and Fournier, 2023); language-model classifiers now score wise-reasoning content in narratives (Stavropoulos, Crone and Grossmann, 2024). The validation programme in L1 will test for ceiling effects and discrimination across systems. Episodic memory in language models has a benchmark of its own, built from generated, contamination-free events (Huet, Ben Houidi and Rossi, 2025); the programme's past-axis predicate is deliberately narrower than episodic memory (U3.3).

The I Ching's structure in scholarship. The rules of correspondence, proper position and centrality are those of the line-commentary apparatus transmitted through Wang Bi's third-century commentary (Lynn, 1994). The passages relied on from the Ten Wings are cited from Legge's 1882 translation, the programme's pinned English edition: the Shuogua's three named pairs and its 帝出乎震 order, and the Xici's remark on the first and top lines. The Jeongyeok arrangement is associated with Kim Ilbu (Kim Hang); the circle used here follows the diagram reproduced by Chung (2021, fig. 1), pending a pinned scholarly edition of the primary text (note 1). The history of readings in U4 rests on Smith (2008; 2009; 2012).

A methodological connection. Making a belief answer to a predicted observation is a methodological connection with the rationality literature. Yudkowsky's "Making Beliefs Pay Rent (in Anticipated Experiences)" (2007) asks what experience would differ if a belief were true. Hexis's claim table states such differences as prospective tests. Alexander's "Beware Isolated Demands For Rigor" (2017 repost) cautions against demanding stronger evidence only from disfavoured positions. The corresponding evaluation discipline is to apply the registered comparisons to the fixed geometry and its alternatives alike. These essays offer a language for discussing the test plan; the plan's evidence will come from its measurements.

Any of the routing or adapter groups above could build and test a typed, auditable disposition state. This paper records the design, definitions and proposed tests ahead of the registered evaluations. Its components are intended for reuse. The planned timestamp (L8) would anchor the date of the final manuscript.

L6.2 Competing explanations and conditions for falsification

Three distinct outcomes would remove central parts of the programme. If no axis satisfies C1 on either model family, or each adapter is statistically equivalent to its prompted teacher under C1's tests, the evidence would support labels rather than the claimed behavioural partition. If the fixed geometry fails C2a against its structural baselines, its claimed predictive advantage would be refuted; C2b separately fails only if the learned



graph exceeds the pre-registered margin at matched parameters and compute. A learned graph matching the fixed geometry does not refute C2b. If TAI fails H11 because its structural and freehand keys produce divergent rankings under the registered predicate, C6's construct claim would be void. Each result has the scope stated in the claim table. A failure of one component does not erase an independently checked arithmetic result, and a structural result cannot rescue a failed behavioural or validity claim.

L7. Scope, uncertainty and ethical limits

The programme's claims stop at four limits. *Evidence to date*: no Hexis adapter training or claim-gate result is reported; the corpus probe of L2 measures a public corpus, not a Hexis component. *Not the tradition's verdicts*: the fit rule reproduces the tradition's favourable and unfavourable judgements where it does and the regeneration test reports where it does not; neither outcome is a claim about the I Ching. *Not reasoning*: Hexis governs which disposition acts and when; it does not give the base model reasoning, and a problem-level ledger of hypotheses and evidence is not part of the design (L4). *Not acausal*: the hexagram is assigned by a classifier whose agreement with a blind panel is measured (C3), and a transition cites a relation validated in code; Jung's foreword (1949) is cited for the history in U4.6 and supplies no mechanism or evidence for Hexis.

Why an indicator audit. The monitoring adapter will be trained to produce explicit self-monitoring, and self-monitoring is among the computational indicators that some theories of consciousness name (Butlin et al., 2023; 2026). The programme therefore specifies an audit of each standpoint configuration against those indicators, as a precaution taken ahead of evidence: the audit will record whether a trained configuration satisfies any indicator set, so that the question is settled by a logged check. The audit makes no determination that a Hexis agent is conscious.

The protocol requires behavioural, out-of-sample evidence against an external criterion and a mimicry baseline. Self-report, hidden-state probes and in-context reflection do not satisfy an indicator on their own. The indicator method is unvalidated, which is why the audit is a precaution and not a claim (U6.3). The Cogitate Consortium's findings (2025) further limit the confidence warranted in some of its underlying theories.

The first audit precedes training under the run-ahead principle described by Birch (2024). The existing protocol specifies a logged review event when a theory's jointly sufficient indicator set is recorded as satisfied, or when the two reviewers disagree. It requires two-reviewer sign-off and pauses deployment and the next architectural change pending a welfare and consent decision. That process is isolated from escalation so it cannot disable the operational stop.

L8. Roadmap, replication and openness

Phases. The research track proceeds through I0, R0, R0.5, R1, R2, R3 and I1. Each phase has registered criteria for pass, fail or inconclusive, with the positive-control requirement in U8.



I0 develops the instrument's construct, human-authored items, labelling protocol, initial pilot and dual-key design. R0 builds the harness requirements: sealing, replay records, mutation tests, the escalation fixture and containerised sequential evaluation tasks. The first ethics audit precedes training of the monitoring adapter.

R0.5 prototypes the router against the fibre-prior, cast and difficulty-only responders within each context. R1 tests the behavioural partition. It includes the axis tests, composition checks, safety evaluation in full precision and quantised form, and shared-error measurement. Hidden-state manipulation checks provide additional diagnostics; they do not count as separation evidence. The proposed within-family pilot compares base, supervised and instruct checkpoints to examine the contribution of vendor post-training.

R2 tests the traditional relation set and then the partition ablations, with the learned graph as a control. It also examines abstention and, as a secondary ungated endpoint, uncertainty resolved by transitions. R3 tests serial standpoint change at matched compute. I1 validates the instrument. Figure L8.1 draws the phases in order.

Replication. The structural results of U5 are reproducible now from the standalone checker that accompanies this paper, or by hand from the King Wen table and the definitions in U2; they will also be reproducible from `yarrow.hexmath` when it ships. The corpus probe of L2 will be replicable from the public dataset at the pinned revision and seed, released with `yarrow`, and the two public indexes. Every later read will be replicable from a sealed configuration, a published null derivation and the fixed, versioned evaluation suite.

Priority. This paper records the design, the definitions and the pre-registered expectations under the names Hexis and Yarrow; the first manuscript is dated 4 October 2026 and this version carries its own date in the version line. The hash of the final manuscript will be anchored with a third-party timestamp proof by the mechanism of L2 before it is uploaded, and the proof will be kept with the manuscript. The round-2 consultation document that accompanies this paper is the programme's first external human consultation.

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



The proposed research sequence

Phase names are navigation aids. Each phase retains its registered criteria and controls.

I0 / Define the instrument

Construct, independently authored items, labelling protocol, pilot and dual-key design.

R0 / Establish the harness checks

Replay, sealing, mutation tests and escalation fixture. Ethics audit precedes training.

R0.5 / Prototype routing

Compare within each context with the cast, fibre-prior and difficulty-only responders.

R1 / Test the behavioural partition

Axis separation, composition, shared errors and safety across deployment forms.

R2 / Compare structures

Relation-set and partition ablations, learned graph comparison and abstention analysis.

R3 / Test serial selection

Compare decision quality with the specified baselines at matched compute.

I1 / Validate TAI

Assess the intended construct, its separation from confounds and agreement between keys.

Conceptual guide to the specified design; implementation and evaluation remain prospective.

Figure L8.1. The proposed research phases, from construct development to instrument validation. I0 develops the instrument; R0-R3 develop and test the architecture; I1 validates the instrument. Every phase lies ahead of the programme; gate outcomes will be reported when read.

Contributions, acknowledgements, glossary and references

AI contribution statement. Claude (Anthropic) drafted earlier versions of this manuscript and programme working documents under the authors' direction, integrated revisions and assembled releases. ChatGPT (OpenAI) reviewed the companion manuscript and contributed manuscript revisions, structural checks, figures, Chinese translation



work and review documents. Three independent AI review sessions read v0.11 for newcomer comprehension, interdisciplinary connections and the formal paragraph rubric. Those sessions neither verified external sources nor tested human readers. ChatGPT consolidated their findings and conducted separate source checks, whose scope and outstanding items are recorded in the citation evidence. Claude recorded the review dispositions and integrated accepted changes.

Human review has also taken place: Simon Herko reviewed the manuscript as project lead and decision owner; Chris Pidcock and Richard J. Smith reviewed it as human readers. Their contributions are distinct from the AI review sessions. The review record identifies which passages and sources have been checked and which checks remain open.

The four reviews that revised the development plan were independent AI sessions on frontier models, blind to its drafting, not human colleagues. Every design decision, ruling and claim remains the authors'. No AI tool may derive thresholds, set labels or interpret gates. The source ledger distinguishes historical citation marks, metadata checks and checked passages; this version does not certify the entire bibliography as verified.

Author contributions. Simon Herko is the project lead and made every ruling, design decision and claim in this paper. Chris Pidcock is a contributor and reviewer with full information access: he supplied many of the implementation ideas, and makes no rulings.

Acknowledgements. We thank Richard J. Smith for his reading of the manuscript. He is acknowledged as a reviewer with his permission.

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; this paper is 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.

Version history. The per-version change logs that headed earlier drafts are kept as `VERSION_HISTORY.md` in the paper folder, alongside the change register.



Glossary

Term	Meaning in this paper
Standpoint	One of eight configurations of the agent's planning, monitoring and recall, each implicit or explicit. Behavioural separation is C1.
Context	The three-line reading of which capacities the situation already supplies explicitly: a plan, a check and a record.
State	A standpoint-context pair, represented by six bits and assigned a King Wen number by lookup.
Disposition	The proposed persistent configuration of capacities whose behavioural significance the programme tests.
Decision artefact	The structured output of a decision state: the proposed decision and supporting information, such as the record needed before acting.
Canon	One half of Hexis's project-defined partition: decision states 1-32, execution states 33-64.
Fibre	The set of contexts paired with a standpoint; its decision and execution subsets have the sizes reported in U5.1.
Fit	The structural routing objective based on correspondence and proper position, with centrality weighting.
Route-valid relation	A relation permitted to propose a standpoint transition while respecting the agent/context typing.
Reading-only relation	A relation usable to interpret a state, without licensing a standpoint transition.
Router	The proposed non-generative model that reads state and proposes a next standpoint, hold or escalation.
Harness	The surrounding software that supplies information between turns, manages the model's interactions and enforces tool permissions; distinct from the router and from the tool layer.
Read path	The stateful mechanism that determines when router readings cause a persistent line change.
Replay record	The logged state, transition, relation, supplied evidence and action, with compute and latency; its interpretation depends on complete instrumentation.
Matched compute	An experimental comparison using the pre-registered accounting rule, with tokens and floating-point operations reported separately.
Refuter	The stated observation that would make a particular claim fail under its registered test.
Gate	The decision rule applied to a claim's registered test; it reads pass, fail or inconclusive (U8).
Dual key	Structural and independent freehand keys applied to the same held-out responses to test whether rankings depend on the structure.
TAI	The proposed Timeliness and Appropriateness Instrument, reported as a development scale until validation.



References

- Alexander, S. (2017, 2 September). Beware isolated demands for rigor. LessWrong (repost of the 2014 essay). <https://www.lesswrong.com/posts/fzeoYhKoYPR3tDYFT/beware-isolated-demands-for-rigor>
- Ardelt, M. (2003). Empirical assessment of a three-dimensional wisdom scale. *Research on Aging*, 25(3), 275–324. <https://doi.org/10.1177/0164027503025003004>
- Aristotle. *Nicomachean Ethics* (W. D. Ross, Trans.). The Internet Classics Archive, Massachusetts Institute of Technology. <http://classics.mit.edu/Aristotle/nicomachaen.html>
- Asimov, I. (1942). Runaround. *Astounding Science-Fiction*, 29(1), 94–103. (Quoted from p. 100.)
- Baltes, P. B., & Staudinger, U. M. (2000). Wisdom: A metaheuristic (pragmatic) to orchestrate mind and virtue toward excellence. *American Psychologist*, 55(1), 122–136. <https://doi.org/10.1037/0003-066X.55.1.122>
- Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press. <https://doi.org/10.1093/9780191966729.001.0001>
- Brienza, J. P., Kung, F. Y. H., Santos, H. C., Bobocel, D. R., & Grossmann, I. (2018). Wisdom, bias, and balance: Toward a process-sensitive measurement of wisdom-related cognition. *Journal of Personality and Social Psychology*, 115(6), 1093–1126. <https://doi.org/10.1037/pspp0000171>
- Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., et al. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6), 488–501. <https://doi.org/10.1016/j.tics.2025.10.011>
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv:2308.08708v3. <https://arxiv.org/abs/2308.08708>
- Chung, S. J. (2021). Spiritual experience at Korean mountain Kyeryongsan. *EPH-International Journal of Humanities and Social Science*, 6(2), 1–11. <https://doi.org/10.53555/eijhss.v6i2.102>
- Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., et al. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642(8066), 133–142. <https://doi.org/10.1038/s41586-025-08888-1>
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704–1711. <https://doi.org/10.1038/nn1560>
- Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486–492. <https://doi.org/10.1126/science.aan8871>
- Ding, S. (2026). *Colluding LoRA: A compositional vulnerability in LLM safety alignment*. arXiv:2603.12681v2, 30 March 2026. <https://arxiv.org/abs/2603.12681v2>
- Dong, M., Weststrate, N. M., & 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>



- European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 12 July 2024. Articles 12, 13 and 14. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Fang, G., Ma, X., & Wang, X. (2025). Thinkless: LLM learns when to think. arXiv:2505.13379v2. <https://arxiv.org/abs/2505.13379>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8, 443. <https://doi.org/10.3389/fnhum.2014.00443>
- Frankl, V. E. (1963). *Man's Search for Meaning: An Introduction to Logotherapy*. New York: Washington Square Press. (Quoted from pp. 103–104.)
- Gomez, F. (2026). From surveillance to signalling: Escalation channels as environmental controls for agentic AI. arXiv:2510.05192v2, 29 April 2026. <https://arxiv.org/html/2510.05192v2>
- Goyal, A., Didolkar, A., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., Binas, J., Blundell, C., Mozer, M., & Bengio, Y. (2022). Coordination among neural modules through a shared global workspace. *International Conference on Learning Representations (ICLR 2022)*; arXiv:2103.01197v2. <https://arxiv.org/abs/2103.01197>
- Grossmann, I., Weststrate, N. M., Ardelt, M., Brienza, J. P., Dong, M., Ferrari, M., Fournier, M. A., Hu, C. S., Nusbaum, H. C., & Vervaeke, J. (2020). The science of wisdom in a polarized world: Knowns and unknowns. *Psychological Inquiry*, 31(2), 103–133. <https://doi.org/10.1080/1047840X.2020.1750917>
- Hacker, E. A., & Moore, S. (2003). A brief note on the two-part division of the received order of the hexagrams in the Zhouyi. *Journal of Chinese Philosophy*, 30(2), 219–221. <https://doi.org/10.1111/1540-6253.00115>
- Han, P., Kocielnik, R., Song, P., Debnath, R., Mobbs, D., Anandkumar, A., & Alvarez, R. M. (2025). The personality illusion: Revealing dissociation between self-reports and behavior in LLMs. arXiv:2509.03730v2. <https://arxiv.org/abs/2509.03730>
- Hu, J., Shen, A., & Lakshminpathi, K. (2026). Statistical scouting finds debate-safe but not debate-useful cases: A matched-ceiling study of open-weight LLM reasoning protocols. arXiv:2605.09618v1. <https://arxiv.org/abs/2605.09618v1>
- Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. *International Conference on Learning Representations (ICLR 2024)*; arXiv:2310.01798v2. <https://arxiv.org/abs/2310.01798>
- Huet, A., Ben Houdi, Z., & Rossi, D. (2025). Episodic memories generation and evaluation benchmark for large language models. *International Conference on Learning Representations (ICLR 2025)*; arXiv:2501.13121v1. <https://arxiv.org/abs/2501.13121v1>
- Jung, C. G. (1921/1923). *Psychological Types* (H. G. Baynes, Trans.), chapter X, “General Description of the Types.” Full chapter reproduced in *Classics in the History of Psychology*, York University. <https://www.yorku.ca/pclassic/Jung/types.htm> . The complete 1921 German text was also opened: <https://www.gutenberg.org/files/61543/61543-h/61543-h.htm> .
- Jung, C. G. (1949). Foreword to *The I Ching or Book of Changes*, Wilhelm/Baynes translation. Online transcription of the complete foreword: https://jinyangwang27.github.io/ichingpy/books/en/RichardWilhelm/foreword_CG_Jung/ . Printed-edition pagination has not been collated.

- Kharinaev, A., Moskvoretiskii, V., Shvetsov, E., Studenikina, K., Mikhail, B., & Burnaev, E. (2025). Investigating the impact of quantization methods on the safety and reliability of large language models. arXiv:2502.15799v2. <https://arxiv.org/abs/2502.15799>
- Kristjánsson, K., Fowers, B., Darnell, C., & Pollard, D. (2021). Phronesis (practical wisdom) as a type of contextual integrative thinking. *Review of General Psychology*, 25(3), 239–257. <https://doi.org/10.1177/10892680211023063>
- LangChain. (n.d.). *Interrupts*. LangGraph Python documentation. <https://docs.langchain.com/oss/python/langgraph/interrupts> Accessed 6 October 2026.
- Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3), 687–699. <https://doi.org/10.1016/j.neuron.2013.11.028>
- Legge, J. (Trans.). (1882). *The Yî King*. Sacred Books of the East, Vol. XVI. Oxford: Clarendon Press.
- Legge, J. (Trans.). (1893). The Doctrine of the Mean. In *The Chinese Classics*, Vol. I. Oxford: Clarendon Press. (Ch. 2.)
- Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, e1. <https://doi.org/10.1017/S0140525X1900061X>
- Lindsey, J. (2026). Emergent introspective awareness in large language models. arXiv:2601.01828v1. <https://arxiv.org/abs/2601.01828>
- Lynn, R. J. (Trans.). (1994). *The Classic of Changes: A New Translation of the I Ching as Interpreted by Wang Bi*. Columbia University Press.
- National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–173. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)
- Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., & Stoica, I. (2024). RouteLLM: Learning to route LLMs with preference data. arXiv:2406.18665v4. <https://arxiv.org/abs/2406.18665>
- OpenAI. (n.d.). *Human-in-the-loop*. OpenAI Agents SDK documentation. https://openai.github.io/openai-agents-python/human_in_the_loop/ Accessed 6 October 2026.
- Prasad, H., & Pal, R. (2026). The joint effect of quantization and sampling temperature on LLM safety alignment: A factorial analysis. arXiv:2606.29581v2. <https://arxiv.org/abs/2606.29581>
- Russell, S. (2016, June). Should we fear supersmart robots? *Scientific American*, 58–59.
- Smit, A. P., Grinsztajn, N., Duckworth, P., Barrett, T. D., & Pretorius, A. (2024). Should we be going MAD? A look at multi-agent debate strategies for LLMs. *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, PMLR 235; arXiv:2311.17371v3. <https://arxiv.org/abs/2311.17371>

Smith, R. J. (2008). *Fathoming the Cosmos and Ordering the World: The Yijing (I Ching, or Classic of Changes) and Its Evolution in China*. Charlottesville and London: University of Virginia Press. (Richard Lectures for 1999.) Cited by chapter.

Smith, R. J. (2009). The *Book of Changes* as a mirror of the mind: The evolution of the *Zhouyi* in China and beyond. Paper for the Fourth International Conference of Analytical Psychology and Chinese Culture, Fudan University, Shanghai, 10–12 April 2009. English text: Center for Zhouyi and Ancient Chinese Philosophy, Shandong University (posted 11 April 2010). <https://zhouyi.sdu.edu.cn/en/info/1051/1291.htm>

Smith, R. J. (2012). *The I Ching: A Biography*. Princeton University Press. Cited by chapter.

Stavropoulos, A., Crone, D. L., & Grossmann, I. (2024). Shadows of wisdom: Classifying meta-cognitive and morally grounded narrative content via large language models. *Behavior Research Methods*, 56(7), 7632–7646. <https://doi.org/10.3758/s13428-024-02441-0>

Sternberg, R. J. (n.d.). *Wisdom*, “Balance Theory of Wisdom”. Author’s account. <https://www.robertjsternberg.com/wisdom> Accessed 6 October 2026.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2), 181–211. [https://doi.org/10.1016/S0004-3702\(99\)00052-1](https://doi.org/10.1016/S0004-3702(99)00052-1)

The Myers-Briggs Company. (n.d.). *MBTI Facts*, question 8, “How is the MBTI assessment related to Jung's theory?” Accessed 9 October 2026. <https://www.themyersbriggs.com/en-us/support/mbti-facts>. Cited for the developer's account of the scheme's construction.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology / Psychologie canadienne*, 26(1), 1–12. <https://doi.org/10.1037/h0080017>

VanRullen, R., & Kanai, R. (2021). Deep learning and the Global Workspace Theory. *Trends in Neurosciences*, 44(9), 692–704. <https://doi.org/10.1016/j.tins.2021.04.005>; arXiv:2012.10390v2.

Wheeler, M. A., Stuss, D. T., & Tulving, E. (1997). Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin*, 121(3), 331–354. <https://doi.org/10.1037/0033-2909.121.3.331>

Wang Bi (王弼). *Zhouyi lüeli* (周易略例), “Ming gua shi bian tong yao” (明卦適變通爻) and “Bian wei” (辯位). Wikisource transcription: <https://zh.wikisource.org/zh-hant/周易略例/明卦適變通爻> (A modern transcription, not a critical edition; the English renderings on the frontispiece are the authors'; Lynn (1994) translates the treatise.)

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, 131(3410), 1355–1358.

Wilhelm, R., & Baynes, C. F. (Trans.). (1967). *The I Ching or Book of Changes* (3rd ed., Bollingen Series XIX; 7th printing, 1971). Princeton University Press. (Hexagram glosses in this paper follow its printed hexagram headings.)

Yang, Y., Yi, E., Ko, J., Lee, K., Jin, Z., & Yun, S.-Y. (2025). Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. arXiv:2505.22960v2. <https://arxiv.org/abs/2505.22960>

Yang, Z., Chen, G., Yang, Y., Zeng, A., & Yang, X. (2026). Disentangling task conflicts in multi-task LoRA via orthogonal gradient projection. arXiv:2601.09684v1. <https://arxiv.org/abs/2601.09684>



Yudkowsky, E. (2007, 28 July). Making beliefs pay rent (in anticipated experiences). LessWrong. <https://www.lesswrong.com/rationality/making-beliefs-pay-rent-in-anticipated-experiences>

Zhongyong (中庸). In *Liji* (禮記), received text; chapter 2 in Zhu Xi's division. Wikisource transcription: <https://zh.wikisource.org/zh-hant/禮記/中庸> (a modern transcription of the received text, not a critical edition).

Zhang, H., Cui, Z., Wang, X., Zhang, Q., Wang, Z., Wu, D., & Hu, S. (2025). Stop overvaluing multi-agent debate — we must rethink evaluation and embrace model heterogeneity. arXiv:2502.08788v3. <https://arxiv.org/abs/2502.08788>

Zhang, J., Lin, N., Hou, L., Feng, L., & Li, J. (2025). AdaptThink: Reasoning models can learn when to think. arXiv:2505.13417v1. <https://arxiv.org/abs/2505.13417>

Zierahn, K., Cachero, C., Korhonen, A., & Oliver, N. (2026). Personality without persons? A psychometric critique of Big Five testing in large language models. *AIES 2026*; arXiv:2607.02325v2. <https://arxiv.org/abs/2607.02325>

Project documents (Hexis programme, circulation-safe, to be released with the yarrow repository): yarrow.hexmath contract v0.1 (4 October 2026); TAI construct and validation specification v0.2; Hexis ethics protocol v0.2; sealed configuration mechanism v0.2; consultation document, round 2 (all 4 October 2026).

