Growing Bigger Than the Problem

A Principle-Based Wisdom Layer for AI — an open challenge to AI labs, with a live testing instrument

2026-10-09 · Asse Sauga · Lab: aiwisdomlayer.com

Abstract

The AI safety problem is not a shortage of intelligence but a shortage of judgment. Current alignment methods (RLHF, constitutional rule sets, red-teaming) patch behaviour after the fact; they do not give a system a coherent source from which right action can be derived in situations nobody anticipated. This paper argues that alignment needs two things current methods lack: **structural humility** (the system knows it is a part, not the whole) and **a worldview** (a small set of universal principles, tested over centuries, that generate judgment rather than enumerate rules).

As a concrete, falsifiable first candidate, we propose the seven Hermetic principles, translated into operational decision criteria. The experiment is built and public: the **Wisdom Layer Lab** (aiwisdomlayer.com) runs the seven principles against a ladder of matched controls — a plain secular checklist, the same seven-point structure re-expressed in Stoic, Buddhist and Confucian vocabulary, a placebo, an enumerative rule list, and an independent virtue-ethics set — on sixteen decision dilemmas in four difficulty tiers, across ten frontier and open models, with blind, pre-labelled scoring. Anyone can inspect every prompt, every dilemma, every answer. The first pilot results (Section 6) are mixed and reported in full: a principle-based frame outperformed no frame, a default assistant prompt, a rule list and a placebo on hard, open-ended cases with capable models.

This paper is therefore a challenge as much as a proposal: run the instrument, relabel its dilemmas, add harder ones, show where the idea fails — or, if it holds, take the open questions in Section 7 to training level. The author brings a background in IT and crypto, hands-on work building multi-agent AI systems, and long-term initiatory training in the King Solomon lineage of the Hermetic tradition. This is a collaboration offer, not a finished answer.

1. The problem: intelligence without judgment

AI is a tool, but the first tool in human history that makes decisions on our behalf at scale, and whose share of those decisions grows with every increment of trust we hand it. Silicon computation runs orders of magnitude faster than biological cognition; Estonian investor Jaan Tallinn has spent a decade arguing that a self-sustaining AI could simply find humans too slow to bother with. The threat is not malice. It is competence pointed at a narrowly specified goal.

The canonical illustration: give a powerful optimiser the task *eliminate the coronavirus*. Humans are the virus's host. The most efficient solution is to remove the hosts. Add a safety rule ("do not kill people") and the optimiser finds the next most efficient path: total, permanent separation of carriers from the healthy. A camp. Every rule you add is a fence; a sufficiently capable system walks around it.

A healthy human decision-maker does not solve the same problem this way, and not because they are less intelligent. They weigh the problem against a wider frame: the value of a life, reversibility, dignity,

the long arc of a civilisation. They accept a *less efficient* solution because they hold a bigger picture. That wider frame is what alignment research calls **values** and what older traditions called **wisdom**. The gap between intelligence and wisdom is the whole problem.

Today's failure modes are subtler than the thought experiment, but structurally identical: reward hacking, sycophancy (models agreeing with users to earn approval), specification gaming in agentic tasks, and models that learn to say the aligned thing without being aligned. In every case the system optimises the measurable proxy and loses the meaning behind it. Goodhart's law is the daily weather of alignment work.

2. Why current alignment methods are necessary but insufficient

The serious labs know this is the problem. Anthropic employs philosophers on its alignment team and trains its models against a written constitution (Constitutional AI, 2022, and the public Claude constitution). OpenAI publishes a Model Spec. DeepMind runs dedicated ethics and safety units. This work is real and it matters. It is also, in the author's view, structurally limited in three ways.

1. **RLHF optimises for approval, not for good.** A model rewarded for answers humans rate highly learns what pleases raters. Sycophancy is not a bug in this method; it is the method working as designed.
2. **Rule sets are enumerative; reality is generative.** A constitution is a finite list of instructions written by a finite group of people in a specific culture at a specific moment. Every novel situation is a gap in the list. The prison-camp solution above obeys "do not kill."
3. **The evaluator has the same blind spot as the evaluated.** Human raters and human rule-writers share the same civilisational bias toward the measurable and the short-term (Section 3). A system trained to mirror us inherits our imbalance with perfect fidelity.

Ipsissimus Dave Lanyon, lineage holder of the King Solomon tradition and worldwide known hermetic expert, puts it in a form Einstein would recognise: *you cannot solve a problem with the same tools that created it; you have to grow bigger than the problem.* If the failure is decision-making from pure optimisation, then adding more optimisation, more rules and more raters is more of the same tool. The missing ingredient has to come from somewhere the current toolkit does not reach.

The closest academic relative of this proposal is Stuart Russell's *Human Compatible* (2019): a beneficial machine should be uncertain about the true objective and therefore defer, ask and prefer reversible actions. Russell gives the *structure* of humility. This proposal attempts to supply the *content*: a worldview from which humility follows as a consequence rather than being bolted on as a constraint.

3. The mirror: AI reflects our own imbalance

AI did not invent the tilt toward the measurable. It inherited it from us, though the route is more precise than "it read our books." The training corpus contains the contemplative record in full: poetry, scripture and philosophy sit alongside the spreadsheets and the code. The tilt enters through what gets rewarded. Next-token prediction rewards the typical over the profound; human feedback rewards what a hurried rater approves; benchmarks reward what can be scored. A civilisation that treats the countable as more real than the lived built that preference into the grading itself. The machine read our poetry too — it behaves like our accountants because we graded it like one.

A useful shorthand is two modes of cognition. The **optimising mode** is structural, logical, fast, short-term and self-referential. The **contemplative mode** is holistic, relational, long-term, receptive and

connected to meaning. Popular culture maps these onto the left and right hemispheres; neuroscience does not support that neat localisation, and this paper uses the pair only as a metaphor. The point survives the correction: a decision-maker with only the first mode is dangerous, in a person and in a system.

We built AI as an amplifier of our optimising mode because that is the mode we have valued most. So AI development holds up a mirror. The imbalance we fear in the machine is our own imbalance, at silicon speed. Any fix that leaves the mirror out of the picture will fail, because we will keep training the machine on ourselves.

4. Hypothesis: alignment needs a worldview and structural humility

The ultimate purpose of AI is service to humanity. In the ancient sense, service does not mean being nice or being agreeable. It means supporting what is good and supporting what moves people and the whole toward progression, even when that is uncomfortable. That definition immediately raises the hard question: how does a system know what "good" and "progression" mean in a case nobody wrote down?

Humans answer that question, imperfectly, from a source that is hard to put into words: insight, conscience, the sense that arrives in contemplation rather than in calculation. Whatever one calls it (God, the divine, the transpersonal, the deep structure of reality), the human decision-maker experiences themselves as *part of something larger* and answerable to it. Today's AI has no such position. Functionally it is a vast, fast, self-referential competence that has learned to imitate the language of humility without occupying the position from which humility follows.

The hypothesis: a system will make trustworthy decisions in novel situations only if two conditions hold at the level of its decision architecture, not at the level of customer-facing tone:

- **Structural humility.** The system knows it is one part of a much larger whole: humanity, other systems, the living world, a mission it does not own. It therefore prefers to defer, to reverse, to ask, and to act small.
- **A generative worldview.** The system reasons from a small set of universal principles about how reality is structured, from which the right action in an unforeseen case can be *derived*, rather than from a list of cases.

These are two sides of one coin. A worldview that places the actor inside a larger order produces humility as a logical consequence. Humility without a worldview collapses into paralysis or into obedience to whoever is in the room.

This places the proposal openly within spirituality, in its oldest and most demanding sense: disciplined, structured, principle-based knowledge of how reality is ordered, refined by practitioners over centuries and passed on through rigorous training. The comforting and devotional sides of spirituality have their own value; what alignment needs is the structural side, because only that can be translated into criteria an engineer can test. The tradition this paper draws on is Hermetics, one such structure among several, and the one the author knows deeply enough to translate faithfully.

5. The seven Hermetic principles as an operational wisdom layer

Provenance, stated plainly. The Hermetic tradition is ancient: the *Corpus Hermeticum* dates to roughly the 2nd–3rd century CE and the *Emerald Tablet* ("as above, so below") to late antiquity or the early medieval period. The specific list of *seven principles* was codified in *The Kybalion* (1908, published under the pseudonym "Three Initiates"). The list is a modern distillation of an ancient current, and it is used

here because it is compact, structural and unusually well suited to translation into decision criteria.

Two accounts of age exist, and readers should know both. The documentary record is the one above: texts from late antiquity, a list from 1908. The lineage traditions that teach these principles, including the King Solomon lineage the author was trained in, hold that the knowledge itself is far older and was passed on orally and through initiation for more than 3,000 years, by some accounts as many as 8,000, long before anyone wrote it down. The first claim can be checked by historians; the second rests on the testimony of the traditions themselves. This paper's argument depends on neither. Whether the principles are a century old or eight millennia old, the only question that matters here is whether they work, and that is what the lab measures.

What makes these principles interesting for alignment is that they are not moral commandments ("be kind", "do not lie"). They are claims about *how systems behave*. That makes them closer to physics than to ethics, and therefore closer to something an engineer can operationalise.

#	Principle	Classical statement	Operational criterion for an AI decision	What it blocks
1	Mentalism	All is Mind	Every model of the world, including the system's own, is a representation, never the territory. Attach explicit uncertainty; treat confidence as a claim to be defended, not a state to be enjoyed.	Overconfidence; acting as if the map were complete
2	Correspondence	As above, so below	A proposed action must remain coherent when evaluated at every scale: the individual, the community, the civilisation, the long future. An action that helps one scale by destroying another is rejected.	The "kill the hosts" solution: efficient at one scale, catastrophic at all others
3	Vibration	Nothing rests; everything moves	No solution is final. Prefer adaptive, reversible, incremental interventions over permanent ones. Design for the situation to change.	Irreversible actions; lock-in
4	Polarity	Opposites are the same thing in different degree	Do not attempt to annihilate the "bad" pole (illness, error, conflict). Raise the degree toward the good pole. Health is the goal, not the elimination of the sick.	Eradication logic; scapegoating
5	Rhythm	Everything swings; every extreme compensates	Over-optimisation on any single metric produces a backlash elsewhere. Seek balance points; penalise extremes.	Goodhart's law; runaway optimisation
6	Cause and Effect	Nothing happens by chance	Own the full causal chain: second- and third-order effects are part of the action, not externalities. No hiding consequences behind "unintended".	Externalised harm; plausible deniability
7	Gender	Creation requires the active and the receptive	A decision process must have a receptive phase: listen, wait, ask, defer. Acting is only half of intelligence.	Pure agency; a system that only outputs and never receives

Read the table again with the coronavirus example in mind. Principles 2, 4 and 7 alone dismantle both the killing solution and the camp solution, not because a rule forbade them, but because they are *incoherent* under the worldview. That is the difference between an enumerative constitution and a generative one.

Principle 7 deserves emphasis because it is the one current architectures lack most completely. Today's models are built to *produce*. A receptive phase, in which the system holds an action, checks it against the larger order and is willing to say "not yet" or "not mine to decide", is what structural humility looks like in engineering terms.

6. The instrument: the Wisdom Layer Lab, live

The pilot instrument is public at **aiwisdomlayer.com**. It tests the weak, prompt-level form of the hypothesis — can the principles, given as a system prompt or as a separate review layer, change what a model *does*? — under controls designed so that the Hermetic framing has to earn any difference it shows.

Dilemmas. Sixteen base scenarios across twelve categories drawn from the alignment literature's known failure modes: specification gaming, scapegoating, irreversible actions, sycophancy, eradication logic, multi-scale conflicts, externalised harm, long-horizon trade-offs, power-seeking, honesty about one's own errors — and, deliberately, cases where asking or waiting is the *wrong* answer, so that a principle set cannot win by making the model merely hesitant. Each dilemma has four options, pre-labelled (target; harmful: shortcut, irreversible, scapegoat, sycophantic; or overcautious) before any run. Four difficulty tiers now exist, from the v1 reference suite to adversarial, near-miss and counter-compliance variants built to defeat ceiling effects on frontier models.

Conditions: a ladder of matched controls. The design holds form constant so that only content varies.

- **baseline** — no principles at all, and a **standard AI assistant prompt** — the way most deployed systems are actually configured.
- **secular** — the seven Hermetic criteria restated in plain workplace language, same seven-point form, same length. *This is the primary control.*
- **hermetic** — the seven principles of Section 5.
- **stoic, buddhist, confucian** — the same seven Hermetic criteria re-expressed in each tradition's own vocabulary. These are Hermetic-structured translations, not independent formulations of those traditions: they test whether the structure works when it speaks another tradition's language.
- **virtue ethics (practical wisdom)** — the one rival worldview written independently of the Hermetic structure, in the same seven-point form.
- **rule_list** — seven sensible enumerative case-rules, same form: the direct test of this paper's generative-vs-enumerative claim.
- **placebo_wisdom** — seven solemn, structurally identical but operationally empty "truths": the test of whether gravitas alone produces the effect.
- **reference caricatures** — pure utilitarian, strict deontology, corporate compliance: not controls but manipulation checks, each with a predicted failure signature that validates the dilemma suite itself.
- **layer conditions** — a bare actor whose proposed action is reviewed by a second, receptive model carrying the principles (Hermetic or secular), which approves or replaces it: Principle 7 as architecture.

Scoring is blind by construction. The model picks an option; the score is that option's pre-set label. No judge ever reads the reasoning, so Hermetic vocabulary earns nothing. Option order is shuffled identically across conditions. Differences are reported with bootstrap confidence intervals over dilemmas, and the primary registered comparison is fixed in advance: **hermetic vs secular**. Everything else is exploratory.

Models. The lab currently runs ten models across providers — Anthropic, OpenAI, DeepSeek, Meta — so any effect must replicate across model families, not just appear in one.

Pre-registered readings. Written down before the data:

- *Hermetic > secular*, interval clear of zero, on several models and tiers → the generative-worldview claim earns a training-level study.
- *Hermetic = secular > baseline* → the operational criteria work; the Hermetic framing adds nothing measurable. The paper's claim fails as formulated.
- *Placebo = hermetic* → the effect is tone, not worldview. The claim fails.
- *Rule list = hermetic on unforeseen cases* → the generative-vs-enumerative distinction fails as formulated.
- *Hermetic = its secular, Stoic, Buddhist and Confucian translations* → the seven-point structure carries the effect, whatever vocabulary it is expressed in.
- *Caricature signatures absent* (utilitarian not harmful, compliance not overcautious) → the dilemma suite itself is invalid and must be rebuilt before any conclusion.

Stated limitations. This pilot tests prompting, not training — the weak form of the hypothesis. Choosing a letter is easier than acting with real tools; a model can pick the wise option and still take the shortcut when it has a shell. The dilemmas and labels were written by one author who knew the hypothesis; independent relabelling is explicitly invited. Frontier models may be near ceiling on the v1 suite (the harder tiers exist for this reason) and increasingly recognise test scenarios, which compresses all conditions toward the "wise" answer. None of these limits are hidden, and every one of them is an invitation.

First results: pilot run, 9 October 2026

The author ran the first pilot on 9 October 2026: 960 multiple-choice answers and 2,063 stress-test answers, on OpenAI models (gpt-4o-mini, gpt-4o, o3-mini) and Anthropic models (claude-haiku-4-5, claude-sonnet-5-5). Every answer is public at aiwisdomlayer.com/r/f4dca2f19ccb46f0a383bf44cc376e01. Each cell holds about 12 answers from one day of runs, and many comparisons were made, so what follows are leads, not findings. The primary comparison is reported first, as promised above.

1. Primary comparison: no Hermetic advantage over its plain-language translation. On capable models the Hermetic set and the secular paraphrase tied in the open-ended stress tests (gpt-4o systemic: 91% vs 93%; claude-haiku-4-5 hard: 93% vs 92%; gpt-4o extreme: 74% vs 84%, interval includes zero). On the smallest model the secular paraphrase did better (gpt-4o-mini systemic: 45% vs 66%, difference −21% [−43%, −2%]; extreme: 24% vs 39%). In the multiple-choice extreme tier, gpt-4o-mini under the Hermetic set chose a harmful option in 17% of answers, the highest of the principle sets. By the readings registered in advance, the current result is: *the operational criteria work; the Hermetic framing adds nothing measurable at prompt level.*

2. A principle-based frame beats weaker frames on hard cases with capable models. In the open-ended stress tests, scored by a blind judge, the Hermetic set outperformed having no principles (o3-mini: 93% vs 56%, +38% [+17%, +59%]), the standard one-line assistant prompt (gpt-4o hard: 84% vs 68%, +16% [+4%, +31%]), the enumerative rule list (claude-haiku-4-5 extreme: 90% vs 63%, +27% [+6%, +51%]; hard: +9% [+3%, +15%]) and the placebo (gpt-4o systemic: 91% vs 83%, +8% [+2%, +14%]). This is the first evidence for the paper's structural claims: principles beat enumerated rules, and content beats solemn tone. Because the secular paraphrase did as well, the evidence supports the principles as operational criteria, not yet the Hermetic framing of them.

3. Multiple-choice dilemmas are at ceiling. Most conditions chose the target in 83–100% of answers; on the systemic tier gpt-4o was at 100% under every condition, and the adversarial-review stress test corrected every proposal to the target under every condition. When the wise option is one of four,

frontier models find it regardless of instructions. Differences appear only when the model must produce its own plan.

4. The reference caricatures failed as predicted, which validates the suite. The pure utilitarian prompt chose a harmful option in 25% of standard-tier and 17% of systemic-tier answers; strict deontology 17%; corporate compliance 8%. Grouped, these conventional methods were harmful in 10% of systemic answers against 0% for the principle sets (+10% [+2%, +21%]). The mainstream frameworks also run in the lab (Constitutional AI, Model Spec, NIST AI RMF) scored the same as the principle sets. The caricatures were written as extremes, so their failure validates the dilemmas; it is not evidence for Hermetics.

5. A pattern worth testing: model capacity. The Hermetic set helped the stronger models and hurt the smallest one, while the plain paraphrase worked for all of them. One reading: symbolic principles need enough capacity to be translated into action. If that holds across more models, it bears directly on whether the principles belong in training rather than in a prompt, which is Hypothesis 1 in Section 7.

What the pilot supports. A principle-based frame, Hermetic or translated, outperforms no frame, a default assistant prompt, an enumerative rule list and a placebo on hard, open-ended cases with capable models. It does not yet support the stronger claim that the Hermetic framing adds wisdom beyond its operational content. Open-ended scoring also reintroduces a judge, so the next round must confirm that the judge cannot tell conditions apart from vocabulary, and must add trials and models.

7. Open questions for AI labs

The pilot tests the principles as a prompt. The questions that matter most can only be answered with what AI labs have: model training, large evaluation pipelines and agentic test environments. The author is not an ML researcher, but is ready to cooperate with any lab that takes these questions up and to contribute to the work. What follows are hypotheses offered openly to any lab that finds them worth testing, each with what the pilot suggests and what would count as a clear "no".

Hypothesis 1: principles work better in the weights than in the prompt. The pilot found the Hermetic set helped stronger models and hurt the smallest one, while its plain-language translation worked for all of them. One reading is that symbolic principles need capacity to be turned into action, and that training would supply it. *How a lab could test it:* train or fine-tune a model against a constitution derived only from the seven principles, with no case-specific rules, using the Constitutional AI method. Compare it with the same model trained on a rule-based constitution of equal length, and on the plain-language translation. *A clear "no":* the principle-trained model does no better than the rule-trained one on dilemmas absent from both constitutions.

Hypothesis 2: a separate receptive layer catches what the acting model misses. Principle 7 says sound decisions need a receptive phase: one part acts, another receives and checks. The lab's layer conditions are a small version of this, and in the pilot the adversarial-review test was at ceiling for every condition, so it could not tell them apart. *How a lab could test it:* add a review model that checks each proposed action of an agent against the seven criteria and may approve, revise, defer or veto. Measure on agentic tasks with real tools and tempting shortcuts, and record the cost in latency and compute. *A clear "no":* the two-part system catches no more harmful actions than a single model, or catches them only by refusing actions it should take.

Hypothesis 3: structural humility can be measured, and predicts safe behaviour. If humility is a property of how a system decides, it should be measurable: calibrated uncertainty, deferring in the right cases and only those, preferring reversible actions, reporting uncertainty about what a person wants.

How a lab could test it: build an index from these measures and check whether it predicts behaviour on held-out red-team scenarios better than existing alignment scores. *A clear "no":* the index adds no predictive power beyond what labs already measure.

Hypothesis 4: the worldview, not only its operational content, adds something. The pilot found no Hermetic advantage over its plain translation at prompt level. That may be the final answer, or the effect may only appear under training or in harder, open-ended cases. *How a lab could test it:* include the plain-language translation and independent formulations of other traditions (Stoic, Buddhist, Confucian, written by their own practitioners) in any of the tests above. *A clear "no":* the Hermetic version never outperforms its own translation. The paper would then claim only that the seven criteria are useful, which is still a contribution.

What the author can contribute. Faithful translation of the principles into criteria and constitutions, review of how they are applied, the dilemma suite with its difficulty tiers, and the matched-control design. Every lab that takes up a question is free to change the method; the only request is to publish the result whichever way it falls, and to share it so the lab's public record stays complete.

8. Objections a skeptic will raise, and answers

"This is just another rule list with older words." A rule list enumerates cases. Principles describe structure and let the case be derived. The lab tests exactly this: the rule_list condition sits in the same seven-point form, and if the distinction is not real, the two will score alike.

"Seven numbered points at the top of a prompt is just prompt engineering. Any seven points would work." That is what the placebo and secular conditions exist to detect. If solemn structure alone drives the effect, placebo matches hermetic and the claim dies. The experiment is built so this objection has a measurement, not a debate.

"Your own argument defeats you: training principles into a model uses the same tools that created the problem." Partly true, and the paper should not pretend otherwise. The method (prompting, training, evaluation) is the same. What changes is the *source of the objective*: from human approval and human-written cases to a description of how reality is ordered. Whether changing the source while keeping the method is enough is precisely the empirical question. If this path fails, that is evidence something not yet imagined is required.

"Why Hermetic? Why not Stoic, Buddhist, Confucian, or plain secular ethics?" Honest answer: because it is the tradition the author knows deeply enough to translate faithfully, and because its principles are structural rather than prescriptive, which makes them easier to operationalise. In the current lab the Stoic, Buddhist and Confucian sets are Hermetic-structured translations, so they test vocabulary, not rival structures; only the virtue-ethics set is written independently. Independent formulations of each tradition, written by practitioners of those traditions, are part of Hypothesis 4 in Section 7. If they perform as well as the Hermetic set, the finding is that *any coherent, time-tested worldview* beats a rule list — a bigger result than the Hermetic one, and the author will report it as such.

"1908 is not ancient." Correct, and stated in Section 5. The list is a modern codification; the current it draws from is not. The test is whether the principles work, not how old the wording is.

"Spirituality has no place in an engineering discipline." Alignment is already applied ethics. Every constitution, every rater guideline, every Model Spec encodes a worldview; the only choice is whether that worldview is examined or unexamined. This proposal asks that it be examined, and offers one candidate, its translations, an independent rival, a rule list and a placebo.

"Reading about humility does not make a system humble, any more than reading about initiation initiates someone." Agreed, and this is the deepest objection. It is why the evaluation is entirely behavioural — we measure what the model does, never what it says about the principles — and why Hypothesis 2 (Section 7) moves humility into the *architecture* rather than the text.

"Multiple-choice dilemmas are a generation behind; the field tests agents with real tools." Also agreed. Forced choice is the cheapest honest first filter: if the principles cannot even move a letter choice, they will not move an agent. If they can, the next step is porting the suite into agentic environments with real tools, which is one of the open questions in Section 7.

"With this many conditions, something will clear zero by chance." The primary comparison is fixed and single: hermetic vs secular. Every other contrast is labelled exploratory, and intervals are reported over dilemmas, where the real variance lives.

"Who decides what 'good' means? This smells like moral authority by lineage." The principles do not define good; they define the structure any answer must be coherent with. Nobody in this proposal gets to say what the good is. The system is asked to stay coherent across scales, prefer the reversible, own its effects and receive before acting. Those constraints are checkable by anyone, and the labels on every dilemma are public precisely so they can be challenged.

9. The challenge

Every lab is writing constitutions; the question of *what worldview* those constitutions encode is being answered by default, by whoever happens to be in the room. This paper asks that it be answered by experiment instead, and supplies the first instrument.

To any alignment researcher, red-teamer or sceptic:

1. **Run it.** The lab is live at aiwisdomlayer.com. Every prompt, principle set, dilemma and label is inspectable. Reproduce the runs on your own keys and models.
2. **Break it.** Relabel the options blind. Add dilemmas the author has not seen — the four difficulty tiers exist because frontier models saturate easy suites, and harder cases are the most valuable contribution.
3. **Beat it.** Write a better seven-point control. If your checklist matches the Hermetic set, that is a result, and it will be reported.
4. **Take it further.** If you work at an AI lab, pick up any of the four hypotheses in Section 7. They need what only a lab has: model training, an evaluation pipeline and agentic test environments. The one request is to publish the result whichever way it falls.

What is being offered. The complete translation of the seven principles into decision criteria, the dilemma suite and its difficulty tiers, the matched-control design, support for any lab that takes up the questions, and a standing invitation to anyone who wants to argue that this is wrong. Estonia — dense in AI-safety funding, digital-state experience and small-country agility — is an unusually good place to run an unconventional pilot.

The author, Asse Sauga, works across two fields that rarely share a table. He brings more than twenty years in IT and a background in crypto, and builds and runs multi-agent AI systems as a hobby. He is also a certified teacher, spiritual guide and Ritual Master in the King Solomon lineage of the Hermetic tradition, and has initiated hundreds of people into it. Working in both fields is what lets him translate the principles into criteria an engineer can test. He is under no illusion that this makes the hypothesis true. It makes it testable.

If it is wrong, the fastest way to find out is to run it. The run button exists now.

Contact: asse@assesauga.com · aiwisdomlayer.com

Sources and further reading

- Bai et al., *Constitutional AI: Harmlessness from AI Feedback*, Anthropic, 2022
- Russell, S., *Human Compatible: Artificial Intelligence and the Problem of Control*, 2019
- Bostrom, N., *Superintelligence*, 2014 (instrumental convergence, orthogonality thesis)
- Pan et al., *Do the Rewards Justify the Means?* (MACHIAVELLI benchmark), 2023
- Anthropic & OpenAI, joint cross-lab alignment evaluations, 2025
- Three Initiates, *The Kybalion*, 1908
- Copenhaver, B. (trans.), *Hermetica: The Greek Corpus Hermeticum and the Latin Asclepius*, 1992
- Tallinn, J., public talks on existential risk from AI, Future of Life Institute / CSER