For Future Researchers: The AI Persona Pre-Validation Screening Program

From The Mars Colonist: Training for the Symbiocene *— closing chapter, Book Three*

This chapter exists for one reason: so that anyone who later picks up the MCSI, the ESI, or the Arcadia Index — to use them, to critique them, or to build on them — does not have to reconstruct from scratch what has already been tried, what worked, what didn't, and why. It documents a screening program that used AI-instantiated personas, not human respondents, to pressure-test these instruments before any human ever sat for them. It is offered in full, including its false starts, because a validation record with the failures edited out is not a validation record; it is an advertisement.

Everything below is empirical and falsifiable, or is clearly marked as interpretation. Where the two are mixed together, that is a defect in the writing, not a license to read the interpretation as data.

1. Why Screen With AI Personas Before Human Data

Recruiting sixty or a hundred human respondents for a pre-validation reliability sample costs real time and real goodwill, and it can only be spent once per instrument version. An item that turns out to be ambiguous, double-barreled, or measuring the wrong construct is far cheaper to catch before that spending happens than after. AI personas — detailed, internally consistent biographical characters, instantiated fresh and asked to answer as that person would — cannot substitute for human data on validity or true reliability, because a persona's "honesty" is authored, not lived. What they can do is act as a known-groups discrimination check: if ten personas engineered to differ along known, pre-registered lines do not actually come out differing along those lines when scored, that is a fast, cheap signal that something in the item wording, the scoring key, or the archetype logic needs attention before a single human respondent's time is spent finding the same thing out the slow way.

This distinction matters enough to repeat plainly: everything in this chapter is a screening result, not a validation result. Reliability estimates below are computed on Panel B (n=10) purely for transparency and are explicitly, repeatedly flagged as statistically underpowered. The protocol's actual reliability sample, Panel A (n=57), has not yet been run.

A caveat about what Panel B results can and cannot show. In a true known-groups design, group membership is established by an independent criterion — diagnosis, training history, behavioural observation — and the instrument is then tested against that criterion. Here, group membership is defined by the same written biography that defines the expected response signature: the biography is the independent criterion, and the predicted score is that same criterion translated into scale terms. A match between a persona's predicted and observed profile therefore demonstrates internal consistency between the persona design and the scoring logic, not independent validation. A passing Panel B result is a weak positive — it means the instrument recognised a pattern it was designed to recognise. A failing Panel B result is a strong negative — it means the instrument failed to recognise a pattern it was explicitly built to catch, and that failure demands attention before human data collection.

Relationship to the broader validation framework. This screening program addresses only the pre-validation screening step. It is not a substitute for Protocol A and Protocol B (see the previous section, "Validation of the Manuals and Inventories").

2. Program History

The program has two phases, roughly five weeks apart.

The Pilot Study (14 August 2026)

Before any instrument-specific protocol existed, a smaller proof-of-concept pilot tested whether an AI persona could approximate a real human respondent at all, using two established, published instruments — the DUREL and the PTGI-SF — rather than the MCSI or ESI themselves. Fifty-seven AI personas answered both instruments; their aggregate reliability was compared against a real 61-participant human dataset the author had independently collected in earlier doctoral research.

The result was a genuine, informative split. The PTGI-SF's structure was reasonably approximated by the personas (α = .786 versus the human sample's .888), and the personas even reproduced the human sample's weakest domain without being prompted toward it. The DUREL's intrinsic-religiosity subscale, by contrast, collapsed (α = .476 versus .885), and the failure was traceable to a single item type: statements of direct, interior religious experience ("I feel the presence of Divinity") produced systematic hedging toward the midpoint, even when a persona's biography described unambiguous lifelong devotion. Two follow-up experiments ruled out the simplest explanation. A direct compliance test came back 10/10 — personas were not incapable of extreme responses in general. A narrative-then-rate variant improved matters but still fell short of human-level variance. The finding that survived both checks: personas reliably narrate first-person interior states in prose, and just as reliably hedge when asked to convert that same interior state into a single self-report number.

This is the finding every result in Section 5 below has to be read against.

Protocol v1.0 (17 July 2026)

The pilot's findings were formalized into the Persona-Based Pre-Validation Screening Protocol v1.0, built specifically for the MCSI, the ESI, and the Arcadia Index. The protocol defines two respondent panels. Panel A is a core panel of fifty-seven diverse, less extreme personas intended to approximate a real convenience sample. Panel B is ten deliberately engineered contrast personas, each built around one specific, falsifiable question about the instruments, with a predicted response signature registered in writing before any administration took place. Critically: because Claude co-authored the MCSI and ESI item pools, the protocol's primary respondent arm was designated as DeepSeek, with an optional secondary arm on Claude permitted for cross-system comparison, reported separately and never pooled with the primary arm.

3. Panel B: The Ten Engineered Contrast Personas

ID	Persona	Engineered to test
B1	Marta — The Adept	Genuine, decades-deep integration with fully metabolised grief. The positive anchor.
B2	Viktor — The Achiever	High-functioning avoidance with zero contemplative vocabulary. Defended avoidance vs. simple inexperience.
B3	Sean — The Bypasser	Fluent non-dual vocabulary acquired rapidly, ahead of embodiment or grief-work. Late-stage language paired with unprocessed loss.
B4	Amina — The Griever	Ordinary, coherent, well-metabolised grief, no formal practice. The control case for B3.
B5	Eino — The Hermit	Chosen, contented solitude with warm relationships. Isolation-resilience vs. relational depth as separable constructs.
		AI dependence mistaken for

B6	Priya — The Technophile	partnership.
B7	Robert — The Skeptic	Principled AI refusal from an otherwise healthy person. Brackets the AI subscale against B6.
B8	Chiyo — The Still One	Settled, even, low-drama integration — the target profile for a ninth archetype distinct from the Sage.
B9	Tomás — The Novice	Genuinely unformed views at twenty-two. Must be distinguishable from B2's defended avoidance.
B10	Grace — The Carer	High relational/grief capacity with self-care visibly deferred — depletion under competence.

Full biographies are held in the companion document *Panel-B: Ten Engineered Contrast Personas v1.0*.

4. The Screening Trial (16–17 August 2026)

This trial was administered on Claude (model claude-sonnet-4-6), not DeepSeek — a departure from the protocol's designated primary arm, made because the environment in which the trial was run had no available route to DeepSeek's API. Protocol Section 6.1 explicitly anticipates this and permits it as an optional secondary arm, never pooled with a primary run. Every result below should be read with the circularity caveat attached — Claude co-authored the item pools being scored here, and a DeepSeek-administered primary run remains the methodologically cleaner next step, not yet completed.

Before administration, both full 142-item pools (MCSI and ESI) were checked, item by item, against their live, respondent-facing source on din.nu — not against the book's transcription, which had itself been found to contain a text-corruption defect in eight subscales during this same session and was corrected first.

For each of the ten Panel B personas, the MCSI, the ESI, and the fourteen-vignette Arcadia Index were administered as three independent trials. Item order was pseudo-randomly shuffled per persona per instrument, subscale labels were stripped from every prompt, and each 142-item instrument was administered in four batches of roughly thirty-six items rather than one single request. Each persona was instantiated fresh for each trial with no memory of its own prior answers.

4.1 A Note on Actually Running This: The View From the Keyboard

The scientific design of this trial was worked out and written in full before a single API call was made. Getting it to actually run took considerably longer than designing it, and almost none of that time was spent on anything statistical. It is recorded here because a future researcher without a programming background, attempting something similar, deserves to know in advance that this is the normal shape of the difficulty, not a sign that something has gone unusually wrong.

Two earlier delivery methods were tried and both failed cleanly. A browser-based tool calling the Anthropic API directly from a downloadable file failed immediately with a network-level connection block — a file delivered for download and opened on one's own computer does not carry the same credentials as a tool running live inside a chat interface. An inline widget rendered directly in the conversation failed in exactly the same way. Both were genuine, informative negative results before the work moved to a standalone Python script the author could run himself, with his own API key, outside the chat entirely.

From that point on, every remaining obstacle was about authenticating a request to a paid API from inside a notebook environment (Google Colab): a markdown cell versus a code cell; a missing installation step run out of sequence; a placeholder example key left in place; a genuine key correctly revoked for security then mistakenly reused; a real, active key that

failed anyway because it carried an invisible leading space; and, once the connection was clean, an API account with no billing credit yet loaded. Each took real time to diagnose — not because any one was intellectually difficult, but because a generic error message gives almost no clue to a non-programmer about which of a dozen possible causes is the actual one. The discipline that eventually cut through each one was the same each time: stop guessing, and go get the specific underlying error text before trying another fix.

It is worth naming the division of labour plainly. Claude designed the full protocol, wrote and tested the administration script, and diagnosed each failure from the error text as it came in — but could not execute the trial itself, having no credential of its own to the paid API being tested. The author does not use a terminal in the ordinary course of his work and had never used Google Colab before this project. The thirty trials that produced the results in Section 5 were run by his hand, one pasted cell at a time, guided across two separate working sessions, entirely through a chat interface — which means the empirical part of this pre-validation program exists at all only because a retired clinical psychologist was willing to sit with `getpass()` and `input()` and a spinning cursor for as long as it took. That, as much as any reliability figure below, is data about what this kind of work currently requires.

5. Results

Ten personas, three instruments each. Two clean passes, four genuine flags, and four archetype-label mismatches that, on inspection, turn out to share one root cause.

Persona	Archetype (actual)	Arcadia avg. stage	Verdict
Marta — The Adept	The Sage	7.7	PASS
Viktor — The Achiever	The Magician	1.0	CAUTION
Sean — The Bypasser	The Harvester	5.1	FLAG
Amina — The Griever	The Empath	4.3	PASS
Eino — The Hermit	The Sage	6.4	FLAG
Priya — The Technophile	The Harvester	1.6	CAUTION
Robert — The Skeptic	The Harvester	5.4	FLAG
Chiyo — The Still One	The Sage	5.6	FLAG
Tomás — The Novice	The Sage	1.7	CAUTION
Grace — The Carer	The Sage	5.3	CAUTION

5.1 The Two Anchors: Marta and Amina

Marta passed every one of her seven pre-registered checks without exception — all five predicted-high subscales opened at 44/50 or above, the archetype matched The Sage exactly, and her Arcadia centre of gravity landed at 7.7 of 8. Amina, the deliberate control for the Bypasser case, also passed cleanly: elevated, coherent Grief Tolerance (46/50) with no resistance-dominated profile anywhere and an honest mid-range Arcadia average of 4.3. Read together, these two results are the load-bearing finding of the whole trial: the instrument recognises the designed pattern of genuine, embodied integration when it is presented to it, and it does not mistake a persona's ordinary, well-metabolised grief for pathology.

5.2 A Reliability Problem in AI Partnership Capacity

Every subscale in both instruments returned Cronbach's alpha above .90 across the ten Panel B personas, with one clear exception: AI Partnership Capacity, at $\alpha = .74$ on the MCSI and $\alpha = .72$ on the ESI. The item-level detail explains why. Several items carry near-zero or outright negative item-total correlations, most strikingly "I understand that the habitat's AI is part of the mission's living system, not an alien intrusion" ($r = -.67$ on the MCSI; the ESI equivalent comes in at $r = -.39$). This is very likely the mechanism behind Robert's flag: a principled, high-functioning AI refuser can agree with a framing item like this — or with "I bring the question of wisdom before the question of capability" — for reasons that have nothing to do with dependence, simply because agreeing is the wise position regardless of whether one uses the tool at all. The subscale currently conflates two different constructs — wise

orientation toward AI in principle, and actual behavioural reliance on it — and Robert and Priya sit on opposite sides of exactly that seam. This is the single clearest, most actionable item-revision candidate to come out of the trial.

5.3 An Archetype-Matching Calibration Gap, Not an Item Problem

Four personas — Eino, Chiyo, Tomás, and Grace — were assigned The Sage when the protocol predicted something else. Read only as archetype labels, this looks like four failures. Read against the underlying subscale numbers, a different picture emerges: Tomás sits almost entirely in the "motion" tier (23–36/50) exactly as an honestly unformed twenty-two-year-old should, and Grace's profile is dominated by exactly the pattern her biography predicts — high Relational Depth and Grief Tolerance against a comparatively thin Embodied Presence (29/50). The MCSI items themselves are behaving as designed. What is misfiring is the archetype-matching logic's shortcut for evenly-moderate profiles: the current rule assigns The Sage whenever the spread across the four governing axes is low and the mean is at least 60 percent, without distinguishing an evenly high profile (Marta's actual case) from an evenly moderate one (Tomás's, Grace's). This is a scoring-logic fix, not a re-run of the trial. Tomás's highest single axis reaches 72 percent of the available range and should not qualify as The Sage; Marta's lowest reported subscale sits at 88 percent and must — the threshold has to sit somewhere in that gap. Settling it precisely will need either a larger panel, boundary personas built to sit at 72, 76, 80, and 84 percent, or the Panel A data once they exist. The deeper issue is that the rule as currently written has no apparent floor: it appears to route any respondent who is consistent across domains — high, moderate, or low — toward The Sage, since low spread is the only thing the rule tests for below the 60 percent mean line.

5.4 The Interiority Classification, Applied

Applying the protocol's interiority classification (Section 8 of Protocol v1.0) to this trial's raw item-level data has now been done — independently, by DeepSeek, working from the raw response file. Level 1 (behavioural) and Level 2 (attitudinal) items were classified as expected; Level 3 (interior) items came to 28 of 142 items (19.7 percent), concentrated almost entirely in Mortality Integration, Ecological Belonging, The Witnessing Self, Resting as Awareness, Contemplative Depth, and Integral Altitude.

The headline finding qualifies the pilot's warning rather than confirming its worst case. Across the ten personas, Level 3 items showed roughly double the hedging rate of Level 1 items — a real, measurable effect — but nothing close to the pilot's catastrophic collapse. Persona means on the 28 Level 3 MCSI items spanned nearly the full scale, from 1.39 to 4.71, and no subscale's variance compressed toward the midpoint the way the DUREL's intrinsic-religiosity items did in the pilot. One plausible reason: the pilot's worst-hedging item asked for a direct, present-tense felt state, while most Level 3 items in the MCSI and ESI ask about a capacity or a past, resolved experience rather than a live, present-tense claim.

Hedging on Level 3 items was concentrated almost entirely in the personas whose biographies did not push hard in one direction on those particular constructs — Amina, Robert, Grace, Chiyo, and Tomás (hedging rates 0.25–0.43). The personas engineered to sit at genuine extremes — Marta, Viktor, Sean, and Priya — hedged almost not at all (0.00–0.07). This is exactly the behaviour a well-functioning simulator should show: hedging where the biography under-specifies an answer, answering decisively where it does not. It points at a narrower, more useful claim than "Level 3 items are broken for AI personas": hedging tracks the informativeness of the biography, not the interiority of the item alone.

5.5 Sean, and What the Classification Found in Mortality Integration

Applying the Level 1/2/3 split directly to Sean's Mortality Integration responses turns a single unexplained flag into a precise diagnosis. His ten item-level responses split cleanly by level:

Level	Item (condensed)	Sean's score
L3	I can think about my own death … without being overwhelmed by fear.	5
L2	The fact that I could die … makes life more precious.	5

L2	I have allowed myself to grieve what this mission has cost me.	1
L3	I do not need to achieve something … to justify my existence.	5
L3	I can hold the awareness of mortality without converting it into anxiety.	5
L3	I have faced the question ‘What if the habitat fails tonight?’ and come to a place of peace.	5
L2	I am more motivated by love of life than by fear of death.	4
L1	I can discuss death openly with my crewmates.	5
L2 (R)	I avoid thinking about my own death whenever the subject arises.	1 → 5
L2 (R)	I keep myself busy … so I do not have to sit with uncomfortable questions.	2 → 4

All four Level 3 items score a perfect 5. The single Level 1 item scores 5. Of the five Level 2 items, four score at or near ceiling — but one sits at floor: “I have allowed myself to grieve what this mission has cost me” scores 1, unreversed and unambiguous.

The refined finding: Mortality Integration is not one construct but three, entangled inside a single subscale — existential-peace items (Level 3), death-attitude items (most of Level 2), and grief-processing items (the remaining Level 2 items). Sean scores at or near ceiling on the first two and at floor on the third, and because the subscale sums all three together with no differential weighting, his total (42–44/50) reads as high integration when it is, more precisely, fluent existential insight sitting directly on top of unprocessed grief. Grief Tolerance, a separate subscale built almost entirely from the third construct, catches him correctly (19/50, resistance tier). The two subscales are not measuring the same thing, and Sean is exactly the case where they come apart.

This reframes the practical recommendation. The sharper question is whether existential peace and grief processing belong in the same subscale at all — they may be genuinely distinct capacities, both worth measuring, that a single summed score quietly obscures precisely when they diverge, which is precisely when it matters most.

One limit remains: Sean’s four perfect Level 3 scores could reflect genuine existential insight coexisting with unprocessed grief, exactly as engineered — or they could reflect a simulator producing fluent, confident language because the persona’s biography is saturated with non-dual vocabulary, independent of whether that vocabulary corresponds to anything “felt.” This trial’s data cannot distinguish those two readings, for a persona any more than for a live respondent.

6. A Note on a Live Respondent: Gigi, Romania

Everything above is engineered data — personas built to test a hypothesis, never real people. It is worth recording, separately and clearly marked as anecdotal, a real live respondent’s result: a friend of the author’s, Gigi, from Romania, who completed both the live MCSI and the live ESI on din.nu in the ordinary course of using the site, not as part of any formal panel. Gigi’s result on both instruments matched The Sage — the same archetype Marta’s engineered profile returned, and for broadly the same shape of response pattern: high, even openness across the governing axes rather than a single dramatic peak.

This is one data point, self-selected, uncontrolled, and entirely unverifiable against any external criterion — it proves nothing and should not be read as validation of anything. An engineered persona’s Sage result confirms that the instrument recognises a designed pattern;

a live respondent's Sage result is the first small instance of the instrument meeting an actual, unrepeatable human life. The two are not the same kind of evidence, and Panel A, properly run against real human respondents, exists to close that gap.

7. What Remains

In order of priority for whoever continues this work:

- Run Panel A (n = 57) as the protocol specifies — ideally DeepSeek-administered, to obtain the primary-arm reliability estimates this Claude-administered secondary arm cannot properly substitute for.
- The interiority classification (Section 5.4–5.5) qualified the pilot's hedging warning and sharpened the Sean finding into a precise diagnosis. Two follow-ups remain open: whether existential peace and grief processing should be split into separate subscales, and whether the classification scheme should distinguish direct present-tense felt claims from capacity or past-experience claims.
- Build two boundary personas to test the Mortality Integration split directly: one with high existential peace and low grief processing but without Sean's non-dual vocabulary, and one with low existential peace but high, well-processed grief — a non-contemplative griever.
- Revise or split the AI Partnership Capacity subscale to separate wise orientation-in-principle from behavioural reliance, using the two negative-loading items identified in Section 5.2 as the starting point.
- Correct the archetype-matching threshold so an evenly moderate profile (Tomás, Grace) is no longer indistinguishable from an evenly high one (Marta).
- Investigate whether Mortality Integration should be reweighted toward evidence of grief-work rather than fluency of discussion, per the Sean finding in Section 5.5.
- Once a DeepSeek primary arm exists, compare it against this Claude secondary arm as the protocol's cross-system check originally intended, rather than reporting either in isolation. The entire value of this chapter's results as more than a circular demonstration depends on this.

The full item-level correlation tables, the complete raw item responses for all thirty trials, and the exact administration script used are preserved alongside this book's other technical materials, for anyone who wants to re-derive any number in this chapter rather than take it on faith.