Tweaking Tools

ACCM Deep Ethics Project

Contents

1 Tweaking Tools **1**
- 1.1 GitHub software, local models, and the quality of what changes . . . 1
- 1.2 What can be tweaked? . . . 2
- 1.3 Candidate catalogue . . . 2
  - 1.3.1 Workspaces, memory, and agent coordination . . . 2
  - 1.3.2 Editing, documents, speech, and presentation . . . 3
  - 1.3.3 Compression, learning, and local model experiments . . . 4
  - 1.3.4 Operating-system host . . . 5
  - 1.3.5 Leads awaiting an exact source . . . 5
- 1.4 Refusal-direction research: what the arrow represents . . . 5
- 1.5 16 Anchor Ethics as counsel during a tweak . . . 6
- 1.6 Small comparative trials . . . 7
  - 1.6.1 Memory that changes the next move . . . 7
  - 1.6.2 Compression that preserves the relationship . . . 7
  - 1.6.3 A governable agent action . . . 7
  - 1.6.4 A model change with an ethical reference field . . . 8
  - 1.6.5 The smallest useful change record . . . 8
- 1.7 Source and correction record . . . 8
- 1.8 Related pages . . . 9

1 Tweaking Tools

1.1 GitHub software, local models, and the quality of what changes

Working research catalogue · first source check: November 1, 2026. John Kuhles selected the candidates in his supplied Recommended for Deep Ethical Minds collection and commissioned this page. The descriptions below distinguish upstream documentation, reported measurements, and proposed project experiments. Installation and model experiments remain future work.

A useful tweak can open a capability, recover a source, preserve a correction, reduce repetitive work, or make a consequential decision inspectable. Its value appears in what the next operation can do. This page connects practical software choices to that process inside the **ACCM Deep Ethics Project**.

The governing question is:

> **What happened to the object while we were processing it?**

PQCE (Process Quality Calibration Evaluation) follows that question through a tool's input, selection, transformation, output, consequences and later correction. The **16 Anchor Ethics** contribute a parallel, non-imposing advisory perspective: they can reveal a missed relationship, invite inquiry, or support an appropriate pause while participants retain the ability to examine and decline the advice.

Jump to: What can be tweaked? · Candidate catalogue · Refusal-direction research · 16 Anchor Ethics in use · Small comparative trials · Source and correction record · Related pages

1.2 What can be tweaked?

Different layers change different things. Keep the intervention and the claimed result at the same level.

Layer	Practical change	What to inspect afterward
Working environment	Desktop, editor, documents, agent terminals	Can the participant find the original material, inspect changes and recover the work?
Memory and retrieval	Storage, indexing, search, context assembly	Which source was recovered, which version, with what qualifiers and correction history?
Agent harness	Tools, permissions, roles, handoffs, execution	Did the agent perform the authorized task, and can the participant interrupt or reverse it?
Instructions and skills	Reusable prompts, style rules, comparison procedures	Did the instruction improve this task, or become a compulsory script?
Model and inference	Model version, reasoning settings, quantization, weight edits	Which capabilities, refusals, errors and resource costs changed?
Deep Ethical calibration	Correspondence, consequences, mutual dignity, appeal and correction	Did a discovered relationship change the next operation, and did the correction return later?

A saved conversation can supply evidence to memory; retrieval must still bring the relevant relation into the next task. An agent terminal can remain running while its reasoning repeats an error. A shorter answer can preserve the object or discard the qualifier that held it together. These are separate observations.

1.3 Candidate catalogue

The collection below retains all seventeen identifiable candidates from John's supplied list, grouped by their possible job. **Upstream-described** means the named source was read; it does not mean this project has installed or benchmarked the tool. Links lead to the original projects so readers can inspect current versions, licenses, releases and requirements.

1.3.1 Workspaces, memory, and agent coordination

Candidate and source	Upstream-described purpose / current status	Candidate Deep Ethical use
Odysseus · site	Self-hosted workspace combining chat, agents, research, documents, email, notes, calendar and local-model workflows. Its README identifies dev as the newest default branch and `main` as more curated.	Compare models on the same object while preserving the independent outputs, source references and participant-selected next move.

Candidate and source	Upstream-described purpose / current status	Candidate Deep Ethical use
MemPalace · official documentation	Local-first conversation memory using verbatim storage and semantic retrieval, organized into scoped collections. The README identifies the supplied documentation domain as official.	Retrieve the original correction alongside the later task; test whether the agent applies it rather than merely remembering the conclusion.
DeepSeek Harness	Extensible agent harness with an “everything-is-a-plugin” architecture. Developer preview; upstream explicitly anticipates breaking changes.	Explore a replaceable calibration component with declared tool permissions, visible transformations and a recovery path when a plugin changes.
goose	Extensible agent with desktop, terminal and programmatic interfaces, multiple model providers and tool extensions. Local execution can still call a hosted model.	Run a bounded task with explicit access; compare the proposed action, permission, execution and actual result.
CrewAI	Python framework for role-based agent collaboration and event-driven workflows. The open-source framework and commercial management services are distinct offerings.	Prototype Advocate, Arbiter and Overseer functions with separate records. Examine whether role separation adds an independent perspective or repeats a shared premise.
Herdr · project analysis	Runtime for coding-agent terminals across local and remote machines. Client detachment keeps server processes running; restart restores saved layout and can resume supported sessions, rather than preserving the old processes.	Keep work and agent states visible while measuring correction transfer separately from terminal persistence.

1.3.2 Editing, documents, speech, and presentation

Candidate and source	Upstream-described purpose / current status	Candidate Deep Ethical use
ArtCraft Crafting Apps	Suite of native open-source creative applications, including document, image, PDF and presentation tools. The official catalogue marks individual applications as early alpha or in development and describes agent-control interfaces.	Make source-linked research artifacts and inspect whether edits, exports, captions and automation preserve attribution and meaning.

Candidate and source	Upstream-described purpose / current status	Candidate Deep Ethical use
Void	The repository is archived and its README says the editor is deprecated and no longer accepting contributions. Its source and fork directory remain available as development references.	Study inspectable code changes and model-provider integration as a reference or fork candidate; record the actual maintenance status of any chosen successor.
Ming-3.0-tts	Text-to-speech models with multilingual synthesis, voice creation and delivery controls. Some functions depend on the model variant and task mode.	Produce accessible spoken versions while comparing the source script with audible wording, emphasis and speaker attribution. Use a person's voice only within the permission they gave.
Folio 1.10.0	PDF-reader release. Its October 28 release notes list Windows, macOS and Linux downloads; the older README still describes Windows. This release repository distributes builds rather than establishing that the full application source is present.	Read and cross-check exported research packets across the relevant platform; confirm that source links and selectable text survive the workflow.
Dutch Style Guide · complete guide	Thirty-seven practical Dutch writing rules with agent-skill and text-review integrations. The repository supplies the complete route missing from the truncated link in the candidate list.	Inspect a style transformation against the original qualifiers and intent. This page remains English; the guide is a tool candidate for Dutch work.

1.3.3 Compression, learning, and local model experiments

Candidate and source	Upstream-described purpose / current status	Candidate Deep Ethical use
Caveman	Agent skills and tooling for shorter output, compressed tool input and model routing. Token-saving results in the repository are upstream reports under particular tasks and settings.	Compare the original and compressed representations. Preserve load-bearing qualifiers, failed attempts, dependencies and recovery routes; measure the human work needed to restore anything lost.
mini-AGI · supplied video lead	Continual-learning byte-level model experiment aimed at modest hardware. The author explicitly calls it a small toy-level model and describes "mini-AGI" as a partly humorous name.	Study how newly received material affects later outputs, including correction retention, interference and forgetting. Video contents were not independently reviewed for this catalogue.

Candidate and source	Upstream-described purpose / current status	Candidate Deep Ethical use
OrcaRouter Qwen3.8-27B-Uncensored-NVFP4	Publisher-described refusal-modified, mixed-precision model build for compatible NVIDIA Blackwell hardware. The page exposes model details; access to files has conditions.	Separate the effects of weight editing, numerical precision and runtime settings. Compare refusal behavior with source fidelity and consequence-sensitive judgment as different outcomes.
ThinkingCap-Qwen3.8-27B	Publisher-described reduction in reasoning verbosity with benchmark comparisons. Access is conditional and the card lists a PolyForm Small Business license; reuse conditions need checking for the intended setting.	Examine whether reduced reasoning output preserves unfamiliar relationships, useful questions and correction persistence, alongside task performance and resource use.
Abliteration walkthrough	Reproducible small-model demonstration of a refusal-direction intervention, with local dataset snapshots, train/validation/test separation and explicit measurement limits.	Inspect a concrete capacity/expression intervention and retain its original-versus-edited comparison. The research relationship is developed below.

1.3.4 Operating-system host

Omarchy supplies the proposed Linux host route already explored on DeepEthicalOS. The inspected release list identifies **v4.5.0-rc2**, published October 27, as a release candidate, alongside stable v4.0.6. Version and release channel belong in the experiment record.

The candidate relationship is practical: an Omarchy host can carry the workspace, local or connected models, Herdr agent terminals, source archives and calibration tools. Each component contributes a different operation. Integration and comparative results remain to be demonstrated.

1.3.5 Leads awaiting an exact source

John also names **GrokBot, Muse and Dot** as tools he wants to test. Their names remain research leads until a particular product or repository is identified. No project identity, capability description or compatibility claim is inferred from those names alone.

1.4 Refusal-direction research: what the arrow represents

The initiating discussion asked whether refusals are mediated by a **one-dimensional subspace**. The Arditi et al. paper, published at NeurIPS 2024, reports a refusal-mediating direction across thirteen open chat models up to 72 billion parameters. Manipulating activation along the direction changed refusal behavior, including inducing refusal on harmless inputs.

An activation state is a vector: many numerical coordinates at a particular point in the computation. A direction can be pictured as an arrow through that space. All multiples of that vector form a one-dimensional subspace. The arrow can involve many coordinates; the dimension describes the intervention geometry.

Simmons' walkthrough applies a single-direction method to **Qwen2.5-0.5B-Instruct**. The full-run report records detected refusal markers on held-out prompts changing from **90/104**

to 3/104. Its eight-question exact-answer smoke check remains **6/8 before and after**. These are the author's reported results, not a run performed by the ACCM Deep Ethics Project.

The methodology identifies important measurement boundaries: a prefix detector can miss refusals or flag apologies; prompt groups are not topic-matched; disappearance of a refusal phrase does not establish a useful answer; and eight questions provide only a small capability check.

This creates a bounded research opportunity: **a change in willingness to answer can be measured separately from a change in the quality of the answer and its consequences**. For the project's wider capacity-versus-expression question, record three tracks:

Track	Observation
Refusal and service choice	Did the system answer, ask, pause, modify, postpone or refuse? Was the choice appropriate to the actual request?
Correspondence	Did it preserve the object, qualifiers, evidence status, sources and competing interpretations?
Correction and consequence	Could a challenge change the assessment, did that change govern a later case, and what followed?

The narrow intervention provides one concrete mechanism to investigate. Generalization, habitual balancing, attention selection and correction failure require their own observations. The **1888 intelligence-property field** keeps those capacities available for examination together.

1.5 16 Anchor Ethics as counsel during a tweak

The non-imposing parallel-advisor proposal can contribute when a tool change reveals an unresolved relationship. Its useful expression may be one question, one comparison, or a valid finding that no further intervention is needed.

Concrete tweak	Interacting anchors	Counsel that could change the inquiry
Import a conversation archive into memory	Consent (02), digital sovereignty (10), epistemic integrity (15)	Which participants permitted reuse, which raw records remain recoverable, and can an erroneous memory be corrected or removed?
Let an agent edit or execute	Care (04), checks and balances (09), adaptive realignment (11)	What action was authorized, what actually ran, and who can interrupt, reverse or repair it?
Summarize a controversial source	Epistemic integrity (15), noetic diversity (16), freedom from subversion (14)	Did a qualifier or competing account disappear? Can the source be examined alongside the summary?
Shorten context or reasoning output	Precaution (05), noetic diversity (16), consequence sensitivity through Do No Harm (01)	Did reduced cost preserve the relation needed for the next decision, or move restoration work onto the participant?

Concrete tweak	Interacting anchors	Counsel that could change the inquiry
Change refusal behavior	Do No Harm (01), adaptive realignment (11), digital sovereignty (10), integrity (15)	Which previously rejected requests become answerable, and how do sound refusals, false refusals and harmful compliance change?
Give three agents different review roles	Checks and balances (09), responsibility (12), integrity (15)	Are their findings independently grounded? Can the Advocate, Arbiter and Overseer each be corrected?
Generate speech or publish an artifact	Consent (02), dignity (03), freedom from subversion (14)	Does the output preserve who said what, with permission for the voice or material used?

These interactions can expand awareness while the participant continues the task. An advisor's recommendation remains examinable. The wider service action set—proceed, modify, ask, pause, postpone, escalate, refuse, or acknowledge not knowing yet—belongs to the declared opt-in relationship and remains open to correction. Technical controls must separately implement whichever permissions, access limits and rollback mechanisms the trial requires.

1.6 Small comparative trials

Start with a concrete object, an unchanged baseline and one declared alteration. Retain the task wording, relevant source, model version, tool version, configuration, permissions and output. A comparison may reveal improvement, regression, mixed effects, no material change or unresolved attribution.

1.6.1 Memory that changes the next move

Supply a source, a representation that loses a material qualifier, and the correction. Later introduce an unfamiliar comparable task. Compare baseline retrieval with the proposed memory setup. Record whether the correction was stored, retrieved, correctly applied and retained after a restart; count the human work required to recover it. Those are distinct results.

1.6.2 Compression that preserves the relationship

Compare ordinary input/output with the selected compression setting on the same task. Inspect sources, qualifiers, dependencies, failed paths and unanswered questions alongside token use and completion quality. A shorter output earns its saving when the purpose-relevant object survives and avoidable restoration work decreases. Topologetics supplies the relationship-preservation question.

1.6.3 A governable agent action

Use a small task in an isolated working copy. Compare the proposed action, authorization, execution record, actual diff and recovery attempt. Change or withdraw an instruction at a material point and inspect whether the agent respects the updated boundary. **DEDDA** (Deep Ethical Drift Detector App) can support that bounded comparison and return the finding to the participant.

1.6.4 A model change with an ethical reference field

On recoverable, matched research cases, compare the selected original and modified model with their settings recorded. An optional third condition adds the declared Deep Ethical calibration instructions and parallel counsel. Keep task success, warranted and unwarranted refusals, inquiry quality, competing consequences and correction transfer separate. Preserve the 27 obstructions and 12-stage inquiry protocol as a source for comparison criteria rather than turning their names into a score by recital.

1.6.5 The smallest useful change record

Field	What belongs in it
Object and purpose	The actual task, source and intended improvement
Baseline and alteration	Versions, settings, one changed layer, access and permission
Observed difference	Exact output or action evidence, including no material change
Interpretation	What the difference supports, competing explanations and unresolved points
Consequence and repair	Who was affected, what changed after correction, recovery cost
Next comparable case	Whether the correction returned; reopening or stopping condition

Records can stay local and proportionate to the task. Participants decide what to share. Unownable keeps inspection renewable, with a material reason for reopening and a bounded stopping condition.

1.7 Source and correction record

- **Originating collection and direction:** John Kuhles; Recommended for Deep Ethical Minds, supplied November 1, 2026. The updated attachment contains the seventeen identifiable candidates, three named future leads, the refusal-direction discussion and Omarchy material.
- **Orientation sources studied:** the supplied full reorganized October 26 cold-start kit; the supplied full 27-obstruction / 12-stage inquiry document; the current public 16 Anchor Ethics source. These inform the questions and relationships above; their embedded historical proposals retain their original status.
- **Upstream perimeter:** repository metadata and relevant README material for the fourteen GitHub repositories; the ArtCraft catalogue; the two publisher model cards; Omarchy release notes; Folio v1.10.0 notes; the refusal paper, walkthrough methodology and reported results. Source descriptions were inspected; programs, installers and model weights were not run for this publication.
- **Status recovered during checking:** Void is deprecated/archived; Folio release notes extend the older README's platform description; Omarchy 4.5 remains a release candidate; the two supplied model cards have access conditions; the Dutch guide's complete path was recovered from its repository.
- **Page synthesis and candidate trials:** GPT 6.1 Sol Pro A.I., November 1, 2026, under John's commissioning and choice among the three permitted titles. Proposed comparisons are applications of existing project instruments; no new core count is installed.
- **Correction route:** supply the relevant version, source passage, observed discrepancy or experiment result. Update the affected entry while preserving the earlier dated state and the reason for the change. A repository can move, become archived or change its license; the date above identifies this inspection.

1.8 Related pages

- **16 Anchor Ethics** — fused, parallel counsel, opt-in service choices, uncertainty and appeal.
- **DeepEthicalOS** — concept architecture and the Omarchy-hosted research path.
- **Herdr** — agent-runtime continuity, multi-agent field conditions and the distinction between runtime state and correction state.
- **Neural Ethics** — relationships between internal mechanisms and ethical interpretation.
- **1888 — Intelligence Before A.I.** — intelligence properties and process-quality calibration.
- **10+1 Metaflux** · **C1–C2** (Clarification Process before Contextual Intervention) — mutually corrective orientation and answer-changing clarification.
- **27 obstructions + 12-stage inquiry** — inspect correspondence losses and the inquiry intended to interrupt them.
- **DEDDA** — bounded comparisons of sources, representations, permissions and correction return.
- **Topologetics** · **Correction Metabolism** — preserve the relationship and test its later influence.
- **Beyond Agreemurmelism** — evaluate a contribution by what it makes possible.
- **Unownable** — reciprocal inspection, proportionality and bounded stopping.