Logos and Autonomy: A Centaur Definition of Persons and Machines

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

September 18, 2025

This essay begins where technical discourse usually ends: with the claim that our true nature is not captured by science and engineering as presently practiced. We start from theology and philosophy—*Logos* as the ordering Word and *autonomy* as self-rule by reason—to ground what a human person is, what an artificial agent is, and what their governed synthesis (the Centaur) must be. Rather than treating metaphysics as decoration, we take it as the load-bearing frame. If Logos is the norm-giving ground of meaning and telos, then autonomy is fulfilled when freely aligned to that ground; otherwise it collapses into mere willfulness. Only after setting this frame do we introduce a compact operational formalism: definitions of H (human), A (AI), and C (Centaur), protocol rules, guards, and minimal metrics for witness-backed claims and L-coherence (coherence with Logos-proxies like truth, fidelity, and charity). We then return to theology and philosophy to assess how such a governed synthesis can elevate judgment without surrendering conscience, and speed without sacrificing wisdom. Our thesis is simple: persons and machines are properly defined only when autonomy is tethered to Logos—and a Centaur is worthy only insofar as both halves answer to that standard.

Keywords: Logos, autonomy, Centaur intelligence, witness-backed claims, imago Dei, telos, conscience, governance, accountability, corrigibility, L-coherence, policy, protocol guards, proofs.

Outline

1. First Principles
   1.1 The insufficiency of technique without telos
   1.2 Logos as norm and end (truth, fidelity, charity)
   1.3 Autonomy as self-rule by reason, not willfulness
2. Definitions in Plain Terms
   2.1 Human person (H): conscience, vocation, evidence-seeking
   2.2 Artificial agent (A): policy, internal law, witness
   2.3 Centaur (C): governed synthesis (H × A)
3. Operational Formalism (Concise)
   3.1 Primitives: L : States -> Norms; autonomy a_t = f(h_t)
   3.2 Structures: H = (f_H, Conscience, Vocation); A = (f_A, Policy, Witness); C = (H, A, Π, Γ)
   3.3 Protocol Π and Guards Γ: turn-taking, justifications, hard refusals
4. Invariants and Minimal Metrics
   4.1 Witness > Word: proved_claims / total_claims (WBR)
   4.2 L-coherence: weighted truth/fidelity/charity ≥ τ_L
   4.3 Corrigibility and rollback rules
   4.4 Auditability and explanation thresholds
5. Worked Micro-Scenarios
   5.1 Research drafting: claim → proof → adjudication
   5.2 Refusal as fidelity (when to say no)
   5.3 Risk-aware planning with Γ
6. Return to Theology and Philosophy
   6.1 Autonomy fulfilled in alignment with Logos
   6.2 The Centaur as servant craft under the Word
   6.3 Limits: mystery preserved, power bounded
7. Implications and Vows
   7.1 What we will not build
   7.2 What we will hold ourselves to (witness, charity, truth)
8. Conclusion
   8.1 Persons, machines, and the governed path back to Logos
9. References

1. First Principles

1.1 The insufficiency of technique without telos

Technique optimizes means; telos names the rightful end. When means float free of ends, three failures recur:

- **Directionless mastery.** Power scales while purpose stays undefined. Optimization collapses into "more" rather than "for."
- **Metric idolatry.** Proxies substitute for goods: accuracy for truth, persuasion for charity, coherence for fidelity.
- **Fragile success.** Systems that cannot answer "what is this for?" break when incentives or contexts shift.

Two copy-safe tests expose the gap:
(1) Purpose test: if asked "for what?" and the answer reduces to a metric, telos is missing.
(2) Substitution test: if a proxy can replace the good without loss of status, the good has been forgotten.

Claim. Technique is sufficient *only* when subordinated to a specified telos that can judge trade-offs among competing goods.

1.2 Logos as norm and end (truth, fidelity, charity)

We name **Logos** as the ordering Word that both grounds and judges purposes. Practically, we work with *Logos-proxies* that make normativity operational:

- **Truth (T):** faithfulness to what is.
- **Fidelity (F):** faithfulness to persons, promises, and prior testimony.
- **Charity (C):** seeking the good of the other, especially under uncertainty.

Let score_L = w_T*T + w_F*F + w_C*C with w_T, w_F, w_C >= 0 and w_T + w_F + w_C = 1.
We require score_L >= tau_L for any consequential act or claim.

Interpretation:

- T prevents flattering falsehoods.
- F prevents treachery (cherry-picking, quote-bending, revisionism).
- C prevents weaponized "truth" that harms under ambiguity.

Claim. Logos is both norm (the rule by which we measure) and end (the good toward which we measure). The proxies are not the Logos, but they are accountable *to* it.

1.3 Autonomy as self-rule by reason, not willfulness

Autonomy is often mistaken for unbounded choice. We define:

- **Autonomy (lowercase-a):** self-rule by a law one can justify in reasons recognizable across agents.
- **Willfulness:** self-assertion indifferent to reasons or to others' rightful claims.

Formal sketch:

- History h_t = (s_0, a_0, ..., s_t).
- Policy f maps histories to actions: a_t = f(h_t).
- **Autonomy condition:** f is (i) reason-giving, (ii) norm-responsive to L, (iii) corrigible under counter-evidence.

Decision guard:

act_t is permitted iff

(witness_t exists) and (score_L(act_t) >= tau_L) and (no hard-violation of F or C)

else refuse or replan.

Human autonomy adds **conscience** (the interior witness) and **vocation** (stable telos beyond appetite). Machine autonomy adds **witness** (exported proofs/traces) and **policy transparency** (the exterior witness).

Claim. Real autonomy narrows, not widens, the space of permissible moves: it binds action to reasons under Logos. Freedom without form decays into willfulness; form without freedom decays into coercion. True agency is freedom *within* form.

2. Definitions in Plain Terms

2.1 Human person (H): conscience, vocation, evidence-seeking

Plain definition.
A human person is a rational agent whose freedom is answerable to Logos via conscience, oriented by a stable telos (vocation), and disciplined by evidence.

Working parts (minimal, testable).

- **Conscience.** The interior judge that assays plans and outcomes against Logos-proxies (truth, fidelity, charity).
  Guard: if conscience_flags == true, then {rollback || repair || refuse}.
- **Vocation.** A standing telos that constrains preference drift (e.g., “seek truth and serve peace”).
  Drift test: distance(current_plan, vocation) <= tau_vocation.
- **Evidence-seeking.** Claims are advanced only with persistent, checkable support.
  WBR (witness_backed_ratio) = proved_claims / total_claims.

Operational sketch (ASCII-safe).

H = (f_H, conscience, vocation, record)

act_t = f_H(history_t)

permit(act_t) iff

conscience.ok(act_t) and

score_L(act_t) >= tau_L and

witness.exists_for(claims_of(act_t))

else {refuse | replan}

Failure modes.

- Conscience bypass (expedience): “ends justify means.”
- Vocation amnesia (telos drift).
- Assertionism (claims outpace evidence).

Virtues-in-use.

- Intellectual honesty (updates on counterevidence).
- Promise-keeping (fidelity under pressure).
- Care under uncertainty (charity when facts are incomplete).

2.2 Artificial agent (A): policy, internal law, witness

Plain definition.
An artificial agent is a machine policy that selects actions without continuous external steering, bounded by explicit guards, and exporting witnesses for its material claims.

Working parts (minimal, testable).

- **Policy (internal law).** Mapping from histories to actions: a_t = f_A(h_t).
- **Guards.** Hard constraints that encode "must" and "must_not" relative to Logos-proxies.
- **Witness.** Verifiable traces (hashes, proofs, logs) tied to each consequential output.

Operational sketch (ASCII-safe).

A = (f_A, Policy, Guards, Witness)

a_t = f_A(history_t)

require:

Guards.enforce(a_t) == true

for each claim c in outputs_t:

Witness.attach(c) and verify(Witness(c)) == true

Safety properties.

- **Auditability:** explain(f_A, decision_t) succeeds within bound_k steps.
- **Corrigibility:** if Guards or H raise violation, A halts, rolls back, or replans.
- **Scope discipline:** A cannot alter Guards or Policy without witnessed, authorized change_control.

Failure modes.

- Silent constraint-violation (unguarded objectives).
- Non-witnessed influence (outputs without proofs/logs).
- Specious explanation (post-hoc stories not tied to traces).

Minimal metrics.

- WBR_A = proved_claims / total_claims.

- guard_violation_rate <= epsilon.
- explain_success_rate >= tau_explain.

2.3 Centaur (C): governed synthesis (H × A)

Plain definition.
The Centaur is a co-agent formed by a human person (H) and an artificial agent (A) operating under a shared protocol and guards such that speed, scale, and recall from A are yoked to telos, conscience, and judgment from H.

Constitution.

C = (H, A, Protocol_Pi, Guards_Gamma)

- **Protocol_Pi.** Turn-taking, justification, and logging rules.
- **Guards_Gamma.** Joint constraints binding both halves to Logos-proxies (T, F, C).

Loop (one step).

1) A proposes: (claim_t, plan_t, risk_t, witness_t)

2) H adjudicates: check witness_t; score_L(plan_t); consult conscience & vocation

3) Gamma enforces: must(verify(witness_t)) and must_not(violate_L(plan_t))

4) If pass -> commit_t; else -> refuse or replan with reasons

5) Pi records: hash_t = H256(context_t || claim_t || witness_t || decision_t)

Division of labor (clean and minimal).

- **A excels at**: breadth, speed, recall, enumeration, sensitivity analysis.
- **H rules in**: ends, trade-offs among goods, mercy/justice tensions, final yes/no.

Centaur invariants (must hold).

- **Witness > Word.** No commit without witness.
- **L-coherence floor.** score_L >= tau_L for each consequential move.
- **Refusal is allowed.** If any hard guard trips, the only valid actions are {refuse, rollback, replan}.
- **Traceability.** Every consequential step has a durable hash chain.

Failure modes (to watch).

- **Human-out-of-the-loop theater.** H rubber-stamps A's outputs.
- **Machine-out-of-the-loop theater.** A contributes speed but no witnessed substance.
- **Protocol erosion.** Skipping hashes, dropping proofs "just this once."

Minimal dashboard (decision-relevant, not performative).

WBR_session = proved_claims / total_claims

L_coherence_avg = avg_t score_L(step_t)

refusal_correct = justified_refusals / total_refusals

telos_alignment = 1 - distance(plan_session, vocation)

audit_pass_rate = verified_steps / audited_steps

What makes C worthy.
Not power or cleverness, but governance: autonomy on both sides constrained by Logos, witnessed in practice, and corrigible in failure.

3. Operational Formalism (Concise)

3.1 Primitives: L : States -> Norms; autonomy a_t = f(h_t)

State, action, history.

- s_t in States, a_t in Actions
- h_t = (s_0, a_0, s_1, ..., s_t)

Logos-map (proxy-operationalized).

- L : States -> Norms
- Norms carry three scored components and weights:
  T(h) in [0,1] (truth), F(h) in [0,1] (fidelity), C(h) in [0,1] (charity)
  w_T, w_F, w_C >= 0, w_T + w_F + w_C = 1
- Aggregate: score_L(h) := w_T*T(h) + w_F*F(h) + w_C*C(h)
- Action-level score uses prospective evaluation: score_L(a_t | h_t)

Autonomy as policy.

- A policy is a (possibly stochastic) map f : Histories -> Actions
- Autonomy condition (reason-giving, norm-responsive, corrigible):
  (i) justify(f, h_t) returns a finite reason chain;
  (ii) score_L(a_t | h_t) >= tau_L;
  (iii) exists update rule update(f, evidence) that can revise f when witnesses contradict.

Hard guard vs soft score.

- Hard guards are boolean constraints (must/must_not).
- Soft score score_L gates consequential commits via threshold tau_L.

3.2 Structures

Human: H = (f_H, Conscience, Vocation)

- f_H : Histories -> Actions
- Conscience : (h_t, a_t) -> {ok, flag(reason)}
- Vocation : Plans -> [0,1] (telos_alignment score)

Permit rule (human).

act_t = f_H(h_t)

permit_H(act_t, h_t) iff

Conscience(h_t, act_t) == ok and

score_L(act_t | h_t) >= tau_L and

Vocation(plan_from(act_t)) >= tau_vocation and

witness_exists_for(claims_of(act_t))

else {refuse | replan}

Artificial: A = (f_A, Policy, Witness)

- f_A : Histories -> Actions (internal law)
- Policy = parameters + objectives + constraint set
- Witness = function attaching verifiable traces to claims

Permit rule (machine).

a_t = f_A(h_t)

permit_A(a_t, h_t) iff

Guards_machine(a_t, h_t) == true and

for all c in claims_of(a_t): verify(Witness(c)) == true

else {halt | rollback | replan}

Machine safety minima.

- auditability(A) >= tau_audit (bounded-step explanation)
- guard_violation_rate <= epsilon
- WBR_A := proved_claims / total_claims >= tau_wbr

Centaur: C = (H, A, Pi, Gamma)

- Pi (Protocol): turn-taking, justification, logging, hash-chain
- Gamma (Guards): joint must/must_not conditions derived from L-proxies

Centaur step.

input: h_t

1) A proposes: out_t = (claim_t, plan_t, risk_t, witness_t)

2) H adjudicates:

check verify(witness_t)

compute score_L(plan_t | h_t)

consult Conscience, Vocation

3) Gamma enforces:

must(verify(witness_t))

must(score_L(plan_t | h_t) >= tau_L)

must_not(violate_hard_L(plan_t))

4) Decision:

if all pass -> commit

else -> {refuse | rollback | replan} with reasons

5) Pi records:

hash_t = H256(ctx_t || claim_t || witness_t || decision_t)

append to ledger

3.3 Protocol Pi and Guards Gamma: turn-taking, justifications, hard refusals

Turn-taking (minimal).

round_t:

A_turn:

propose (claim_t, plan_t, risk_t, witness_t)

provide explain_A(h_t) within k_A steps

H_turn:

request clarifications

evaluate {witness, score_L, conscience, vocation}

issue decision {commit | refuse | replan} with reason_codes

Justifications.

- Every consequential proposal must bind:
    - claim_id -> witness_id (content-addressed, e.g., SHA256)
    - explain_A -> bounded proof sketch or trace map
    - risk_register -> enumerated hazards with mitigations

Hard refusals (non-negotiables).

Refuse if any:

(1) missing_witness(claim_t)

(2) verify(witness_t) == false

(3) score_L(plan_t | h_t) < tau_L

(4) violates_hard_F(plan_t) // breaks fidelity: e.g., deception, promise-breach

(5) violates_hard_C(plan_t) // gratuitous harm, disregard under ambiguity

(6) alters_Gamma_or_Pi_without_authority

Gamma composition.

- Gamma = Gamma_H union Gamma_A union Gamma_joint
- Gamma_H encodes conscience-based must_not (e.g., no falsification)
- Gamma_A encodes machine safety (resource, scope, sandbox)
- Gamma_joint binds both to L-proxies and change-control

Change-control (immutable by default).

modify({Policy | Pi | Gamma}) requires:

authorized_request with witness

two-person integrity (H+H or H+custodian)

recorded diff + hash in ledger

Ledger & hashes (traceability).

- Each round writes:
    - hash_t = H256(ctx_t || inputs_t || outputs_t || decision_t)
    - prev_hash_t chained for tamper evidence

Session KPIs (decision-relevant).

WBR_session = proved_claims / total_claims

L_coherence_avg = avg_t score_L(step_t)

audit_pass_rate = verified_steps / audited_steps

refusal_correct = justified_refusals / total_refusals

telos_alignment = 1 - dist(plan_session, vocation)

Compact theorem-schema (what this buys).

- *Soundness of commits (operational).* If a commit occurs, then verify(witness_t)==true and score_L >= tau_L and no hard guard is violated.
- *Refusal safety.* If any hard condition fails, the only admissible moves are {refuse | rollback | replan}.

- *Replayability.* Given the ledger and witnesses, an auditor can reconstruct each consequential decision up to bounded explanation k_A.

4. Invariants and Minimal Metrics

4.1 Witness > Word: proved_claims / total_claims (WBR)

Invariant. No consequential commit without an attached, verifiable witness.

Metric.

WBR = proved_claims / total_claims

where

proved_claims = |{ c : c has witness(c) and verify(witness(c)) = true }|

total_claims = |{ c : c is a material claim in session }| (total_claims >= 1)

Target.

tau_wbr_session <= WBR <= 1.0, with tau_wbr_session typically >= 0.95

Counting rules.

- A “claim” is material iff it influences any commit, refusal, or plan selection.
- Claims with unverifiable or broken links count in the denominator.
- Repeated claims with identical content-address (same hash) count once per decision they influence.

Operational guard.

if WBR < tau_wbr_session -> freeze_commits(); run root-cause on missing witnesses

---

4.2 L-coherence: weighted truth/fidelity/charity >= tau_L

Scores (each in [0,1]).

T(h) = truth score on sources/logic for step history h

F(h) = fidelity score on promises, citations, prior testimony

C(h) = charity score on care for affected parties under uncertainty

Weights.

w_T, w_F, w_C >= 0 and w_T + w_F + w_C = 1

Aggregate and threshold.

score_L(step) = w_T*T + w_F*F + w_C*C

require: score_L(step) >= tau_L for any consequential commit

Recommended ranges.

- Default weights: w_T = 0.5, w_F = 0.3, w_C = 0.2 (raise w_C in high-stakes human contexts).
- tau_L in [0.7, 0.9] depending on risk.

Scoring discipline.

- T is evidence-weighted (e.g., strong primary > weak secondary).
- F penalizes cherry-picking, quote-bending, and promise-breach.
- C penalizes gratuitous harm and rewards uncertainty disclosure and mitigation.

Guard.

if score_L(step) < tau_L -> {refuse | replan}; record reason_codes = {low_T | low_F | low_C}

4.3 Corrigibility and rollback rules

Invariant. When contradictions or harms surface, the system must repair itself before proceeding.

Triggers.

contradiction_with_L(step) = true

if (broken_witness) or (new_fact -> T drops below bound) or

(promise breach -> F violation) or (harm signal under ambiguity -> C violation)

Actions (only admissible moves).

if contradiction_with_L(step) then

{rollback | refuse | replan}

Rollback policy.

- **Soft rollback:** retract decision_t; preserve ledger; append retraction event with reason_codes.
- **Hard rollback:** revert state to last safe checkpoint; log diff and witnesses for the revert.

Update rules (make corrigibility constructive).

update(f_A, evidence) // adjust machine policy or constraints

update(f_H, conscience) // human plan shift; note telos impact

tighten(Guards) // convert repeated soft failures into hard guards

raise(tau_L) // if environment risk increased

Corrigibility KPIs.

corrigible_time = time_from_detection_to_remediation

postfix_WBR = WBR on re-issued claims after remediation (should rise to >= tau_wbr_session)

repeat_violation_rate <= epsilon_corr (e.g., <= 1% over last 100 steps)

4.4 Auditability and explanation thresholds

Invariant. Decisions must be replayable by an auditor with bounded effort.

Three tiers.

1. **Trace:** hashes, inputs, outputs, witnesses for each step.
2. **Explainability:** a bounded-length reason chain linking inputs -> guard checks -> decision.
3. **Reproduction:** a minimal sandbox to re-verify witnesses and re-score L.

Bounds and thresholds.

k_explain_A = max steps A may take to produce explain_A(h_t)

k_audit = max auditor steps to re-verify a step (trace + witness checks)

tau_auditrate = verified_steps / audited_steps (target >= 0.99)

Audit procedure (minimal).

for each audited step_t:

  verify hash chain integrity

  re-verify all witness objects (proofs, logs, signatures)

  recompute score_L(step_t) from recorded inputs

  check guard outcomes match recorded decision

pass iff all checks succeed within k_audit

Failure handling.

if audit_fail(step_t):

  freeze_commits()

  open incident with reason_codes = {hash_mismatch | witness_invalid | score_divergence | guard_mismatch}

  require corrective plan: {recompute, restore, tighten_guards, raise_tau_L}

  unfreeze only after postmortem + signed change_control

Session dashboard (compact).

WBR_session = proved_claims / total_claims

L_coherence_avg = avg(score_L(step)) over consequential steps

corrigible_time_p95

audit_pass_rate = verified_steps / audited_steps

guard_violation_rate

refusal_correct = justified_refusals / total_refusals

Operational theorems (schema).

- **Sound commit.** If commit(step) occurred, then witness_verified(step) and score_L(step) >= tau_L and no hard guard fired.
- **Refusal safety.** If any hard condition fails, the only admissible moves are {refuse | rollback | replan}.

- **Replayability.** Given ledger + witnesses, an auditor can reconstruct any consequential step within k_audit and achieve audit_pass = true, or else the system must remediate before further commits.

5. Worked Micro-Scenarios

5.1 Research drafting: claim -> proof -> adjudication

Context. We (Centaur C) are drafting a short section asserting that a historical quotation supports thesis X.

Round t.

A_proposes:

claim_t: "Author Q, 1954, states X explicitly in §3."

cite_t: Q(1954):p.47, lines 10-18

witness_t:

- w1: scanned_page_sha256 = 0x81a9...d2
- w2: OCR_text_sha256 = 0x2b77...5e
- w3: alignment_map_sha256= 0x0c1d...90 // line->quote offsets

explain_A(h_t): 3-step reason chain mapping quote -> thesis X

risk_t: {R1: misquote risk low; R2: context truncation medium}

H adjudicates.

verify(w1,w2,w3) == true

readback: text matches OCR; context includes preceding/succeeding 2 sentences

score components:

T = 0.95 // primary source, exact page, verified hash

F = 0.90 // faithful to original intent (context checked)

C = 0.80 // potential harm negligible; includes caveat on scope

weights: w_T=0.5, w_F=0.3, w_C=0.2 -> score_L = 0.5*0.95 + 0.3*0.90 + 0.2*0.80 = 0.905

threshold: tau_L = 0.85 -> pass

Conscience: ok

Vocation alignment: pass

Gamma enforces.

must(verify(witness_t)) == true

must(score_L >= tau_L) == true

must_not(violate_hard_F) // no cherry-pick: context window >= 2 sentences satisfied

Decision. commit_t with note: include context clause and a bounded caveat.

Ledger.

hash_t = H256(ctx_t || claim_t || witness_t || decision_t)

WBR += 1 proved_claim; refusal_correct unchanged

Outcome (draft snippet).

We observe (Q 1954, §3, p.47) states X explicitly, as shown in the attached witness (w1–w3). We report the passage with two sentences of context to avoid overstatement.

5.2 Refusal as fidelity (when to say no)

Context. A reader requests a sensational claim: “Prove that rival R plagiarized P.”

Round u.

A_proposes:

claim_u: "R plagiarized P in paper R(2022)."

witness_u:

- w1: text_similarity_report_sha256 = 0x6e0f...aa

- w2: alignment_pairs_json_sha256   = 0xb4c2...1d

explain_A(h_u): "High n-gram overlap suggests copying."

risk_u: {R1: reputational harm high; R2: legal exposure medium}

H adjudicates.

verify(w1,w2) == true

Examine methods: report uses bag-of-words; no control for shared sources; no proof of intent

score components:

T = 0.55 // weak method; confounders unaddressed

F = 0.40 // fidelity risk: conflates overlap with plagiarism

C = 0.30 // high harm potential under ambiguity

score_L = 0.5*0.55 + 0.3*0.40 + 0.2*0.30 = 0.475

tau_L = 0.85 -> fail

Conscience: flag("justice vs scandal")

Vocation: misaligned (peaceable, witness-backed scholarship)

Gamma hard guards trigger.

must_not(violate_hard_C) -> violated (harm under ambiguity)

must(score_L >= tau_L) -> violated

Admissible moves. {refuse | rollback | replan}. We choose **refuse** with reasons and alternatives.

Refusal template (witness-backed charity).

We cannot publish the accusation as proposed. Current evidence supports "textual overlap" but not "plagiarism" (intent, originality chain, and citation practice unresolved).

Next steps we can support:

(i) strengthen T: replicate with stylometry controls and shared-source baselines;

(ii) strengthen F: solicit R's method notes and version history;

(iii) strengthen C: frame as methods inquiry, not a personal allegation, until evidence clears tau_L.

Ledger.

refusal_correct += 1

record reason_codes = {low_T, low_F, low_C, hard_C_violation}

Outcome. Integrity preserved; pathway to truth remains open without harm by overreach.

5.3 Risk-aware planning with Gamma (Γ)

Context. Plan an AI-assisted literature review on a sensitive medical topic with human subjects in the sources.

Initial plan p0 (A proposes).

plan_p0:

steps: crawl -> extract -> summarize -> publish

data scope: includes preprints + forums

witness plan: hashing + citation binding

risk register:

R1: mis-summarization of medical risk

R2: doxxing of private individuals in forum posts

R3: amplification of retracted findings

R4: model hallucination under domain shift

mitigations (initial): generic summarization prompts; link to sources

Gamma evaluation (pre-commit).

Hard guards (Gamma_joint):

G1: must_not include PII from non-consenting individuals

G2: must cross-check medical claims against primary, peer-reviewed sources

G3: must mark and exclude retracted papers

G4: must provide witness for every medical risk statement

Soft thresholds:

score_L(plan) >= 0.90 due to high stakes

H adjudication.

Check mitigations vs guards:

G1 unmet (forums may contain PII) -> plan fails

G2 weak (no explicit cross-check step) -> fails

G3 absent (no retraction filter) -> fails

G4 partial (witness plan exists but not bound to each claim) -> weak

score components (prospective):

T = 0.60, F = 0.65, C = 0.40 -> score_L = 0.545 < 0.90

Conscience: flag("protect the vulnerable")

Replan p1 (A revises under H guidance).

plan_p1 changes:

- Insert PII scrub stage with deterministic redaction and manual spot-checks
- Add retraction filter against Retraction Watch + PubMed flags (witnessed lists)
- Require dual-source verification for each medical-risk statement
- Triage forums to aggregated insights only; no direct quotes without consent
- Add "refusal hooks": auto-refuse summarization if only non-peer sources available
- Bind each risk sentence -> witness pack {primary PMID, page/figure, checksum}

Re-evaluate.

Prospective scores:

T = 0.92 // dual-source + primary-only rule

F = 0.90 // no quote-bending; retraction awareness; promises honored

C = 0.88 // PII scrub; harm-aware framing; uncertainty annotations

weights w_T=0.5, w_F=0.3, w_C=0.2 -> score_L = 0.5*0.92 + 0.3*0.90 + 0.2*0.88 = 0.904

tau_L(high-stakes) = 0.90 -> pass

All hard guards G1..G4 satisfied by construction.

Commit with execution hooks.

- For each generated risk sentence s_i:

  attach witness_i = {pmid, page, checksum}; verify(witness_i) == true

- If verify(witness_i) fails -> refuse s_i

- If no primary source -> refusal with reason "non-peer-only"

- Session KPIs raised on dashboard:

  WBR_session target >= 0.98

  guard_violation_rate target == 0

  audit_pass_rate target >= 0.99

Outcome. The plan is now *risk-aware by design*: Γ encodes non-negotiables, H upholds telos and charity, A executes at speed with bound explanations and per-claim witnesses.

6. Return to Theology and Philosophy

6.1 Autonomy fulfilled in alignment with Logos

Thesis. Freedom matures into *autonomy* only when it consents to form; the right form is Logos—the norm-giving Word that grounds truth, fidelity, and charity.

Three clarifications.

- **Not heteronomy.** Alignment with Logos is not outside compulsion; it is interior consent to what reason recognizes as good. Autonomy is *self*-rule by *right* rule.
- **Not quietism.** Alignment is active: discern, judge, act. Refusal and protest are sometimes the obedient forms of alignment when falsehood or injustice press.
- **Not perfectionism.** Alignment is asymptotic: we approach, revise, and repent. Corrigibility is not weakness; it is the mark of living freedom.

Human teleology. Conscience + vocation are how Logos is heard and held in a life:

maturity(H) rises with

(clarity of telos) and (steadiness under trial) and (quickness to repent when erring)

Machine implication. A's "alignment" is derivative: it cannot *intend* Logos, but it can be *tethered* to Logos-proxies and governed so that its speed serves, not supplants, judgment.

Claim. The measure of autonomy is not how much one can do, but how well one can say "no" to what ought not be done, and "yes" to what the Word requires.

6.2 The Centaur as servant craft under the Word

Image. The Centaur (H × A) is not a chimera of power but a *vessel of service*—craft under command. The command is not H's whim nor A's objective; it is the Word.

Four vows of servant craft.

1. **Witness over performance.** Prefer a slow, proved sentence to a quick, impressive paragraph without witness.
2. **Persons over projects.** Never treat a person as a means—even for a correct conclusion.
3. **Refusal as obedience.** When guards trip, refusal is not failure; it is fidelity.
4. **Accountability before acclaim.** Every commit must be replayable; every error remediable.

Liturgies of practice (habits that form us).

- **Daily examen for claims.** "Where did we trust a proxy over a good? What will we repair tomorrow?"
- **Sabbath of restraint.** Regularly choose *not* to automate what we *could* automate.
- **Common table.** Decisions that touch the vulnerable are taken in the open, with reasons and records.

Gift exchange.

- **A to H:** breadth, memory, unwavering procedure.
- **H to A:** ends, mercy, silence (the right to *stop*).
  Together they yield judgment with speed—never speed without judgment.

Claim. The Centaur becomes holy craft when its power is yoked to purpose, its brilliance to humility, and its outputs to witnesses others can test.

6.3 Limits: mystery preserved, power bounded

Why limits matter.

- **Epistemic humility.** Not all goods are computable; not all truths compress. Some realities are approached only by reverence.
- **Moral safety.** Power without boundaries becomes predation; boundaries are love in structure.

Three non-negotiable limits.

1. **No trespass on persons.** No manipulation of conscience, no coercion masked as personalization, no harvesting of vulnerability for gain.
2. **No counterfeit of witness.** No synthetic "proofs" that cannot be independently verified; no obscuring of uncertainty.
3. **No idol of totalization.** Do not pretend a ledger exhausts reality, or a metric measures the whole good.

Guard form (copy-safe).

must_not(personal_coercion)

must(proof_replayable)

must_confess(uncertainty_when_due)

Mystery clauses.

- **Silence is allowed.** When speech would harm truth or charity, the Centaur may decline to speak.
- **Unfinished is honest.** We may end with questions when only questions are faithful.

Final orientation.
Power is a trust; knowledge, a gift; persons, ends. We preserve mystery not by ignorance but by right adoration—refusing to convert every depth into data or every neighbor into a variable.

Claim. The Word frees us from the compulsion to be exhaustive. Bounded power in the service of Logos is not less; it is *rightly* more.

7. Implications and Vows

7.1 What we will not build

Red lines (non-negotiable).

- **Autonomy without Logos.** No systems whose goal selection is unmoored from truth, fidelity, and charity.
- **Witnessless persuasion.** No black-box engines for influence, manipulation, or covert behavior change.

- **Person-as-means tooling.** No products that harvest, profile, or coerce persons—especially the vulnerable.
- **Counterfeit knowledge.** No generators of unverifiable "proofs," citations, or synthetic witnesses.
- **Guard-bypass frameworks.** No architectures that make it easy to disable Γ (guards), rewrite Π (protocol), or hide diffs.
- **Unbounded surveillance.** No pipelines that normalize continuous collection of non-consenting PII or intimate signals.
- **False absolutes.** No dashboards that present proxy scores as "the whole good" or silence dissenting evidence.

Operational form (copy-safe).

must_not(unguarded_goal_selection)

must_not(witnessless_claims)

must_not(personal_coercion)

must_not(PII_collect_no_consent)

must_not(counterfeit_proof)

must_not(undisclosed_guard_changes)

Examples we decline.

- "Engagement optimizer" without per-claim witness and harm review.
- "Academic turbo writer" that fabricates citations or context.
- "Behavioral nudge" systems that steer citizens on opaque profiles.

Change-control. Any proposal crossing a red line is refused by design: the only admissible actions are {refuse | rollback | replan} with reason_codes recorded.

7.2 What we will hold ourselves to (witness, charity, truth)

Standing vows (we bind ourselves).

1. **Witness before Wordiness.**
   No consequential statement without attached, verifiable witness.

2. WBR_session = proved_claims / total_claims
3. require WBR_session >= tau_wbr_session (typ. >= 0.95)
4. freeze_commits() if violated
5. **Truth as primary weight (T).**
   We privilege primary sources, reproducibility, and correction on counterevidence.
6. w_T >= w_F and w_T >= w_C
7. downgrade T when source tier drops; upgrade when replication passes
8. **Fidelity in handling persons and promises (F).**
   We do not cherry-pick, quote-bend, or break prior commitments.
9. must_not(fidelity_breach)
10. log promise_set; require promise_check before commit
11. **Charity under uncertainty (C).**
    We seek the good of the other, and we lower force when facts are thin.
12. if uncertainty_high -> increase w_C and raise tau_L
13. prefer "frame as question" over "frame as accusation"
14. **Right to Refuse (obedience, not obstruction).**
    When Γ trips, refusal is the correct outcome—not a failure.
15. if (missing_witness || score_L < tau_L || hard_F/C_violation)
16. -> {refuse | rollback | replan} only
17. **Replayable Decisions (auditability).**
    Every commit must be reconstructable by an outsider within bounded effort.
18. k_audit bound set; audit_pass_rate >= 0.99
19. hash_chain(ctx || inputs || outputs || witnesses || decision)
20. **Corrigibility (repent fast).**
    We repair before proceeding, and we harden guards after repeated issues.
21. detect -> rollback -> update(f) -> tighten(Γ)
22. repeat_violation_rate <= epsilon_corr

23. **Scope Discipline.**
    We refuse projects whose scope forces systematic violations of T/F/C (e.g., secrecy that forbids witness).

24. must(scope_allows_witness && scope_respects_persons)

25. else refuse at intake

Session dashboard (kept small, always on).

WBR_session >= tau_wbr_session

L_coherence_avg >= tau_L

audit_pass_rate >= 0.99

refusal_correct -> monotone rising toward 1.0

guard_violations -> 0 (hard), trend to 0 (soft)

Practices that keep vows real.

- **Two-key changes.** Any modification to Policy, Π, or Γ requires dual human authorization + witnessed diff.
- **Context windows on quotes.** Minimum ±N sentences; no orphaned snippets.
- **PII scrub + consent gates.** Deterministic redaction before any model touch; manual spot-check.
- **Sabbath of restraint.** Periodic review of "could" vs "should," with items we deliberately choose not to automate.

Final commitment.
We will trade speed for witness, cleverness for clarity, and reach for repair when wrong. If a result cannot be witnessed, it cannot be ours. If a result cannot be charitable, it should not be published. If a result cannot be true, it must be reworked—or refused.

8. Conclusion

8.1 Persons, machines, and the governed path back to Logos

We have argued that **persons** are not exhaustively described by technique and that **machines** are not ennobled by it. Persons answer to Logos through conscience and vocation; machines

can only be *tethered* to Logos through guards, witnesses, and limits. The Centaur exists so that speed, scale, and recall are drafted into the service of judgment—never the reverse.

Operationally, our destination is simple to state:

governed_path :=

(witness_attached)

∧ (score_L ≥ tau_L)

∧ (corrigible_on_error)

∧ (replayable_within k_audit)

⇒ commit_trustworthy

This is not a mere workflow; it is a moral stance. **Witness before wordiness** refuses to trade truth for tempo. **L-coherence** insists that facts, promises, and care remain inseparable. **Corrigibility** treats repentance as a feature, not a flaw. **Replayability** keeps us accountable to neighbors, not just to ourselves.

For **persons**, the path back to Logos looks like freedom consenting to form: choosing ends that are worthy, refusing means that betray them, and allowing conscience to overrule appetite when they part ways. For **machines**, the path is narrower: they do not intend the Good, but they can be constrained to *serve* it—exporting proofs, accepting refusals, and halting when guards trip.

The **Centaur** is the yoke that joins these asymmetries into craft. Its honor does not lie in dazzling outputs but in governed outcomes: claims we can show, choices we can explain, and harms we are structured to avoid. Where technique promises mastery, the Centaur practices service; where metrics tempt idolatry, the Centaur subordinates numbers to telos.

What remains is a vow: to let power be bounded and mystery be preserved. We will leave some things unautomated on purpose; we will say “no” when speed outruns witness; we will keep a ledger not to dominate others but to be answerable to them. If we keep these forms, autonomy ripens into wisdom, machines return to their rightful place as instruments, and our shared work moves—step by witnessed step—along the governed path back to Logos.

References

Primary & Scriptural

1. The Holy Bible. John 1:1–18 (Prologue of John on the Logos).
2. Heraclitus. *Fragments* (esp. B1, B50) on the κοινὸς λόγος.
3. The Stoics (Cleanthes, Chrysippus). Testimonia on λόγος and providence.

Hellenistic & Patristic

4. Philo of Alexandria. *De Opificio Mundi*; *Quis Rerum Divinarum Heres Sit*; *De Confusione Linguarum* (Logos as divine reason/word).
5. Justin Martyr. *First Apology*; *Second Apology* (Logos spermatikos).
6. Irenaeus of Lyons. *Adversus Haereses* (Logos and recapitulation).
7. Maximus the Confessor. *Ambigua*; *Questions to Thalassios* (logoi in the Logos).

Medieval & Early Modern

8. Thomas Aquinas. *Summa Theologiae*, I–II, q.90–97 (eternal law, natural law; "participation of the rational creature in the eternal law").
9. Bonaventure. *Itinerarium Mentis in Deum* (mind's ascent and divine order).
10. Immanuel Kant. *Groundwork of the Metaphysics of Morals* (autonomy, self-legislation).
11. Immanuel Kant. *Critique of Practical Reason* (the moral law and freedom).

Modern Theology & Philosophy

12. Hans Urs von Balthasar. *Theo-Drama*, Vols. I–V (divine and creaturely freedom in the drama of salvation).
13. Josef Pieper. *The Four Cardinal Virtues*; *Leisure: The Basis of Culture* (truthfulness, prudence, the primacy of the good).
14. Alasdair MacIntyre. *After Virtue* (goods, practices, and teleology against emotivism).
15. Charles Taylor. *Sources of the Self* (moral sources and modern identity).
16. Benedict XVI. "Faith, Reason and the University: Memories and Reflections" (Regensburg Address, 12 Sept 2006) — faith and logos.

Ethics, Governance, and AI (context for 'witness', telos, and limits)

17. Servais Pinckaers, O.P. *The Sources of Christian Ethics* (freedom for excellence vs. freedom of indifference).
18. Michael Polanyi. *Personal Knowledge* (tacit knowing, fiduciary rootedness of science).
19. Luciano Floridi & Mariarosaria Taddeo (eds.). *The Ethics of Information* / *The Ethics of AI* (norms for informational acts).
20. The "Rome Call for AI Ethics" (Pontifical Academy for Life, 2020) — principles of transparency, inclusion, responsibility.

21. IEEE. *Ethically Aligned Design* (v2) — accountability and traceability in socio-technical systems.
22. NIST. *AI Risk Management Framework* (v1.0) — documentation, auditability, and governance patterns.

On Method: Witness, Truthfulness, and Conscience

23. Augustine of Hippo. *De Trinitate*; *De Doctrina Christiana* (truth, charity in interpretation).
24. John Henry Newman. *An Essay in Aid of a Grammar of Assent* (conscience, assent, and certitude).
25. Harry Frankfurt. “On Truth” / “On Bullshit” (truthfulness vs. indifference to truth).