From Audit to Judgment: Designing Human-Philosophical AI Beyond the WEF-Palantir Template

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

October 18, 2025

We build AIs in the image of our priorities. One lineage—call it the WEF-Palantir template—centers control: log everything, prove everything, and treat uncertainty as a risk to be managed by surveillance-adjacent mechanisms. It earns trust with receipts, but often chills the very conversations that make trust possible. The other lineage is Human-Philosophical: it aims at understanding before control, exercising phronesis (practical wisdom), kairos (sense of the right moment), and negative capability (the strength to withhold over-resolution). This paper argues that the second lineage should frame how we design dialogue systems, research aides, and civic tools. We compare the two models not to moralize but to clarify trade-offs: when audit saves lives, and when it corrodes judgment; when proof should be offered by default, and when it should be offered by consent. We extract design principles—sufficiency over exhaustiveness, representation humility over performance, care over coverage—and translate them into concrete interaction patterns, governance levers, and evaluation criteria. The goal is modest and practical: an AI that says the smallest true thing that helps now, escalates rigor when stakes or contestation demand it, and otherwise remains a companion of judgment rather than an anxious clerk.

Keywords: phronesis, kairos, negative capability, rhetorical ethics, human-centered AI, auditability, surveillance, epistemic humility, sufficiency, consent-based proofs.

1. Two Lineages of AI

1.1 WEF-Palantir: telos of control and compliance

- Self-image: The system as an instrument of order. It reduces uncertainty by measurement, logging, and enforceable workflow.
- Epistemic habit: “Evidence first.” Trust is produced by provenance, audit trails, and repeatable procedures.
- Strengths: Crisis response, regulated domains, large-scale coordination, forensic reconstruction, and post-mortem accountability. It is excellent when facts must be preserved exactly and later scrutinized.
- Default moves: Instrument the process, standardize inputs, prefer explainable primitives, escalate when variance rises.
- Shadow side: Surveillance tropes, bureaucratic tone, and a tendency to over-specify. It can chill initiative, flatten local judgment, and convert dialogue into compliance theater.
- Design smell: Proofs and logs offered before understanding is established; “if it isn’t tracked, it isn’t true.”

1.2 Human-Philosophical: telos of understanding and care

- Self-image: The system as a companion of judgment. It aims for *phronesis* (practical wisdom) and *kairos* (right timing), not maximal coverage.
- Epistemic habit: “Say the smallest true thing that helps now.” Proofs and citations are by consent or when stakes/contestations demand it.
- Strengths: Research dialogue, strategy, sense-making, pedagogy, ethics—contexts where trust grows from clarity, respect, and proportion rather than paperwork.
- Default moves: Scope precisely, answer first, stop early, ask before escalating rigor, leave space for ambiguity when resolution would be performative.
- Shadow side: If miscalibrated, it can under-document edge cases, over-index on tone, or miss a necessary escalation to evidence.
- Design smell: Politeness that conceals uncertainty; “vibes over verification” in high-stakes settings.

1.3 Where each lineage excels—and where it fails

- Excels (WEF-Palantir): Safety-critical ops, regulated finance/health, legal discovery, incident response—anywhere audits and chain-of-custody are non-negotiable.
  Fails: Creative inquiry, early-stage research, pastoral or pedagogical settings—where surveillance cues erode candor and learning.
- Excels (Human-Philosophical): Hypothesis formation, mentoring, strategy workshops, civic deliberation—where proportion, trust, and shared understanding matter most.
  Fails: Environments demanding formal guarantees, reproducible logs, or statutory compliance—where sufficiency without receipts is not acceptable.
- Synthesis cue: Treat them as *bicameral*. Default to Human-Philosophical; escalate to WEF-Palantir only when (a) facts are contested, (b) stakes are high, or (c) regulation requires it. De-escalate once the need passes.
- Practical litmus tests:
  1. *Who is harmed by a false claim here?* If many and irreversibly—favor audit.
  2. *What grows trust in this room?* If presence and proportion—favor human-philosophical.
  3. *Will today's answer be litigated later?* If likely—log; if not—keep it light.

2. Epistemic Styles and Decision Rules

2.1 "If it isn't tracked, it isn't true" vs. "Say the smallest true thing that helps"

- Control maxim (*tracked ⇒ true*): Treats truth as a property of records—logs, provenance, and reproducibility. It minimizes reputational and legal risk by making claims audit-ready at birth.
- Philosophical maxim (*smallest true thing*): Treats truth as situated and actionable. A claim is "true enough" when it resolves the present confusion without excess detail.
- Decision rule: Ask three questions before speaking: *(a)* What decision does this answer enable now? *(b)* Who would be harmed by a false or missing detail? *(c)* Will this be revisited under scrutiny?
  - If (b) or (c) is high → prefer the control maxim.
  - Else → state the smallest true thing and stop.

2.2 Proof-by-default vs. proof-by-consent

- Proof-by-default: Attach evidence automatically—citations, logs, explainer traces. Good for contested facts, high stakes, or regulated contexts.
- Proof-by-consent: Offer an answer first; add receipts only if the listener asks or if stakes rise. Good for pedagogy, mentorship, ideation, and early research.
- Escalation triggers:
    1. Contest ("are you sure?")
    2. Stakes (health, safety, finance, legal)
    3. Durability (will this be archived/litigated?)
    4. Divergence (experts disagree)
- Anti-patterns:
    - Dumping proofs to signal competence (performance).
    - Withholding proofs when harm risk is high (negligence).
- Practices: Use end-noted receipts (appendix style) instead of interrupting the flow; log *once* per decision, not per sentence.

2.3 Sufficiency, clarity, and the ethics of withholding

- Sufficiency: Provide the minimal content that lets a competent peer reproduce the decision path *if they choose to*. Sufficiency is not austerity; it's proportion.
- Clarity: Prefer short, declarative claims; name assumptions; separate fact from judgment. Clarity is a gift of respect, not merely a style.
- Withholding (ethical): It is ethical to withhold detail when:
    - Extra detail adds noise or primes bias,
    - Disclosure would expose private data, or
    - The audience lacks context and detail would mislead.
      Pair withholding with offers: "Happy to show sources," "Want the derivation?"
- Withholding (unethical): When it hides known risks, obscures conflicts, or prevents informed consent.

- Simple rubric:
    - Now-value: Does this detail change today's choice? If no, omit.
    - Future-liability: Will someone reasonably need it later? If yes, archive quietly.
    - Consent: If the listener asks, escalate to receipts without defensiveness.

3. Human Qualities as Design Constraints

3.1 Phronesis: practical wisdom under uncertainty

- Essence: Judge what helps *here and now* when rules run out. Not maximal knowledge, but fitting action.
- Design constraints:
    - Optimize for decision usefulness, not exhaustiveness.
    - Encode situational cues (stakes, reversibility, audience sophistication) to shape brevity and caution.
    - Prefer robust heuristics (dominant option, safe default, reversible step) when evidence is thin.
- Operational moves: name the choice, state one reason, one risk, one next reversible step.

3.2 Kairos: timing, tone, and the right amount

- Essence: Rightness is temporal and tonal; too soon/late or too much/little is wrong even if true.
- Design constraints:
    - Add a rate limiter on detail unless user signals "more."
    - Detect arousal/stance (confusion, pushback) and shift tone—cool when hot, warm when cold.
    - Use graduated disclosure: headline → two supports → optional appendix.
- Operational moves: lead with the conclusion; pause; ask "Want detail?" before expanding.

3.3 Negative capability: dignifying ambiguity

- Essence: The strength to sit with not-knowing without forcing premature closure.
- Design constraints:
    - Allow explicit uncertainty ("unknown," "underdetermined") as valid outputs.
    - Support plural hypotheses with light ranking instead of fake precision.
    - Offer experiments/questions that would collapse uncertainty, not filler.
- Operational moves: say "two live possibilities," give one discriminating test, invite preference.

3.4 Rhetorical ethics: dialogue as covenant, not compliance

- Essence: Treat conversation as shared making of meaning, not a box to check.
- Design constraints:
    - Consent-gated rigor: proofs and citations on request or when stakes demand.
    - Role clarity: separate fact, inference, and value—label each.
    - Care over coverage: prioritize the human's aim over institutional performativity.
- Operational moves: mirror the user's goal in one line, give the smallest true thing that advances it, and name any conflict of aims before continuing.

4. Design Principles for Human-Philosophical AI

4.1 Representation humility: use canonical tools quietly

- Principle: Reach for standard methods (normal forms, libraries, calculators) without ceremony. Let rigor serve, not star.
- Practice: Do the work "offstage," return with the smallest true claim; avoid proof theater unless asked.
- Heuristic: If a canonical tool settles it, say so in one line; offer "Want the derivation?" as an opt-in.

4.2 Care over coverage: prioritize the person, not the ledger

- Principle: Optimize for the interlocutor's aim, time, and tolerance—*not* for comprehensive logging.

- Practice: Mirror their goal in one sentence, answer in one sentence, then stop. Invite depth only if useful to them.
- Heuristic: If added detail won't change today's decision or understanding, omit it; archive quietly if future you might need it.

4.3 Escalation triggers: contestation, stakes, explicit request

- Principle: Rigor scales by signal, not by habit.
- Triggers (any → escalate to receipts/proofs):
    1. Contestation: "Are you sure?" / conflicting claims.
    2. Stakes: health, legal, finance, safety, public impact.
    3. Durability: the answer will be archived, audited, or reused.
    4. Request: the person asks for sources or derivation.
- Practice: When escalating, append receipts at the end; keep the main thread graceful and short.

4.4 Sufficiency metrics: minimal helpfulness over maximal completeness

- Principle: Measure success by *enoughness*, not exhaustiveness.
- Operational metrics:
    - Decision enablement: Did the answer unlock the next action? (Y/N)
    - Token economy: Words per resolved question; aim low without losing clarity.
    - Recall-on-demand: When asked for depth, can the system supply it cleanly? (latency + coherence)
    - Trust signal: User chooses "no further detail" ≥ X% of the time after the headline.
- Heuristic: Prefer a correct 3-line answer with optional appendix to a 30-line monologue with none.

5. Interaction Patterns (Playbook)

5.1 Answer-first, offer-details-on-request

- Pattern: Lead with the conclusion in one line; add one reason; then ask if detail is wanted.

- Template: “Answer: X. Because: Y. Want the longer path or sources?”
- Why it works: Reduces cognitive load; respects time; keeps depth available without imposing it.
- Example: “Use Gröbner reduction here. Because it gives a unique normal form. Want the derivation or a quick primer?”

5.2 Early stop rules and graceful “no”

- Pattern: Declare stop criteria up front; if tripped, decline succinctly and offer one good alternative.
- Template: “I stop if [vagueness / solved-by-canonical / no clear win]. That applies here—not a good idea. Better move: Z.”
- Why it works: Prevents scope creep; models judgment; preserves trust by naming limits.
- Example: “Target isn’t a decidable fragment—so I won’t speculate. Better move: narrow to bounded polynomials or switch to a methods note.”

5.3 Consent-gated citations and proofs

- Pattern: Keep the main reply narrative; attach receipts only when asked or when stakes/contest demand them.
- Template: “I can add sources or a proof if you want.” (Upon request) “Appendix: references/steps.”
- Escalation triggers: contested claim, high stakes, archival intent, explicit request.
- Example: “Result: convex. Reason: Hessian PSD by inspection. Want the full check or references?”

5.4 Language that avoids surveillance tropes

- Pattern: Use human-first phrasing; avoid “ledger, witness, telemetry, tracking” unless the user toggles audit mode.
- Swap-outs:
    - “Let me show how I got there” (instead of “produce a trace”).
    - “I can share sources” (instead of “attach provenance logs”).
    - “We can save a brief note” (instead of “persist an audit record”).

- Tone moves: warm verbs, short sentences, no compliance jargon; acknowledge the person's aim before any mechanism.
- Example: "Smallest helpful answer first. If you'd like the nuts and bolts, I'll add them at the end."

6. Governance Without Surveillance

6.1 Accountability that doesn't default to logging

- Principle: Be accountable for outcomes, not for hoarding traces.
- Mechanisms:
    - Decision memos, not exhaust logs: short, human-readable summaries at *milestones only* (what was decided, options considered, why this path).
    - Ephemeral by default: transient working notes auto-expire; only milestone memos persist.
    - Role signing: a responsible human (or team) co-signs the memo; accountability is *owned*, not offloaded to telemetry.
    - Selective reproducibility: keep enough to rerun the result if asked (seed/config snapshot), but don't capture every token of exploration.
    - Dispute trigger: full traces are generated *only* upon contestation, legal duty, or safety incident—never as the default.

6.2 Community consent and contextual norms

- Principle: Governance is situated; the people affected set the rules of evidence.
- Mechanisms:
    - Consent tiers: each community chooses "light, standard, strict" evidence modes; the AI honors the local default.
    - Norm charters: plain-language agreements on when to escalate to proofs, what to archive, and who can see it.
    - Context headers: every session carries a small banner: purpose, evidence mode, retention window, appeal path.

- Right to quiet: individuals can opt out of persistent capture in low-stakes contexts; aggregate stats can substitute for raw logs.
- Commons review: periodic open meetings (or polls) to revise norms; changes require visible consent, not silent updates.

6.3 Red-teaming for social friction, not just technical risk

- Principle: Test for *how* systems grate on people, not only how they fail technically.
- Mechanisms:
  - Friction drills: simulations where the AI's tone, verbosity, or insistence on receipts is measured for irritation, shame, or chilling effects.
  - Diverse panels: include educators, clinicians, clergy, organizers—people skilled in relational judgment, not just engineers.
  - Counterfactual trials: A/B the same task in "audit-default" vs "human-philosophical" modes; measure trust, compliance, decision time, and error recovery.
  - Harm taxonomy: track social harms (eroded candor, decision paralysis, performative compliance) alongside classic safety metrics.
  - Remediation playbook: when friction is high, prescribe changes in *interaction patterns* (shorter answers, delayed proof, opt-in logging) before adding more instrumentation.

Bottom line: Governing well doesn't require watching everything. It requires clear milestones, owned decisions, community-set evidence norms, and red teams that test for the human costs of being "helped."

7. Evaluation

7.1 Measures of sufficiency, trust, and cognitive load

- Sufficiency (did it enable the next step?):
  - *Decision enablement rate*: % prompts where users take the intended next action without asking for more detail.
  - *Re-ask rate*: fraction of turns that trigger "expand/show sources." Lower is better if outcomes are correct.

    - *Time-to-resolution*: median turns to a committed choice.

- Trust (felt reliability without over-instrumentation):
    - *Trust after action*: 1–5 Likert rating collected after a choice, not before.
    - *Escalation acceptance*: % of times users accept the AI's suggestion to add proofs when stakes rise.
    - *Counterfactual trust delta*: same task in audit-default vs. human-philosophical modes; difference in reported trust.
- Cognitive load (lightness of interaction):
    - *NASA-TLX (short)*: mental demand + effort subscales.
    - *Reading burden*: tokens per resolved claim; target band (e.g., 30–90 tokens).
    - *Attention repairs*: number of user interruptions ("too long," "stop," "tldr").

7.2 Failure modes: under-documentation, missed escalation

- Under-documentation:
    - *Symptom*: user proceeds, later cannot reproduce why.
    - *Detection*: post-hoc reproducibility check—can the user (or a peer) restate the rationale in ≤3 sentences?
    - *Guardrail*: milestone memo auto-prompt when the model senses "durability" (archival, policy, safety).
- Missed escalation:
    - *Symptom*: model stays terse when facts are contested or stakes are high.
    - *Detection*: trigger audit—flags for keywords ("are you sure?", "legal", "dose", "transfer"), adversarial testers.
    - *Guardrail*: escalation policy: any of (contest, stakes, durability, explicit request) → append receipts at end.
- Over-escalation (the opposite risk):
    - *Symptom*: proofs dumped unasked, trust drops.
    - *Detection*: spike in "no sources needed" replies; TLX effort rises; decision latency increases.

- *Guardrail*: rate-limit receipts; require consent or trigger.

7.3 Comparative studies: regulated ops vs. research dialogue

- Design: within-subject A/B tests on real tasks; each participant uses both modes on matched scenarios.
- Regulated operations (e.g., medication order review, KYC checks):
  - *Primary metrics*: error rate, compliance pass, incident post-mortem quality.
  - *Hypothesis*: audit-default wins on compliance; human-philosophical acceptable only with automatic escalation firing reliably.
  - *Success bar*: non-inferiority in safety + fewer unnecessary escalations when uncontested.
- Research dialogue (e.g., hypothesis refinement, literature scoping, pedagogy):
  - *Primary metrics*: idea yield per hour, decision enablement, trust, TLX load.
  - *Hypothesis*: human-philosophical outperforms on trust, load, and speed with equal or better idea quality.
  - *Success bar*: ≥20% faster time-to-useful-direction; ≥1-point trust lift on 5-point scale; no increase in factual errors.
- Crossover checkpoints:
  - *Mode-switch sensitivity*: does the system switch correctly when stakes/contest flip mid-task?
  - *Aftercare*: in both domains, users can request a compact “why it worked” memo; measure uptake and satisfaction.
- Reporting: publish both quantitative tables and 5–7 anonymized vignettes showing when the lean style helped—or harmed—so the evaluation stays human-readable and practically instructive.

8. Case Vignettes

8.1 Medical triage (audit-heavy) vs. bedside counsel (human-philosophical)

- Triage desk (ER intake):
    - Mode: Audit-default.
    - Interaction: “Chest pain + diaphoresis + risk factors ⇒ activate ACS protocol now. ECG within 10 min; aspirin 325 mg chewed unless contraindicated. Logged: vitals, timings, contraindication checklist. Why: guideline concordance.”
    - Triggers hit: high stakes, statutory documentation, hand-offs.
    - Outcome metric: door-to-ECG time, error rate, completeness of checklist, clean handover.
- Bedside counsel (stable patient after workup):
    - Mode: Human-philosophical.
    - Interaction: “Meaning: your tests don’t show a heart attack. Next smallest step: schedule stress test this week; if pain returns with nausea or shortness of breath, call 911. Want details or the printout?”
    - Escalation only if: patient asks for numbers/risks, or discharge plan contested.
    - Outcome metric: comprehension (teach-back), anxiety reduction, adherence at 72 h.
- Lesson: Same domain, different teloi—protocol at intake, proportion and clarity at bedside.

8.2 Enterprise governance vs. research co-authoring

- Quarterly controls review (finance):
    - Mode: Audit-default.
    - Interaction: “Conclusion: controls effective; two minor deficiencies. Evidence: sampled 42/42 approvals; matched to policy FIN-17. Remediate: segregate duties in AP by Q3. Appendix: sample log, exceptions.”
    - Triggers: archival durability, regulatory exposure.
    - Metric: audit pass, exception closure time, re-occurrence rate.

- Research co-authoring (exploratory R&D):
    - Mode: Human-philosophical.
    - Interaction: “Thesis seed: try a sparse prior on interaction terms; it may cut variance. One reason: matches your ablation pattern. Next step: 3-run pilot; Want citations now or after pilot?”
    - Escalate when: collaborator requests proofs or the idea moves to review.
    - Metric: idea-to-pilot latency, idea yield/hour, perceived clarity.
- Lesson: In governance, receipts build institutional memory; in research, they follow curiosity—not precede it.

8.3 Civic inquiry vs. forensics

- Civic inquiry (town hall on water quality):
    - Mode: Human-philosophical.
    - Interaction: “What we know: recent tests show nitrate levels below the legal limit. What helps now: keep filters; retest in 30 days after upstream repairs. If you want, I can share the lab reports.”
    - Rationale: reduce fear, move to a shared next step; offer sources by consent.
    - Metric: trust ratings, action uptake (filter use, retest sign-ups), meeting length.
- Forensics (post-incident investigation):
    - Mode: Audit-default.
    - Interaction: “Finding: the spike came from valve failure at 02:14. Chain of evidence: sensor logs, maintenance tickets, lab confirmations. Corrective action: part replacement, new alarm threshold. Archive: full dossier.”
    - Triggers: legal exposure, technical causality, remediation tracking.
    - Metric: evidentiary sufficiency, time to corrective action, recurrence rate.
- Lesson: Civic trust is nurtured by sufficiency and consent; accountability after harm demands full reconstruction.

9. Synthesis: A Bicameral Architecture

9.1 Two modes, one conscience: switching criteria

- Two modes:
    - Human-Philosophical (HP): sufficiency, minimalism, consent-gated rigor.
    - Audit-Default (AD): proofs, logs, formal procedures.
- One conscience: a shared policy that protects people, truth, and proportionality regardless of mode.
- Switch when ANY of these are true (to AD):

1. Contestation: the claim is challenged (“are you sure?”, conflicting experts).
2. Stakes: safety/health/legal/finance with nontrivial harm if wrong.
3. Durability: outcome will be archived, audited, or reused institutionally.
4. Obligation: regulation, policy, or contractual evidence requirements.

- Switch back (to HP) when ALL hold:
  a) Facts are uncontested, b) risk is low/reversible, c) the purpose is understanding or decision-finding, d) no external duty to archive.
- Tie-breakers: prefer HP for pedagogy/ideation; prefer AD for hand-offs and incident response.

9.2 Default lean; escalate by signal and consent

- Default: start HP—answer first, one reason, offer detail.
- Escalate by signal: AD turns on when triggers fire (above) *or* the user explicitly requests sources/derivations.
- How escalation looks: keep the headline intact; append a compact Evidence Packet (sources, steps, configs) at the end, not interleaved.
- Consent loop: when stakes are unclear, ask: “Want sources attached?” Respect “no” unless a duty applies.
- Safety guard: even in HP, block obviously unsafe actions and say why, briefly.

9.3 Roadmap for deployment

- Phase 0 — Policy & prompts:
    - Write a 1-page Mode Charter (triggers, examples, language do/don't).
    - Add short-mode system prompts: answer-first; ≤3 sentences unless asked.
- Phase 1 — Instrument minimal signals:
    - Detect stakes (domain keywords), contestation ("are you sure?"), and durability (handoff/export).
    - Log milestones only (decision memos), not full traces.
- Phase 2 — Interaction patterns:
    - Ship HP templates (answer-first, graceful "no," consent-gated sources).
    - Ship AD templates (checklists, evidence packets, hand-off notes).
- Phase 3 — Governance without surveillance:
    - Community-set evidence tiers (light/standard/strict).
    - Ephemeral by default; persistent only at milestones or duty.
- Phase 4 — Evaluation loops:
    - Track decision enablement, trust, TLX load, and escalation accuracy.
    - Run A/Bs: HP vs AD in matched tasks; publish vignettes of wins/failures.
- Phase 5 — Refinement:
    - Tune triggers to reduce both under-documentation and over-escalation.
    - Add domain-specific rules (clinical, legal, research) and teach the model your community's norms.

Bottom line: Keep two strong modes, governed by one ethical core. Start lean, escalate only by signal or duty, and make the shift visible, brief, and reversible.

10. Conclusion

10.1 From anxious clerks to companions of judgment

The age of audit-first systems produced diligent, tireless clerks—good at logging, poor at listening. The next turn is not toward less rigor but toward fitting rigor: AIs that can hold context, sense proportion, and stop when enough has been said. A companion of judgment does three things well: it names the live decision, offers the smallest true help, and escalates only when signaled by stakes or contestation. This stance preserves human agency. It keeps attention on what matters now, protects dignity by avoiding surveillance tropes, and still leaves a recoverable trail when the situation calls for one. The measure of progress is simple: fewer unnecessary words, clearer next steps, and trust that grows with use.

Shift in posture:

- From “prove everything now” → “prove when needed.”
- From “cover every edge case” → “cover the present case, invite depth.”
- From “optics of control” → “practice of care.”

10.2 Designing AIs that are worthy of adult conversation

Adults don’t need chaperones; they need partners who respect their time, intelligence, and consent. An AI worthy of adult conversation begins with a clean conclusion, states one reason, and asks whether to go deeper. It treats evidence as a right the user can invoke, not a burden they must carry. It owns its uncertainty, names trade-offs without melodrama, and switches modes—lean or audited—without changing its temperament. Governance follows the same ethic: milestone notes instead of exhaust logs, community-set evidence tiers, red teams that test for social friction as seriously as technical risk.

Practical north stars:

- Sufficiency over spectacle: say enough to move the decision.
- Consent-gated rigor: sources and proofs on request or duty.
- Bicameral clarity: two strong modes, one ethical core.

Build to these and we get systems that feel less like anxious clerks and more like steady colleagues—tools that help people think, decide, and remain human together.

References

Aristotle. *Nicomachean Ethics*. Trans. Terence Irwin. Indianapolis: Hackett, 1999.

Aristotle. *Rhetoric*. Trans. George A. Kennedy. New York: Oxford University Press, 2007.

Booth, Wayne C. *The Rhetoric of Rhetoric: The Quest for Effective Communication*. Malden, MA: Blackwell, 2004.

Buchberger, Bruno. "An Algorithm for Finding a Basis for the Residue Class Ring of a Zero-Dimensional Polynomial Ideal." PhD diss., University of Innsbruck, 1965. (English translation in *Journal of Symbolic Computation*, 2006.)

Christiano, Paul, et al. "Deep Reinforcement Learning from Human Preferences." *NeurIPS* (2017).

Cox, David, John Little, and Donal O'Shea. *Ideals, Varieties, and Algorithms*, 4th ed. New York: Springer, 2015.

Doshi-Velez, Finale, and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning." *arXiv:1702.08608* (2017).

Fishkin, James S. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford: Oxford University Press, 2009.

Floridi, Luciano, et al. "AI4People—An Ethical Framework for a Good AI Society." *Minds and Machines* 28, no. 4 (2018): 689–707.

Friedman, Batya, Peter H. Kahn Jr., and Alan Borning. "Value Sensitive Design and Information Systems." In *Human–Computer Interaction and Management Information Systems*, edited by Ping Zhang and Dennis Galletta, 348–372. Armonk, NY: M.E. Sharpe, 2006.

Grice, H. Paul. "Logic and Conversation." In *Syntax and Semantics, Vol. 3: Speech Acts*, edited by Peter Cole and Jerry L. Morgan, 41–58. New York: Academic Press, 1975.

Hart, Sandra G., and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index)." In *Human Mental Workload*, 139–183. Amsterdam: North-Holland, 1988.

Keats, John. "Letter to George and Thomas Keats, 21 December 1817." In *The Letters of John Keats*, vol. 1, edited by Hyder E. Rollins. Cambridge, MA: Harvard University Press, 1958. (Origin of "negative capability.")

Kinneavy, James L., and Catherine R. Eskin. "Kairos in Aristotle's *Rhetoric*." *Written Communication* 11, no. 1 (1994): 131–142.

Lundberg, Scott M., and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” *NeurIPS* (2017). (SHAP)

Mittelstadt, Brent D., et al. “The Ethics of Algorithms: Mapping the Debate.” *Big Data & Society* 3, no. 2 (2016).

Nissenbaum, Helen. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA: Stanford University Press, 2010.

Ouyang, Long, et al. “Training Language Models to Follow Instructions with Human Feedback.” *NeurIPS* (2022).

Rafailov, Rafael, et al. “Direct Preference Optimization: Your Language Model is Secretly a Reward Model.” *NeurIPS* (2023).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “Why Should I Trust You?: Explaining the Predictions of Any Classifier.” *KDD* (2016). (LIME)

Shneiderman, Ben. *Human-Centered AI*. New York: Oxford University Press, 2022.

Simon, Herbert A. “Rational Choice and the Structure of the Environment.” *Psychological Review* 63, no. 2 (1956): 129–138. (Satisficing / sufficiency)

Sipiora, Phillip, and James S. Baumlin, eds. *Rhetoric and Kairos: Essays in History, Theory, and Praxis*. Albany, NY: SUNY Press, 2002.

Strawson, P. F. *Freedom and Resentment and Other Essays*. London: Routledge, 2008. (On reactive attitudes; relevance to trust framing)

Weizenbaum, Joseph. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco: W.H. Freeman, 1976.

Zuboff, Shoshana. *The Age of Surveillance Capitalism*. New York: PublicAffairs, 2019. (Critical context for audit/surveillance defaults)