Doctoral Advanced Module — AI Systems in the Pursuit Discipline

The discipline's doctrine now runs in two renderings — human and machine — and the machine rendering makes claims that can be checked against the market. This seminar turns the AI layer into a research object: what it encodes, how to evaluate what it produces, and where the human stays Accountable.

Level: doctoral (L3 Architect / L4 Networker) · **Format:** one 3-hour seminar per week, five sessions · **Prerequisites:** DL 901 (design and measurement); DL 903 (empirical study design) recommended · **Companions:** doctrine/08, align/concept-to-system-map.md, modules/doctoral/research-agenda.md · **Stack-agnostic promise:** the machine rendering is studied as *concepts* — what each piece claims, and how you would test the claim. No lesson requires operating any tool (doctrine/08).

This module is one of the three C.A.S.E.-adjacent advanced modules the degree crosswalk promises at the doctoral tier (alongside advanced capture management and partnerships-and-ecosystems). Where DL 903 studies the discipline's *empirical machinery* — the composites, the calibration monitors, the golden-thread walk — this seminar goes one level deeper into the **machine rendering itself**: the versioned package of rules, formulas, playbooks, prompts, and retrieval corpus that encodes the doctrine, and the evaluation methodology a researcher must build to decide whether the machine's recommendations can be trusted.

Course description

The pursuit discipline now has an unusual property: its doctrine is written twice. Once for humans (this curriculum, the `doctrine/` files) and once for machines (the DNA corpus — a versioned package the concept-to-system map keeps in lockstep). The two renderings are supposed to say the same thing — "one doctrine, two renderings" — and that claim, like every claim in the discipline, can be tested.

The seminar has three jobs, one per part of the module:

1. **Read the machine rendering as a research object.** What does each kind of artifact — a rule, a formula, a playbook, a prompt, a retrieval corpus — *claim* about the world? The answer is more subtle than it looks: a formula claims its weights are right, a prompt claims a persona instantiates reliably, a retrieval corpus claims its material is faithful and non-fabricated.
2. **Build the evaluation methodology.** How does a researcher decide whether an agent's capture recommendations are any good? The seminar assembles the evaluation program — calibration, discrimination, decision-fidelity, retrieval quality, reproducibility — that turns "the AI made a

recommendation" into "the AI's recommendation is trustworthy enough to act on."

3. **Study the human-accountability frontier.** The most consequential governance design in the machine side is the boundary between what runs automatically and what a human stays Accountable for. The Dream Gate is where that boundary is exercised — a decision structure with trigger, criteria, owner, and a mode that can be advisory or blocking. The Gate is a research object in its own right.

The module is deliberately **stack-agnostic** in the strongest sense: the machine rendering is described as a *conceptual architecture* (rules / formulas / playbooks / prompts / retrieval corpus), not as any product's files. A student who completes this seminar and then meets a completely different implementation of the same ideas — built by a different team in a different decade — would recognize every claim and every test.

Learning outcomes

By the end of this module, a student can:

1. **Explain**, at concept level, the architecture of the machine rendering — the five artifact families, what each claims, and what each hides.
2. **Translate** any piece of the machine rendering into a falsifiable research question about the discipline.
3. **Design** an evaluation study of a machine-produced recommendation — calibration, discrimination, decision-fidelity, retrieval quality — using the DL 901/DL 902 toolkit.
4. **Analyze** the human-accountability frontier — the Gate as a decision structure, advisory vs. blocking mode, the Accountable boundary — and design the study that tests whether the boundary is drawn correctly.
5. **Defend** a seminar research design that sits on the research agenda and is dissertation-ready.

Session 1 — One doctrine, two renderings: the machine as a research object

The claim. The doctrine is now encoded in a versioned, machine-readable package, and that encoding makes testable claims about the world. "One doctrine, two renderings" is a *maintained* claim — the human and machine sides are kept in lockstep by a versioning discipline — and a maintained claim can be verified or falsified.

The architecture, at concept level. The machine rendering has five artifact families, each a different kind of claim:

- **Rules** — decision conditions. A rule says *when a condition holds, take this*

action (a bid/no-bid disposition keyed to a rubric, a pricing floor enforced per contract type, a compliance gate). A rule claims its conditions are the right ones and its threshold is set correctly.

- **Formulas** — arithmetic. A formula computes a number from inputs (a pWin composite, a price-to-win estimate, a readiness index). A formula claims its inputs matter, its weights are right, and its output means what it says.
- **Playbooks** — sequenced procedures. A playbook orders the work of a stage (a capture plan, a proposal-production sequence, a compliance-matrix build). A playbook claims that following the sequence produces the intended result.
- **Prompts** — parameterized instructions that instantiate a persona or a stage operator (a capture-manager persona, a bid/no-bid analyst, a compliance officer, a red-team critic). A prompt claims that the persona it instantiates behaves consistently and in role.
- **Retrieval corpus** — curated material the machine retrieves from at inference time (past performance, agency win patterns, the doctrine itself). A corpus claims its material is faithful, current, and non-fabricated.

Each artifact is a *rendering of a durable concept* (doctrine/08): the concept column never changes, the artifact column is versioned. That is the architecture's central design bet, and it is a bet a researcher can examine.

The hidden assumptions. That the encoding is faithful to the concept (a machine rule can drift from the doctrine it renders); that the versioning discipline actually keeps human and machine in lockstep; that the five artifact families are the right decomposition; that "versioned" means "honest" rather than merely "tagged."

Design exercise. Take one doctrine chapter (e.g., doctrine/05, scoring) and map what each artifact family would encode for it — one rule, one formula, one playbook step, one prompt, one corpus entry. Then rank the artifacts by *checkability*: which claims could a researcher actually test against the market?

Reading. doctrine/08 in full; align/concept-to-system-map.md (the seam that keeps the renderings in lockstep); modules/doctoral/research-agenda.md §2 (the data environment).

Session 2 — Rules and formulas: the arithmetic of the machine

The claim. The machine's rules and formulas — the pWin composite, the bid/no-bid rubric, the pricing floor — encode the discipline's arithmetic, and the arithmetic predicts.

The machinery, at concept level. The pWin composite is a weighted sum of factor scores (past performance, technical capability, price competitiveness, incumbent/relationship advantage, team strength, customer knowledge, strategic

alignment), normalized to [0,1], with recommendation thresholds keyed to the score. The bid/no-bid rubric maps factor conditions to a disposition (bid / bid-with-conditions / no-bid). The pricing floor enforces a margin minimum per contract type. Each is a measurement instrument *and* a decision rule — and the discipline's DL 903 already teaches how to treat a composite as an instrument.

The hidden assumptions. That the seven factors are the right ones; that the weights are right (or that one set serves all agencies and funder families); that factor scores are assigned reliably by different scorers; that the thresholds sit at the right points; that a formula's output means what the recommendation says it means.

The research questions. These are the factor-validation (Thread B) and prediction-calibration (Thread A) questions, now aimed at the *encoded* arithmetic:

- Is the composite calibrated — and in which bands is it over- or under-confident?
- Do the hand-set factor weights match what the data show, and do they hold across agencies?
- Do two scorers score the same opportunity similarly — or is the composite measuring the scorer?
- Are the recommendation thresholds set so the *dispositions* (bid / no-bid) discriminate outcomes, not just the continuous score?

Design exercise. Teams design the **encoded-formula evaluation**: the data (predicted factor scores and dispositions joined to terminal outcomes for a sample of pursued bids), the model (logistic regression of won/lost on the factors), the measurement checks (inter-rater reliability of factor scoring, calibration statistics — Brier, ECE, reliability diagrams), and the survivorship mitigation. The deliverable is a design that answers "is the machine's arithmetic honest?" with data.

Reading. doctrine/05; DL 903 Sessions 2–3 (the composite and the calibration question); DL 901 Unit 4 (measurement).

Session 3 — Playbooks, prompts, and the retrieval corpus: the language of the machine

The claim. The machine's prose artifacts — playbooks that sequence work, prompts that instantiate personas, a retrieval corpus that grounds outputs — shape what an agent produces, and that shaping is measurable.

The machinery, at concept level. Playbooks are structured procedures with named phases (a 14-phase capture arc, a color-team sequence). Prompts are parameterized instructions: a *stage operator* prompt runs one pipeline stage; a *persona* prompt instantiates a role (capture manager, bid/no-bid analyst, compliance officer); a *critic* prompt instantiates an adversarial reader (pink team, red team, black hat, gold team). The retrieval corpus is a curated store the

machine draws on to ground its output — past performance, agency win patterns, doctrine text — and it carries the discipline's integrity standard: documented fact separated from clearly-labeled illustrative material, nothing invented attributed to a real organization.

The hidden assumptions. That prompts instantiate their personas reliably (the same prompt in the same situation produces the same role behavior); that retrieval returns the right material (and that what is retrieved is what is used); that the corpus is faithful and non-fabricated; that a playbook's sequence is what produces its outcome.

The research questions. A distinct thread of evaluation opens here, about *language artifacts* rather than arithmetic:

- **Persona reliability.** Does a capture-manager prompt produce role-consistent behavior across runs and situations — and how would you measure role-consistency?
- **Prompt sensitivity.** When one element of a prompt changes (the evidence it is told to weigh, the persona it instantiates), does the output change in the intended direction — or does wording noise dominate?
- **Retrieval quality.** Does the retrieval corpus return material that is relevant (precision), complete (recall), and faithful to its sources? Information-retrieval evaluation — precision, recall, normalized discounted cumulative gain (nDCG), mean reciprocal rank (MRR) — supplies the measurement vocabulary, at concept level.
- **Output fidelity.** When the machine produces a capture recommendation, is the reasoning it states actually the reasoning behind the recommendation — or a plausible after-the-fact story?

Design exercise. Teams design the **prose-artifact evaluation study**: a content-analysis protocol (two coders scoring machine outputs against a codebook, with inter-rater reliability — DL 901's measurement discipline), a retrieval-quality evaluation (relevant vs. retrieved material, judged against a labeled set), and a prompt-ablation logic (change one element, measure the output difference, preregistered before seeing outputs).

Reading. align/concept-to-system-map.md (the persona/critic/stage rows — the machine's role structure); DL 901 Unit 4 (reliability); the non-fabrication standard from the dissertation sequence (course/doctoral-dissertation-sequence.md Stage 4).

Session 4 — Evaluating machine outputs: the validation research program

The claim. An agent's capture recommendations can be validated — and the validation record is the research program the discipline needs. The question is not "does the AI work?" but "under what conditions, against what benchmark,

and with what observation floor can we say a machine-produced recommendation is trustworthy?"

The machinery, at concept level. Evaluation has three layers:

- **Calibration** — are the machine's probability estimates honest? (A 0.7 recommendation should win roughly 70% of the time; reliability diagrams, Brier score, expected calibration error localize the answer.)
- **Discrimination** — do the machine's recommendations separate good opportunities from bad ones? (Do pursued-then-recommended opportunities resolve better than pursued-then-warned ones?)
- **Decision-fidelity** — do the humans who act on the machine do *better* than the humans who do not? (The comparison is agent vs. human vs. agent-plus-human; the ethical and empirical heart of the program.)

The evaluation substrate is the **prediction-and-outcome record** — predicted recommendation at decision time joined to the terminal outcome. The discipline's **golden-thread verification** (the staged walk of the whole pipeline on a canary company) is its reproducibility pattern: a whole-pipeline evaluation that can be re-run, whose output is a documented artifact.

The hidden assumptions. That a green golden-thread run certifies the pipeline (a stage labeled "pass" can mean "the check ran," not "the outcome was right"); that the evaluation metrics capture what matters; that the observation floor is adequate to separate signal from noise; that "withdrawn" really is a non-win.

The research questions. This is the evaluation-program thread:

- **The validation framework question.** What would a rigorous evaluation of an AI-assisted pursuit pipeline look like — a framework with outcome-based stage criteria instead of health checks? (The golden-thread is the discipline's answer; a dissertation can design the better one.)
- **The benchmark question.** Against what should machine recommendations be judged — expert judgment, the composite alone, a human-plus-machine ensemble?
- **The observation-floor question.** How many terminal outcomes must accumulate before a validation claim about the machine is trustworthy?
- **The drift question.** Does the machine's calibration drift as the market, the corpus, or its own usage changes — and does the drift predictably precede a problem?

Design exercise. Teams design the **recommendation-evaluation study**: the data (predicted recommendations + outcomes + the version of the machine), the headline statistics (calibration and discrimination measures), the benchmark comparison, the survivorship mitigation, and the preregistered decision rule for "this machine is not yet trustworthy."

Reading. research-agenda Threads A and E; DL 902 Movement 3 (calibration) and Movement 2 (logistic regression); DL 903 Session 6 (golden-thread verification as a research pattern).

Session 5 — The human-accountability frontier: the Dream Gate as a research object

The claim. The boundary between what runs automatically and what a human stays Accountable for is the discipline's most consequential governance design — and the Dream Gate is where that boundary is exercised. The Gate is not just infrastructure; it is a decision structure worth studying as carefully as the pWin composite.

The machinery, at concept level. A gate is a decision structure with four parts: a **trigger** (an event that fires it), **criteria** (written in advance), a **decision** (approve / change / reject), and an **owner** (the human or role accountable). The Gate has a **mode**: *advisory* (the machine recommends, the human decides) or *blocking* (the machine waits for the human). The discipline's Accountable boundary — the ADR 0016 idea that humans stay Accountable for irreversible gates — draws the line between what an agent runs and what a person owns. The Gate delivers one canonical approval card across channels, so every approval is a recorded, auditable decision.

The hidden assumptions. That the boundary is drawn in the right place (that the right decisions are the human-owned ones); that human oversight actually improves outcomes (rather than adding friction or rubber-stamping); that gate *mode* changes decisions (that a blocking gate behaves differently from an advisory one); that gate discipline — criteria written before the work — is worth its cost.

The research questions. This is the decision-and-governance thread (Thread D) aimed at the machine's human frontier:

- Does **gate mode** change decisions? A field-experiment or quasi-experimental question: when the same class of decision moves from advisory to blocking (or back), do approval patterns and outcomes change?
- Does **human gating** improve portfolio outcomes — do organizations that gate honestly outperform organizations that delegate more to automation?
- Where is the **Accountable boundary** drawn correctly — which decisions are safe to automate, and which must stay human, on evidence rather than assertion?
- Does the **record of gate decisions** (the approval cards) predict outcomes — is a "change" decision a leading indicator of a pursuit that needed work?

Design exercise. Teams design the **gate-evaluation study**: the data (gate records — trigger, criteria, decision, mode, owner, outcome — over a defined window), the design (comparative case or difference-in-differences around a mode change), the measurement (approval rates, change rates, terminal outcomes), and the ethical protocol (human decisions are sensitive organizational data — consent, confidentiality, provenance).

Reading. doctrine/04 (gates and governance) in full; the Accountable boundary concept as rendered in align/concept-to-system-map.md §4; research-agenda Thread D.

Research-question bank

Each question is dissertation-sized: researchable, feasible, and a contribution the field needs.

1. **Calibration of the machine rendering.** Is the machine's probability-of-win estimate calibrated — and in which bands is it over- or under-confident, and does calibration hold across agencies and funder families?
2. **Factor validity of the encoded composite.** Do the seven factor weights match what the data show, and do they hold across agencies — or is a single weight set a fiction?
3. **Reliability of factor scoring.** Do two experienced scorers (or two runs of the machine) score the same opportunity similarly — or is the composite measuring the scorer?
4. **Persona reliability.** Does a prompt instantiate its persona consistently — and how would a researcher measure role-consistency across situations and runs?
5. **Retrieval quality.** Does the retrieval corpus return material that is relevant, complete, and faithful — and does retrieval quality predict output quality?
6. **The validation framework.** What would a rigorous evaluation of an AI-assisted pursuit pipeline look like — a framework with outcome-based stage criteria instead of health checks?
7. **The benchmark question.** Do machine recommendations, human judgment, or a human-plus-machine ensemble calibrate and discriminate best — and under what conditions?
8. **Gate mode and outcomes.** Does moving a decision class from advisory to blocking (or back) change approval patterns and terminal outcomes?
9. **The Accountable boundary.** Which decisions can be safely automated, and which must stay human — and is the discipline's current boundary drawn in the right place?
10. **Drift and the versioning discipline.** Does the machine's calibration drift as the market, the corpus, or its own usage changes — and does the versioning discipline's lockstep claim hold?

Reading list

The module is stack-agnostic; the reading is concept-first. Named concepts and schools, not pirated material:

- **The doctrine, as the durable object the machine renders:** doctrine/03 (the pipeline), doctrine/04 (gates), doctrine/05 (scoring), doctrine/08 (the two-layer rule). The machine is studied as a *rendering* of these concepts.

- **The seam:** `align/concept-to-system-map.md` — the correspondence table that maps each durable concept to its versioned machine encoding; and the DNA corpus index (`align/dna-corpus-full-index.md`) for the artifact inventory.
- **The empirical-design posture:** DL 901 (measurement and design), DL 902 (statistics and calibration), DL 903 (the machinery as a research object) — the methods this module assumes and deepens.
- **Judgment and calibration:** the research program on judgment under uncertainty (Kahneman and Tversky's heuristics-and-biases program); the statistical concepts of calibration — Brier score, expected calibration error, reliability diagrams; the winner's-curse and reference-class forecasting traditions.
- **Information-retrieval evaluation:** the measurement concepts of precision, recall, normalized discounted cumulative gain (nDCG), and mean reciprocal rank (MRR) — the vocabulary for judging whether retrieval returns the right material.
- **Automation and oversight:** the human-in-the-loop / human-on-the-loop oversight concepts, and the literature on automation trust and reliance — the scholarly frame for the Accountable boundary.
- **Reproducibility:** the golden-thread verification pattern as a reproducible-research pattern, and the replication benchmark from the dissertation sequence.

Dissertation tie-in

This module feeds three places on the research agenda at once. The **validation-framework question** is the methodology contribution — the dissertation that designs "how to know an AI-assisted pursuit pipeline works" with outcome-based stage criteria. The **calibration and factor questions** are Threads A and B — the evidence that the machine's arithmetic is honest. The **gate and Accountable-boundary questions** are Thread D — the evidence that the human frontier is drawn correctly. A dissertation from this module typically lands on one: a validation framework, a calibration study of the machine rendering, or a gate-evaluation study.

The seminar in two sentences

The discipline now has a machine rendering that makes checkable claims — and the machine's recommendations can only be trusted if somebody builds the evaluation program that tests them. This seminar trains the researchers who will: read the machine rendering as a research object, design the studies that validate its outputs, and test where the human must stay Accountable. **The AI layer is a claim about the doctrine — and every claim can be tested.**

Reflection questions

1. Which claim of the machine rendering do you most suspect is *not* true — a factor weight, a persona's reliability, a gate's boundary — and what would your study show if your suspicion were right?
2. For a machine recommendation to be trustworthy, what is the single study you would want to see first — and can you design it with the data you could obtain?
3. If your study is published and a capture executive reads it, what is the one sentence of guidance about trusting the AI layer they should be able to take away?