

The Irreducible Officer

Purpose, Accountability, and AI-Enabled Strategic Judgment

I. The Performance Standard Has Changed

Strategic decisions are increasingly built from AI-shaped inputs. Officers will use AI directly, but they will also inherit its work indirectly in intelligence reports, analytic summaries, planning tools, and staff processes that have already sorted, summarized, and framed what they see. AI already does that work fluently, often so fluently that the framing becomes invisible and a finished assessment can be genuinely strong while the judgment inside it belongs largely to the machine. For most knowledge work, that is a convenience. For an institution whose job is to certify judgment, it is a problem. The finished product, the thing faculty have always read to find the reasoning, no longer reliably contains it.

Picture two officers handed the same hard problem. The first works it alone, reviews the material, builds a frame, and fights the assumptions until the argument closes. The second defines the problem, then drives a set of AI agents against it, assigning them different roles: adversary behavior, alliance dynamics, historical precedent, domestic politics, and other pressures on the decision. She sets them against each other, moderates the disagreement, and synthesizes the result. Her product is faster, and by most measures stronger. Read the two papers cold and you would rank hers higher.

Now ask which officer understands the problem. The papers will not tell you by themselves. AI did not lower the bar for strategic judgment; it raised it, then hid whether the officer cleared it. The second officer may have done the more demanding work, directing machine capability toward a purpose she owns. Or the machine may have handed her a frame she never examined, and she dressed it in the vocabulary of someone who had. The person is still present. The ownership may not be.

The policy questions are real, and NWC should get them right. The college has to decide where AI is allowed, how students disclose it, and which systems NWC trusts with which material. But a school can settle all of that and still certify the wrong thing. Disclosure tells faculty that a student used AI. It does not tell them whether the judgment in the work is theirs. The harder task sits downstream of policy. NWC has to prepare officers who can use machine speed while keeping purpose, reliance, and accountability attached to human judgment.

None of this is new to NWC faculty. They have long read the finished paper against the work around it. Seminar challenge, revision, assumption audits, oral defense, and their feel for the student can all expose a borrowed argument.

But those checks are best at catching the officer who cannot defend the work. The harder case is the officer who does the old work well, producing careful, defensible, unaided analysis that would have passed by prior standards. In the classroom, he looks finished. In practice, he may be a step slow, behind peers, subordinates, and adversaries who pair similar judgment with stronger command of the machine. Certifying the first and missing the second is the gap a good school is built to close.

NWC can lead here. The same shift that weakens the finished product as evidence is what gives the college its chance to build what the rest of professional military education does not yet have: a way to teach and assess

strategic judgment when the work is done with AI rather than around it. The finished product can no longer answer the question that matters. How does a faculty member tell the officer who owns the frame from the one who inherited it?

II. What the Problem Looks Like

Return to the second officer. Her method — define the problem, drive a team of agents against it, and synthesize the result — is the standard NWC should prepare officers to meet. Human judgment supplies purpose, context, and accountability. Machine capability expands what one officer can search, simulate, compare, and test. The work is teaching officers to direct that workflow: to set purpose, test frames, calibrate reliance, expose failure modes, and defend the judgment as their own.

So is it a problem that her product is genuinely superior?

Only conditionally. Whether her method remains an exercise of judgment or becomes a substitute for it is exactly what assessment now has to determine. The failure modes below are where that distinction breaks down.

Frame capture occurs when the model supplies the first plausible frame and the student never achieves enough distance to revise it. The danger is not that the frame is obviously wrong; it may be entirely reasonable. It may simply define the problem too narrowly, privilege one set of interests over others, assume a theory of adversary behavior, or treat a structural constraint as fixed when it should be contested. Once accepted, later revisions improve the answer inside the wrong boundary. The frame capture is invisible inside the final product.

Fluency substitution is easier to miss. AI produces the tone of analytic maturity (balanced paragraphs, caveats in the right places, the measured voice of a considered judgment) and the student mistakes well-ordered language for well-owned reasoning. Researchers studying AI's effects on student cognition have introduced the concept of epistemic confinement to describe this condition: an illusion of competence while operating entirely within AI-constructed analytical boundaries, where the student believes they are thinking independently while the frame doing the work was never theirs (Chow et al., 2026). In strategy, fluency substitutes for deciding. A paragraph that balances every consideration may never identify which risk actually matters most.

Premature synthesis appears when a student asks AI to connect material before doing enough work to know what should be connected. The output links sources, themes, and concepts in ways that feel coherent. But if the student cannot reconstruct why those connections matter, the synthesis belongs to the model. The student has bypassed the developmental struggle of forming a mental map and inherited one instead.

Uncalibrated reliance begins with a reasonable impulse: AI is useful, the output is confident, and parts of the task feel tedious. The problem is that AI performance is uneven in ways that are not always visible from the outside. Tasks that look similar may differ significantly in whether AI helps or hurts. Appropriate reliance requires the student to identify which part of the task they are delegating, what evidence would justify that delegation, and what independent checks are required before discovering the error downstream.

Invisible delegation occurs when the student does not notice which parts of the work they have handed over. Asking for "feedback" may delegate criteria. Asking for "a better structure" may delegate the argument. Asking for "counterarguments" may delegate the range of imaginable objections. Asking for "a more strategic version" may delegate the meaning of strategic. The language of assistance hides the transfer of judgment. The student believes they are working; the model has already done the framing.

Institutional monoculture is the class-level version of the same problem. When many students use similar systems, similar prompts, and similar defaults, the range of strategic frames available to a seminar narrows. AI can create a surface appearance of diversity (different arguments, different structures, different evidence) while reproducing common assumptions at the level of problem definition. Research into the diversity of AI-generated ideas finds that language models aggregate knowledge into a unified distribution in ways that human cognition does not: people exhibit knowledge partitioning, each occupying a distinct semantic region, in ways that independent AI samples do not replicate (Deng, Brucks & Toubia, 2026). Post-training alignment compounds this further, compressing the distribution of outputs toward the statistical center (Murthy, Ullman & Hu, 2024). Pedagogical research on LLM integration in higher education frames this compression as epistemic narrowing: the constraining of students' exposure to diverse, ambiguous, or contested knowledge by tools optimized for convergence and fluency (Vendrell & Johnston, 2026). In a war college seminar, that compression raises a serious possibility: the same systems that help students produce stronger work may also narrow the range of strategic imagination the seminar is meant to develop.

Responsibility laundering is the final danger. A recommendation becomes easier to defend because the model generated it, or easier to soften because the model's language distributes agency. The analysis lands with the apparent weight of objectivity. But AI does not become accountable for the recommendation. The human remains responsible for the final judgment. When the recommendation proves wrong, shaped by assumptions no one examined and optimizing toward a target no one explicitly chose, the question of who chose the frame does not have a satisfying answer.

These failure modes define the standard the second officer's method has to meet. Used well, her method is judgment exercised through a more powerful workflow. The officer using AI well can explain the purpose the agents were serving, the frame that organized their work, the reliance decisions made across uneven outputs, and the judgment she remains prepared to defend.

Faculty can assess how the student represented the problem, how they used or refused AI support, and whether the discipline transfers when the case changes. Frame, reliance, transfer. Those are the observable practices that show whether purpose and accountability stayed with the human.

III. What Finished Work Can No Longer Carry

The finished product still matters. But if the artifact now carries less of the evidence, faculty need to know how much less, and what has to carry the rest. A paper can show structure, balance, strategic vocabulary, and clean prose while leaving the student's actual contribution unclear. A product can be better than an unaided version and still leave faculty unsure who owned the purpose, the frame, the reliance decisions, and the final judgment.

Bastani et al. provide a useful warning. Students using unscaffolded AI tutors improved during supported practice, then performed 17 percent below students without access when the support was removed (Bastani et al., 2025). Tool-assisted performance is real performance. NWC still needs to know whether students have built both the human foundation and the AI-enabled practice: whether they can reason without the scaffold when needed, and whether they can direct the scaffold when it is available.

That is the assessment shift. Faculty need evidence of ownership inside AI-enabled work: purpose expressed through frame, reliance decisions the student can defend, accountability for the final judgment, and transfer to a changed case.

IV. Purpose as the Irreducible Human Act

An AI system can do a great deal of useful work inside a strategic problem. It can generate alternatives, surface assumptions, identify internal contradictions, simulate adversarial objections, and accelerate drafting. What it cannot do is choose the purpose. Selecting what should count as progress, what risks are acceptable, what ends deserve pursuit is a prior act that precedes any optimization. In any AI-enabled workflow, it belongs to a human being who remains accountable for the choice.

AI systems can optimize, rank, recommend, and work toward goals. But someone outside the system sets those goals, accepts them, or lets them govern the work. When the human fails to provide a purpose, provides one too vaguely, or accepts the system's inferred purpose without noticing, a default can govern the work. That is still not the same as authorizing the purpose. The system can operate inside those commitments, but it cannot authorize them. Nor can it absorb accountability for what it produces under their direction. One careful account of AI's normative commitments puts the failure plainly: misspecified values, divergent objectives across stakeholders, and the treatment of optimization as a justification for action are all failures that occur before the system runs, in the specification of what the system is for (Laufer, Gilbert & Nissenbaum, 2023).

At NWC, the practical version of this is the strict prompt. When faculty give students a precisely bounded question to answer, faculty have already made the hard strategic choices. The student is executing inside a structure someone else built, which happens to be exactly the structure AI is best at working inside. What gets bypassed is the part that matters most: the work of deciding what problem to solve, and why, and against what standard.

The implication runs the other way. The student has to frame the problem before the analytic structure arrives, because deciding what problem to solve, why it matters, and what standard should govern the answer is exactly the judgment AI cannot make. The framing is where the judgment lives: in the determination of what the situation requires, whose interests are implicated, what assumptions are doing work, and what kind of answer would actually matter. That is the intellectual work, not a preliminary step before it.

Future AI systems may generate better problem definitions, compare frames more rigorously, and identify strategic errors that humans miss. But a system capable of generating its own problem frames is still generating them toward some purpose, against some signal, in pursuit of some objective that was embedded in its design or inferred from its context. The oracle can only be an oracle if someone has already resolved what winning means. A system capable of reframing may relocate the human obligation, but it cannot eliminate it. As AI becomes more capable of manipulating frames, the purpose-definition requirement becomes less visible, not less real. The more capable AI becomes at absorbing what used to be visible human work, the harder it becomes to locate the human judgment that authorized it, and the more important it becomes to be able to find it.

Purpose-definition also depends on situated judgment. The person responsible for the work has to read local context, tacit institutional knowledge, shifting constraints, and the unease that something in the official framing is wrong. A model may process some of those signals, but it cannot be accountable for what they mean. That judgment is not infallible; it carries biases that can produce creativity or error. But the obligation it carries is one a model cannot assume. The output still has to be read back against a contested world. The same event can mean different things to different actors. Stated positions may be performative, incentives may be hidden, and relevant information may exist only as interpersonal signal or institutional practice that no dataset contains. Situating an output in that world is a human act. AI cannot perform it on behalf of the person who will be held accountable for the result.

The human who defines what the system works toward remains accountable for what the system produces. Purpose and ownership travel together, and frame literacy is the discipline that keeps that bond visible.

V. Appropriate Reliance as a Teachable Competency

The threat runs in two directions simultaneously. AI can perform fluently enough to supply a frame before the student has claimed one, and it can perform unevenly enough that reliance on a confident output leads the work off course. The second of those failures is less discussed and equally consequential.

The educational target is specific. Students need to predict, with reasonable accuracy, when AI performs well for a given type of task and when it does not, and calibrate their use accordingly.

AI improved performance on some tasks and degraded it on others, and the boundary was not obvious in advance. Tasks that looked similar from the outside differed in whether AI helped or hurt, a finding Dell'Acqua et al. describe as the jagged technological frontier (Dell'Acqua et al., 2023). Future leaders will operate along that frontier in every AI-enabled workflow, facing systems capable enough to invite reliance and uneven enough to make reliance dangerous. The educational response is pattern recognition: learning to identify what kind of task is in front of you, where models tend to be strong, where they tend to fail, and what independent checks are required before the output governs the work.

The distinction between trust and reliance matters here. Trust is a subjective disposition, a feeling of confidence that a system will perform well. Reliance is the observable act of accepting its output and acting on it. Appropriate reliance is harder and more discriminating: accepting support when it is warranted, refusing or verifying when it is not, and being able to give an account of the difference (Raees & Papangelis, 2026). A student who trusts AI generally has learned almost nothing transferable. One who has developed a disciplined account of when and why to rely, across task types, risk levels, and domains, has developed something real. Faculty can teach, model, and assess that competency.

The practical implication extends to how students interact with systems, not only whether they use them. Users who can only inspect an AI answer are still downstream of the frame. Users who can change the task, revise the context, reset the criteria, and design the checks are exercising the judgment the assignment is meant to build. The goal is students who can shape AI systems, specifying what they are asking the system to do, why, and what would count as a satisfactory result.

Instructors and students need a simple diagnostic for where human judgment is operating in a workflow. Minimal-context prompting inherits the model's frame almost entirely. Structured prompting with explicit purpose and evaluation criteria moves the human contribution upstream. Reusable workflows with defined review steps, evaluator loops that surface disagreement, and institutional systems built on faculty judgment move it further still. Each step is a step toward greater explicitness about what the human is doing, why, and what they are accountable for.

VI. Friction as Developmental Design

Some work looks inefficient because it is waste; some because it is how judgment forms (Ceccarelli, 2024; Collins, Brown & Newman, 1989). The difference matters enormously for educational design, and AI is very good at removing both kinds without distinguishing between them.

Friction worth removing is real and abundant. Formatting, search, repetitive drafting, clerical assembly: these consume time without building judgment. AI can eliminate them and free student and faculty attention for the

work that actually matters, a gain the design should capture.

Friction worth protecting is less obvious but more important. The struggle to define a problem before having a structure handed to you. The first failed attempt to connect ends, ways, and means that reveals an incoherent argument. The discomfort of defending a claim in seminar that turns out not to survive challenge. The revision that matters not because the final sentence is better but because the student has discovered what the argument actually is. These are developmental events. If AI removes them too early, supplying the first frame before the student has struggled to form one or synthesizing sources before the student has built the mental map to evaluate those connections, it produces a more polished artifact and a weaker thinker.

The aviation automation record is the relevant precedent. Decades of flight-deck automation improved operations measurably: safer flights, more efficient procedures, reduced crew workload. The same period produced skill erosion, mode confusion, and a systematic reluctance to intervene when automation failed. The mechanism was not carelessness. Studies of experienced pilots found that those with more glass-cockpit hours showed measurably reduced manual flight skills and less effective instrument crosscheck: the automation had absorbed the practice that built the underlying competency (Young, Fanjoy & Suckow, 2006). Separately, detailed documentation of incidents on highly automated aircraft showed that experienced pilots failed to track what the automation was doing not from inattention but because the system's behavior had become opaque: the automation was acting in ways the pilots had not commanded and could not predict (Sarter & Woods, 1997). The industry's response was not to reduce automation. It was to design the gap back into training, building deliberate practice for the moments when automation is unavailable, misleading, or wrong.

The PME equivalent is deliberate exposure to AI failure. Students should meet systems that help, systems that tempt them forward too quickly, systems that are partially wrong, and moments when the system is unavailable. They learn when to lean on the tool, when to slow down, and when to intervene by practicing those distinctions before speed forces the choice.

The design principle for NWC follows the same logic. The institution should identify the forms of effort that build strategic judgment and design AI use around them, protecting the friction that matters and removing the friction that merely consumes time. That requires faculty judgment and explicit design. If instructors do not decide which friction matters, AI will decide by default. The same principle has emerged independently in higher-education pedagogy: preserving productive struggle before AI engagement, and sequencing AI-mediated with AI-free phases, are now foundational design requirements for learning environments where AI is present (Vendrell & Johnston, 2026).

The design principle is no garden paths. Good assignments require students to own a frame before they can answer — problems where the template answer is wrong, or where multiple coherent frames exist and the student must defend a choice among them. Those problems force the exploration that builds judgment more reliably than word-count requirements or disclosure policies.

VII. Accountability Is Structural

AI can compress the work before a decision. It cannot own what follows. A system that did not choose the purpose cannot answer for the consequences of pursuing it.

That matters most in national security work, where AI can make a recommendation look settled before the human deliberation behind it is visible. AI can accelerate staff work, but command responsibility cannot be transferred (Andres, 2026). In AI-enabled strategic decision games, Andres describes conviction-shaped output. Players can

produce confident assessments, clear recommendations, and decisive proposals even when the workflow has compressed or bypassed the deliberation that normally earns conviction.

The professional weak point is the officer accepting the system's confidence without owning the reasoning. Red teams, structured analytic techniques, seminar challenge, and institutional review exist because human judgment is fallible. Those checks only work when a human remains answerable for the result.

NWC is preparing officers for organizations that need to see a human own the decision. AI can accelerate analysis, surface options, and model consequences. A human recommendation brings context with it. Experience, incentives, reputation, and the way a person answers when pressed all travel with the recommendation. A commander can weigh those signals. A model carries none of them and can still sound equally confident. It can support a judgment, but it cannot provide the human presence that makes accountability clear to subordinates, partners, or commanders who will live with the result.

First-person ownership matters in the classroom. The student should be able to say why they accepted an output, why they rejected one, and why they remain accountable for the recommendation despite the system's contribution. Students are practicing the accountability structure their professional roles will require.

Frame literacy is the discipline of directing machine capability toward a purpose the human has genuinely owned. A student who owns the frame and uses AI to pressure-test it can exercise more rigorous judgment with AI than alone, while remaining answerable for what the system produced.

VIII. Assessment That Makes Ownership Visible

If finished artifacts carry less evidentiary weight, assessment has to make ownership visible inside the work. The question is whether the student can account for purpose through frame, the reliance decisions they made, the judgment they exercised, and the way that discipline travels when the case changes.

Purpose through frame, reliance, accountability, and transfer. Together they describe what capable AI-enabled strategic judgment looks like in practice. A student who can define the purpose, express it through a defensible frame, calibrate AI support, remain answerable for the judgment, and carry that discipline into a changed case gives faculty evidence the finished artifact cannot provide alone.

Frame evidence makes the student's starting point explicit. Before submitting the final product, the student names the problem frame, the key assumptions, the criteria for success, the evidence standard, and the role AI played in the work. This should be short and specific: a page, not a portfolio. The questions are: why this problem, why this frame, what would change it? The student who answers those questions under questioning, and revises under challenge rather than retreating to the artifact's language, has owned the frame. The one who cannot has not.

Reliance evidence shows what the student did with AI. Students identify which outputs they accepted, which they modified, which they rejected, which they verified independently, and which they withheld AI from entirely, and why in each case. A short oral defense tests whether the student genuinely owns those choices rather than recording them after the fact. Research on oral assessment finds that compared to static written response, it provides a substantially richer picture of student understanding, precisely because it allows the assessor to probe explanations and observe how students reason under follow-up (Theobald, 2021). A student who can defend a reliance decision under questioning has exercised it. A student who cannot has merely disclosed it.

Transfer evidence tests whether the discipline travels. Faculty give students a polished but misframed AI-generated strategic assessment and ask them to diagnose the hidden frame, expose the assumptions, identify

the missing evidence, and articulate the failure point. The student then defends the critique orally and converts it into something reusable — a rubric, checklist, red-team protocol, or after-action note. That final step connects individual learning to institutional learning. The student produces an artifact that another student or instructor could use. The assessment reveals whether the student's judgment has become a transferable practice or remains a one-time response.

Assessments that generate this evidence work for the same reason good assignments always have. Problems where the template answer is wrong, or where the student must defend a choice among coherent frames, put a student genuinely in the work rather than pattern-matching its surface.

Recent work on authentic assessment in AI-mediated learning contexts argues for the same shift from the design side: authenticity cannot be enforced through detection; it has to be redesigned into the structure of the task (Perkins, Roe & Furze, 2024; Mollick & Mollick, 2023). The shift is from what students know to how they apply knowledge, make judgment, and justify choices with AI in the loop. Process transparency (prompts, iterations, rationale) and oral defense make thinking visible in ways that finished artifacts cannot. The same redesign has been independently theorized in higher-education AI pedagogy: aligning assessment with intended cognition, rather than surface output quality, is the necessary response when fluency no longer signals understanding (Vendrell & Johnston, 2026).

This approach increases assessment burden on faculty, and that is a real cost. It is one reason the pilot described in the next section starts small and builds shared artifacts (rubrics, flawed-assessment libraries, oral defense criteria) so that burden distributes over time rather than multiplying independently for every instructor.

IX. A Foundation Pilot

The pilot is a foundation layer, not the full future state. It tests whether students can own a problem frame before they scale judgment through AI. Later exercises should ask students and seminar teams to design, direct, and evaluate multi-agent workflows, using varied expertise and machine speed to test more frames than any one officer could test alone. This first pilot asks whether the human obligation is visible before that complexity is added.

If AI-enabled command requires officers to use speed without surrendering ownership, the pilot gives NWC a way to practice that behavior before operational speed makes the cost real.

The pilot runs on a framework NWC already teaches. The National Security Strategy Primer gives students five elements of strategic logic. They analyze the strategic situation, define desired ends, identify or develop means, design ways, and assess costs and risks. The Primer also makes clear that the work is iterative; assumptions, interests, political aims, and reassessment shape the whole process. The pilot adds no new vocabulary. It uses that one and asks a harder question of it: in an AI-enabled workflow, which parts of the strategic situation and frame stay the officer's to own, and how would a faculty member tell?

The sequence runs inside a single existing assignment where framing is the central demand.

Step one: unaided problem frame. Students produce a short problem frame without AI. They identify the strategic problem, key assumptions, relevant actors, desired ends, possible ways and means, risks, and evidence needed. Faculty score it on completion, not product quality, preserving the developmental friction of initial framing: the work of forming a mental map before receiving one.

Step two: AI challenge. Students bring their initial frame to an AI system with a specific task: identify the assumptions I may have missed, generate alternative frames for this problem, role-play a skeptical faculty member, surface risks or blind spots I have not named. Students record what they accepted, what they rejected, what changed, and why in each case.

Step three: misframed AI assessment. Students receive a polished AI-generated strategic assessment that is competent on its own terms and wrong for the strategic problem. Faculty choose the flaw at the level of frame. The answer might over-optimize for one success criterion, treat one constraint as decisive too early, assume away adversary adaptation, or import the wrong lesson from analogy. Hallucination or factual error would be easier to detect. The harder failure is a frame problem. The analysis is internally coherent and fails because it is grounded in the wrong understanding of the situation.

Step four: diagnosis and revision. Students identify the hidden frame, the assumptions that produced it, the evidence it suppressed, and the failure point. They revise and produce a short final recommendation that reflects their corrected understanding.

Step five: oral defense. Faculty ask students to explain what AI got wrong in the flawed assessment, what AI made easier in their own process, where reliance was appropriate and where they refused it, what changed between their first frame and their final one, and what evidence would change the recommendation. The oral defense is where frame ownership becomes visible, or does not.

Faculty should evaluate the pilot by what they can now observe: who can explain purpose through frame, who can use AI to pressure-test a chosen frame, who can recover from a flawed output under pressure, and who can defend the final judgment. Those observations tell NWC whether the foundation is strong enough to support more complex AI-enabled work later: students designing workflows, testing competing frames, and defending judgment when the system gives them more capability than structure.

X. The Institutional Opportunity

NWC is unusually well positioned to take this seriously. Its graduates will work in national security environments where uncalibrated reliance, invisible delegation, and responsibility laundering are not academic risks. The classroom is the lower-stakes environment where the habits get built: where faculty can engineer controlled failures, let students experience the seductive fluency of a wrong answer, and teach the discipline of slowing down to expose the frame before it governs the work.

NWC faculty already bring much of the strategic judgment this requires. They know how to spot thin reasoning, how to ask the question that surfaces a hidden assumption, and when a student is performing sophistication rather than owning it. The harder task is joining that judgment to AI fluency: enough command of current systems to build and direct AI-enabled workflows, see where they help and fail, and defend reliance decisions under strategic scrutiny. Some faculty may already be near that standard; others can get there with support from colleagues and practitioners working near the edge of current practice. The institutional aim is to build that combined capacity inside the faculty, so the people assessing students can also recognize, model, and improve the work. Making that tacit judgment and emerging AI fluency explicit, designed, and transferable is the work that remains. NWC can do that through prompts, rubrics, flawed-assessment libraries, oral defense criteria, and faculty development sequences that have instructors diagnose the same AI output and compare what they notice. That is how the institution converts faculty judgment and faculty learning into durable institutional assets.

This is the ordinary work of a serious educational institution operating in an AI-enabled environment. The NWC curriculum already exports judgment through graduates, faculty scholarship, seminar practice, wargames, and professional networks. If NWC develops a rigorous pedagogy for AI-enabled strategic reasoning, documents it, teaches it to new faculty, revises it as the technology changes, and shares it with other PME institutions, it will have built something more useful than another AI policy. It has given faculty a way to teach, observe, and improve judgment in the environment their graduates are already entering.

Exported frames carry their assumptions invisibly. A prompt or workflow that embeds a particular theory of adversary behavior, a particular evidence standard, or a particular definition of strategic success will reproduce that frame at scale without the open argument that should accompany institutional guidance. Institutional AI-enabled teaching tools need to surface the assumptions they carry, to be traceable, revisable, and faculty-governed in the same way the assessment design asks students to be.

NWC's graduates will serve in organizations already operating inside AI-enabled decision environments, and many will help lead organizations increasingly shaped by those environments for the next twenty years. Many institutions are managing a policy question about whether and how students may use AI. PME frameworks to date have concentrated there: acceptable-use policy, classification tiers, faculty literacy training, and the infrastructure of responsible adoption (Smith, 2025). That governance work is necessary groundwork, but it leaves the institution stuck at the permission layer while the harder pedagogical work waits. NWC has the specific mission, the faculty depth, and the operational stakes to build a pedagogy: a transferable, rigorous account of what AI-enabled strategic leadership requires and how to teach it. If NWC does that work, it will build a serious model of responsible AI-enabled leadership in PME that other institutions can inspect, adapt, and improve.

XI. Conclusion

The first officer worked alone.

He struggled through the problem, built his frame from scratch, and produced a strategic approach that reflects the effort of that construction. The second built a team of agents, directed their inquiry against competing hypotheses, moderated the disagreement, and synthesized a result. Her product, by most measures, is better.

The stronger product matters. It still leaves the professional question: can either officer account for the purpose the work was pursuing and stand behind the judgment it produced? Did they choose the purpose? If the work is challenged, if a hidden assumption surfaces, if the recommendation proves wrong under changed conditions, can they explain what they chose, why, and on what basis they remain accountable for it?

The second officer orchestrating a set of agents is practicing exactly that competency, provided she defined what those agents were working toward, directed them against that purpose, calibrated her reliance across the uneven terrain of what each model does well, and can defend the result under pressure. That is what frame ownership looks like at scale. An officer who has not built that foundation first produces the same workflow and a different outcome: frame capture, responsibility laundering, uncalibrated reliance compounded across every agent in the loop.

NWC graduates must be able to direct AI-enabled systems toward a plainly owned purpose, calibrate reliance across the jagged frontier of what those systems actually do well, and stand behind the judgment under questioning because the purpose was theirs. That is the standard the operational environment will require. NWC's task is to make it teachable, observable, and repeatable.

Frame the problem. Calibrate the tool. Refuse the garden path. Own the decision.

References

- Andres, R. B. (2026). *AI and leadership: Preparing commanders for machine-speed war* [Unpublished manuscript]. U.S. National War College.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakc■, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122. <<https://doi.org/10.1073/pnas.2422633122>>
- Ceccarelli, G. (2024). Apprenticeship was the point. *Meditations on Tech*. <<https://www.meditationsontech.com/p/apprenticeship-was-the-point>>
- Chow, W. W., Peng, S., Atiq, A., Truong, V., & Guo, M. (2026). "AI enhanced my critical thinking": Investigating the paradox of student perceptions and cognitive offloading in GenAI use. *Pacific Journal of Technology Enhanced Learning*. <<https://ojs.aut.ac.nz/pjtel/article/view/246>>
- Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Lawrence Erlbaum Associates.
- Dell'Acqua, F., McFowland, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Working Paper. <<https://www.hbs.edu/faculty/Pages/item.aspx?num=64700>>
- Deng, Y., Brucks, M., & Toubia, O. (2026). Examining and addressing barriers to diversity in LLM-generated ideas. arXiv preprint arXiv:2602.20408. <<https://arxiv.org/abs/2602.20408>>
- Laufer, B., Gilbert, T. K., & Nissenbaum, H. (2023). Optimization's neglected normative commitments. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. arXiv preprint arXiv:2305.17465. <<https://arxiv.org/abs/2305.17465>>
- Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. arXiv preprint arXiv:2306.10052. <<https://arxiv.org/abs/2306.10052>>
- Murthy, S. K., Ullman, T., & Hu, J. (2024). One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. arXiv preprint arXiv:2411.04427. <<https://arxiv.org/abs/2411.04427>>
- Perkins, M., Roe, J., & Furze, L. (2024). The AI Assessment Scale revisited: A framework for educational assessment. arXiv preprint arXiv:2412.09029. <<https://arxiv.org/abs/2412.09029>>
- Raees, M., & Papangelis, K. (2026). From trust to appropriate reliance: Measurement constructs in human-AI decision-making. arXiv preprint arXiv:2604.23896. <<https://arxiv.org/abs/2604.23896>>
- Sarter, N. B., & Woods, D. D. (1997). Team play with a powerful and independent agent: Operational experiences and automation surprises on the Airbus A-320. *Human Factors, 39*(4), 553–569. <<https://journals.sagepub.com/doi/10.1518/001872097778667997>>
- Smith, B. (2025). Educating the AI-ready warfighter: A framework for ethical integration in Air Force professional military education. *Wild Blue Yonder*. <<https://www.airuniversity.af.edu/Wild-Blue-Yonder/Articles/Article-Display/Article/4219340/educating-the-ai-ready-warfighter-a-framework-for-ethical-integration-in-air>>

-fo/>

Theobald, A. S. (2021). Oral exams: A more meaningful assessment of students' understanding. *Journal of Statistics and Data Science Education, 29*(2), 156–159.

<<https://www.tandfonline.com/doi/full/10.1080/26939169.2021.1914527>>

Vendrell, M., & Johnston, S.-K. (2026). Scaffolding critical thinking with generative AI: Design principles for integrating large language models in higher education. *Computers and Education: Artificial Intelligence, 10*, 100572. <<https://doi.org/10.1016/j.caeai.2026.100572>>

Young, J. P., Fanjoy, R. O., & Suckow, M. W. (2006). Impact of glass cockpit flight training on manual flying skills. *Journal of Aviation/Aerospace Education & Research, 15*(2).

<<http://commons.erau.edu/jaaer/vol15/iss2/5/>>