No. 67
November 2026

The Mitchell Forum

Human Backstop Not Included: What CCA Concepts Must Measure

by Joshua Wilson, PhD

About the Forum

The Forum presents innovative concepts and thought-provoking insight from air and space experts here in the United States and across the globe. In order to afford publishing opportunities for thoughtful perspectives, Mitchell Institute's Forum provides high visibility to writing efforts on issues spanning technology and operational concepts, defense policy and strategy, and unique interpretations of changing geopolitical relationships.

The opinions, conclusions, and recommendations expressed or implied in this article are those of the author and do not necessarily reflect the official policy or position of the U.S. Government, Department of Defense/War, the U.S. Air Force, the U.S. Space Force, or the Mitchell Institute for Aerospace Studies.

Introduction

The case for collaborative combat aircraft (CCA) rests on three propositions: affordable mass, distributed risk, and continued human accountability for what the machines do. The first two are being engineered with enormous care. The third is treated largely as a given, even though it may be the most consequential assumption in the concept. Publicly visible measures of sustained human control or other failsafe modes remain incomplete.

The implicit premise is straightforward. When autonomy behaves unexpectedly, degrades, loses its link, or produces a recommendation the situation no longer supports, a human operator will detect the deviation, understand what is happening, and intervene effectively enough to preserve the mission. Different concepts of operation will assign different fallback tasks to the human; some systems will revert to autonomous contingency behaviors rather than direct human control. The bottom line is that wherever an operator is expected to recognize, judge, re-task, or recover the team, those tasks require a demonstrated competence. Further, the more capable the automation, the harder that human competence is to maintain: the routine cases are handled by the machine, so what reaches the human is, by selection, the unusual, the ambiguous, and the consequential.

Commanding autonomous combat teammates requires these supervisory competencies, which are not fully captured by traditional flight-leadership qualifications. They can vary across operators and be affected by individual factors, such as the amount of supervisory demand each person can manage effectively. These competencies may deteriorate with a prolonged reliance on automation or insufficient degraded-mode practice, even when conventional flying proficiency and medical qualification remain intact. If true, the human fallback assumed in CCA concepts cannot be validated

through airframe performance tests alone. It requires its own task standard, repeated measures of performance, individual baselines, and a sustainment cost. This is a proposition to test—not an assertion that no CCA human-factors research exists.

Publicly available program descriptions have not yet established a fleet-wide, recurrent proficiency standard for that competence or a transparent estimate of what sustaining it will cost. This is not a marginal omission in the ledger; it is an unpriced dependency at the center of the enterprise. It persists for a simple reason: the airpower enterprise has continuous visibility into their machines but almost none into the humans who are supposed to catch them when they fail.

In another recent Mitchell Forum paper, Maj Gen Thomas Taverney, USAF (Ret.), examines when human judgment must remain in the decision loop for military AI, and he defines meaningful human control as ensuring that personnel have "sufficient information, authority, time, and understanding to challenge or reject AI-generated recommendations."[1] That framework establishes where the human must stand. This paper takes up the question that comes after it: whether the human standing there will be able to act. Information, authority, time, and understanding are necessary, but they presume the operator has maintained their readiness to detect, judge, and intervene. Every safeguard in a human-control architecture runs through that person. Human capability must therefore be engineered and sustained as deliberately as the software it supervises.

As mentioned before, the operator-to-CCA ratio may also have a significant effect. Under comparable mission conditions, operators will not all show the same relationship between the number of autonomous teammates supervised and the quality of their decisions. Some may sustain excellent anomaly detection and re-tasking until a threshold is crossed; others may plateau at a lower level, even when the aircraft count is reduced. If these differences persist across repeated trials, a single fleetwide staffing ratio will conceal meaningful variations in capability. The ratio must be established empirically for the actual tasks, autonomy behaviors, and mission environment needed in a peer-level engagement.

The Irony Is Forty Years Old

In 1983, cognitive psychologist Lisanne Bainbridge published a five-page paper in the journal Automatica that became one of the foundational works in human factors.[2] Her argument was simple and structurally inconvenient: automating most of a task leaves the human responsible for two functions—monitoring the automation, which people perform poorly over time, and taking over when it fails, which requires skills that can decay precisely because the automation has displaced their routine use.

The paradox becomes most acute at the moment of takeover. Manual intervention is required because something has already gone wrong. The operator must therefore recognize an abnormal condition, diagnose it, and execute an atypical response under time pressure. In other words, the system demands unusually high human performance at the exact moment its normal mode of operation has done the least to preserve it.

Four decades of research have elaborated on this mechanism. Researchers described the out-of-the-loop performance problem in 1995, linking it to skill loss, degraded situational awareness, vigilance failures, complacency, and the shift from active to passive information processing.[3] Further study later showed that automation can produce mental underload rather than simply relieve workload.[4] The demand curve, therefore, runs in both

directions: prolonged periods of insufficient engagement can be followed, with little warning, by acute overload.

This is not a speculative edge case. It is a durable and repeatedly observed problem in human-automation research.[5]

Current thinking about military AI restates the irony without resolving it. If AI is right 99.9 percent of the time, warfighters will grow comfortable and begin to assume every decision it makes is correct. How can warfighters recognize when the edge case happens in sufficient time and have the awareness to intervene when needed?[6] In his Mitchell Forum paper, Gen Taverney proposed a framework that assigns an on-the-loop role to human operators—supervising and intervening by exception—across a growing share of missions. That is precisely the role Bainbridge described. While research organizations are beginning to measure the dimensions of human-CCA supervision, the unanswered question is how to turn this work into an operational, recurrent standard for the person expected to act when an exception arises.

What Actually Decays

For CCA planners, there is at least one study that initially appears reassuring. In 2014, sixteen airline pilots were asked to operate in a 747-400 simulator at varying levels of automation. Instrument scanning and manual control were reasonably well retained. Stick-and-rudder performance, in other words, held up.[7]

The more important vulnerability was cognitive. Pilots often detected failures, yet failed to cross-check other instruments, diagnose the underlying problem, or prevent the consequences of an unresolved malfunction. Retention of the cognitive skills required for manual flight depended in part on how actively pilots had remained engaged while supervising the automation. The danger is not the pilot whose hands have gone rusty; It is the pilot who can still fly the aircraft but has lost track of what it is doing and why.

That distinction maps directly onto the CCA fallback problem. The critical question is not whether a fighter pilot will forget how to fly the aircraft: it is whether an operator supervising four, eight, or twelve autonomous teammates can recognize when their collective behavior no longer fits the situation, determine why, and then re-plan on a compressed timeline. This skill is supervisory cognition, not aircraft handling—and it is precisely the category of performance most vulnerable to passive engagement with automation.

A Different Kind of Flight Leadership

Traditional flight leadership depends on a capability that is easily overlooked: a wingman is a trained aviator with independent judgment. A flight lead communicates intent, assigns responsibilities, and coordinates with another human who can interpret ambiguity, notice something the lead missed, disagree when necessary, and adapt to an unexpected development. The wingman is not merely another aircraft. The wingman is another decision-maker.

Collaborative combat aircraft alter that relationship. Autonomy may perform selected tasks with exceptional consistency and speed, but an autonomous teammate does not necessarily share a human aviator's tactical understanding, improvisational judgment, or appreciation of mission-level risk. Even when its action is correct within its programmed limits, that action may no longer fit a changing tactical situation. The flight lead must know not only what the aircraft has been tasked to do, but also the conditions under which the automation's interpretation of that task may become unreliable.

This is not traditional flight leadership with a larger formation. It is a partly different supervisory skill: translating commander's intent into machine-executable tasks, anticipating the limits of multiple autonomous agents, calibrating trust without constant second-guessing, prioritizing attention among competing alerts, and reallocating tasks when a teammate behaves unexpectedly. It requires deciding when not to intervene as much as deciding when to step in. And unlike a human wingman, a machine may be unable to explain its reasoning in terms the flight lead can use at the decisive moment.

The distinction in a flight leadership role may be especially important when the human is also flying a crewed aircraft under threat. The supervisory task competes directly with aircraft control, tactical communication, threat recognition, and crew coordination. A controller on the ground may face a different allocation of attention and authority. Both roles deserve analysis; neither should be treated as interchangeable simply because both command CCAs. The relevant metric is not the number of autonomous aircraft displayed on a screen, but the number and complexity of decisions the human must own.

That raises a question the training enterprise should answer before endeavoring to transition: does excellence as a flight lead of human wingmen predict excellence in directing autonomous teammates? Some knowledge, skills, and abilities will transfer. Others will need adaptation, and some may prove new enough to require separate selection, instruction, and proficiency standards. An experienced fighter pilot may still be the ideal CCA commander, but flight hours, tactical qualifications, and a conventional checkride do not, by themselves, validate skills the pilot has not yet been required to demonstrate.

Evidence Is Emerging—But a Standard Is Not Yet Established

The Air Force is not ignoring human-machine teaming. Its experimental effort already includes simulator-based operator studies. Publicly announced flight activity has also put crewed-aircraft coordination with autonomous platforms into practice. A particularly relevant 2026 Air Force Research Laboratory (AFRL) study examined how 15 participants—including fighter pilots and weapons systems officers—selected assets while supervising four simulated CCAs.[8] While it found that observed task-performance differences between professional backgrounds were not reliably distinct in this small sample, the study explicitly cautioned that this did not prove the same competencies transferred across roles or missions.

That result makes the present argument more precise, not less urgent. It shows that operator behavior can be measured in CCA-relevant scenarios and that assumptions about prior aviation experience warrant empirical testing. It does not establish how reliably a controller will detect a rare failure after months of uneventful supervision, how quickly that person can diagnose and recover a degraded team, or how frequently that skill must be exercised to remain current. Demonstrating competent tasking during a short experiment is different from establishing a sustained, deployable fallback capability.

Contingency on Paper Is Not Capability

The military is very good at specifying contingencies. It is less consistent at converting every contingency into a rehearsed and measurable human capability. Millennium Challenge 2002 is often invoked here. In this U.S. Joint Forces Command war game, a red force team led by retired Marine Lt Gen Paul Van Riper used unconventional, low-technology

tactics to inflict heavy simulated losses on the blue fleet in the opening days of conflict. Although the game's relevance to the present situation is tenuous, one of its lessons remains relevant: Blue's difficulties arose not simply from executing known procedures poorly, but from encountering conditions that had been assumed away and therefore insufficiently rehearsed. Plans existed. Practiced responses did not.

Degraded-mode CCA supervision still risks being treated more as a specified contingency than a demonstrated, currency-maintained human capability. Some of the underlying tasks are already being explored experimentally. The remaining distance is from a successful demonstration to a documented recurrent standard, with known training requirements and a cost the Air Force can plan against.

Skills Persist Where a Checkride Forces Them

In 2016, the Department of Transportation Inspector General reported that the FAA could not determine how often airline pilots flew manually.[9] The agency had also failed to ensure that carrier training programs adequately addressed manual flying skills. The regulatory response has largely relied on safety alerts and advisory material encouraging additional hand-flying, including its Safety Alert for Operators memorandum on manual flight operations proficiency, SAFO 17007.[10] The distinction matters, though, as encouragement is not a proficiency standard.

Aviation already understands the converse. Instrument approaches to minimums, engine-out procedures, unusual-attitude recoveries—these skills persist because the consequences of losing automation at a critical phase of flight are taken seriously enough to be written into recurrent evaluation. What is tested is trained. What must remain current is given a currency requirement. Skills that are merely encouraged occupy a different category altogether.

No publicly described, fleet-wide equivalent of a degraded-mode CCA checkride yet establishes an evaluated maneuver, a validated currency interval, and an agreed performance standard for the specific intervention tasks assigned to different operators. That is not to claim that the Air Force is doing no training or research: it is to distinguish exploratory human-machine teaming work from an enduring readiness requirement around which squadrons, training capacity, and force structure can be planned.

What Measurement Would Require

The gap is tractable, and the work required to close it is modest relative to the programs that depend on the answer. It comes down to four requirements for measurement, each building on the one before.

Define the fallback task concretely. Replace the abstract requirement for the human to remain in control with an explicit sequence of observable actions that a CCA controller must execute when autonomy degrades. This includes detecting the anomaly, characterizing it, re-tasking or recalling assets, reallocating effort, and recovering as much of the mission as possible. If the task cannot be written as an evaluable sequence, it cannot be trained or costed with precision.

Establish the decay curve. Measure fallback performance as a function of time since last practice, operating context, and team size. The scaling question is not academic. An O-3 Captain supervising sixteen autonomous aircraft would, by aircraft count, oversee roughly a squadron's worth of teammates.[11] But aircraft count alone is not the true burden. One aircraft may perform several autonomous functions, and some failures demand attention while others can be handled by the system. The experiment must measure how many consequential supervisory decisions remain

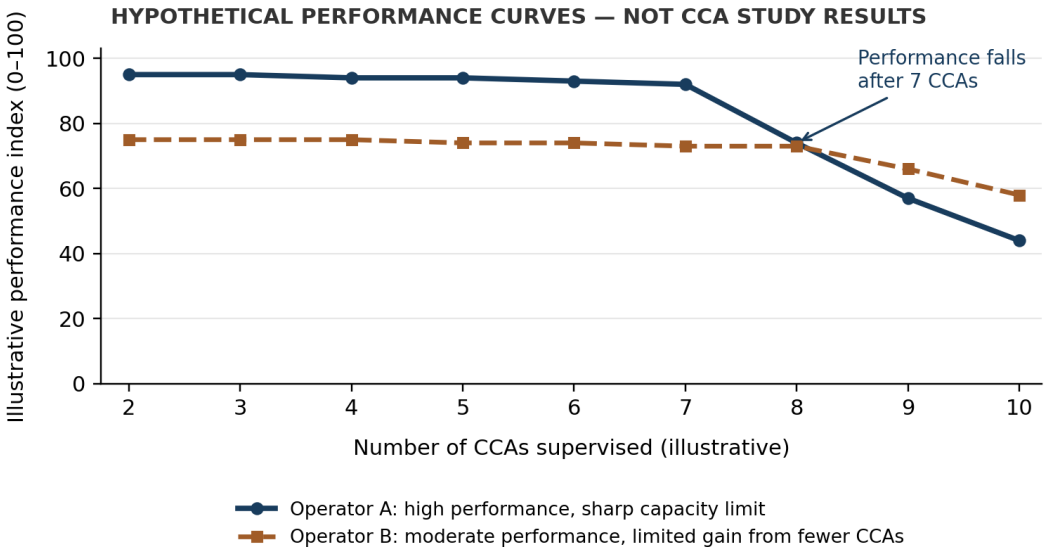


Figure 1: Two hypothetical operator response curves. Operator A demonstrates high proficiency through seven CCAs but declines beyond that span; Operator B remains moderately proficient across smaller spans, with little benefit from reducing the number supervised. All plotted values are invented for illustration, not empirical CCA findings. Aircraft count alone is an incomplete measure of supervisory demand.

with the human and how that burden changes as the team grows. There is little reason to assume the training needed to preserve fallback competence remains constant as supervisory span expands.

Instrument the supervision, not merely the outcome. Research suggests that retention is related to the operator's level of engagement when automation is functioning normally.[12] If the same relationship holds in CCA control, routine supervisory behavior may provide a leading indicator of later takeover competence. That is a testable proposition, and it is measurable in the simulator today.

Build the individual baseline. Decay curves are almost certainly not uniform across operators. Fallback performance should be measured against each controller's own baseline, not only against a fleet standard, so that currency intervals can eventually be set according to the needs of the individual rather than the average. A standard that is right on average is wrong for most of the people it governs.

Capacity Is Not Proficiency

Imagine one operator who supervises seven CCAs exceptionally well, but whose anomaly detection and re-tasking deteriorate when an eighth is added. Now imagine another whose performance is only moderate with eight CCAs and barely improves when the number falls to four.[13] The first operator may have reached a supervisory capacity limit; the second may have a proficiency gap, an interface problem, or another limitation that reducing the number of aircraft does not remedy. These numbers and response patterns are illustrative, not established CCA findings, but they expose a consequential distinction: proficiency at a given task and tolerance for increasing supervisory demand are not interchangeable. A fleet-average human-to-CCA ratio could conceal these two very different operational problems—which demand two different remedies.

No particular CCA limit has been established, and the hypothesis of the illustrative scenario in figure 1 has yet to be tested and proved. But there is reason to expect real variation. Studies have found

that spatial ability and perceived attentional control influenced how operators benefited from automation assistance in a simulated multi-vehicle task.[14] Other work has shown that estimates of multi-vehicle supervisory capacity change substantially when time lost to interaction and recovery of situational awareness is considered.[15] Together with the established individual differences in vulnerability to fatigue and cognitive degradation, the findings make a uniform performance profile an assumption requiring evidence, not a safe default. Moreover, aircraft count is only a proxy for demand: eight teammates operating predictably may require less attention than three that simultaneously generate ambiguous or time-critical decisions.

A useful evaluation would map each operator's performance across increasing and decreasing supervisory loads, while keeping mission conditions comparable and accounting for the decisions the automation actually leaves to the human. It should identify where anomaly detection, diagnosis, judgment, intervention timing, and situational awareness begin to change—and whether reducing demand restores performance. Those measures should be repeated after instruction, practice, and extended periods of reliable automation. The specific test is whether the same increase in supervisory span produces differences in performance across operators under matched conditions, and whether reversing that increase restores performance. The result would be a defensible individual operating envelope, not a presumed universal ratio. That envelope would inform training and task allocation without replacing common minimum safety standards.

Importantly, simulator hours, evaluation events, currency intervals, instructor requirements, and the training-pipeline consequences that follow from them should be treated as program costs rather than externalities. The number belongs in the estimate where the dependency belongs.

These measurements also have an obvious home in the governance structures now being put forward. One proposed feature is a set of governance gates that an AI-enabled capability must pass before advancing to a higher level of autonomy, including a test of the effectiveness of human intervention mechanisms.[16] That criterion requires more than a successful software response or a one-time human demonstration. It requires evidence that the human part of the mechanism works under realistic time pressures and remains effective after ordinary periods of low-demand supervision. A recurrent operator standard could give the gate measurable meaning.

Four Courses of Action the Air Force Should Implement

Course of action one: Define the autonomous team leadership task. Air Combat Command, Air Education and Training Command, the test community, and operational CCA units should jointly produce a task analysis for each planned supervisory role. This need not assume that every fighter pilot will directly pilot a failed CCA; it must begin with the responsibilities actually assigned by the concept of operation for a particular mission. At a minimum, autonomous team leadership tasks should enumerate what the operator must recognize, which decisions are retained by the machine, what authorities permit re-tasking or termination, and how the operator should respond to the loss of link, degraded sensors, conflicting mission priorities, or an autonomous action that is tactically inappropriate but technically permitted. The first deliverable is an observable sequence of decisions and actions, not another broad statement that

humans remain accountable. Subject-matter experts should distinguish tasks transferred from conventional flight leadership from tasks unique to autonomous command. That distinction becomes the basis for syllabi and evaluations.

Course of action two: Measure fallback performance in mission-representative simulation. The Air Force Research Laboratory and operational test organizations should run repeated trials that vary team composition, supervisory span, link quality, automation reliability, and time since the last degraded-mode event. At least some scenarios should begin after a long interval of uneventful, apparently reliable automation; announcing every malfunction in advance would defeat the point. These efforts should measure time to detect an anomaly, accuracy of diagnosis, quality of the subsequent decision, false interventions, correct re-tasking or recall, mission loss, and the time needed to recover useful situational awareness. Next, it is important to record workload, but not to confuse a workload questionnaire or a biometric signal with operational proficiency. Efforts should vary supervisory load within each operator as well as across operators, so that the experiment can distinguish an individual workload limit from a proficiency shortfall that persists even when workload is reduced. Where possible, they should also compare operators with different aviation backgrounds and repeat each person's trials over time. The deliverable is a defensible performance distribution and evidence about which conditions predict dangerous decline instead of establishing an arbitrary fleetwide threshold selected before the data exist.

Course of action three: Turn the results into training and sustainment. Once the fallback task has a defensible standard, Air Education and Training Command, working with Air Combat Command, should evaluate whether routine autonomous-team supervision preserves it or whether periodic scenario injections are needed. They should compare fixed refresher intervals with a performance-based approach in which each operator's documented trend determines additional practice. This should begin with simulator and debrief data already collected for training: decisions, response latency, event recognition, and recovery outcomes. Trainers and sustainers should explore physiological sensing as a possible early warning source, but only if it predicts independently measured performance beyond what ordinary behavioral data provide. Data should be visible first to the operator, with clearly defined access controls and a separation between developmental coaching and medical qualification decisions. A useful output would be a proposed currency event, a training remediation process, and a privacy policy that aviators can understand before participating.

Course of action four: Put human fallback on the acquisition balance sheet. CCA program leadership, requirements authorities, and training commands should require each operating concept to identify the human competencies on which it depends and the evidence supporting those assumptions at defined decision points. These efforts should include pricing the recurrent simulator time, instructor capacity, training devices, assessment personnel, syllabus revisions, operational scheduling, and remediation needed to preserve those skills. The output should be a human-readiness annex that travels with the system's test evidence and life-cycle cost estimate. If a concept assumes one controller can supervise a certain formation, it should show the demonstrated conditions under which that ratio holds, how performance varies among qualified operators, and the additional support required when it does not. If reliable human fallback proves too costly or slow, designers can instead revise the autonomy, alerting methods, or degraded-mode behavior. Either response is more honest than carrying an untested assumption forward without a number.

An initial effort need not wait for a complete operational CCA fleet. Existing high-fidelity simulations can begin to establish the task analysis and baseline today, test surprise events and repeated supervisory exposure, and finally evaluate retention and remediation across subsequent sessions. A small, structured effort could identify which measures deserve larger validation before they become qualification standards. Crucially, the measure of success is not simply a faster intervention. Sometimes the safest choice is allowing a sound autonomous contingency to continue. The standard should reward correct judgment within the human's actual authority, not intervention for its own sake.

The Aircraft Is Watched Continuously; The Operator Is Checked Once a Year

Measuring fallback competence runs into a paradoxical problem: the enterprise of uninhabited vehicles has almost no continuous visibility into the human operators at all. The contrast with how closely these programs monitor their machines is stark.

Modern military aircraft report on themselves constantly. Health and usage monitoring systems on rotorcraft record vibration, loads, and exceedances on every flight. Department of Defense policy on condition-based maintenance directs the services to perform maintenance based on evidence of need rather than on fixed intervals.[17] Fifth-generation aircraft were designed around prognostics and health management, which attempt to predict a failure before it happens. A gearbox is not inspected once a year and presumed sound in between. Performance data is trended, and a maintainer can watch the curve bend long before the part fails.

The operator gets a different model. Aeromedical readiness rests on the periodic flight physical, the annual health assessment, and surveillance of the medical record in between. That architecture was built to answer one question at discrete points in time: is this person fit to fly in general? For disease, it answers that question reasonably well. It was never designed to answer the question CCA concepts now depend on: how is this operator's supervisory performance trending, and how close is it to the edge right now during flight?

The failure modes described earlier in this paper do not appear in a physical. Vigilance decrement, out-of-the-loop drift, and the slow erosion of critical cognitive skills produce no lab value and no abnormal vital sign. A controller can pass an annual physical in excellent health while sitting months into the decay curve. The current system cannot see that. This is not because flight surgeons are inattentive but because the clinic has no instruments able to measure it. Readiness, in the sense that matters for human CCA operators, is a performance state. Performance states are observed in the cockpit and the simulator, not the exam room.

Averages make the blindness worse. Human performance research has shown for decades that operators differ widely, but individuals consistently, in how they respond to fatigue, sleep loss, and sustained monitoring. For example, vulnerability to sleep deprivation behaves like a trait: some people degrade quickly and others barely at all, and that pattern holds across repeated exposures.[18] There is little reason to expect supervisory vigilance to be more uniform. A currency interval set to the fleet average will be too long for some operators and needlessly short for others. An enterprise that knows only the average does not know which of its operators is near the edge.

What would continuous visibility actually look like? Not a new medical examination, and not a camera in the cockpit feeding a supervisor's screen. It would

look more like the maintenance data the enterprise already collects on its aircraft: a baseline established for each operator in the simulator, supervisory decisions and recovery performance logged in ordinary training, and a trend showing whether detection, diagnosis, and re-planning are holding, improving, or sliding. The airline-pilot research cited earlier suggests that how an operator engages while automation is functioning normally may predict later takeover competence. If that relationship holds for CCA supervision, behavioral traces from recurring training could supply part of a leading indicator. What is not yet publicly evident is a standardized system for keeping and interpreting that performance history across the operator's career.

The Air Force has already adopted the right vocabulary. Air Combat Command leads an Optimized Human Weapon System effort across major commands, and performance specialists now work alongside flying squadrons.[19] Much of the publicly visible emphasis has been on physical attributes: neck and back injury, G tolerance, conditioning, and retention. That work is valuable and overdue. But the human capability on which CCA concepts lean hardest includes supervisory cognition, and it has no comparably established, fleet-wide performance-trending standard in the public record. No one would accept a maintenance concept for a new airframe built only on a periodic inspection and a logbook review. The point is not to treat a person like a gearbox; it is to recognize that the enterprise already understands the advantage of identifying degradation before failure and should develop an ethically governed analog for operational cognitive competence.

The same asymmetry runs through the current governance literature. Proposed frameworks require models to be "continuously evaluated for bias, drift, reliability, adversarial manipulation, & operational effectiveness," and are premised on platforms engineered to "degrade predictably, transparently, and monotonically" rather than erratically.[20] Those are the right requirements for the machine. Equivalent precision about the human intervention role is less visible in the acquisition and readiness discussion. The model is monitored for drift. The operator needs an evidence-based method for recognizing and responding to drift in performance as well.

A Medical Model Built to Select Out

The periodic-screening model is not an oversight; It was built deliberately for a particular set of conditions. Military aviation medicine matured when candidates were plentiful relative to cockpits and the overriding concern was the safety of each sortie. Under those conditions, selection is the rational strategy: set high standards, screen hard at entry, and remove anyone who falls below them. Each disqualification cost the Air Force relatively little, because another qualified candidate was waiting. The flight surgeon's central decision became binary: qualified or not.

Those conditions no longer hold. According to recent analysis by the Mitchell Institute, the Air Force's fighter pilot shortfall grew from about 200 in 2006 to roughly 1,000 by 2017 and stood at about 1,150 in early 2025.[21] RAND estimated that producing a single new basic-qualified F-16 pilot cost about $5.6 million and an F-22 pilot nearly $11 million, in 2018 dollars.[22] Moreover, these costs are incurred before the years of experience that make an aviator genuinely valuable. When every trained operator is scarce and expensive, a medical model optimized for removal is optimized for the wrong problem.

The selection model also produces a predictable behavioral response. When the most likely consequence of disclosing a problem is losing flying status, aviators learn

not to disclose. In a 2023 survey of 264 U.S. military pilots, 72 percent reported a history of health care avoidance and 42.5 percent reported misrepresenting or withholding information on a health questionnaire.[23] The sample was self-selected, and the figures should be read with that caution, but the direction matches what I saw as a psychologist sitting across from aviators who were weighing what to say. The phrase "lie to fly" exists for a reason. A readiness model whose main input is voluntary self-report, and whose main output is grounding, is structurally designed to receive bad data.

CCA concepts sharpen the problem. The supervisory controller of an autonomous team is a high-investment operator whose value grows with experience. That is the person who knows how the team behaves when it is behaving well, and who therefore notices first when it is not. Losing that person to a binary decision is expensive. Keeping that person in the cockpit while they conceal a problem is dangerous. The CCA era needs a different objective: sustainment. The aim is to keep trained operators performing in their optimum range for as long as possible, and to detect drift early enough that the correction is small.

Sustainment is not a lower standard. The standards that govern who may fly do not need to move. What changes is that a second layer is added beneath them, aimed at keeping qualified operators qualified rather than discovering at the next physical, or the next mishap, that they no longer are. Trending a gearbox does not lower the airworthiness standard. It makes the standard easier to meet, and it makes failures rarer and cheaper.

Visibility, Not Scrutiny

Any proposal to monitor operators more closely will be heard by the Air Force's aviators as one more way to be grounded. If continuous readiness data flows into the same up-or-out pipeline, the "lie to fly" problem will simply migrate from the medical questionnaire to the sensor. Operators are resourceful, and they will learn to defeat the sensor.

The distinction that matters is that between scrutiny and visibility. Scrutiny is gathering the information an institution needs to decide whether someone should be removed. Visibility is gathering the information an institution needs in order to keep someone performing. The two can draw on similar data, but they are different systems with different owners and different consequences. Visibility must be built into the design, not promised in a policy memo. A few principles follow.

Establishing a model of visibility requires determining data ownership and retention, in writing. Who holds the data, how long it is kept, and who may query it should be settled before the first sensor is issued, not negotiated after the first incident.

The operator must see the data first. Readiness data should reach the operator before it reaches anyone else, and in a form the operator can act on: before a mission, during the mission when appropriate, and in the debrief afterward. A tool aircrew use to manage their own performance is a tool they will protect. A tool used on the aircrew is one they will work around.

This data should be advisory by default. What travels beyond the crew should be at the trend level and advisory. A flight surgeon who receives a sustainment dashboard should be under no obligation to act on it, and a flag on that dashboard should carry no automatic consequence for flying status.

Sustainment data should be separated from the disqualification process except for a small number of safety thresholds defined in advance and published to those being monitored. Operators can accept a rule they can read. They will not accept discretion they cannot see.

Done this way, visibility addresses the disclosure problem instead of worsening it. It replaces a contest over self-report with a shared, objective picture that the operator and the flight surgeon both see. If the governance is wrong, the effort will fail no matter how good the sensors are.

Augmented Cognition, the Second Time

Visibility and sustainment describe what the institution should know about its operators and what it should do with that knowledge. The final step is what the system should do for the operator in the moment. That idea has been tried before, and its history is instructive.

Beginning in 2001, DARPA's Augmented Cognition program set out to build systems that sensed an operator's cognitive state in real time and adapted to it: throttling information during overload, redirecting attention, and shifting tasks between sensory channels. Industry teams applied the approach to the Tomahawk weapon control system, a soldier system, a Marine expeditionary vehicle, and a Joint Unmanned Combat Air Systems (J-UCAS) control station. That last testbed was an operator station for an unmanned combat aircraft, which means the program was asking an early version of the CCA question twenty years ago. Laboratory results were striking. The program was later renamed and wound down without fielding an operational system. That should not be viewed as a referendum on the idea moving forward today. Augmented cognition was not wrong. It was early.

Augmented cognition's own record reveals what it lacked. Three limits stand out: sensors that could not leave the laboratory, models that could not cope with individual differences, and no operational demand urgent enough to force the question. The program leaned heavily on large or complex laboratory instruments. It also used sensors that competed for limited space on the operators' heads, took time to attach and verify, and suffered electromagnetic interference that operational environments would only make worse. Finally, its first-phase integration experiment found that most of the cognitive-state measures it tested worked for some participants and not for others.[24]

Each of those limits can now, at least potentially, be mitigated. Eye tracking, cardiac sensing, and respiratory sensing can now be built into equipment aircrew already wear, such as the helmet, the harness, and the flight suit, without modifying the aircraft. Machine learning methods can build individual models instead of one-size-fits-all gauges, learning a specific operator's baseline and how that operator drifts across repeated sessions. Most importantly, a compelling use case now exists. J-UCAS was only a concept demonstration. CCAs are being bought, and force structure is being designed around a human supervisor.

The right intervention is rarely more information. In my own years crewing the CH-53E, what most often separated the moments when the crew was performing well from the moments when it was not was the amount of information arriving, not its accuracy. A correct fuel call delivered in the middle of a critical radio exchange is accurate and relevant, but badly timed, costing the crew attention it does not have. An augmentation system that understood the operator's state would have held the call for a few seconds and then explained why it had. That is a small behavior, but it is the kind that cannot be designed without knowing the individual operator, and it is the kind that decides whether a human supervising twelve autonomous teammates stays ahead of the team or falls behind it.

The second attempt at augmented cognition should differ in purpose as well as in tools. The first generation of augmented cognition was largely about protecting the

operator from too much information. The CCA problem runs in both directions: overload when the autonomous team degrades, and underload during the long stretches when it works as intended. An augmentation system built only to offload work will accelerate the decay this paper describes. It will make the human's job easier right up to the moment the human is needed. The governing principle should instead be minimum-effective intervention: the least assistance that keeps the operator in the optimum performance range. Sometimes that means stepping in early, before performance has visibly slipped. Sometimes it means withholding help or handing a task back to keep supervisory skill exercised. Success should be measured not only by how the mission goes today but also by whether the operator can still take over tomorrow.

Why This Must Happen Now

None of this is an argument against collaborative combat aircraft. The operational logic is compelling, and the strategic pressure to increase the Air Force's combat capacity is real. But the human dependency must be characterized before force structure hardens around assumptions. This sequencing matters because force-structure decisions are already being made that will be difficult to unwind later. Once aircraft are bought, squadrons are stood up, and manning tables are written against a CONOPS that assumes a cost-free human backstop, identifying the criteria needed to measure operator CCA competency is late to need. Discovering that fallback competence requires substantially more training, staffing, or currency is no longer a minor program adjustment. It becomes a correction to a structure already built on an unverified premise.

The courses of action explored in this paper are not a separate agenda from the CCA program. They address the human half of it. Visibility tells the enterprise where each operator actually stands. Sustainment keeps the operators it has already paid for in the cockpit and performing. Augmentation, executed with restraint, keeps the human engaged enough to remain an effective backstop rather than a nominal one. Each of these is an engineering problem of the same kind the Air Force is already solving for the aircraft: define the requirement, instrument it, model it, and maintain it to a standard.

National policy now requires that AI adopted for national security be "controllable," defined as the ability to monitor a system and take corrective action as needed.[25] That ability does not live in the software; it lives in the person doing the monitoring, and a system is controllable only if its controller is ready.

The machines are getting that treatment, as they should. The operators are still getting a model designed for a different era, a different supply of pilots, and a different job. If airpower is built around a human supervising a team of autonomous aircraft, then the human cannot be an assumption in that system. The human is a subsystem, and it deserves the same visibility, the same maintenance philosophy, and the same investment in getting the interface right.

The autonomy is moving through testing and increasingly explicit standards. The human half of the team needs an equally explicit readiness standard. Until fallback competence is defined, measured, maintained, and priced, the Air Force cannot fully know what it is buying—only what it is assuming the human will still be able to do when the machines need help most. That leaves too much to hope. And you don't win on hope. ✪

Endnotes

1 Thomas D. Taverney, *Machine Learning & Artificial Intelligence in Military Operations:* The Imperative of Human Judgment (Arlington, VA: The Mitchell Institute for Aerospace Studies, 2026).
2 Lisanne Bainbridge, "Ironies of Automation," *Automatica* 19, no. 6, 1983, pp. 775–779.
3 Mica R. Endsley and Esin O. Kiris, "The Out-of-the-Loop Performance Problem and Level of Control in Automation," *Human Factors* 37, no. 2, 1995, pp. 381–394.
4 Mark S. Young and Neville A. Stanton, "Malleable Attentional Resources Theory: A New Explanation for the Effects of Mental Underload on Performance," *Human Factors* 44, no. 3, 2002, pp. 365–375.
5 Barry Strauch, "Ironies of Automation: Still Unresolved After All These Years," IEEE Transactions on Human-Machine Systems 48, no. 5, 2017, pp. 419–433.
6 Taverney, *Machine Learning & Artificial Intelligence in Military Operations*. The 99.9 percent observation and the question about recognizing the edge case in time are his.
7 Stephen M. Casner, Richard W. Geven, Matthias P. Recker, and Jonathan W. Schooler, "The Retention of Manual Flying Skills in the Automated Cockpit," *Human Factors* 56, no. 8, 2014, PP. 1506–1516.
8 Jacob B. Driggs, Megan B. Morris, Mandy H. Ugolini, and Christopher W. Myers, "An Exploration of Statistically Derived Individual Differences in Collaborative Combat Aircraft Operators," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2026.
9 U.S. Department of Transportation, Office of Inspector General, *Enhanced FAA Oversight Could Reduce Hazards Associated With Increased Use of Flight Deck Automation*, AV-2016-013 (January 7, 2016).
10 Federal Aviation Administration, *Safety Alert for Operators 17007: Manual Flight Operations Proficiency* (May 4, 2017).
11 According to the standards of the author's own service experience.
12 Casner et al., "Retention of Manual Flying Skills."
13 Seven and eight CCAs in the example are illustrative counts, not proposed limits.
14 Julia L. Wright, Jessie Y. C. Chen, and Michael J. Barnes, "Human-Automation Interaction for Multiple Robot Control: The Effect of Varying Automation Assistance and Individual Differences on Operator Performance," *Ergonomics* 61, no. 8, 2018, pp. 1033–1045.
15 M. L. Cummings and P. J. Mitchell, "Predicting Controller Capacity in Supervisory Control of Multiple UAVs," IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans 38, no. 2, 2008, pp. 451–460.
16 Taverney, *Machine Learning & Artificial Intelligence in Military Operations*. The governance-gate construct, including the effectiveness of human intervention mechanisms as a gate criterion, is his.
17 Department of Defense, DoD Instruction 4151.22, "Condition Based Maintenance Plus (CBM+) for Materiel Maintenance."
18 Hans P. A. Van Dongen, Maurice D. Baynard, Greg Maislin, and David F. Dinges, "Systematic Interindividual Differences in Neurobehavioral Impairment from Sleep Loss: Evidence of Trait-Like Differential Vulnerability," *Sleep* 27, no. 3, 2004, pp. 423–433.
19 Briel Cook, "Advancing Medical Readiness for Enhanced Lethality," Air Combat Command Public Affairs, April 18, 2025.
20 Taverney, *Machine Learning & Artificial Intelligence in Military Operations*. Both quoted phrases are from his proposed framework.
21 Heather R. Penney, *Want Combat Airpower? Then Fix the Air Force Pilot Crisis* (Arlington, VA: The Mitchell Institute for Aerospace Studies, 2025).
22 Michael G. Mattock et al., *The Relative Cost-Effectiveness of Retaining Versus Accessing Air Force Pilots*, RR-2415-AF (Santa Monica, CA: RAND Corporation, 2019).
23 William R. Hoffman et al., "Self-Reported Health Care Avoidance Behavior in U.S. Military Pilots Related to Fear for Loss of Flying Status," *Military Medicine* 188, nos. 3–4, 2023.
24 Mark St. John, David A. Kobus, Jeffrey G. Morrison, and Dylan Schmorrow, "Overview of the DARPA Augmented Cognition Technical Integration Experiment," *International Journal of Human-Computer Interaction* 17, no. 2, 2004, pp. 131–149.
25 The White House, National Security Presidential Memorandum/NSPM-11, June 5, 2026. The memorandum directs that AI adopted for national security be "reliable, robust, steerable, and controllable," and defines controllability as "the ability to monitor the operation and outcomes of a system and take corrective action as needed."

About The Mitchell Institute

The Mitchell Institute educates the general public about aerospace power's contribution to America's global interests, informs policy and budget deliberations, and cultivates the next generation of thought leaders to exploit the advantages of operating in air, space, and cyberspace.

Forum Submissions and Downloads

For more information about submitting papers or ideas to the Forum, or for media inquiries, email our publications team at forum.mitchellaerospacepower@afa.org

Copies of Forum papers can be downloaded under the publications tab on the Mitchell Institute website at https://www.mitchellaerospacepower.org

About the Author

Joshua Wilson, PhD, is a licensed psychologist and former Air Force officer whose work focuses on human performance in human-machine teams. He served in the Marine Corps from 1999 to 2004 as a CH-53E crew chief, including combat flights in Iraq in 2003, and in the Air Force from 2012 to 2019 as a psychologist, attaining the rank of major with assignments at Lackland, Hurlburt Field, and Edwards Air Force Base. He collaborates with Dunedain Systems on human-performance and AI research, including a submitted Department of the Navy SBIR proposal. His current interests include using AI as a force multiplier in military aviation and health care. He holds a doctorate in counseling psychology from the University of Oklahoma..