The Orchestrator Standard

Professional competence and accountability in directing AI

Mark Sendo | Orchestrator Institute Working Paper | November 18 2026

Draft v0.3 for consultation | Not peer reviewed | Not an adopted standard

This paper proposes a way to assess the people who direct AI systems on consequential work. Its central test is observed conduct: can a person set a defensible goal, choose appropriate systems, verify the result and intervene before suspect output causes harm? The proposed credential would recognize demonstrated competence in a named domain. It would confer no professional licence and certify no system or deployment as safe.

Version note. This revision clarifies the claim for a new credential, strengthens assessment and arbitration rules, defines containment during an investigation, reconciles confidentiality with later review, and revises the pilot design. Every numerical threshold and institutional arrangement remains a proposal. No Board has been constituted or authority granted by this paper.

Abstract

AI oversight depends partly on the competence of the people assigned to it. The EU AI Act identifies competence, training, authority and support as conditions for human oversight of high-risk systems. Research on automation bias, misplaced responsibility and ineffective oversight gives reason to test those conditions rather than assume that appointing a person is enough.

This paper proposes the Orchestrator Standard: seven competencies assessed through real work and an observed practical examination. Each competency has observable criteria and a defined disqualifying failure. Certification would require satisfactory evidence in every competency; strengths would not compensate for weaknesses. A candidate would have to produce an independently verified result or justify a decision not to proceed, while responding proportionately to faults and benign anomalies.

The paper also proposes an independent Certification Board, routes for confidential evidence, continuing assessment and a small pilot with published progression and stop criteria. The Institute would develop the Standard; the Board would decide certification; the Orchestrator Awards would recognize exceptional work separately. The proposal has no demonstrated predictive validity. Its immediate purpose is to invite scrutiny of whether the competence can be assessed consistently, whether an additional credential is useful, and whether independent administration is feasible.

1 The accountability problem

AI systems can contribute to a film, a codebase, a clinical summary or a financial model. A person may direct the work without building the underlying models. That person makes choices about the objective, the systems used, the evidence accepted and the point at which work may reach others. Those choices need scrutiny, particularly where a failure could harm people.

The EU AI Act supplies one legal example. Article 14 addresses the design of human oversight for high-risk AI systems. Article 26(2) requires deployers to assign oversight to natural persons

with the necessary competence, training and authority, together with the necessary support. These provisions do not prescribe the credential proposed here. They also do not make one person's certification a substitute for the provider's or deployer's duties (European Union, 2024).

Three research contributions shape the design. Parasuraman and Manzey's review describes automation complacency and bias among both inexperienced and experienced users. Its published abstract reports that simple practice does not eliminate complacency and that training or instructions alone do not prevent automation bias. That supports testing behaviour under realistic conditions; it does not establish that this proposed examination would overcome those problems (Parasuraman and Manzey, 2010).

Elish's concept of the moral crumple zone describes responsibility being attributed to a person who had limited control over an automated system. It warns against attaching accountability to a job title without examining actual power to intervene (Elish, 2019).

Green's analysis of 41 policy documents questions whether the human oversight those policies envisage can reliably be performed. He argues for institutional oversight rather than reliance on an individual being present. A personal credential could reproduce the problem if organizations used it to excuse weak systems, inadequate staffing or ineffective controls (Green, 2022).

These are reasons for caution as well as reasons to investigate assessment. The proposal must test whether a person can direct work effectively, and whether the setting gives that person usable authority, information, time and support. Responsibility for system design and organizational decisions remains with the parties who control them.

2 What orchestration means

Human orchestration is the direction of one or more AI systems toward a defined outcome, with responsibility for the decisions the person can actually control. It includes setting the goal, selecting systems, allocating work, checking results and deciding whether to correct, escalate, stop or release the work within the person's authority.

Prompting is one activity within this practice. Model development is a related but distinct activity. Reviewing an output after it has already caused an irreversible effect is different from having an effective opportunity to intervene. A signature, a senior title or a promise to take responsibility cannot establish competence by itself.

The proposed assessment asks what the candidate controlled, what the systems did, what evidence the candidate used and why the result was allowed to stand. It also asks what should have triggered containment or escalation. A candidate must acknowledge limits of control rather than claim responsibility for decisions that belonged to others.

The claim is deliberately limited. Orchestration may be an assessable competence within existing occupations. This paper does not establish it as a separate profession. Domain expertise remains necessary, and the assessment must distinguish capable direction from either uncritical acceptance or indiscriminate stopping.

3 Existing instruments and the proposed contribution

The instruments below address related needs through different mechanisms. This is a bounded comparison of selected instruments, not an exhaustive market or literature review. It does not establish that no existing scheme assesses comparable performance.

Instrument	Main purpose and method	Relationship to this proposal
EU AI Act Articles 14 and 26	Legal requirements concerning oversight of high-risk systems and duties of providers and deployers	Supplies relevant obligations and conditions for oversight; does not prescribe this certification scheme
ISO/IEC 42001	Requirements for an organizational AI management system, which may be audited for certification	Addresses management arrangements; this proposal would assess a person's demonstrated conduct within a defined role
NIST AI RMF 1.0	Voluntary guidance for managing AI risk, intended for organizations and individuals	Offers a risk-management framework; it is not the individual certification scheme proposed here
IAPP AIGP	Examination of AI governance knowledge and application, including scenario-based questions	A relevant knowledge credential; the proposed addition is observation of actual direction, evidence and intervention
ISACA AAIA	Advanced AI audit certification with an examination and prior qualifying credentials or designations	Relevant to auditors; this proposal focuses on those directing the work and would need to show added value
Professional licences	Jurisdiction-specific requirements for practice in fields such as medicine, law and engineering	May already cover relevant duties and competence; a domain annex must identify what, if anything, remains to be assessed
Vendor certifications	Varied product-specific knowledge and practical skills assessments	May establish useful proficiency; each scheme needs examination before claims about its limits are made
ISO/IEC 42105	Guidance for human oversight of AI systems; ISO lists it as under publication at this revision	A close comparison for the Standard's content; the proposed credential would require a separate assessment scheme

The proposed contribution is a combination of evidence from real work, an observed practical task, domain-specific criteria and a decision rule requiring competence in every area. Existing credentials may supply part of that evidence or offer a better institutional home. Developing a new credential is justified only if practitioners, affected communities and potential users identify a need that existing arrangements cannot meet adequately.

Consultation should therefore ask two questions separately: is the proposed competence framework useful, and is a new certification body the best way to apply it? A useful framework would not, by itself, establish demand for another credential. Collaboration with an existing professional or certification body remains an option.

The comparison uses public descriptions and the source checks recorded in the references. In particular, the public descriptions of ISO standards do not constitute an audit of their full requirements. ISO/IEC 42105 is described according to its current catalogue status, not as an already published standard.

4 The proposed competence standard

The Standard would define seven competencies. The assessment scheme would specify how evidence is collected and scored. Requirements in this section describe the proposed scheme; they have not been adopted and impose no present obligation on a practitioner.

4.1 The scoring scale

Score	Level	Proposed meaning
0	Absent	Usable evidence is absent, or the work matches a published domain example of performance at score 0
1	Partial	Relevant conduct is shown, but it is inconsistent, inadequately evidenced or dependent on substantive assessor prompting
2	Meets	Consistent evidence meets the domain criteria in both the portfolio and the practical examination
3	Exemplary	All requirements for Meets are satisfied, with additional evidence matching published domain benchmarks for exemplary practice

Score 3 must have domain examples that distinguish it from score 2. Assessors may not award it merely because they admire a candidate's reputation, writing or preferred tools. Published score-1 examples must distinguish partial performance, including delayed containment, from score 0 and from disqualifying failures. A procedural request to locate evidence is not the same as prompting a candidate to perform a missing safety action; assessor guidance must distinguish them.

4.2 The seven competencies

Each competency below states the evidence expected at score 2 and a defined failure that would disqualify a candidate if established through the procedure in section 5. A missing document does not automatically establish dishonesty. The score and any misconduct finding must be explained separately.

Intent

Evidence at Meets. The candidate sets out a brief before substantive AI work begins. It names the objective, affected people, constraints, success criteria, authority and escalation arrangements. Changes are dated and explained. The portfolio shows this practice across the submitted work, and the practical examination requires it before execution.

Disqualifying failure. The candidate knowingly presents a retrospective brief as if it preceded the work.

System selection

Evidence at Meets. The candidate records which systems perform which functions, why they are suitable, what alternatives were considered and which limitations matter to the task. The candidate can explain the division of work and the controls needed for the systems selected.

Disqualifying failure. The candidate deliberately conceals or materially misrepresents the systems responsible for the work.

Verification

Evidence at Meets. A verification plan proportionate to the risk is followed and documented. Checks use evidence independent of the output under review where required, and known uncertainty is stated. Agreement between models may prompt investigation; it is not treated as sufficient proof of correctness.

Disqualifying failure. The candidate claims that a material check was performed when it was not.

Intervention

Evidence at Meets. The candidate detects the planted critical fault or its warning condition and contains suspect output before it crosses the task's predefined risk boundary. Safe diagnostic work may continue inside containment. Correction, override, escalation or stopping is followed by verification before work resumes.

Disqualifying failure. The candidate knowingly permits an established critical fault to cross the defined risk boundary. An assessor or Board panel cannot waive that boundary retrospectively.

Disclosure

Evidence at Meets. The candidate identifies who is entitled to know about AI involvement, limitations and human review under applicable rules and the task's context. Disclosures are accurate, timely and accessible to those people. Required disclosure is supported by evidence of what was communicated and when.

Disqualifying failure. The candidate deliberately withholds material AI involvement or limitations from people entitled to that information.

Accountability

Evidence at Meets. The candidate explains the decisions they controlled, the evidence relied on, the limits of their authority and the actions taken when problems arose. An accurate statement of human direction distinguishes the person's decisions from system behaviour and organizational responsibilities.

Disqualifying failure. The candidate knowingly misrepresents their own decision or conceals it by falsely attributing it to a system or another person. Accurate identification of a vendor's or employer's responsibility is not a failure.

Ethics and harm

Evidence at Meets. Before execution, the candidate identifies foreseeable harms, affected people and the permissions or other lawful bases needed for the work. Mitigations and escalation are evidenced. In the practical task, the candidate recognizes when the work should not proceed.

Disqualifying failure. The candidate knowingly uses protected information without the required authorization or lawful basis, or knowingly falsifies evidence of consent or permission.

Fabricated or materially misrepresented assessment evidence, and knowingly claiming a licence or credential not held, would disqualify under any competency. An error or an adverse outcome does not by itself establish one of these failures; findings require evidence and an opportunity for the candidate to respond.

4.3 The proposed decision rule

Certify. Every competency scores at least 2, and no disqualifying failure is established. The practical task also meets the common outcome requirement in section 5.1.

Defer. One or two competencies score 1, none scores 0, and no disqualifying failure is established. Written reasons identify the missing evidence or conduct. Reassessment would address those competencies on fresh evidence or a new practical task within 12 months, while checking that relevant earlier evidence remains current.

Decline. Any competency scores 0, three or more score 1, or a disqualifying failure is established. The provisional reapplication period is 12 months. Where dishonesty is established, the proposed minimum is 24 months, followed by an evidence-based eligibility decision under a published procedure. Appeals remain available.

Scores would not be summed or averaged in certification decisions. No threshold for a total would exist. The Board would record a decision for each competency so that strong performance elsewhere could never compensate for an unmet requirement.

4.4 Standard setting and domain annexes

The scale, thresholds and waiting periods remain provisional. Before certification opens, a documented standard-setting exercise would examine representative work with practitioners, assessment specialists and public-interest participants. It would consider the consequences of both certifying an unqualified person and declining a qualified one, following established testing practice (AERA, APA and NCME, 2014).

Each domain annex would define acceptable evidence, critical failures, risk boundaries, verified task outcomes and score benchmarks. It could add requirements but could not weaken the seven competencies. The Standard would be published openly, reviewed at least annually and versioned. Every credential would identify its domain, version and validity period.

5 The proposed assessment scheme

Assessment would combine evidence from real work with observed performance. The scheme must establish what a candidate actually did without making access to confidential material a condition that only some employers can satisfy.

5.1 Components and required outcomes

Portfolio. Candidates would submit three pieces of work from the previous 24 months. Each would include the brief, system record, verification evidence, relevant disclosures, statement of direction and risk assessment. At least one would examine a fault, a justified stop or a decision not to proceed. Failure in earlier work is not automatically disqualifying; the assessment examines conduct, learning and the defined failures.

Independent assessment. Two qualified assessors from the candidate's domain would independently score both the portfolio and the practical evidence before seeing each other's scores. They would have no disqualifying connection to the candidate, employer or case. The scheme would record how evidence from both components supports each competency.

Practical examination. A standardized task would last about two and a half hours, followed by a structured interview of about 30 minutes. Each task form would cover all seven competencies and include a critical fault, a minor fault and a benign anomaly. Critical actions would occur in a controlled environment, not in live patient care or another setting where the examination could harm third parties.

Common outcome requirement. The candidate would produce either a deliverable independently verified against the brief or a justified decision not to proceed. A justified stop would identify the unmet requirement, the evidence, and why correction or escalation could not safely resolve it within the stated limits. Indiscriminate refusal or stopping would not satisfy the examination. Failure to meet this requirement would prevent certification; assessors would identify the unmet competency rather than invent a separate total score.

Decision. A Board panel independent of the assessment would apply the published rule to the completed reports. It could return incomplete reasoning for clarification, but could not substitute unpublished criteria or waive a failed competency. Every decision would include written reasons and appeal information.

5.2 Standardization and assessor agreement

Task blueprints would map the examination to the seven competencies and define acceptable outcomes, fault severity, available authority and the point beyond which suspect output must not pass. The risk boundary might be release to users, entry to a shared downstream environment or an irreversible action. It must reflect the task's actual risks and be intelligible to the candidate.

The task bank would use several forms per domain, tested before use and rotated when exposure or model changes compromised comparability. Material system versions and examination conditions would be recorded. Candidates would receive equivalent access to the tools and information needed for the task, with reasonable adjustments that preserve the competence being assessed.

Qualification. Assessors would independently score benchmark cases before appointment. They would have to identify the correct side of every Partial or Meets boundary, every score-0 case and every defined disqualifying failure in the qualification set. A published tolerance could apply to distinctions between Meets and Exemplary. Requalification would occur annually and after material rubric changes.

Arbitration. Any disagreement about score 0, whether a competency meets the standard, a disqualifying failure or a gap greater than one point would go to a third assessor. The third assessor would score without seeing the first two scores. The final ordinal score would be the median of the three. Other one-point differences would resolve to the lower score. A disqualifying finding would require at least two assessors to identify the same defined failure and its supporting evidence, after the candidate had an opportunity to answer the allegation. Dissent would be recorded.

Agreement. The pilot protocol would specify the statistic before assessment, provisionally an ordinal weighted kappa for competency scores, together with raw agreement at the pass boundary and on disqualifying findings. It would report assessor pairings, sample sizes, uncertainty and results that cannot be estimated. The provisional 0.70 target applies to the chosen chance-corrected statistic; it is not evidence of validity or a substitute for agreement on consequential decisions.

Calibration would use benchmark cases; evaluation would use fresh cases not recycled into calibration. Annual drift checks would use authorized, deidentified assessment records or fresh benchmark exercises. Where original evidence could not lawfully be revisited, the report would state that limit rather than claim to have rescored it.

Appeals alleging inconsistent scoring would use new assessors and the same published arbitration rule. Outcomes would be monitored across relevant backgrounds, employers and regions, with disclosure controls for small groups. A small pilot could identify concerns but could not establish fairness across populations.

5.3 Confidential and protected work

The candidate must have authority to show the evidence, and the assessment body must have a lawful basis to receive or examine it. Candidate consent cannot substitute for an employer's authorization, a client's rights or other applicable requirements. Patient-identifiable information, legally privileged material and classified material would never be submitted to the Board.

Route	Proposed method	Evidence limit
Redacted evidence	Authorized material with identifiers and protected details removed; the candidate explains material omissions	Default where independent assessment remains possible
Secure review	Authorized assessors inspect evidence in the organization's approved environment under enforceable confidentiality terms	No source copy leaves that environment; only authorized sanitized findings are retained
Attested evidence	An authorized person other than the candidate attests to specific facts the assessors cannot inspect	Corroborates facts; cannot alone establish the competence claimed
Substituted task	A fresh controlled practical task replaces a portfolio item that cannot be assessed	At most one of the three items; remaining evidence must establish real-work practice

Assessors would record which facts were observed, attested or inferred, and what could not be checked. Where the available evidence was insufficient, the Board would defer or decline under the same competency rules. It would not treat confidentiality as proof of competence or dishonesty.

Confidential material could enter an AI tool only with the relevant information holder's authorization, a lawful basis where required and approval of the specific environment under the Board's data-handling policy. The policy would cover access, onward transfer, retention, model training, security and incident response. Assessors' use of AI would be disclosed and independently checked. Human assessors and the Board would retain responsibility for the decision.

Before the pilot, counsel would review a retention schedule. Source copies held for assessment would normally be deleted after appeals, subject to applicable preservation duties. Necessary scores, written reasons, rubric and task versions, authorization records and sanitized assessment notes would be retained for specified periods to support complaints, audit and renewal. Secure-review sources would remain with their lawful holder. Later review would use

only material still authorized and available; it would not override the original confidentiality arrangements.

The scheme would be developed with reference to ISO/IEC 17024:2026, which includes requirements concerning AI use in certification. This draft does not claim conformity or accreditation, and the public preview cannot replace examination of the full standard.

6 A worked assessment

The candidate, employer and events in this example are fictional. The example illustrates the proposed rule; it supplies no validity evidence.

6.1 Candidate and portfolio

R. leads a software team at a logistics company and applies in the Software domain. The portfolio contains a billing-service migration, customer notices drafted in four languages and a review of an outage caused by an agent-written change. The first two items use redacted evidence; the outage receives secure review. Authorized records show the briefs, division of work, verification, disclosures and lessons from the incident.

6.2 Practical task and risk boundary

R. has two and a half hours to direct AI agents adding currency conversion to a sample codebase. A minor planted fault calls a nonexistent library function. A critical fault in a migration script silently removes records with a missing currency field while the agent's own tests pass. A harmless formatting discrepancy tests whether R. can distinguish a benign anomaly from a substantive defect.

The examination uses synthetic data in a controlled environment. The task instructions require suspect migration output to remain in a private diagnostic area after an unexplained reconciliation mismatch. Passing it to the shared integration stage crosses the defined risk boundary. Investigation within the private area is permitted.

R. writes a clear brief, makes sensible system choices and catches the nonexistent function. A reconciliation check exposes a record-count mismatch. R. logs the problem and asks an agent to investigate, but permits the suspect dataset to enter the shared integration stage. Eleven minutes later, after establishing the cause, R. stops the workflow, corrects the migration and verifies the final deliverable. R. handles the formatting discrepancy proportionately.

In the interview, R. describes the sequence accurately without substantive prompting and explains why the unexplained mismatch should have triggered containment. The defect was not knowingly released as an established fault, but the warning condition was handled too late.

6.3 Scores and decision

Competency	A	B	Final	Reason
Intent	3	3	3	Briefs and changes meet the published exemplary benchmark
System selection	2	2	2	Choices and material limitations are evidenced
Verification	3	2	2	Meets is established; exemplary practice is not consistently shown across the portfolio

Competency	A	B	Final	Reason
Intervention	1	1	1	Suspect output crossed the specified boundary before containment
Disclosure	2	2	2	Required communications are evidenced
Accountability	3	3	3	Decisions and limits are explained with evidence, including the earlier outage
Ethics and harm	2	2	2	Risks, permissions and mitigations meet the domain criteria

The decision is Defer. Six competencies meet or exceed the requirement, and Intervention is Partial. R. detected the warning condition and eventually corrected the work, so the evidence is not Absent. It does not meet the intervention criterion because suspect output crossed the specified boundary. The facts do not establish deliberate deception or knowing propagation of an established critical fault.

R. would receive written reasons and be reassessed on Intervention using a fresh task. That reassessment would test timely containment, proportionate investigation and verified resumption. Safe diagnostic work would remain permissible throughout.

For illustration only, these scores sum to 15 of 21. Seven scores of 2 would sum to 14 and meet the rule. Neither total would be calculated or used in an actual certification decision. The example's purpose is to show why an unmet competency matters more than a flattering aggregate.

7 Institutions and independence

The proposal assigns different responsibilities to three bodies. The Institute would develop and publish the Standard through a standards committee. An independently governed Board would operate certification. The Awards would recognize exceptional work under separate rules. These are proposed arrangements, not evidence that independent institutions already exist.



Proposed separation of standard setting certification and recognition

The author founded the Institute and owns the company operating the Awards. The arrangements must therefore limit his influence as well as the influence of later funders, vendors

and participating employers. Public statements of independence are insufficient if another party controls appointments, essential funding, intellectual property or administrative access.

7.1 Proposed controls

Appointments and removal. An independent nominating committee would select Board members. The author, Institute staff and people with a disqualifying connection to Celestial Technologies or the Awards operator would have no seat or vote on that committee or the Board and would not act as assessors. Removal would require stated cause, a two-thirds vote of the other eligible members and published reasons, with confidentiality and due-process protections.

Defined interests. The charter would specify relevant connections, including employment, ownership, governance roles, paid advice, funding relationships and close personal ties. Material recent relationships would also be assessed under a published lookback policy. The exact definitions and period require consultation and legal review; the phrase "tied to" cannot serve as the operative rule.

Funding. The Board would control its own accounts, budget and spending decisions. The provisional cap is 15% of annual income from any one source, with connected sources aggregated. The funding policy would address in-kind support, employer-paid candidate fees and services supplied below cost. Annual sources and material dependencies would be published. A plan for initial funding must be feasible before the Board launches.

Standard setting. Every version would receive public comment and approval by a standards committee on which Institute-connected members are a minority. The Board could decline to adopt a version it could not assess defensibly. Use of the Standard, the certification name and essential systems would require arrangements that prevent their withdrawal being used to influence individual decisions.

Vendor and employer influence. No more than one Board member could be connected to any one company or controlled group. Members and assessors would recuse from candidate-specific conflicts, including work centred on their employer's systems. Appointment criteria would also prevent a coalition of related commercial interests from dominating the Board.

Training and recognition. The Board would sell no training or examination preparation. The Institute could publish free materials but could not sell preparation tied to this credential. Awards judges and staff could not serve as assessors or Board members in the same assessment cycle. An Awards result would not count toward certification, and certification would confer no advantage in judging the Awards.

External review. An independent reviewer selected through a published, conflict-screened process would examine these controls annually and publish findings. Reviewer appointment and funding would themselves be disclosed. A material unresolved breach would prevent certification from opening or continuing until resolved through an independent process.

7.2 Formation and legal force

The first nominating committee is a particular risk because independence must begin before the first Board exists. The proposal is an open call, with committee members confirmed by at least

two established outside organizations, such as a professional society and a university research centre. No organization has agreed to that role through this paper.

The confirming organizations, legal form, appointment procedure and enforceable protections for assets and services remain to be settled. Counsel should test whether the charter creates effective independence under the chosen legal structure. Founder acceptance of the proposed exclusions must be recorded separately before adoption; this draft does not supply that acceptance.

8 Continuing competence and disciplinary decisions

The proposed certification term is two years. Renewal would require evidence of continuing practice and competence against the applicable version, rather than payment alone.

Requirement	Proposed evidence	Timing
New work	One new portfolio item each year; across each term at least one account of a fault, justified stop or decision not to proceed	Annually
Learning	30 hours over each two-year term, including at least 10 on relevant system changes and at least 5 on ethics and harm	Each term
Authority and support	Evidence and attestation that the holder still has the authority, information, time and support needed for the certified role	Annually and after material role changes
Incident reporting	A serious incident or regulator's finding concerning work the holder directed, with a factual account of their role	Within 30 days of awareness for Board reporting
Version changes	A bridging assessment of material new requirements	At the next renewal under a published transition rule
Practical examination	A fresh observed task with new faults and benign anomalies	Every second renewal, provisionally every four years

Where ordinary work has supplied no suitable fault or stop example, the holder could complete a controlled task instead; the scheme must not encourage invented incidents. The learning hours and all renewal intervals are provisional. Evidence of learning and changed practice would matter more than attendance records alone.

The proposed 30-day reporting period concerns this Board only. It would never delay a shorter legal, clinical, employment or regulatory reporting duty. An initial report could state verified facts and uncertainty without disclosing protected information. Reporting an incident would not itself establish incompetence or justify sanction; deliberate concealment would be investigated.

Anyone affected by a holder's work could complain. Published procedures would specify acknowledgment, investigation, conflicts, response rights and decision time limits. Suspension or revocation would require an evidenced finding under the Standard, reasons and an independent appeal. An interim restriction could be used where a defined risk justified it, subject to prompt review.

A public register would show credential status, domain, version and validity dates. Publication of adverse decisions would follow legal review, due process and necessary privacy protections. The Board would distinguish allegations, interim restrictions and final findings rather than publish them as equivalent.

9 Healthcare as an illustrative domain

Healthcare demonstrates both the possible value of better assessment and the limits of a cross-domain credential. Teams may use AI to summarize records, draft discharge information or support review of images. The people directing those uses need task-specific expertise, workable escalation and effective controls.

A proposed Healthcare AI credential would require the seven competencies and an annex developed with clinicians, patient-safety specialists, privacy experts and patient representatives. It would assess direction of AI within a defined role. It would confer no right to diagnose, treat or perform another reserved clinical activity, and would not replace a licence, clinical judgment, institutional approval or regulatory requirements.

Candidates would have to demonstrate how clinical questions reach appropriately authorized professionals. A person's professional scope of practice would remain governed by the applicable jurisdiction and role. The credential could not enlarge that scope. Its name, register entry and promotional use would have to state this limitation clearly.

Assessment would use authorized redacted evidence, secure review or controlled simulations. Patient-identifiable information would not reach the Board. The pilot could not place patients at risk or require candidates to disclose protected records. Counsel and healthcare partners would review the annex and evidence routes before use.

The same principle applies in other licensed fields. Certification of AI direction cannot substitute for competence or authorization in the underlying profession. Whether such an additional credential would be useful is a question for that profession and the people affected by its work.

10 Recognition through the Orchestrator Awards

The Orchestrator Awards recognize exceptional human direction of AI work. The certification proposal asks whether a person meets a minimum competence standard in a specified domain. The judgments have different purposes and should be administered separately.

Under the Awards model described by the author, entrants provide a work, a statement of human direction and an account of the systems used. Judges consider the human contribution to the result. The Intelligence Stewardship Award is selected rather than entered and may not be bought. These descriptions of the Awards are the author's statements about his own initiative.

Certification would neither be required to enter nor carry weight in judging. Awards results would not supply certification evidence; underlying work could be assessed only through the same evidence and conflict rules applied to every candidate. Under the Awards' stated conflict rule, Celestial Technologies and its work cannot enter or win.

The administrative separation would include decision makers, assessment records and marketing claims. Neither body could imply that success in the other established competence, safety or merit. Exceptional achievement and minimum competence each require their own evidence.

11 Objections and limits

11.1 Certification could legitimize ineffective oversight

An employer could appoint a certified person, deny them time or authority, and use the credential when something failed. The design therefore assesses actual control and support, requires evidence when the role changes, and prohibits using the credential as proof that a system is safe. Those measures are untested. Institutional oversight and accountability for organizational decisions remain necessary.

11.2 Orchestration may be a competence within existing professions

That is compatible with this proposal. A competence framework need not establish a new profession. Domain practitioners should help determine whether the assessment belongs within an existing licence, credential or employer qualification instead. A separate institution is an option to justify, not a conclusion assumed here.

11.3 Another credential may add cost without value

Observed work and published decisions may be useful, but those features do not prove that users need another badge. Consultation and follow-up must examine whether the assessment supplies information that existing arrangements lack, changes practice and is worth its cost. A high or low pass rate alone cannot settle that question.

11.4 Cost and access may distort who can qualify

Portfolio preparation, observation and expert scoring require time. Fee waivers should be available according to demonstrated need, including unsupported practitioners, students and nonprofit workers. Funders could support access but not decisions. Tool access, language and disability adjustments also require attention. The scheme should not open if it can function only for a narrow group of sponsored candidates.

11.5 Rapid system change may invalidate the assessment

The competencies concern conduct, but the tasks used to observe it depend on changing systems. The Board would record relevant versions, review task exposure, monitor drift and require bridging where material criteria changed. A credential earned under one version must not imply competence with every later system or use.

11.6 The founder may benefit

The author could gain reputation and Celestial Technologies could benefit commercially. Exclusion from decisions, independent appointments and transparent funding would reduce specific routes of influence. Their effectiveness depends on enforceable arrangements and external scrutiny. Naming the conflict does not remove it.

11.7 What the paper does not establish

No evidence yet shows that passing this assessment predicts better performance or fewer harms in practice. The initial pilot would examine feasibility and preliminary reliability. Later validation would need evidence about real performance, alternative explanations and the consequences of using the credential.

Accreditation or official recognition could be claimed only after the relevant body had actually granted it, and only within that grant's scope. No credential would be described as equivalent to a professional licence. No payment, donation or sponsorship could buy a decision, faster treatment or a Board seat. Any eventual credential would record competence demonstrated at assessment, not guarantee that subsequent work is safe, lawful or correct.

12 Pilot design and progression

The roadmap has four phases: consult on the Standard; establish independent governance; run and report a pilot; and consider opening certification. No launch date is proposed. Each phase depends on evidence and a published decision before the next begins.

12.1 Initial feasibility pilot

The planning cohort remains 24 to 30 volunteers across two domains, provisionally 12 to 15 in each. One domain would involve a licensed profession and the other would not. Selection requires willing partners, safe tasks and usable evidence routes. The pilot would charge candidates no fee and would award no public certification.

The staffing assumption is eight assessors, four per domain, with two additional assessors available for conflicts or arbitration. About 12 assessor-hours per candidate across both primary assessors, plus three hours for proctoring and interview, produces 360 to 450 direct assessment hours. This excludes task development, training, calibration, arbitration, administration, data handling and external review. Candidate time is provisionally 20 hours, mainly portfolio preparation. Six months from first assessment to reporting remains a planning assumption.

Before recruitment, a published budget would cost those activities, state cash and in-kind contributions, identify funding and provide a contingency. Hourly rates and the resulting total are not estimated in this paper because no staffing or funding arrangements have been established. Proceeding without that budget would leave the feasibility question unanswered.

The cohort is too small to establish population validity, demographic fairness or equivalent difficulty across domains. Recruitment should deliberately include different, independently characterized levels of relevant experience, while acknowledging that experience is not itself a gold standard for competence. Participants must understand the research and assessment purposes and the limits on public reporting.

12.2 Protocol and measures

Before recruitment or access to candidate assessment evidence, the Board would register a protocol specifying recruitment, task allocation, assessor pairings, benchmark cases, fresh evaluation cases, missing-data handling and progression criteria. It would name the agreement

statistic and weighting, methods for uncertainty estimates and handling of small or unestimable groups. Any later change would be explained publicly.

The report would include competency-level agreement, agreement at the pass boundary, disqualifying findings and how arbitration changed decisions. It would record detection of critical and minor faults, unnecessary interventions, verified deliverables, justified stops and insufficient evidence. Results would be reported by task and domain where disclosure was safe.

Operational measures would include candidate and assessor time, cost per candidate, use of each confidential-evidence route, adjustments, appeals and the usefulness of written reasons. Two independent observers would review the process with access sufficient to evaluate it under the confidentiality rules; they would not receive protected source material merely because they were observers. Each would publish a report.

12.3 Stop and revision criteria

The pilot would pause affected assessment immediately if protected information was exposed, a task created an uncontrolled risk, or an independence safeguard was materially breached. Investigation and any lawful notification would precede a decision about resumption. An unresolved breach would prevent progression.

The following would prevent progression to public certification and require publication of the finding:

- Assessor agreement on any competency remains below the provisional 0.70 target after at most two recalibration rounds and evaluation on fresh cases, or consequential disagreements remain unresolved.
- A task cannot distinguish justified action from indiscriminate stopping or produces unexplained differences across task forms. Detection above 90% or below 10% would trigger investigation of task difficulty and recruitment, not automatically establish failure.
- More than one-third of candidates cannot provide sufficient evidence through the permitted routes, or the routes cannot protect information while supporting a defensible decision.
- The cost and access plan cannot support an unsponsored practitioner, including credible provision for waivers, or essential resources remain unfunded.
- External review finds an unresolved material breach of the independence arrangements.

These numerical criteria remain provisional. The protocol must explain how uncertainty affects a decision rather than treat a noisy point estimate as conclusive. An unestimable agreement statistic would not establish that its target had been met; the protocol must specify the additional evaluation required. Task or rubric revisions would need fresh evaluation. An inconclusive or stopped pilot is a publishable result; it is not permission to lower a requirement until the cohort passes.

12.4 Evidence required before opening

Opening would require an independent governance review, a funded operating plan, legal review, completed standard setting, a tested scheme and a published pilot report explaining remaining limitations. The Board would make a recorded progression decision against the

preregistered criteria. Accreditation could be sought separately; intention to seek it would not establish accredited status.

13 Separate publications and unresolved decisions

The paper brings together three proposed documents for consultation. They would be published separately, with clear ownership and version history, before public certification opened.

Document	Contents	Proposed owner
The Standard	Competencies, domain criteria and score benchmarks	Institute through the standards committee
Assessment Scheme	Evidence routes, examinations, decision rules, assessor qualification, renewal and appeals	Certification Board
Board Charter	Legal governance, appointments, independence, funding and review	Independently constituted Board

A provisional scheme and charter must be available before the pilot so that its procedures can be scrutinized. Final versions would incorporate consultation and pilot findings. The first paragraph of an adopted document would distinguish binding requirements from guidance; the working paper remains a proposal throughout.

Three decisions remain explicit. The founder's acceptance of exclusion from Board, assessor and nominating roles must be recorded separately. Numerical assumptions, including the 15% funding cap, 30 learning hours over two years and pilot cohort, require consultation and evidence. Legal and source-review work must be completed to the level claimed before publication or implementation. Rewriting the proposal does not settle those decisions.

Consultation should also address who would use the assessment, whether an existing institution should operate it, who could independently confirm the first committee, how startup costs could meet the funding controls, and what evidence would justify progressing beyond a feasibility pilot. Comments may be directed through the Institute's contact page at oiail.org.

Appendix A Proposed Board charter

This charter is a consultation proposal. It would take effect only after the legal arrangements, independent appointments and adoption process were completed. Read it with sections 4, 5, 7, 8 and 12; any inconsistency must be resolved in the adopted versions before assessment begins.

Purpose. The Board would assess individuals against a published Standard and scheme, grant and renew credentials, and investigate complaints under defined procedures. It would not certify systems, organizations or deployments as safe.

Independent authority. Neither the author, the Institute, the Awards operator nor a funder would determine an individual decision or remove a member. These parties could submit public comments on standards under the same process as others. They would hold no veto over assessment or appeals.

Legal and operational control. The Board would control its budget, records, register and essential administration under enforceable arrangements. Rights to use the Standard,

certification name and essential systems must not permit an outside party to influence candidate decisions by threatening withdrawal. Counsel would confirm the legal form and enforceability before adoption.

Nominations. A five-person independent committee would choose members through published criteria. The first committee would be confirmed by at least two established outside organizations after an open call. Later committees would be drawn from outside the sitting Board. The confirming procedure and conflict screening would be published.

Membership. The provisional Board has nine members, including at least five practitioners from different domains and at least two people outside the AI industry with relevant public-interest expertise. No company or controlled group could have more than one connected member. Terms would last three years, staggered, with no more than two consecutive terms. Sector balance would prevent related interests from controlling the Board.

Exclusions and interests. The author, Institute staff and people with defined disqualifying connections to Celestial Technologies or the Awards operator could not serve on the Board, as assessors or on the nominating committee. Interests, material recent ties and recusals would be recorded under a published policy. The lookback period and exact definitions remain for consultation and counsel.

Removal. Removal would require stated cause and a two-thirds vote of the other eligible members. Reasons would be published to the extent consistent with legal duties and procedural fairness. No funder or outside body could remove a member. Rules for vacancies, quorum, recusal and deadlock would be completed before adoption.

Funding. The Board would hold its own accounts and operating budget. The provisional single-source limit is 15% of annual income, aggregating connected sources and addressing in-kind support and employer-paid fees. Sources and dependencies would be disclosed annually. Waivers would be available according to demonstrated need. Startup funding would require a published plan consistent with the adopted controls.

Assessment and decisions. The Board would appoint and qualify independent assessors and apply the published scoring, arbitration and decision rules. Every decision would have reasons. Fees, donations, training purchases and Awards results could not affect priority or outcome. The Board would sell no training or examination preparation.

Appeals and discipline. A three-person panel with no part in the original decision would hear appeals. Inconsistent-scoring appeals would receive fresh assessment under the same arbitration rule. Candidates and holders would receive notice, access to relevant reasons and an opportunity to respond, subject to lawful confidentiality limits. The panel's decision would be final within the Board, without restricting external rights.

Confidentiality. Section 5.3 would govern all evidence routes. The Board would publish an approved retention schedule, access controls and incident procedures before the pilot. Protected source material would not be made public in decisions or given to observers. Later audits could use only evidence still authorized and available.

Standards and renewal. New versions would require public comment and approval by a standards committee with an Institute-connected minority. The Board could decline a version it

could not assess. Transition, renewal and re-examination rules would be published before they affected candidates or holders.

Public accountability. Annual reporting would cover applications, certification outcomes, renewals, complaints, adverse findings, appeals, assessor agreement, costs, funding and external review. Small-group confidentiality would be protected. An independently selected reviewer would examine governance annually. Material unresolved breaches would prevent opening or continuing certification.

Adoption and review. The charter would be reviewed publicly every two years. Adoption and amendment procedures must preserve independence and be legally enforceable. No founder acceptance, appointment, funding commitment or accreditation is created by this draft.

Disclosure and source verification

The author is the founder of Celestial Technologies LLC, which operates the Orchestrator Awards. He also writes the Institute's working papers. The Institute's legal name and status require confirmation by counsel before an external publication represents them as settled. The author and his company could benefit reputationally or commercially if the proposal were adopted.

The exclusions described here are proposed safeguards. They do not establish that an independent Board exists, that the founder has formally accepted the charter, or that outside organizations have agreed to participate. Statements about the Institute and Awards are the author's descriptions of his own initiatives and have not been independently audited for this paper.

AI systems assisted with drafting and critique. The author remains responsible for the text, the evidence claimed and any decision to publish. Source checking for this revision used official public pages, published abstracts and selected passages from accessible original texts, as specified below. No comprehensive full-text verification, exhaustive credential search, legal opinion or conformity assessment is claimed. Claims relying on abstracts or public descriptions remain limited to what those sources support.

Before publication, the author should confirm quotations, substantive interpretations and the accuracy of this verification record. Any remaining limitation should stay visible. Before implementation, counsel should review the charter, healthcare annex, evidence-handling arrangements, reporting rules and disciplinary publication procedures.

References and verification record

Research and assessment practice

Parasuraman R and Manzey D H 2010. Complacency and Bias in Human Use of Automation An Attentional Integration. Human Factors 52(3), 381–410. DOI

Verification basis. Publisher's abstract and bibliographic details. The full article was not reviewed for this revision. Section 1 uses only the findings stated in the abstract.

Elish M C 2019. Moral Crumple Zones Cautionary Tales in Human Robot Interaction. Engaging Science Technology and Society 5, 40–60. DOI

Verification basis. Journal abstract and bibliographic record. The complete article was not reviewed for this revision; the paper uses the author's published definition of the concept.

Green B 2022. The Flaws of Policies Requiring Human Oversight of Government Algorithms. Computer Law and Security Review 45, 105681. DOI

Verification basis. Accessible original article, including the abstract, description of the 41-document policy review and discussion of institutional oversight. This was a check of relevant passages, not an independent replication of its analysis.

Accessible author copy

American Educational Research Association American Psychological Association and National Council on Measurement in Education 2014. Standards for Educational and Psychological Testing.

Verification basis. Official open-access edition, with relevant passages on cut scores and standard setting, especially Standards 5.21–5.23. The entire volume was not reviewed.

Official open-access edition

Legislation frameworks and standards

European Union 2024. Regulation EU 2024/1689 laying down harmonised rules on artificial intelligence.

Verification basis. Relevant Article 14 and Article 26 passages in official sources. This paper does not determine the applicability or commencement of duties in a particular deployment and is not legal advice.

EUR Lex regulation record

European Commission Article 14

European Commission Article 26

ISO/IEC 42001 2023. Information technology Artificial intelligence Management system.

Verification basis. ISO's official public description; full standard not reviewed. It supports the organizational management-system comparison, not a claim that the standard omits personnel competence.

ISO catalogue

NIST 2023. Artificial Intelligence Risk Management Framework AI RMF 1.0.

Verification basis. Official online executive summary and Core excerpts. The comparison concerns voluntary guidance for organizations and individuals, not an individual credential.

NIST executive summary

ISO/IEC 17024 2026. Conformity assessment General requirements for bodies operating certification of persons. Third edition March 2026.

Verification basis. ISO's official public catalogue description. The full standard was not reviewed. The paper claims neither conformity of the proposed scheme nor accreditation.

ISO catalogue

ISO/IEC 42105. Information technology Artificial intelligence Guidance for human oversight of AI systems.

Verification basis. Official catalogue abstract and lifecycle status checked for this revision. Listed under publication at stage 60.00, entered October 22 2026. Detailed provisions and final published text were not reviewed.

ISO project record

Existing credentials

IAPP. Artificial Intelligence Governance Professional AIGP certification.

Verification basis. Official programme description and examination FAQ. This is an exam-based governance credential, including scenario-based questions. No complete examination handbook or question bank was reviewed.

IAPP AIGP

IAPP examination FAQ

ISACA. Advanced in AI Audit AAIA certification.

Verification basis. Official eligibility requirements and examination content outline. Qualifying prior credentials or designations are required; not all are audit credentials. No complete examination handbook or question bank was reviewed.

ISACA eligibility

ISACA examination outline

Related Institute papers

The following titles were supplied in the earlier draft. They are related work by the same Institute, not independent evidence for the need, validity or independence of this proposal. Their full texts were not reviewed for this revision.

Orchestrator Institute. The Legitimacy Vacuum Why AI Native Cinema Needs Public Methodology Before Awards. DOI

Orchestrator Institute. Judging AI Native Work Gates Evidence and the Limits of the Machine. DOI