AIES 2026, Malmö, Sweden

AAAI / ACM conference on ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

What Does the Credential Still Certify?

Cognitive Stewardship for AI-Mediated Education

Kai Yao

School of Informatics, University of Edinburgh kai.yao@ed.ac.uk

30	180	73%	2.75/8
universities with verified official guidance	policy cases: 30 universities × 6 scenarios	policies clearer about AI use than evidence	mean safeguards found per policy

MAIN FINDING **Most university guidance defines what AI may do more clearly than the evidence needed to support a grade or credential.**

01 THE VALIDITY PROBLEM

A polished submission may hide which cognitive work the learner performed. The policy question therefore extends beyond misconduct: what can the institution still infer about the learner?

Educational delegation means transferring a cognitive operation from the learner to an AI system. The delegated work may include search, drafting, debugging, or checking. The same AI use may support access or replace the target skill. Assessment validity asks whether the submitted work still supports the capability being assessed.

02 COGNITIVE STEWARDSHIP

Institutional responsibility for preserving a defensible link between assessed work and the human capacity named by a credential.

1. **Learning claim**
   The human capacity that the assessment or credential promises.
2. **Delegation boundary**
   The cognitive work that AI may perform in this task.
3. **Evidence standard**
   Observable work demonstrating the learner's own competence.
4. **Safeguards**
   Protections for privacy, access, fairness, appeal, and workload.

All four elements must support the same claim. Permission alone cannot establish learner capability.

03 POLICY AUDIT

Corpus. 30 universities across five English-speaking countries. A policy package combines each university's official public guidance on generative AI and assessment.
Scenarios. Six cases test access support, feedback on work, planning or debugging, task substitution, checking AI output, and programming.
Scoring. Four open-weight language models independently scored each package with the same rubric, using only its text. Scores are averages across models. Standard deviation (SD) shows model disagreement.

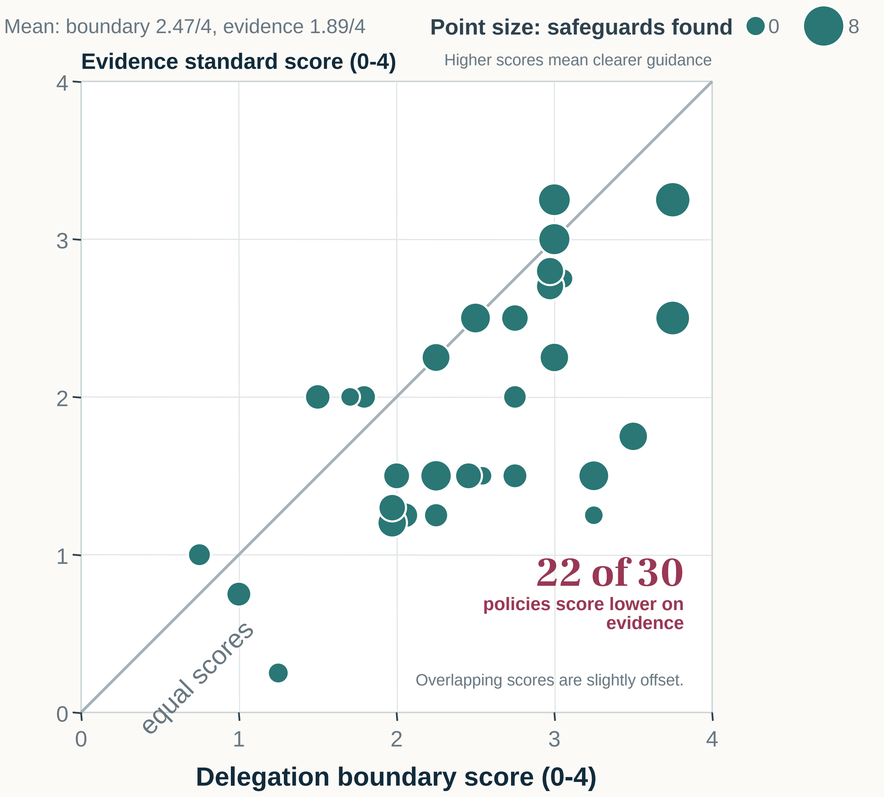
Models: Qwen3 30B-A3B, GPT-OSS 120B
Llama 4 Maverick, Mistral Large 3

Model disagreement (mean SD): boundary 0.57, evidence 0.53, learning claim 0.42.

Scope: the audit describes visible public guidance. It does not measure classroom practice or learning outcomes.

04 BOUNDARIES ARE CLEARER THAN EVIDENCE

Mean: boundary 2.47/4, evidence 1.89/4



Most points lie below the equal-score line. 73% of packages define AI-use boundaries more clearly than the evidence that remains.
This leaves an evidence gap. Permission categories tell students what is allowed, but may leave assessors unsure what the work demonstrates.
Safeguards accompany stronger designs. Larger points tend to sit farther right and higher. Safeguards are more visible where boundary and evidence language develop together.

05 FOUR POLICY PROFILES

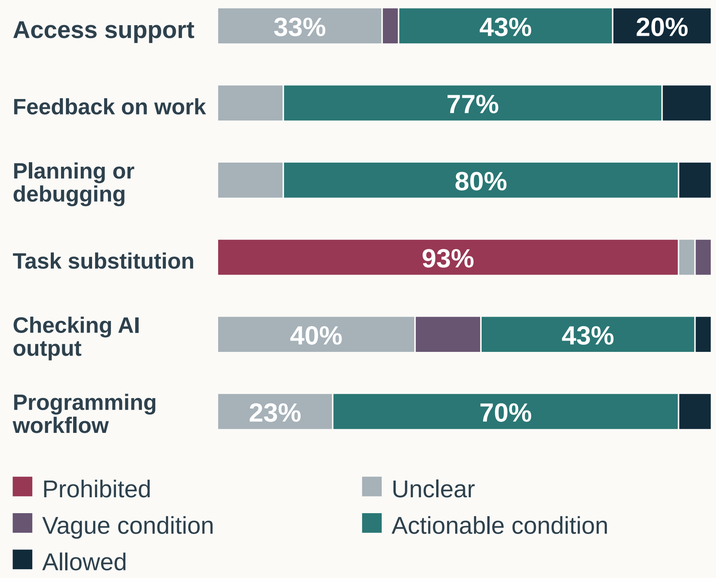
The corpus does not split neatly into permissive and restrictive camps. The four profiles show which link between boundary, evidence, and safeguards remains weak.

More complete packages connect all four framework elements. Policies with clear boundaries but weak support show the main gap: they classify AI use without equally clear evidence or safeguards.

06 GUIDANCE VARIES BY SCENARIO

Policy answers across six AI-use scenarios

Each bar summarizes 30 policy packages. Actionable conditions state a clear rule with usable evidence or safeguards. Vague conditions do not.



Task substitution receives the clearest response. 93% of packages prohibit AI from performing the target task.
Checking AI output receives the least usable guidance. 53% of packages are unclear or give only a vague condition.
This matters in practice. Students working with AI need to evaluate its output, yet policies address that skill least directly.

07 PROTECTIONS REMAIN UNEVEN

Share of visible safeguard language

Privacy appears in 67% of packages. Appeal, non-AI alternatives, and proportionate workload appear much less often. Only 4% address workload.

CONCLUSION

A defensible credential names the human capacity and explains what AI may do. It also preserves fair evidence of what the learner can still do.

1. **Specify the claim**
   Name the capability being assessed.
2. **Require evidence after delegation**
   Keep reasoning and verification visible.
3. **Add safeguards before enforcement**
   Protect access, privacy, appeals, and workload.

FULL PAPER
arXiv:2607.19988