

Peer Review

Review of: "Decision Support or High-Risk? Classifying AI Workforce-Intelligence Systems Under the EU AI Act"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper asks whether external workforce-intelligence platforms fall inside the high-risk employment category of Annex III of the EU AI Act or count as decision support (Abstract, Section 1). These platforms turn public labour-market signals (job postings, competitor hiring, leadership moves, skill-demand shifts) into aggregate guidance about roles and populations. The central thesis is that classification turns on architecture, deployment discipline, and use, not on data type alone, and that "it is only aggregate data" is insufficient (Section 1, Section 9). The research question is framed in design-science terms, following Hevner et al. and Peffers et al. (Section 1).

The artefacts are a taxonomy of four archetypes crossed on two dimensions, data subject and proximity to a decision about a person: A individual scoring, B internal aggregate analytics, C external talent-market intelligence, D workforce planning and AI-exposure modelling (Section 3, Table 1). Each archetype is worked through Article 6 and Annex III, with three tensions named: the "materially influencing" test, the profiling proviso, and the GDPR applying independently (Section 4). Five architectural properties are then set out (Section 5), followed by a five-layer reference architecture and an eight-question checklist (Section 6, Table 2), and a practice layer covering vendor assurance, effectiveness and bias testing, monitoring, and an indicative RACI (Section 7, Table 3). The demonstration is a worked illustration of one generic system in Section 6. Evaluation is deferred to future work under the FEDS framework (Section 1, Section 8.5), and the author states that the paper is analytical and design-oriented, not empirical or doctrinal (Section 8.4).

2. Whether the conclusion holds on the data

The part of this paper I would most want colleagues to read is the two-dimension taxonomy in Section 3, Table 1 (data subject x decision proximity). It gives a clear vocabulary for a class of systems the author calls under-analysed. I also value that Section 4 does not take the easy route: the author says the favourable classification of C and D is conditional, and names three tensions honestly, including that a recommendation to reduce a function "eventually resolves into decisions about the people in it". Table 2 can be used as is, since each control is a question, tied to a provision the paper cites, with evidence to retain. The remark at the end of Section 5 that the five properties are "necessary but not sufficient" sets the right frame for it.

I read this as a design-science paper, so for me the question is less whether a dataset supports a result and more whether the argument is careful and the artefacts can be checked. On that reading, no deployed or vendor system is assessed. The only demonstration is the generic hypothetical worked illustration in Section 6, a system mapping automation exposure across R&D role families. Evaluation is explicitly deferred to FEDS, expert interviews, or a Delphi study, and case studies (Section 8.5), and Section 8.4 acknowledges that the architecture reflects "a designer's vantage rather than a survey of deployed systems". So the claim that classification turns on architecture is argued, not yet tested. I would add one plain sentence near the conclusion saying what is demonstrated and what is deferred.

The argument is strongest where it stays conditional and turns the reasoning into questions that someone can answer and document. It rests on unstated judgement at one point. In Section 6, the author writes that items bearing on profiling and on material influence "require judgement that should be exercised with qualified legal input". These are exactly the items that decide C and D. It would help to say a little more about what that judgement involves, for example which facts a reviewer would weigh, so that the hardest calls are not left entirely to the reader.

Two tensions in the paper's own text deserve a second look. First, Archetype C inputs include "executive and leadership movements" (Section 1) and "leadership moves" (Table 1, row C), yet the same row says "no identifiable individuals", and Section 4 lists "scoring named individuals in a competitor's organisation" as a profiling trigger. Leadership moves concern named people, so it would help to say how this input passes the aggregation boundary. Second, Section 7.5 says Article 26(7) requires employers to inform worker representatives "before a high-risk AI system is put into service", then says this obligation

is "engaged even where a system is self-classified as non-high-risk". As written, these two sentences pull against each other. Separating legal obligation from recommended practice would resolve it.

3. Reproducibility

Here, reproducibility is about the instrument, since no data or code is involved. The question I ask is whether another expert could apply the framework to the same system and reach the same classification. A good part of it is already operational. Each of the five properties is stated "as requirements a designer can implement and an auditor can verify" (Section 5), and Table 2 lists evidence per control, for example, re-identification test results and replay logs.

What I would still define is the aggregation boundary. It relies on "minimum aggregation thresholds" (Section 5.1, Section 6 Layer 3), but I searched the text for a numeric threshold or a named anonymity criterion and did not find one. A rule such as a minimum group size would help, and so would a test behind "trivial re-identification" (Section 5.1, Table 2 row 1). I would also clarify the reproduction claim. Section 5.2 promises that "a given input and configuration reproduce a given output", yet Layer 3 admits language-model components (Section 6). Section 5.2 does mention pinning pipeline and model versions and confining non-determinism to a bounded part, but it does not say how outputs of the language-model layer are replayed.

The check I miss is an inter-rater one. Interviews, a Delphi study, and case studies appear only in the research agenda (Section 8.5), and I found no agreement statistic for the framework itself. Section 7.2 recommends "human-machine agreement studies" for systems, so a small version for the framework would fit well. Two or three independent raters could classify a handful of publicly described products with Table 1 and Table 2, and the author could report the agreement and the items where they split. I think this would show whether the framework gives the same answer in different hands.

4. What specifically to fix

1. In Table 1, row C (and Section 1), "leadership moves" and "executive and leadership movements" are listed as inputs, while the same row says "no identifiable individuals". I would either remove these inputs from row C or explain how named-person signals are aggregated before modelling. If the author chooses the second route, it would help to add a line to Table 2 for it.
2. In Section 5.1, Section 6 (Layer 3), and Table 2 row 1, it would help to give an operational rule for the aggregation threshold and for "trivial re-identification". An example value with its rationale would

be enough, as long as an auditor can use it to check pass or fail.

3. In Table 2, row 4, the explain-only control is anchored to "Art. 6(3)(a), (c)". Section 2.3 lists the four conditions in this order: narrow procedural task, improving a completed human activity, detecting decision-making patterns, preparatory task. Section 4 argues for C and D "as a preparatory or advisory task". Could the author confirm which lettered conditions this control is meant to map to, and align Table 2 with Section 4?
4. In Section 7.5, the paragraph on worker representatives has two sentences that sit uneasily together. One refers to the moment "before a high-risk AI system is put into service", and the other says this obligation is "engaged even where a system is self-classified as non-high-risk". I would reconcile them and mark which part is a legal obligation as the author reads it and which part is recommended practice.
5. In Section 5.2, Section 6 (Layer 3), and Table 2 row 3, Section 5.2 already mentions pinning model versions. It would help to add how the language-model layer is made replayable beyond that, for example, decoding settings and logged prompts and outputs. Without this, "fixed-seed reproduction" covers only the deterministic layers.
6. Table 3 and Section 7.5 differ on ownership. Section 7.5 says the AI governance lead "owns the classification decision", while in Table 3 the business owner is Accountable for both the pre-purchase classification memo and the Article 6(4) assessment, with the governance lead Responsible. I would align the text and the table. Also, worker representatives appear in Table 3 only as Informed of bias results. Since Section 7.5 calls consultation a deployment gate, listing them as Consulted on the classification memo would match the text.
7. For the worked illustration in Section 6 and the discussion in Section 8.5, I would add a small demonstration beyond the hypothetical system. Apply Table 1 and Table 2 to two or three publicly described products, ideally with two raters, and report where they agree and where they split. This is the largest item on my list, and it could also be a short appendix.
8. In the first paragraph of Section 1, the paper is dated Aug 5, 2026, and says the bulk of the high-risk regime "becomes applicable on 2 August 2026". A tense update would keep the timeline current.

5. Recommendation

Minor revisions. The author has built a carefully designed and clearly written design-science framework (taxonomy, five properties, reference architecture, checklist, practice layer) that practitioners can use now (Sections 3 to 7), and the limits are stated openly in Section 8.4. The fixes are mostly local. I would revisit

the leadership-moves input in Table 1, add an operational aggregation threshold (Section 5.1), give the explain-only control in Table 2 a clearer anchor, and tighten the Article 26(7) wording (Section 7.5). The ownership mismatch in Table 3 also needs aligning. One larger addition I would encourage is a small inter-rater demonstration on real products, which would turn the argument into evidence.

Declarations

Potential competing interests: No potential competing interests to declare.