

Peer Review

Review of: "Differentiating Conscious Human Entities from Autonomous AI Agents: A Heart Rate Variability (HRV) Framework for Human Authentication in the Emerging AI Economy"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper proposes heart rate variability, read without contact by a 77 GHz FMCW radar, as the physiological anchor for telling a live person at a workstation apart from an autonomous software agent. Its central conceptual move, which I take to be the main change in version 2, is to separate Human Presence Verification (HPV: a living person occupies the sensing zone) from Human Contribution Verification (HCV: that person actually produced the work). HRV is assigned to HPV only, and §3.5 explicitly scopes it as a liveness anchor rather than a stable identity biometric. On top of this, the author specifies a four-layer architecture (§4): radar sensing, an IBI pipeline taken from the Massive Radio Sensing work of Xu et al. [10], a 47-feature ensemble of Random Forest, RBF-SVM, and a two-layer LSTM with fixed voting weights (§4.3.2), and a Proof-of-Humanity (PoH) credential issued every 60 seconds as an ES256-signed JWT (§4.5). A threat model with six attack rows (Table 2), a privacy mapping to GDPR, HIPAA, CCPA, and the EU AI Act (Table 4), application domains (§6.2), and pseudocode (Appendix A) complete the design.

No new measurement is reported. The only empirical support is Xu et al.'s published sensing accuracy, reproduced in Table 3 (for example, Table 3, row Mean IBI Error, column Seated: 24.78 ms), which the paper reads against its own requirement of a mean IBI error below 30 ms (Table 1, row Accuracy). The abstract closes by saying that Xu et al.'s evaluation "confirms that contactless HRV sensing achieves inter-beat interval estimation errors as low as 15.47 ms, sufficient to distinguish live human physiological

rhythms from synthetic or replayed signals." I read the paper as a design proposal with a feasibility argument, and I review it on those terms.

Some things here are done well, and I want to say so plainly. The HPV/HCV split is useful and clearly argued; §5.1 is honest that adversary A_2 (the colluding human) defeats any purely physiological scheme. Table 2 now carries a "Basis for Risk Classification" column, and its caption admits that no row rests on original trials. §7.2 and item 3 of §9 state that validation of the full pipeline is future work and call it "the most significant open gap." The numbers copied from Xu et al. are consistent across the abstract, §1, §7.1, and Table 3 (15.47 ms supine, 24.78 ms seated, stability index 1.70 ms seated). That internal care is not universal in position papers.

2. Whether the conclusion holds on the data

2.1 The data that exist answer a different question. Xu et al. measured how closely radar-derived IBIs track a reference. They did not, as far as the paper reports, test whether radar HRV separates a live person from a replayed or synthesised signal. Nothing in Table 3 speaks to the false-accept or false-reject rate of the liveness classifier, and §9 item 3 concedes that FAR, FRR, and AUC-ROC are unknown. So the last sentence of the abstract, and the phrase in §10 that the approach is "technically feasible," claim more than the evidence supports. The 30 ms target itself (Table 1, row Accuracy) is given no derivation: I could not find any argument linking 30 ms of IBI error to a particular separability of live and spoofed signals.

2.2 The sensing resolution contradicts the liveness rationale. §2.1 grounds liveness in "the fine-grained temporal structure of HRV (sub-millisecond variation patterns)." The sensor the paper relies on has a mean IBI error of 24.78 ms seated (Table 3, row Mean IBI Error, column Seated). Structure below one millisecond cannot be observed through an error roughly 25 times larger than a millisecond. Either the liveness argument must move to features that survive a 25 ms error, or the sub-millisecond claim should go. The same tension affects the orienting-reflex test (§5.3.2): the paper never states the expected size of the "transient IBI lengthening," so I cannot tell whether it is detectable above a 24.78 ms error on a single beat.

2.3 The classifier is described as if it had been trained. §4.3.2 gives exact hyperparameters (500 trees, depth 12, $C = 10$, dropout 0.3) and says "The ensemble weighting was determined by held-out cross-validation." Cross-validation on what? There is no dataset in the paper, and §7.2 says liveness classification has not been validated. The weights 0.45 / 0.25 / 0.30 therefore cannot have come from cross-validation, or else data exist that the paper does not report. Both readings need fixing. The training

problem is also harder than the text suggests: the classes are live, replay, synthetic, and null (§4.4.1), so a training set needs labelled attack recordings, and §4.3.2 says models are "trained within the secure enclave," where only one worker's 300 resting IBIs from enrollment (§4.3.3) are available.

2.4 The policy threshold cannot be met by the specified score. This is the clearest technical error I found, and I checked it twice. Appendix A step 4d defines $\text{human_score} = \text{sigmoid}(w_{\text{RF}}\text{RF_prob} + w_{\text{SVM}}\text{SVM_prob} + w_{\text{LSTM}}\text{LSTM_prob})$. With the §4.3.2 weights, $0.45 + 0.25 + 0.30 = 1.00$, so the weighted sum of three probabilities lies in $[0, 1]$. The logistic sigmoid maps $[0, 1]$ onto $[0.5, 0.731]$ ($\text{sigmoid}(0) = 0.5$; $\text{sigmoid}(1) = 1/(1 + e^{-1}) \approx 0.731$). §4.5.3, Level 3, rejects any token with a `human_score` below a threshold given as "e.g., 0.85." Under the pseudocode as written, every token from every live human fails. And even a fully absent person scores at least 0.5 unless step 4c fires. The prose in §4.3.2 describes a plain soft-voting average, which would not have this problem, so the fix is simple, but the two specifications disagree today.

2.5 The relay-attack argument does not work physically. Table 2 (row Relay Attack) and §5.3.3 say that relay hardware adds latency "detectable within the 200 ms challenge-response timing window" because the delay is "bounded below by the speed-of-light constraint." In 200 ms, light travels about $0.2 \times 3 \times 10^8 \text{ m/s} = 6 \times 10^7 \text{ m}$, i.e., 60,000 km. A timing window of that size excludes no relay anywhere on Earth. The radar's own ranging operates on nanosecond round trips, but the challenge-response channel is judged on the physiological timescale of seconds (1–5 s in §4.3.3, 1–3 s in §5.3.2). The "Low–Medium" rating for this row has no basis as stated.

2.6 Replay and synthesis are rated "Low" on arguments that do not cover the obvious attacks. A replay of a genuine recording (Table 2, row Signal Replay) preserves every nonlinear and RSA feature, since it is a real human signal. Its detection therefore rests entirely on the orienting reflex, whose repeatability the paper does not examine. Orienting responses habituate with repetition, and the schedule in §4.3.3 issues a challenge on average every 8 minutes, i.e., about 7.5 per hour ($60 / 8$), across a working day. §5.3.2 also calls the tone "imperceptible" while relying on it to elicit an orienting reflex, which requires the stimulus to be perceived. For synthesis (row Synthetic Injection), reference [19] is a paper on surrogate data for nonlinearity testing. It shows that nonlinearity can be distinguished from linear Gaussian surrogates. It does not show that a generator trained on real HRV, or a mechanical actuator driven by recorded thorax motion (which carries RSA with it), is detectable. The replay row's bound, "(response window / mean IBI period)," is also vacuous as written: for a response window of several seconds and resting IBIs of around

one second, the ratio exceeds 1, so it bounds nothing. I may be misreading what the author intends by this expression; if so, it needs a derivation.

2.7 An unsourced number. Table 2, row Colluding Human (A_2), gives an "Estimated ~15% residual evasion when combined with cognitive engagement probes and AI-content detection (see Section 5.4)." §5.4 contains no estimate and no source. Given that the caption says risks are grounded in literature or theory, this figure should either be derived or removed.

2.8 The credential certifies presence but is used to certify authorship. §3.4 insists that presence is "necessary but not sufficient" for contribution. Yet `involvement_pct` is defined in §4.5.1 and Appendix A step 4e as the fraction of seconds with a detected IBI, which is presence and nothing else, and §4.5.3 then uses $\text{involvement_pct} \geq 0.90$ "for a session claiming full human authorship." This quietly reintroduces the conflation the paper was revised to remove. The same problem appears at the formal level: §3.2 defines $C: (S, \Phi) \rightarrow \{H, D, X\}$, but a physiological channel cannot separate H from X, since a human is present in both. The HCV layer (§5.4) is three short paragraphs with no specification comparable to §4.

2.9 The token is not bound to any work product. The claims listed in §4.5.1 are `session_id`, `iat/exp`, `device_cert_fingerprint`, `hrv_commitment`, three scores, `nonce`, and `sig`. None of them is a hash of the deliverable. A valid PoH token from any session in which a person sat at the desk can therefore be attached to any document, including one an agent wrote elsewhere. For the use cases in §6.2 (filings, peer reviews, copyright evidence), this is the central gap, and it is not listed in Table 2 or §9.

2.10 Some cryptographic claims are internally inconsistent. §4.5.4 claims forward secrecy through ephemeral session keys, but the token is signed with the long-term device attestation key (§4.5.1, `sig`; Appendix A step 6), and no session key appears anywhere in the protocol. §4.5.4 also says `hrv_commitment` enables audit verification "by the original device," while §4.3.3 and Appendix A step 7 discard the raw IBIs; once they are gone, nobody can recompute the hash, so the commitment verifies nothing. Replay protection by `nonce` (§4.5.4, Time-binding) requires the relying party to store seen nonces, which is not specified.

2.11 Regulatory mapping. I am not a lawyer and cannot verify every article, but several mappings in Table 4 deserve a second look. Art. 10 of the EU AI Act concerns data and data governance for training sets rather than purpose limitation (row Purpose Limitation). Consent as the lawful basis in an employment relationship is widely regarded as problematic because of the power imbalance, which weakens row Informed Consent. When identity binding is off, HRV used for liveness is arguably health data rather than biometric data "for the purpose of uniquely identifying" a person (§8.1), which changes

the Art. 9(2) analysis. HIPAA applies to covered entities and their associates, and most employers are neither. And the EU AI Act's prohibition on emotion recognition in the workplace is directly relevant to a system that reads autonomic state at work, but it is not discussed.

3. Reproducibility

Because there is no experiment, the usual reproducibility items (data, code, seeds, compute) have nothing to attach to. No code, no simulation, and no dataset are released, and none is described. What remains is whether the design could be reimplemented from the text. Largely, it can, and the specificity of §4.3 and §4.5 is a genuine strength. Several parameters are, however, inconsistent or ambiguous, and an implementer would have to guess.

Feature count. §4.3.1 states 12 time-domain features and lists mean IBI, SDNN, RMSSD, pNN50, pNN20, SDDSD, mean HR, SDRR, CV-IBI, NNx (x = 20, 50, 80), and triangular index. If NNx counts as three features, that is 13; if it counts as one, 11. SDNN and SDRR are the same quantity under two names, and mean HR is a function of mean IBI. The nonlinear list includes multiscale entropy at scale 1, which equals SampEn. The abbreviation RR for recurrence rate collides with R-R intervals. The stated total of 47 (12 + 8 + 15 + 12) should be reconciled with the actual list.

Window and timing parameters. §4.3.3 says IBI estimation "runs continuously at 1 Hz," but IBIs arrive once per beat, not once per second; "300-sample windows (approximately 5 minutes)" assumes 60 beats per minute. The same section yields a classification "every 2.5 minutes," while §4.5.2 and Appendix A step 4 draw a new human_score from the 300-IBI window every 60 seconds. §4.5.1 hashes a "60-second IBI window"; Appendix A step 4f hashes ibi_buffer[-60:], which is 60 beats. VLF power (§4.3.1) is barely estimable from a 5-minute window, and the paper should say how it is computed.

Challenge-response window. Three different values: 1–5 s (§4.3.3), 1–3 s (§5.3.2), 200 ms (Table 2 row Relay Attack, §5.3.3). Appendix A step 3b uses window = 5 s and a 150 ms tone, within the 100–200 ms of §5.3.2. How response.is_reactive is decided (threshold, beats examined) is not given, and challenge_pass_rate is never computed from challenge_results in Appendix A.

Pseudocode gaps. In Appendix A, step 4c jumps to step 5 when liveness fails, leaving involvement_pct and hrv_commitment undefined in that branch. The token assembled in step 5 omits device_cert_fingerprint, which §4.5.1 lists as a claim.

Table 3 provenance. The caption says "All values are sourced from Xu et al. Table I," yet the Clinical column contains approximations ("~22", "15-20", "~13", "~8"). I could not check these against Xu et al. from

this text. If they are read off a figure, say so; if they are in Table I, quote them exactly. The Auth. Relevance entry for pNN50 ("Adequate for frequency-band liveness test") pairs a time-domain metric with a frequency-domain test. The abstract mentions "30 healthy participants," a figure that does not appear anywhere in the body.

Hardware. The radar model (TI AWR6843AOP, 77 GHz, 4 GHz bandwidth, 12 virtual channels, §4.2) is named. The claim of token generation under 15 ms on a Cortex-A72 (§4.5.2) is described as "expected"; §7.2 reports 312 ms per inference on a Raspberry Pi 4B from Xu et al. The text should state that the 15 ms figure covers signing only and is not measured. ICNIRP compliance (Table 1) is asserted without emitted power or duty cycle.

4. *What specifically to fix*

1. **Abstract, last sentence; §10, final paragraph.** Rewrite so that Xu et al. are cited for IBI accuracy only. Their results do not show that radar HRV distinguishes live from replayed or synthetic signals, and the abstract should say the liveness discrimination is untested.
2. **Appendix A step 4d versus §4.3.2 and §4.5.3.** Remove the sigmoid, or change the threshold. As written, the score lies in [0.5, 0.731] and can never reach the example threshold of 0.85.
3. **§4.3.2, sentence on held-out cross-validation.** Delete it, or report the data, split, and metrics it refers to. Also, say what data the in-enclave training would use, and where labelled replay, synthetic, and null examples come from.
4. **§2.1, "sub-millisecond variation patterns".** Reconcile with the 24.78 ms seated IBI error in Table 3, or drop it.
5. **Table 2, row Relay Attack, and §5.3.3.** The speed-of-light argument fails at a 200 ms window (about 60,000 km of light travel). Replace it with a real mechanism or mark the row as unmitigated.
6. **Table 2, rows Signal Replay and Synthetic Injection; §5.3.2.** Discuss habituation of the orienting reflex at roughly 7.5 challenges per hour, the expected magnitude of the IBI response relative to a 24.78 ms sensing error, and the contradiction of an "imperceptible" stimulus that must be perceived. Either derive the evasion bound in the Replay row or remove it. Narrow what [19] is cited for.
7. **Table 2, row Colluding Human.** The "~15% residual evasion" figure needs a source or a derivation. Otherwise, delete it.
8. **§4.5.1 claims list.** Add a binding to the work product (e.g., a hash of the deliverable or of a signed interaction log). Add token transplantation to Table 2 as an attack.

9. **§4.5.3, Level 3.** Stop calling $\text{involvement_pct} \geq 0.90$ evidence of "full human authorship"; it is a presence fraction (Appendix A step 4e). Rename the field or define a separate HCV score.
10. **§3.2.** The codomain $\{H, D, X\}$ is not identifiable from Φ . Split C into an HPV output from Φ and an HCV output from S plus probes, which is what §3.4 already argues in prose.
11. **§4.5.4, Forward secrecy and Minimal disclosure.** Define the session key or remove the forward-secrecy claim. Explain how hrv_commitment can be audited after raw IBIs are discarded, or drop the audit claim.
12. **§4.3.1.** Recount the time-domain features, remove the SDNN/SDRR duplicate and MSE scale 1 (identical to SampEn), rename recurrence rate.
13. **§4.3.3, §5.3.2, Table 2, Appendix A.** One value for the challenge-response window. One definition of the classification cadence (60 s or 2.5 min), and state IBI windows in beats rather than seconds.
14. **Table 3.** Quote the clinical column exactly from Xu et al. or state that it is approximated from a figure; fix the pNN50 relevance entry.
15. **Table 4 and §8.1.** Revisit the AI Act Art. 10 mapping, employee consent as a lawful basis, HIPAA applicability, and the AI Act's rule on emotion recognition at work.
16. **References.** [6] (the GPT-2 report) does not support the claim about brittle CAPTCHAs, and [13] (Kubios software) does not support the claim that HRV structure cannot be synthesised. For [16], volume 8(1) of IEEE TIFS does not match the year 2020; please check the citation.
17. **A small experiment would change the paper's standing.** Even a pilot on public HRV recordings would help (live versus replayed versus surrogate series, with Gaussian noise of 24.78 ms added to mimic the radar), reporting FAR and FRR for the proposed ensemble. It would turn §7 from a borrowed number into evidence.

5. Recommendation

Major revisions. The problem is real, and the HPV/HCV distinction is a worthwhile contribution; version 2 is candid about the absence of original validation. But the paper still claims more than it shows. The abstract presents IBI accuracy as proof of liveness discrimination, the scoring rule in Appendix A cannot pass the policy in §4.5.3, the relay defence is physically wrong, the credential is not bound to the work it certifies, and the classifier is described as cross-validated without data. Each of these can be fixed in the text, and a small simulation on public recordings would give the feasibility claim something of its own to stand on.

Declarations

Potential competing interests: No potential competing interests to declare.