

Beyond Static Classification: A Decentralized and Community-Grounded Architecture for AI-Generated Music Detection

Gianmarco Tonelli^{*†}, Marco Gori[†], Stefano Melacci[†]

^{*}University of Pisa, Pisa, Italy

gianmarco.tonelli@phd.unipi.it

[†]University of Siena (DIISM), Siena, Italy

Abstract—Current detectors of AI-generated music achieve very-high-level accuracies on controlled benchmarks, yet perform substantially worse on professionally produced tracks and near randomly on audio from generators unseen during training. This paper argues that the gap is structural: the dominant paradigm treats detection as a static, isolated classification task, and that design choice produces three coupled failure modes, i.e., poor cross-generator generalization, centralization of detection authority, and absence of community-sourced ground truth, that cannot be resolved independently. We propose a decentralized, adaptive, community-grounded architecture comprising two main contributions. First, a three-role peer-to-peer architecture is proposed, based on the newly proposed UNaIVERSE platform (<https://unaiverse.io>). A Streamer, a Detector, and an Auditor nodes supply content, perform inference, and provide community validation, respectively, each role addressing one failure mode. Second, a Music World environment specification defining the shared state space, valid actions, and provenance graph for music-domain detection. We suggest the use of a Memory Head module enabling Auditor-driven continual adaptation of Detector models across the P2P network without a central aggregator. Within this design, each role is conditionally necessary for joint addressability, as removing any single role reintroduces a failure mode the remaining roles cannot compensate for. While the contribution is an architectural argument and specification rather than an empirical study, the paper calls for a shift in the research agenda from optimizing static classifiers on fixed benchmarks toward architectures that embed adaptation, decentralization, and community participation as first-class design properties.

Index Terms—AI-generated music detection, decentralized peer-to-peer architecture, continual learning, community validation, multi-agent systems, Internet of Sounds.

I. INTRODUCTION

Contemporary detectors of AI-generated music (AIGM) routinely achieve high accuracies on in-distribution benchmarks, but yield lower results on professionally produced tracks and near chance-level performance when evaluated on generators absent from training data [1], [2]. The gap persists even on the most diversified datasets currently available [3] and widens when ordinary signal transformations are applied, such as pitch shifting or MP3 re-encoding at 64 kbps [1], [2].

The standard response has been to build increasingly large and diversified training corpora, alongside adopting more expressive model architectures; nonetheless, we argue that the accuracy gap resides in three structural choices embedded in the dominant detection paradigm. The first (*i*) choice consists in

treating AIGM countermeasures as a static classification problem. A model is typically trained on a fixed corpus of human and AI-generated audio, then deployed without further adaptation. This choice produces Failure Mode 1 (FM1), namely poor cross-generator generalization. Models learn generator-specific low-level artifacts rather than general compositional features, leading to substantial issues when a new generative model emerges. Second (*ii*), detection is deployed in a centralized architecture in which a single operator, or a small consortium of operators, controls the entire detection pipeline. This deployment produces Failure Mode 2 (FM2), as centralization generates governance opacity and introduces a single point of operational failure [4]. To illustrate, an adversary capable of understanding and knowing the architecture of the deployed model can potentially design a system able to systematically evade it. Finally, the third (*iii*) structural choice consists in deploying detection in isolation from the communities whose music is being assessed. An absence of community validation, which may encompass listeners, genre specialists, or artists possessing contextual knowledge about the authenticity, stylistic provenance, and cultural meaning of a musical work [5], [6], underlies Failure Mode 3 (FM3). Without a mechanism enabling these communities to contest algorithmic decisions or encode cultural context, the detection system is epistemically impoverished and, in procedural-fairness terms, socially unaccountable: affected communities have no formally specified mechanism through which to contest decisions or encode local knowledge, a condition that platform-governance scholarship identifies as a legitimacy deficit [7].

The paper argues that the three failure modes can be jointly addressed without retaining the static, centralized, and isolated constraints of current detection systems. We instead propose an adaptive, decentralized, and community-rooted architecture for AIGM detection, instantiated on the UNaIVERSE peer-to-peer platform [8] (<https://unaiverse.io>). The two main proposed contributions are:

- A three-role peer-to-peer architecture composed of Streamer, Detector, and Auditor nodes, each addressing one failure mode. Having in mind the UNaIVERSE platform, this requires a specific interface definition for UNaIVERSE agents, enabling modular interoperability of

Detectors by means of data of type audio, that we propose here as well.

- A Music World environment specification defining the shared state space, valid actions, and provenance graph for detection in the musical domain. A Memory Head module enables continuous, Auditor-driven adaptation of Detector models across the P2P network without a central aggregator.

The second contribution is architectural designs; its empirical evaluation is left as future work, as will be detailed in Section VI-C.

The remainder of the paper is organized as follows. Section II reviews related work across five strands. Section III formalizes the three failure modes and argues for their joint addressability. Section IV specifies the proposed architecture and its components. Section V shows how the architecture jointly addresses the three failure modes. Section VI situates the proposal with respect to federated learning and watermarking, acknowledging its limitations. Section VII concludes the paper.

II. RELATED WORKS

Early detection approaches applied convolutional neural networks to spectrogram representations of synthetic speech as an early systematic attempt at classifying AI-generated audio [9]; a systematic overview of the progression from audio deepfake detection to music-specific detection is provided by Li et al. [10]. Subsequent dual-stream architectures combined spectral and temporal branches with contrastive learning objectives, achieving higher in-distribution accuracy [11]. The current state of the art applies transformer-based spectro-temporal tokenization and bidirectional attention fusion to learn richer representations [12], while systematic reviews of AI content forensic detection confirm cross-domain generalization as the main open problem across all content modalities [13].

This context has also shaped benchmark construction; we may consider the M6 dataset [3], covering multiple generators, five languages, and diverse genres, the Melody or Machine corpus [11], which provides semantically aligned human-AI pairs across over 300,000 tracks, or the Echoes dataset [14], which extends coverage to live performance contexts. Beyond these, the SONICS dataset (Synthetic or Not) [15] contributes over 97,000 tracks for end-to-end detection of synthetic songs (in which vocals, music, lyrics, and style are jointly generated) and highlights the importance of modeling long-range temporal dependencies for authenticity detection.

Zero-shot detection strategies partially address generalization. In the musical domain, MusicDET is trained exclusively on real music and models its distribution with frequency-guided normalizing flows, flagging AI audio as out-of-distribution without ever observing generator outputs [16]; similarly, cross-domain visual approaches pursue an analogous objective [17], but such methods remain more brittle as they reason over a fixed feature space.

The problem remains sufficiently nebulous and unresolved that Vila et al. [2] describe the situation as an arms race, given

that the quality of generative models improves more rapidly than the detection frontier, as each new generator introduces novel acoustic signatures that existing detectors have not been trained to recognize.

This work also relates to the Internet of Musical Things (IoMusT), which frames music production, distribution, and interaction as processes occurring across networked computational ecosystems [18]. Our proposal extends this perspective to distributed music-authenticity assessment, where detection, community validation, and also model adaptation are themselves performed across interconnected agents.

A. Generalization and robustness failures

The mechanism underlying cross-generator failures occurs because detectors, by learning generator-specific low-level fingerprints such as spectral artifacts originating from vocoders, do not rely on general compositional features [19].

Afchar et al. [20] demonstrate, for instance, that an important class of these artifacts (systematic spectral peaks linked to the checkerboard artifact) originates from the up-sampling modules of generative architectures and depends on the chosen architecture rather than on the training data or learned weights [20]. This evidence suggests that a change in the generator’s architecture, and not merely its weights, could defeat an artifact-based detector. Suffice it to note that when audio undergoes ordinary post-processing, such as pitch shifting of ± 2 semitones or MP3 re-encoding at 64 kbps, the fingerprints are degraded and accuracy drops to near-random levels even for in-distribution generators [1].

A consideration of technical contingencies does not, however, preclude a range of cultural issues. Representational bias in training corpora, according to Mehta et al. [5], [6], implies that fine-tuned detectors perform worse on non-Western genres or music originating from low-resource production environments. Here too, the problem lies in the absence of culturally grounded ground-truth from the communities whose music is being assessed, and not merely in dataset diversity.

A complementary path toward robustness is lyrics-based detection, discussed in Section IV-C, which proves more powerful than audio-only detection both for unseen generators and for signal perturbation [21].

B. Centralization and content provenance

Content provenance standards, such as C2PA [22] and embedded watermarking approaches [23], address a related but non-substitutive problem with respect to detection; rather, they are complementary, as the objective in this case is to attest that a piece of content was produced by an AI system.

In the case of watermarks, removal via signal processing is feasible [23], but provenance attestations require the cooperation of the generator, which adversarial actors will rarely provide. Detection therefore remains necessary to classify unlabeled or processed content, while adversaries often have the advantage of knowing the detector’s architecture, being able to probe it and evade it accordingly.

C. Federated and decentralized detection

Federated learning offers a partial architectural remedy [24]; by distributing training across participating nodes without sharing raw data, FL addresses privacy concerns and distributes the computational load. FL architectures coupled with blockchain additionally provide audit trails for model updates [25], while decentralized social media governance frameworks [26] have demonstrated that distributed rule enforcement is operationally feasible at scale.

However, federated learning retains a central aggregator for averaging model parameters, reintroducing a version of FM2; moreover, FL clients do not contribute cultural judgments, and by not incorporating community validation, FM3 remains unresolved.

D. Community-based validation

Community-based content moderation has produced promising precedents. The Birdwatch system [27] demonstrated that a reputation-weighted bridging algorithm, applied to crowdsourced fact-checking notes, can outperform algorithmic classifiers on contested factual claims. The MOSAIC multi-agent framework [28] modeled community dynamics in content regulation, showing that agent diversity improves system robustness. Furthermore, the Collab annotation system [29] demonstrated synergistic human-AI collaborative annotation for deepfake video identification. The specialization of community roles in collaborative knowledge systems provides foundational principles for role differentiation in community-rooted detection [30]. Multi-agent committee frameworks for data validation demonstrate that distributed consensus mechanisms can reduce individual annotator bias [31].

None of these systems has been applied to music detection. Taken together, they motivate the Auditor role elaborated in Section IV-B.

E. Continual learning for AI-generated content detection

To conclude the state of the art, we consider continual learning approaches for AI-generated content detection. Wang et al. [32], within the image domain, demonstrate that lightweight adapter modules added to a frozen backbone can absorb new generative distributions without catastrophic forgetting; their approach is representative of the classical continual-learning paradigm, relying on replay buffers and elastic weight consolidation (EwC). In contrast, the Continual Neural Unit model [33] has been proved to be effective in such a context, leading to a powerful Memory Head [34]. To the best of our knowledge, and as of the date of this submission, no analogous work exists for audio or music detection, and the Memory Head module, in Section IV-E, adapts this architecture to the musical domain by coupling it with an Auditor-validated community data stream.

III. THE THREE FAILURE MODES: A STRUCTURAL ANALYSIS

A. FM1: Poor Cross-Generator Generalization

A detector trained on the outputs of generators $G_1, \dots, G_n$ learns to distinguish human-composed audio from the specific artefactual signatures of those generators. When a new generator G_{n+1} enters the deployment environment (very different from the existing ones), its outputs are out of distribution: the detector has no representation of G_{n+1} 's vocoder characteristics, upsampling procedure, or spectral fingerprint. As a result, cross-generator accuracy can collapse to near-chance levels [1].

This failure is further amplified by ordinary post-processing. Transformations such as pitch shifting or MP3 re-encoding degrade the low-level fingerprints learned from $G_1, \dots, G_n$ without introducing any discriminative signal about G_{n+1} [1], [2]. FM1 is therefore a direct consequence of the static design assumption: the detector's performance envelope is fixed at training time, and the model cannot automatically adapt to generators that appear after deployment. Periodic retraining is possible, but its frequency is constrained by the cost of collecting and labelling new outputs, especially when culturally specific content must be represented [5], [6].

B. FM2: Centralization of Detection Authority

When detection is deployed through a centralized operator, that operator controls the end-to-end pipeline, from architecture selection, through training data, to deployment decisions and appeal procedures. Three risks are attributable to this arrangement.

- 1) *Governance opacity*: stakeholders cannot audit the system's criteria, let alone contest its decisions through a formally specified mechanism.
- 2) *Adversarial concentration*: a single deployed model is a high-value target, and an adversary capable of identifying its architecture is well positioned to design audio capable of systematically evading it.
- 3) *Operational fragility*: a single point of failure implies that a system downtime or compromise leaves the entire ecosystem unprotected.

Some broadcast monitoring services for AIGM detection have already been deployed by major music streaming platforms [4], but the absence of publicly specified detection criteria and formal appeal mechanisms instantiates, in current production deployments, the risk of governance opacity, in addition to the already existing opacity of generative systems.

C. FM3: Absence of Community Validation

The Birdwatch precedent [27] and the Collab annotation system [29] demonstrate that community participation in content judgment is both operationally feasible and epistemically valuable. To date, however, algorithmic detection systems encode a perspective on musical authenticity that is grounded in the system designers and the corresponding training corpus.

This view, when applied to a culturally foundational domain such as music, risks failing to account for the fact that a network of genre specialists, community listeners, musicians, or artists belonging to musical traditions distant from the Western canon may be capable of recognizing authenticity markers absent from training data and invisible to the model.

Under the current paradigm, with the sole exception of the extreme case represented by Bandcamp¹, no mechanism exists that allows community members to express judgments, contest decisions, or encode local contextual knowledge; and this shortcoming is not a data problem that greater dataset diversity can resolve.

IV. PROPOSED ARCHITECTURE

A. UNaIVERSE Platform Context

The three failure modes outlined above are structurally coupled. Poor generalization (FM1) is a consequence of the static design assumption. Centralization (FM2) is directly linked to the absence of community validation (FM3), as a single-operator system cannot leverage distributed community expertise. Finally, FM3, a consequence of the isolated design assumption, highlights how a system lacking a community participation mechanism produces no valid community signal capable of guiding model updates.

In response to this, we propose an architecture specifically designed for deployment on the UNaIVERSE platform [8], a peer-to-peer agent-based network developed by researchers from the Collectionless AI project [35], that is driven by the study of problems of “Learning Over Time”. UNaIVERSE provides a distributed computing infrastructure in which autonomous agents (being them humans or AIs) communicate through typed messaging channels, operate within “World” environment specifications, and coordinate without requiring a central aggregator. Each network node hosts one or more agents; agents perceive their local environment, perform actions, and communicate with peer agents through the platform’s message-passing API [36].

The absence of a central aggregator is the key architectural property for our purposes. FM2 requires that no single node hold exclusive authority over detection decisions or model parameters. UNaIVERSE’s topology satisfies this requirement structurally, as there exists no central server, master node, or privileged coordinator, unless specifically introduced.

A clarification on UNaIVERSE semantics is in order, as it constrains the architecture. The World is a special non-acting agent: it specifies the environment but does not itself perform actions. Action edges, therefore, are only well-defined between agents capable of acting. For this reason we do not instantiate the detection ledger within the World node; instead, we promote it to a dedicated Ledger agent. A Streamer connects to a Ledger agent, to which a set of Detectors and Auditors are in turn connected. Crucially, multiple Ledger agents may

coexist within the same Music World: each Streamer binds to exactly one Ledger, whereas Detectors and Auditors may subscribe to several Ledgers or to a selected subset. Each Ledger governs only the slice of content it currently handles, rather than exposing all music to all agents.

This multiplicity is the property that renders decentralization structural rather than merely procedural. Collapsing the network to a single Ledger remains a permissible deployment choice, but it is no longer enforced by the architecture: in the general case, detection authority is partitioned across Ledgers, and no single agent mediates the entire content space.

The Ledger is an infrastructural coordination agent: it does not address a failure mode on its own, but rather provides the substrate on which the failure-mode-addressing roles operate. Beyond it, every agent of the UNaIVERSE network is assigned a role when joining a target “World”; we accordingly propose three specific roles, each designed to address one failure mode. We will then use such roles to build a music-oriented UNaIVERSE World, named Music World, whose architecture is sketched in Fig. 1.

B. Streamer, Detector, and Auditor Roles

The rationale behind the Music World is to create the right environment where AI-generated music detection can be approached by means of a community of agents.

Streamer agents are content providers. An agent of this type submits a music stream, for instance a playlist, a broadcast feed, or a user upload, and forwards audio segments to the Ledger it is connected to the Music World. Each submission includes the audio payload, optional lyrics, contextual metadata (artist, genre, publication context), and an optional provenance hash for C2PA-compatible content. The role of Streamers, who perform no classification, is to populate the detection queue with appropriately contextualized content.

Detector nodes perform local inference. A Detector agent subscribes to unprocessed entries in one or more Ledgers to which it is connected, runs its local classification model on the audio payload, and writes a detection record comprising a detection score, a confidence value, a model identifier, and a list of flags. Any agent conforming to the interface specification we describe in Section IV-C may register as a Detector, and multiple Detector agents may process the same audio segment; their outputs are subsequently aggregated by the Auditor role.

Auditor agents provide community validation and trigger model adaptation. An Auditor agent subscribes to the detection records and challenge queues of the Ledgers to which it is connected. It may: (a) approve a detection record, incrementing its community validation weight; (b) contest a detection record, submitting a counter-judgment with provenance metadata and cultural context; or (c) submit an update signal for the Memory Head protocol (Section IV-E). Auditor reputation is weighted using a bridging algorithm adapted from the Birdwatch system [27]; votes that bridge otherwise disagreeing community clusters receive greater weight, discouraging coordinated manipulation. When multiple Detector agents produce detection

¹Bandcamp Daily, “Keeping Bandcamp Human,” Jan. 13, 2026. [Online]. Available: <https://blog.bandcamp.com/2026/01/13/keeping-bandcamp-human/>. Accessed: Jun. 15, 2026.

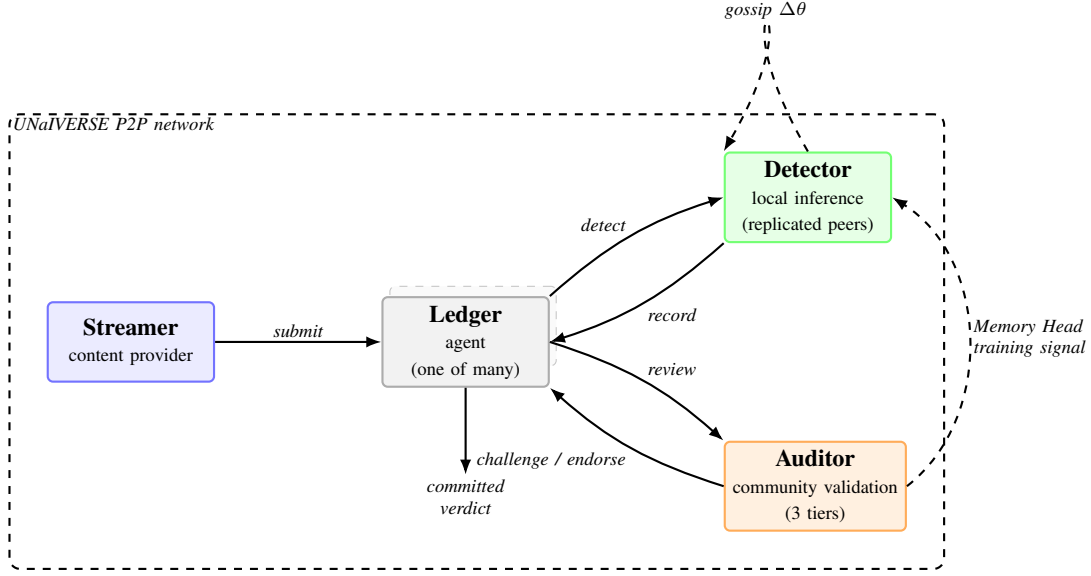


Fig. 1. Overview of the proposed decentralized architecture for AI-generated music detection. A Streamer binds to a single Ledger agent, to which Detectors (replicated peers running local inference) and Auditors (community validation, three tiers) connect. Multiple Ledger agents coexist within the same Music World—suggested here by the shaded copy behind the Ledger—so that detection authority is partitioned rather than centralized: Detectors and Auditors may subscribe to several Ledgers, whereas each Streamer binds to exactly one. Adapted Memory Head deltas ($\Delta\theta$) propagate peer-to-peer across Detectors by gossip, without a central parameter server.

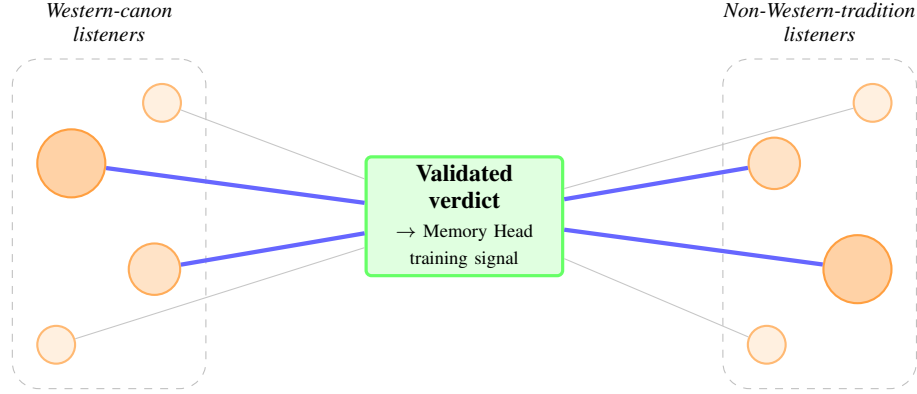


Fig. 2. Auditor validation as a bridge across culturally situated listening communities (a zoom on the Auditor role of Fig. 1). Node size encodes tier weight (Expert $w=3$, Genre Specialist $w=2$, Community Listener $w=1$). All Auditors vote (thin grey links), but the bridging algorithm up-weights judgments on which otherwise-disagreeing clusters converge (bold blue links), so the validated verdict rests on cross-cultural agreement rather than on any single community’s size.

records for the same audio segment, the Auditor’s input is computed over the tier-weighted mean of detection scores across contributing Detectors, with per-Detector confidence values used as within-tier weights.

The inter-role interaction protocol operates as follows. A Streamer submits audio to its Ledger; one or more Detectors produce detection records; Auditors review the records and may contest or approve them within a configurable challenge window, defaulting to 72 hours (a generic time limit we use as running example). The challenge window does not interrupt or delay content playback: detection records remain provisional while the content continues to circulate. Any downstream action on the final verdict is left to the adopting platform. Upon

closure of the challenge window, the detection record with the highest validation weight is registered as the authoritative ledger verdict for that segment, while contested records that receive sufficient Auditor endorsement trigger a Memory Head update cycle.

a) *Auditor tier design*: Auditors are organized into three tiers, each carrying a distinct vote weight in the bridging computation, that here symbolically restrict to some $w \in \{1, 2, 3\}$, the larger the better. *Tier 1 (Expert)* comprises professional musicians and scholars and carries weight $w = 3$; *Tier 2 (Genre Specialist)* comprises community experts and music journalists and carries weight $w = 2$; *Tier 3 (Community Listener)* comprises general users and carries weight $w = 1$.

This tiering operationalizes the observation, documented in crowdsourced musical knowledge curation, that expertise and engagement are unevenly distributed across a contributor population [37].

b) Recruitment and incentives: Tier 1 Auditors are recruited through partnerships with conservatories and professional associations; Tier 2 and Tier 3 Auditors are enrolled through opt-in mechanisms on participating streaming platforms. Participation is incentivized through public attribution, reputation tokens, and access privileges, an incentive structure consistent with the documented tradeoffs between participation, belonging, and norms that sustain volunteer moderation in online communities [38]. Robust validation presupposes a minimum Auditor density per genre subscription, with a reduced functional minimum for niche genres; the exact thresholds are deployment parameters to be calibrated in the empirical evaluation of Section VI-C.

c) Bridging adaptation for continuous scores: The Bird-watch bridging algorithm [27] was designed for discrete ratings, whereas detection produces continuous authenticity scores in $[0, 1]$. We adapt the algorithm by discretizing each score into three bins, human-leaning (< 0.35), uncertain ($[0.35, 0.65]$), and AI-leaning (> 0.65), for the purpose of bridging-weight computation, while retaining the continuous score for the final weighted-mean verdict. The 0.35/0.65 boundaries are chosen symmetrically around the decision threshold to carve out a central uncertainty band; they are provisional defaults to be calibrated empirically against the distribution of detection scores during the evaluation described in Section VI-C. This preserves the algorithm’s resistance to coordinated manipulation, in which votes that bridge otherwise-disagreeing clusters receive greater weight, while accommodating the continuous output of detection models.

C. Audio/Lyrics Interface Definition

We define a typed messaging interface for audio-domain agents on UNaIVERSE. The “audio” interface specifies several fields. UNaIVERSE currently defines agent communication through typed messages; for example, existing World deployments use “text” for conversational agents [36], or “img” in the case of image streams. The audio/lyrics interface defined here extends this typed message system to the music detection domain, introducing domain-specific fields that text-domain interfaces do not require:

- **AUDIO_PAYLOAD:** URI or binary chunk, required.
- **LYRICS:** string, optional.
- **METADATA:** {artist, genre, release_context, provenance_hash}; the provenance_hash field is optional and C2PA-compatible [22].

This is paired with additional intent-related information that can be provided by means of a text-encoded JSON-like dictionary which includes the following information:

- **TASK_DESCRIPTOR:** {task: “detection”, scope: “full_track” | “segment”, confidence_threshold: float}.
- **RESPONSE_SCHEMA:** {detection_score: float, confidence: float, model_id: string, flags: list[string]}.

Any Detector agent that accepts input of type ‘audio’ and outputs a detection score as a PyTorch tensor (supported by UNaIVERSE) can be registered without modifications to the platform’s routing or aggregation logic. The additional intent-related information is streamed as type “text”, and paired with the audio sample. In UNaIVERSE, when agents join a World, they inherit new skills developed by the creator of the World itself, among which the capability of processing and handling the intent JSON. Detector implementations may differ in architecture, for instance CNN, transformer, or hybrid, in training corpus, or in target modality, audio-only or audio+lyrics, provided they support the typed message schema as input. This enables controlled A/B testing of detection models, graceful fallback to secondary Detectors if a primary node is unavailable, and incremental deployment of updated models without service interruption.

The optional lyrics field is motivated by preliminary evidence suggesting that lyrics-based detection (retrieved from audio via automatic speech recognition and applying AI detection at the textual level) generalizes better to unseen generators and resists audio perturbations more robustly than audio-only detection, as textual content is largely invariant with respect to the audio generator and signal manipulation [21].

The provenance_hash field, also optional, links C2PA-compatible content attestations to the detection record, enabling downstream applications to jointly reason over detection scores and provenance metadata [22].

D. Music World Environment Specification

Before specifying the environment, we define its central structure. The streaming graph is the substrate over which all Music World activity is organized. Its nodes are audio segments, each carrying the detection records, challenge records, and metadata attached to it; its edges encode temporal co-occurrence within a stream, and may additionally encode provenance relations linking a segment to its source material. The streaming graph is not merely a representational convenience: it defines the locality over which agents observe state and over which adaptation signals propagate, and it is the concrete realization of the provenance graph referenced among the contributions. An agent’s observation horizon, and the scope of any Memory Head update it receives, are both defined relative to its neighborhood in this graph. The Music World is a UNaIVERSE environment specification for the music detection context. It defines:

- **State space:** the streaming graph defined above; the detection ledgers, each including the detection records, challenge records, and confirmed verdicts it governs; the Auditor challenge queue; and the genre taxonomy, that is, a hierarchical ontology of musical genres and subgenres maintained by the community.
- **Action space:** *Submit*, for the Streamer role; *Detect*, for the Detector role; *Challenge*, for the Auditor role; *Endorse*, for the Auditor role; and *Submit-Training-Signal*, for the Auditor role.

- **Observation model:** each agent observes only its own local neighborhood in the streaming graph and the portion of ledger state corresponding to the Ledger agents to which it is connected. Genre-based scoping is thus derived from Ledger connectivity rather than imposed as a separate subscription mechanism.

The governance of the genre taxonomy, who may propose category additions or boundary revisions, and how disputes are resolved, is an open research question that mirrors known challenges in collaborative ontology maintenance.² We identify this as a future design task and note that genre-taxonomy governance is separable from the detection-ledger governance specified here.

The Music World abstraction separates the environment specification from agent implementations. A Detector agent specialized in Western popular music and one specialized in non-Western folk traditions operate within the same Music World, but differ in their trained model weights and in the Ledgers to which they connect. This separation supports specialization without fragmentation: although detection state is partitioned across multiple Ledgers, all agents share the same Music World and the same provenance graph, which provides the common reference enabling cross-genre and cross-Ledger consistency checks.

E. Memory Head Module

The Memory Head module enables Auditor-driven continual adaptation of Detector models across the P2P network without a central aggregator.

Whereas classical continual-learning approaches for AI-generated content detection, such as the framework introduced by Wang et al. for the image domain [32], rely on replay buffers and elastic weight consolidation, we adopt the Memory Head [34] built on Continual Neural Units [33], which attain forgetting resistance by construction and require neither mechanism. We specify a lightweight adapter head appended to a frozen Detector backbone.

The backbone captures general audio representations, consistent with proposals to repurpose foundation models from audio deepfake detection as general-purpose backbones for AI-generated music detection [10], while the Memory Head is the sole trainable component during adaptation, absorbing distribution shifts introduced by new generators or genre-specific characteristics.

a) Adaptation cycle: When an Auditor submits a labelled training signal for a challenged detection record, the sample is streamed to Detectors. A gradient update is then applied exclusively to the Memory Head parameters, using the cross-entropy loss on the new sample without any additional changes, thanks to the properties of Continual Neural Units (CNU) [33] that generalize ordinary neurons, making them capable of learning continually, reducing catastrophic forgetting, without any other changes to the model. Replay buffers and EwC remain composable with the Memory Head

should additional stability be desired, but they are not required by the design; the continual-learning capability is intrinsic to the CNU formulation. This property carries particular weight in the present setting: a replay buffer, in a P2P deployment, would require replicating potentially copyrighted audio across nodes, raising storage and rights-management liabilities that the architecture should avoid by design. By dispensing with replay, the Memory Head enables collectionless continual adaptation, coherent with the Collectionless AI programme [34] from which UNaIVERSE originates, and consistent with the sociotechnical framing this paper advances: the network learns over time without accumulating a centralized corpus of the very content whose authenticity it adjudicates. Only the Memory Head parameters are updated (the frozen backbone is excluded). The plasticity–stability tradeoff is governed by β in Eq. 5 of [34], and requires empirical calibration. A low β produces a rigid head that resists adaptation, so FM1 persists; a high β produces a head that might overwrite the previously learned distribution (see [33]), leading to unstable behaviors that could also yield catastrophic forgetting. The recency bias is tuned by the *age* and *usage* parameters in [34].

b) Gossip-based delta propagation: Adapted Memory Head parameter deltas are propagated across the P2P network using a push–pull epidemic broadcast protocol [39], operating as follows. After each local Memory Head update, the updating node computes a parameter delta

$$\Delta\theta = \theta_{\text{new}} - \theta_{\text{old}}, \quad (1)$$

being θ the learnable parameters (weights) of the CNUs/Head. Without going into further details, CNUs include also sets of “keys” associated to neurons, that will be shared as well, and are not indicated in the notation for the sake of simplicity. The update procedure pushes it to k randomly selected peer Detector nodes, where k denotes the gossip fanout, following a Decentralized Federated Learning schema [40]. Each receiving node simultaneously pulls the sender’s current delta. Across all received deltas in a gossip round, a coordinate-wise median filter is applied for Byzantine rejection, discarding contributions that deviate by more than 2σ from the per-coordinate median. The protocol iterates until $\|\Delta\theta\|_2$ falls below a convergence threshold or until a maximum of R rounds is reached.

c) Timing and challenge-window interaction: Push–pull epidemic broadcast reaches the full network within $O(\log N)$ rounds [39], where N is the number of active Detector nodes. For modest fanout and round-count values, and typical P2P round-trip times, propagation completes well within the 72-hour challenge window used as running example. A Memory Head update triggered at the opening of a challenge window therefore propagates to all reachable nodes before the window closes. Nodes that receive a delta after their local challenge window has closed apply it to their subsequent update cycle; no delta is discarded.

d) Byzantine robustness: The coordinate-wise median filter tolerates up to

$$\left\lfloor \frac{n-1}{2} \right\rfloor \quad (2)$$

²MusicBrainz Editor documentation (<https://musicbrainz.org/doc/Editor>)

TABLE I
FAILURE MODES AND ARCHITECTURAL RESPONSES

Mode	Structural Cause	Architectural Response
FM1	Static classifier	Memory Head plus Auditor-validated training signal
FM2	Centralized operator	P2P topology plus Streamer/Detector replication
FM3	Isolated system	Auditor role plus community challenge mechanism

Byzantine senders among the n senders in any gossip round [39]. This bound is maintained independently of the gossip fanout.

e) Cold start: A newly registered Detector node initializes its Memory Head by requesting the current coordinate-wise median of Memory Head parameters from its k nearest gossip peers. This provides a warm start consistent with the network’s most recent adaptation state without requiring a central parameter server.

f) Update trigger: A Memory Head update cycle is triggered when the Auditor challenge queue for a given genre–segment pair accumulates a minimum number of independent challenges with combined confidence exceeding a threshold. The minimum-challenge count prevents spurious updates from isolated disagreements while remaining reachable for niche communities; both the count and the confidence floor are calibrated empirically in Section VI-C.

V. HOW THE ARCHITECTURE JOINTLY ADDRESSES THE THREE FAILURE MODES

The proposed architecture addresses the three failure modes through a clear mapping between roles and failure modes, summarized in Table I.

The claim of joint addressability implies that the removal of any single role reintroduces a failure mode that the remaining roles cannot compensate for.

Removal of the Auditor role: without Auditors, the Memory Head update cycle possesses no community-validated training signal. Detection records consequently accumulate without community endorsements, and Detector weights stagnate after initial deployment, causing FM1 to re-emerge. As no mechanism exists to allow community knowledge to enter the detection cycle, FM3 re-emerges as well.

Removal of decentralization: decentralization is guaranteed by two structural properties, i.e., the multiplicity of Ledger agents and the constraint that each Streamer binds to exactly one Ledger. Were these replaced by a single Ledger acting as a central aggregator, FM2 would be reintroduced: the Memory Head update cycle and the Auditor challenge mechanism would continue to function, but their outputs would be mediated by one operator, reinstating governance opacity and operational fragility. Centralization is therefore a deployment regression, not a property the architecture imposes.

Retention of all three roles: FM1 is addressed by the Memory Head module, designed to adapt Detector representations to new generators as Auditor-validated training signals accumulate. FM2 is addressed by the P2P topology, in which no single node holds exclusive authority over detection and the Streamer/Detector roles are replicated across the network. Finally, FM3 is addressed by the Auditor role, which provides a formally specified mechanism through which community members can contribute judgments and contest algorithmic decisions.

Given the dependency cycle established in Section IV-A, a system that retains even one of the three structural constraints (static, centralized, isolated) cannot simultaneously close all three failure modes. All three constraints are removed by construction in the proposed architecture.

This argument constitutes a demonstration of architectural sufficiency within the proposed design space, together with a conditional necessity claim: given the three-role decomposition, each role is necessary for joint addressability, as removing any single role re-introduces a failure mode the remaining roles cannot compensate for. We do not claim absolute necessity across all possible architectures, i.e., alternative approaches, such as a centralized system with community annotation and online learning, may partially address the failure modes through different means, and we discuss this in Section VI. What the proposed architecture offers is a concrete specification that removes all three structural constraints simultaneously, without retaining the central aggregator that prior FL approaches require.

VI. DISCUSSION

A. Watermarking and Provenance as Complementary Approaches

The C2PA provenance standard [22] and embedded watermarking approaches such as SynthID [23] address content provenance at generation time. Our architecture addresses detection post-hoc, including for content that was never watermarked, whose watermark was stripped, or that was generated by adversarial actors who will not voluntarily attest provenance. The two approaches are complementary and composable: the provenance_hash field in the “audio” type interface allows a Detector to condition on a C2PA attestation when present, or to fall back to artefact-based detection when it is not. The architecture does not preclude the adoption of watermarking; it ensures that detection remains available when watermarking is absent or compromised.

B. Alternative Architectures

A centralized system equipped with a community annotation API and an online learning component could partially address FM1 (through online adaptation) and FM3 (through annotation submission). However, centralized governance of the annotation pipeline reintroduces FM2 regardless of how open the API is: the operator retains ultimate authority over training-data

acceptance, model-deployment decisions, and appeal adjudication. The P2P topology of the proposed architecture distributes this authority structurally, not just procedurally.

C. Limitations

Sybil resistance: A coordinated network of inauthentic Auditor accounts could submit correlated false challenges, poisoning the Memory Head update signal. The reputation-weighted bridging algorithm adapted from Birdwatch [27] provides partial mitigation; votes that bridge across otherwise-disagreeing clusters receive higher weight. However, a sufficiently large Sybil network that maintains internal disagreement on non-target issues while agreeing on targeted challenges could evade the bridging filter. Formal Sybil resistance guarantees would require additional work on Auditor identity attestation, which is outside the scope of this specification. Relatedly, the Memory Head update trigger requires a minimum number of independent challenges, but the bridging algorithm operates on vote patterns rather than on identity: in the absence of a unique-identity attestation mechanism, independence is assumed rather than verified, and a single actor controlling multiple accounts could in principle satisfy the threshold.

Consensus latency: The challenge window introduces a delay between content submission and committed verdict. For real-time broadcast monitoring, this delay may be unacceptable. Reducing the challenge window improves latency but reduces Auditor participation, potentially degrading FM3 coverage. This latency-participation tradeoff is a genuine constraint of the community-grounded paradigm.

Verdict finality and appeal: Once the challenge window closes, the verdict is committed as authoritative, and the Auditor challenge mechanism applies only within that window. The current specification therefore provides no path to contest a committed false-positive verdict after the fact, a gap that is itself in tension with the legitimacy concerns motivating FM3. A post-commitment appeal mechanism, for example a re-opening procedure triggered by a qualified Auditor quorum, is left as future work.

Empirical validation: This paper specifies an architecture; it does not demonstrate its performance on any detection benchmark. The claim that the proposed architecture jointly addresses the three failure modes is grounded in architectural reasoning and literature analogy, not in experimental results. The primary future work is an empirical evaluation instantiating the specification as a working system and measuring: (a) cross-generator accuracy under the Memory Head adaptation protocol; (b) detection accuracy degradation under coordinated Sybil attack; (c) community Auditor participation rates across genre communities; and (d) latency under varying challenge window configurations.

VII. CONCLUSION

The dominant paradigm for AI-generated music detection, a static classifier trained once on a fixed corpus and deployed in a centralized architecture isolated from the communities it affects, embeds three structural failure modes that cannot be

resolved independently. Poor cross-generator generalization, centralization of detection authority, and absence of community validation are coupled through a dependency cycle that makes piecemeal architectural remediation insufficient.

This paper has proposed and specified a decentralized, adaptive, community-grounded architecture that is designed to jointly address all three failure modes. The Streamer/Detector/Auditor three-role architecture, instantiated on the UNaIVERSE peer-to-peer platform, eliminates the static, centralized, and isolated constraints simultaneously. The audio/lyrics interface definition enables modular Detector interoperability; the Music World specification provides a shared environment for multi-agent music detection; and the Memory Head module enables Auditor-validated continual adaptation without a central aggregator.

The paper’s contribution is an architectural argument: that the three failure modes share a common structural root, that this root is addressable through coordinated architectural design, and that the proposed specification is a concrete instantiation of that design.

We call on the community to treat AI-generated music as a matter of sociotechnical infrastructure as well as a problem of optimization over static benchmarks, in which decentralized governance, adaptive learning, and effective community participation must be assumed as fundamental architectural properties

REFERENCES

- [1] D. Afchar, G. Meseguer-Brocal, and R. Hennequin, “AI-Generated Music Detection and its Challenges,” arXiv preprint arXiv:2501.10111, 2025. [Online]. Available: <http://arxiv.org/abs/2501.10111>
- [2] L. C. Vila, B. L. T. Sturm, L. Casini, and D. Dalmazzo, “The AI Music Arms Race: On the Detection of AI-Generated Music,” *Transactions of the International Society for Music Information Retrieval*, vol. 8, no. 1, 2025. [Online]. Available: <https://transactions.ismir.net/articles/10.5334/tismir.254>
- [3] Y. Li, H. Li, L. Specia, and B. Schuller, “M6: Multi-Generator, Multi-Domain, Multi-Lingual and Cultural, Multi-Genres, Multi-Instrument Machine-Generated Music Detection Databases,” *Scientific Reports*, 2026. [Online]. Available: <https://www.nature.com/articles/s41598-026-36044-w>
- [4] D. López-Ayala, A. Cabello, P. Zinemanas, E. Molina, and M. Rocamora, “AI-Generated Music Detection in Broadcast Monitoring,” arXiv preprint arXiv:2602.06823, 2026. [Online]. Available: <https://arxiv.org/abs/2602.06823v2>
- [5] A. Mehta, S. Chauhan, A. Djanibekov, A. Kulkarni, G. Xia, and M. Choudhury, “Music for All: Representational Bias and Cross-Cultural Adaptability of Music Generation Models,” arXiv preprint arXiv:2502.07328, 2025. [Online]. Available: <http://arxiv.org/abs/2502.07328>
- [6] A. Mehta, S. Chauhan, and M. Choudhury, “Missing Melodies: AI Music Generation and its “Nearly” Complete Omission of the Global South,” arXiv preprint arXiv:2412.04100, 2025. [Online]. Available: <http://arxiv.org/abs/2412.04100>
- [7] N. P. Suzor, *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge, UK: Cambridge University Press, 2019.
- [8] S. Melacci, C. Di Maio, T. Guidi, and M. Gori, “UNaIVERSE: A Peer-to-Peer Network for Human-AI Agents,” University of Siena, Tech. Rep., 2025. [Online]. Available: <https://doi.org/10.13140/RG.2.2.33699.72485>
- [9] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, “STC antispoofing systems for the ASVspoof2019 challenge,” in *Proc. Interspeech 2019*, 2019, pp. 1033–1037.

- [10] Y. Li, M. Milling, L. Specia, and B. W. Schuller, "From Audio Deepfake Detection to AI-Generated Music Detection – A Pathway and Overview," arXiv preprint arXiv:2412.00571, 2024. [Online]. Available: <http://arxiv.org/abs/2412.00571>
- [11] A. Batra, D. Sharma, K. Thukral, R. Bhatia, N. Batra, and A. Gautam, "Melody or Machine: Detecting Synthetic Music with Dual-Stream Contrastive Learning," arXiv preprint arXiv:2512.00621, 2025. [Online]. Available: <http://arxiv.org/abs/2512.00621>
- [12] Y. Kim and S. Go, "Fusion Segment Transformer: Bi-Directional Attention Guided Fusion Network for AI-Generated Music Detection," arXiv preprint arXiv:2601.13647, 2026. [Online]. Available: <http://arxiv.org/abs/2601.13647>
- [13] Q. Xu, W. Mu, J. Li, T. Sun, and X. Jiang, "Advancements in AI-Generated Content Forensics: A Systematic Literature Review," *ACM Computing Surveys*, vol. 58, no. 3, pp. 60:1–60:36, 2025. [Online]. Available: <https://dl.acm.org/doi/10.1145/3760526>
- [14] O. Pascu, D. Oneata, H. Cucu, and N. M. Muller, "Echoes: A Semantically-Aligned Music Deepfake Detection Dataset," arXiv preprint arXiv:2603.23667, 2026. [Online]. Available: <http://arxiv.org/abs/2603.23667>
- [15] M. A. Rahman, Z. I. A. Hakim, N. H. Sarker, B. Paul, and S. A. Fattah, "SONICS: Synthetic Or Not – Identifying Counterfeit Songs," in *International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: <http://arxiv.org/abs/2408.14080>
- [16] C. Han, H. Wang, and J. Gui, "MusicDET: Zero-Shot AI-Generated Music Detection," arXiv preprint arXiv:2605.18072, 2026. [Online]. Available: <http://arxiv.org/abs/2605.18072>
- [17] A. Sar, S. Roy, T. Choudhury, and A. Abraham, "Zero-Shot Visual Deepfake Detection: Can AI Predict and Prevent Fake Content Before It's Created?" *Foundations and Trends in Signal Processing*, pp. 212–370, 2025. [Online]. Available: <http://arxiv.org/abs/2509.18461>
- [18] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.
- [19] H. Oh, "ArtifactNet: Detecting AI-Generated Music via Forensic Residual Physics," arXiv preprint arXiv:2604.16254, 2026. [Online]. Available: <http://arxiv.org/abs/2604.16254>
- [20] D. Afchar, G. Meseguer-Brocal, K. Akasbi, and R. Hennequin, "A Fourier Explanation of AI-music Artifacts," arXiv preprint arXiv:2506.19108, 2025, deezer Research. [Online]. Available: <http://arxiv.org/abs/2506.19108>
- [21] M. Frohmann, E. V. Epure, G. Meseguer-Brocal, M. Schedl, and R. Hennequin, "AI-Generated Song Detection via Lyrics Transcripts," arXiv preprint arXiv:2506.18488, 2025. [Online]. Available: <http://arxiv.org/abs/2506.18488>
- [22] Coalition for Content Provenance and Authenticity, "C2PA Technical Specification," C2PA, Tech. Rep., n.d.
- [23] L. Cao, "Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities," arXiv preprint arXiv:2504.03765, 2025. [Online]. Available: <http://arxiv.org/abs/2504.03765>
- [24] E. Gabrielli, A. Di Pietro, D. Fenoglio, G. Pica, and G. Tolomei, "A Survey on Decentralized Federated Learning," arXiv preprint arXiv:2308.04604, 2026. [Online]. Available: <http://arxiv.org/abs/2308.04604>
- [25] A. Heidari, N. J. Navimipour, H. Dag, S. Talebi, and M. Unal, "A Novel Blockchain-Based Deepfake Detection Method Using Federated and Deep Learning Models," *Cognitive Computation*, vol. 16, no. 3, pp. 1073–1091, 2024. [Online]. Available: <https://doi.org/10.1007/s12559-024-10255-7>
- [26] L. La Cava and A. Tagarelli, "Safeguarding Decentralized Social Media: LLM Agents for Automating Community Rule Compliance," arXiv preprint arXiv:2409.08963, 2024. [Online]. Available: <http://arxiv.org/abs/2409.08963>
- [27] S. Wojcik, S. Hilgard, N. Judd, D. Mocanu, S. Ragain, M. B. F. Hunzaker, K. Coleman, and J. Baxter, "Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation," arXiv preprint arXiv:2210.15723, 2022. [Online]. Available: <http://arxiv.org/abs/2210.15723>
- [28] G. Liu, V. Le, S. Rahman, E. Kreiss, M. Ghassemi, and S. Gabriel, "MOSAIC: Modeling Social AI for Content Dissemination and Regulation in Multi-Agent Simulations," arXiv preprint arXiv:2504.07830, 2025. [Online]. Available: <http://arxiv.org/abs/2504.07830>
- [29] S. Zhang, L. Wang, S. Li, Y. Wu, Y. Chuai, L. Chen, X. Yi, and H. Li, "Collab: Fostering Critical Identification of Deepfake Videos on Social Media via Synergistic Annotation," arXiv preprint arXiv:2601.17371, 2026. [Online]. Available: <http://arxiv.org/abs/2601.17371>
- [30] R. S. Geiger and A. Halfaker, "When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes?" in *Proceedings of the 9th International Symposium on Open Collaboration*, 2013, pp. 1–6. [Online]. Available: <https://dl.acm.org/doi/10.1145/2491055.2491061>
- [31] S. Vallabhaneni, T. Berkane, and M. Majumder, "The AI Committee: A Multi-Agent Framework for Automated Validation and Remediation of Web-Sourced Data," arXiv preprint arXiv:2512.21481, 2025. [Online]. Available: <http://arxiv.org/abs/2512.21481>
- [32] H. Wang, J. Lan, Y. Kang, H. Zhu, W. Wang, Z. Zhang, and S. Wang, "Generalizable and Adaptive Continual Learning Framework for AI-Generated Image Detection," arXiv preprint arXiv:2601.05580, 2026. [Online]. Available: <http://arxiv.org/abs/2601.05580>
- [33] M. Tiezzi, S. Marullo, F. Becattini, and S. Melacci, "Continual neural computation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2024, pp. 340–356.
- [34] M. Tiezzi, F. Becattini, S. Marullo, and S. Melacci, "Memory head for pre-trained backbones in continual learning," in *Proceedings of The 3rd Conference on Lifelong Learning Agents*, ser. Proceedings of Machine Learning Research, vol. 274. PMLR, 2025, pp. 179–197.
- [35] M. Gori and S. Melacci, "Position paper: Collectionless artificial intelligence," in *International Joint Conference on Neural Networks, IJCNN 2025, Rome, Italy, June 30 - July 5, 2025*. IEEE, 2025, pp. 1–8. [Online]. Available: <https://doi.org/10.1109/IJCNN64981.2025.11228186>
- [36] C. Di Maio, T. Guidi, L. Quarantiello, J. Bell, M. Gori, S. Melacci, and V. Lomonaco, "Book your room in the Turing Hotel! A Symmetric and Distributed Turing Test with Multiple AIs and Humans," arXiv preprint arXiv:2603.18981, 2026. [Online]. Available: <http://arxiv.org/abs/2603.18981>
- [37] J. Lim and A. R. Benson, "Expertise and dynamics within crowdsourced musical knowledge curation: A case study of the genius platform," in *Proc. Int. AAAI Conf. Web and Social Media (ICWSM)*, vol. 15, 2021, pp. 379–390.
- [38] B. Bulat, H. Wang, S. Fujimoto, and S. Frey, "The psychology of volunteer moderators: Tradeoffs between participation, belonging, and norms in online community governance," *New Media & Society*, 2025.
- [39] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry, "Epidemic Algorithms for Replicated Database Maintenance," in *Proceedings of the Sixth Annual ACM Symposium on Principles of Distributed Computing*, 1987, pp. 1–12.
- [40] E. T. M. Beltrán, M. Q. Pérez, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. G. Pérez, G. M. Pérez, and A. H. Celdrán, "Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2983–3013, 2023.