Embedding Amnesia and the Partial Ear of Machine Listening

Matt Sandler*
Feed Media Group
matt@feed.fm

Abstract

Musical embeddings have been made the substrate of choice for machine listening, facilitating discovery, creation, and curation on contemporary music platforms. Yet every learned representation is necessarily one of forgetting, as it reflects a model whose "ear" hears only part of the story, coding certain features while muting others. Features which go beyond the waveform — linguistic meaning, liturgical function, seasonal significance, accessibility of rights, or cultural intelligibility — are neither mistakes nor gaps that more training will fill, but rather are intrinsic absences in the representational space. In response to curator insights from a field-deployed music retrieval system, we explore embedding amnesia as a theoretical framework to understand what a learned representation leaves behind. We situate this concept in the context of media theory, algorithmic criticality, and musical ontology to claim that the primary question of the generative era is not what AI can produce but what it can understand.

1 INTRODUCTION

A song retrieves, at high cosine similarity, a parodic cover of itself. Same chord progression, same instrumentation, same vocal register; one is a sincere ballad, the other a comedy-variety-show send-up that lifts the source in order to mock it. A curator marks the cover as not-a-fit for a romantic playlist without listening past the first verse. The parody is constituted by its citational distance from the source — a relation the encoder cannot read, because the relation is held in the listener's recognition, not in the audio. The embedding has done what it was trained to do: it has found something that sounds nearby. What it cannot encode is the social act that turns the cover into commentary. The generative turn this conference convenes around — AI-driven systems reshaping music creation, performance, listening, and value under conditions of *hyper-reproduction* — is built atop substrates of this kind, and the substrate's blindnesses propagate into every one of those four registers.

This paper is about what is missing from the encoding rather than what is present in it. Contemporary music platforms run on learned audio representations — joint embeddings that map waveforms (and sometimes text) into a shared vector space, where retrieval, recommendation, captioning, generation, and increasingly composition proceed by nearest-neighbor lookup or by sampling from a learned manifold. Parker and Dockray (2023) trace the emergence of *machine listening* as a constellation of technical practices that aspire to "cover all possible sounds"; Sterne (2022) asks whether this listening is listening at all, and answers that it is the conversion of sound into data and the application of machine learning to that data — a process structurally distinct from human audition. These embeddings are the substrate of the generative turn; they underwrite the tracks now spawned daily through prompt-driven interfaces and govern how those tracks are surfaced, ranked, recombined, and offered (Drott, 2018, 2024). The question of what embeddings carry is no longer a narrow engineering question. It is a question about the conditions of musical experience in a hyper-reproductive age.

We name this condition **embedding amnesia**: the structural underdetermination of certain features by a learned audio representation, because the binding context for those features is not present in the signal the encoder takes. Linguistic meaning, liturgical function, seasonal significance, accessibility of rights, cultural intelligibility — these are not bugs in the embedding. They are not gaps that more contrastive pairs will close in the way a representation-quality gap would close. Not in the sense that no engineering work helps, but in the sense that the binding condition is external to the waveform, and external bindings cannot be authoritatively recovered from internal ones. A retrieval system built on this substrate inherits these commitments wholesale; a generative system built on the same substrate inherits them too. The sincere song and the parodic cover are acoustically near. They are socially incompatible. The encoder cannot make this difference authoritative, because the difference is held in the listener's recognition rather than in the audio.

We develop this claim in three movements. First (Section 2), we operationalize embedding amnesia as a specific theoretical concept, distinct from related notions such as model bias, data sparsity, or distribution shift, and we identify four loci where it manifests in deployed music systems. Second (Section 3), we name the practitioner standpoint from which the loci are visible to us, and offer three illustrative scenarios drawn from deployed retrieval practice. Third (Section 4), we situate the concept in three theoretical traditions — media theory, algorithmic criticality, and musical ontology — to show that the question of what an embedding forgets is continuous with longer histories of substrate, automation, and ontological commitment. We then turn (Section 5) to the generative era proper: what does it mean for music creation, not just music retrieval, when the representational substrate is amnesiac? Section 6 develops a constructive turn, organized around the brief-as-score reading of curatorial labor we read as native to the AIMC programme. We close (Section 7) with a reframing of the field's primary question. The challenge of AI in music is not the production of more music. It is the cultivation of a representational ear that can hear what curators, listeners, and traditions have always heard.

*Corresponding author: matt@feed.fm.

Contributions. This paper contributes (1) the concept of *embedding amnesia* as structural underdetermination, distinct from bias, data sparsity, and modality gap; (2) a four-locus typology with a 2×2 axis (in-signal/extra-signal, stable/relational) that predicts what kinds of intervention address what kinds of forgetting; (3) theoretical readings across media theory, algorithmic criticality, and musical ontology; and (4) the brief-as-score reading of curatorial work, with design and evaluation orientations for AI music systems under hyper-reproductive conditions.

2 EMBEDDING AMNESIA: THE CONCEPT

2.1 A definition

We define **embedding amnesia** as the structural absence of attributes from a learned representation, such that the absence is not remediable by additional training within the same representational paradigm. Three commitments distinguish the concept.

First, embedding amnesia is *about the substrate, not the model*. A particular checkpoint — CLAP (Wu et al., 2023), MERT (Li et al., 2024), or any of the joint audio-text and self-supervised audio encoders that share their input regime — is not the unit of analysis; the audio-as-input regime is. When we say a representation forgets, we mean that the choice of input modality (a waveform, possibly with paired captions) places limits on what kinds of attributes can be learned, and that these limits hold across all training runs that respect the modality.

Second, embedding amnesia is *not bias* in the canonical fairness sense. Bias as the field uses it concerns features that the model encodes but encodes asymmetrically across groups. Amnesia concerns features that the model cannot encode at all. The two often co-occur but answer different diagnostic questions. A bias audit asks: did the model learn this attribute fairly? An amnesia audit asks: was this attribute available to be learned in the first place?

Third, embedding amnesia is *not data sparsity*. Data sparsity assumes the model has not yet seen enough examples; the implicit promise is that scale will solve it. Amnesia is the position that scaling audio data is constitutively underdetermined for attributes that are not derivable from audio. The language of a track, the date of its liturgical use, the rights status of its constituent samples, the social register an ornament signals — these are not waiting for one more data run to emerge from the spectrogram. They live in metadata, calendars, legal databases, and cultural memory. Even were every track tagged exhaustively, the binding decision still depends on the listener's relational position — per occasion, per deployment, per brief — which arrives at the moment of encounter; the substrate cannot absorb a relation that is constituted in use.

2.2 Distinction from prior diagnoses

Several prior diagnoses describe related but distinct phenomena. The *semantic gap* names the distance between low-level features and the concepts listeners actually use; the term entered multimedia retrieval through content-based image search (Smeulders et al., 2000) and was carried into music informatics at the MTG (Herrera et al., 2005; Celma et al., 2006). It is a representation-quality problem, partly closed by contrastive learning over pairs and by larger encoders. The *glass ceiling* of MFCC-based timbre similarity (Aucouturier and Pachet, 2004) is an asymptote within a feature class. Both can in principle be approached by better encoding of audio. Embedding amnesia is the position that, at the limit of better audio encoding, certain attributes remain absent because they were not in the audio to begin with. The *modality gap* (Liang et al., 2022) names the geometric separation between text and audio embeddings even within a contrastively-trained joint space; this is a learning artifact reducible by better objectives. Embedding amnesia would persist if the modality gap were closed, because the categories that bind musical fit-for-use — rights, function, register — are not text either, in the sense of being in the captions paired with audio at training time. The text modality re-introduces some of what audio leaves out, but it relocates the absence rather than dissolving it. The nearest relative is Sturm's diagnosis of flagship MIR tasks — genre recognition, emotion recognition — as ill-defined problems: systems rewarded for reproducing ground-truth labels can succeed as "horses," responding to confounded correlates of categories whose membership criteria were never in the signal (Sturm, 2013, 2014). Amnesia offers one account of why such categories resist acoustic definition: their binding criteria are extra-signal and relational — what counts as Christmas music, or as "world music," is settled by communities and markets, not by waveforms — so a system given only audio can track them only by proxy.

2.3 Will multimodality solve it?

The natural reply to embedding amnesia is the engineering reply: pair audio with text, metadata, and structured ontologies, and amnesia closes. The reply is partly right. A multimodal model with adequate annotation can encode some of what audio cannot — language identifiers, genre tags, basic mood vocabulary. Even here the burden falls on collecting data that accurately describes musical context and listeners — the contextual facets Schedl et al. (2014) identify as MIR's persistent hard case, harder to acquire and quicker to perish than content features. It is partly wrong because the loci we develop below are not modalities of the work; they are relational facts about its place in a social, legal, and temporal world. Liturgical season is a calendrical relation between a track and a community. Rights clearance is a contractual relation varying by jurisdiction. Cultural intelligibility is not annotatable in the way "major key" is annotatable, because it is constituted in use. Multimodal closure absorbs some forgotten content; the remainder is what Born (2010) calls the social plane of mediation, which is not absorbed by adding a vector.

2.4 Four loci

Across the deployed retrieval contexts we know from practice (see Section 3), embedding amnesia manifests in at least four loci. We name them not as exhaustive but as illustrative; each is a place where curator judgment routinely supplements cosine ranking, and each is a place where additional contrastive training does not change the underlying constraint.

(i) Linguistic meaning. Audio embeddings encode a sense of vocal timbre, phonetic texture, and prosody, but they do not encode language identity, semantic content, or lyrical reference in any reliable way. A listener who wants pop in one language and is offered a high-cosine acoustic neighbor in another reads the recommendation as not-a-fit on a basis the encoder has no access to.

(ii) Liturgical, seasonal, and ritual function. Music carries calendar. Christmas songs are Christmas songs not because of any acoustic property unique to them — jingle bells, after all, can decorate a summer pop song with no calendrical implication — but because of a social agreement about when and why they are heard. The same is true of devotional music, of music for funerals, of children's repertoire, of music for celebration or mourning or work. A concrete case: the trio from Elgar's *Pomp and Circumstance* March No. 1 is, in the United States, a graduation processional, and in Britain, "Land of Hope and Glory," a patriotic anthem with its own occasions — one waveform, two incompatible ritual bindings, neither recoverable from the signal. The waveform contains some traces of these uses, but the binding of waveform to occasion is held in communal agreement and its metadata traces, not in the signal.

(iii) Cultural intelligibility. A track is heard within a cultural register. A protest song, a wedding song, a club anthem, a parodic cover — each carries a load of in-group/outgroup cues, ironic distance, citational reference, that listeners read fluently. The parody case is the sharpest: "Weird Al" Yankovic's *Eat It* is acoustically engineered to sit as close as possible to Michael Jackson's *Beat It* — same groove, same guitar timbres, same vocal register — because the joke depends on the proximity. Whatever an encoder learns about the pair, maximal nearness is the correct acoustic answer and the wrong curatorial one. Encoders capture some surface correlates of register, but the meaning that holds the register together is communal and historical. The substrate's grasp on this remains partial; what is partial is not an artifact of the present generation of encoders but a structural feature of the audio-only modality.

(iv) Rights, clearance, and availability. The fourth locus is qualitatively different from the first three. Whether a track may be served depends on licensing, regional availability, sample clearance, and performance rights — institutional availability constraints, legal-textual artifacts that downstream systems must consult regardless of how the embedding is built. We include rights here not because the substrate should encode it, but because curator judgment routinely encodes it alongside the other three in deployed practice, and because it makes vivid that the binding constraints on retrieval are not all of one kind.

These four loci are not ranked, and they are not exhaustive. They are markers of a wider structural condition: any representation built from one modality leaves what is not available in that modality outside its grasp, and the leaving-outside is constitutive.

A useful image: the encoder is a microphone; the curator is a sensor with a life. The microphone has a frequency response, a polar pattern, a noise floor; what falls outside it is not in the recording. The curator hears with legal awareness, generational context, calendar, brand fluency, and community position, activated selectively against the brief. Different sensors, different ears.

2.5 What the typology predicts

Table 1 crosses the two axes implicit in the four loci. The typology matters because it predicts what kinds of intervention can address what kinds of forgetting. Stable extra-signal facts (rights) are addressable by lookups against authoritative external sources. Relational extra-signal facts (calendar, function) are addressable by maintained metadata that tracks occasion. Relational in-signal facts (cultural intelligibility, register) are not addressable in any straightforward way, because they require the listening community as an interpretive layer. Conflating all four under a single "add another vector" fix is the failure of imagination this paper is about.

The axes are older than the embedding era. Schedl et al. (2014) partition MIR's subject matter into music content and music context, user properties and user context; the extra-signal loci live in their contextual quadrants, and none of the absences catalogued here awaited embeddings to exist. What the embedding era changes is that a substrate built from content alone is now asked to stand in for all four.

3 STANDPOINT AND FIELD OBSERVATION

The four loci above are not theoretical entities. We name them from a particular vantage on deployed retrieval practice, and the vantage is part of the argument. We write from inside a non-generative ML practice that operates without large-scale behavioral signal — not the platform-streaming context most ML criticism addresses, but a register in which curatorial judgment is the primary signal that says whether a recommendation lands. In platform streaming the binding decision is largely taken by skips and listen-throughs, and the agnostic skip absorbs every reason for it. In the register we describe, curators name the reason, and the named reasons are what Section 2 catalogues. The standpoint conditions what we can see and what we miss; we treat it as a feature of the work, in the practitioner-positioned spirit Sturm et al. (2024) call for and Seaver (2022) models from the engineering side.

To be explicit about method: this is a practitioner-theoretical reflection, and no corpus is analyzed. The scenarios below are stylized composites — situated observations from practice, with identifying detail removed — offered as diagnostic exposures of representational mismatch rather than as frequency claims. An empirical companion would require a locus-labelled corpus of curator rejections; Section 6.1 sketches what such a benchmark would look like.

Three scenarios, one per extra-signal or relational locus. *Calendar:* a summer brief retrieves an acoustic neighbor whose post-chorus quotes a holiday standard — close in groove, impossible by season. *Cultural intelligibility:* a brief asking for a sincere love song retrieves a parodic cover of one; the parody is constituted by citation the encoder cannot read. *Language and listener-in-room:* a brief asking for energetic background music with no English vocals retrieves a track whose second verse competes with on-screen dialogue. In each case the encoder has done what it was trained to do; what is left over is what the curator's ear holds.

4 THEORETICAL READING

To frame embedding amnesia as more than a list of complaints — to argue that the partial ear of machine listening has theoretical and historical depth — we situate it across three disciplinary readings: media theory, algorithmic criticality, and musical ontology. Each recovers different stakes for the same finding.

Table 1: The four loci, organized by whether the binding constraint is recoverable from audio (*in-signal*) or external to it (*extra-signal*), and by whether it is a stable property of the work or a relational fact about its use. Mitigations are not solutions; they indicate the kind of intervention available, not the closure of the gap.

Locus	In audio?	Stable / relational	Example failure	Available mitigation
Linguistic meaning	in-signal	relational	High-cosine acoustic neighbor in a language the listener does not want or read as devotional	Language ID + tagged metadata; relational binding still left to interpretation in use
Liturgical / seasonal function	extra-signal	relational	Summer-pop seed retrieves a Christmas standard with aligned timbre and groove	Calendar metadata, season tags, occasion-aware filters
Cultural intelligibility	in-signal	relational	Parodic cover read by encoder as nearest neighbor of its sincere source	No straightforward mitigation; requires the listening community as interpretive layer
Rights, clearance, availability	extra-signal	stable	Acoustically near track is unlicensed for the requested territory or use	Authoritative external lookup against contractual / statutory sources

4.1 Media Theory: The Substrate Speaks

Friedrich Kittler argued that "media determine our situation" (Kittler, 1999): the apparatus through which sound, image, or text passes is not neutral. The phonograph does not record sound; it records a particular subset, defined by what its diaphragm can transduce, what its needle can inscribe, what its wax can hold. What falls outside this subset is gone. The recording is the event minus what the medium could not carry.

The audio embedding is the latest entry in this lineage. It is a transduction layer converting a waveform into a vector with learned correspondences. Like Kittler's gramophone, it determines its situation by what it can and cannot pick up. The four loci we have named are the room-tone above the embedding's frequency: never going to be in the signal, and the signal is what the encoder takes. Contemporary machine listening is not listening to music; it is listening to what the encoder permits to count as music, which is not the same.

This reading does not require nostalgia for some pre-digital intelligibility. Every medium has its substrate logic; the gramophone forgets differently than the score, the score differently than live performance. What matters is the discipline of asking, of any new substrate: what does this one carry, and what does it leave outside?

Vilém Flusser's philosophy of technical apparatuses sharpens the point from a complementary direction. For Flusser, the camera is not a tool but an apparatus executing a *program*: the space of possible photographs is inscribed in the apparatus in advance, and the photographer who accepts its categories operates as a *functionary* of that program (Flusser, 2000). The trained encoder is an apparatus in exactly this sense. Its program — input modality, architecture, contrastive objective — fixes the space of possible hearings before any track arrives, and a system that accepts cosine ranking as the space of musical relation has become the program's functionary. Where Kittler names what the substrate cannot inscribe, Flusser names a second, subtler failure: mistaking the program's space of outputs for the space of the possible. His counter-move is instructive — freedom, for Flusser, is *playing against the apparatus*. The curator who overrides a nearest neighbor for reasons the encoder cannot register is playing against the apparatus in this precise sense, a figure we return to in Section 6.1.

4.2 Algorithmic Criticality: Producing without Understanding

The second reading concerns what happens when an amnesiac substrate is asked to scale. Bender et al. (2021) argue that large language models, trained on vast corpora to predict plausible continuations of text, generate fluent output without comprehension — "stochastic parrots" that approximate the surface form of language without being grounded in the meaning that language conveys. The resulting systems are powerful, deployed widely, and held by their builders to be reasoning engines; they are also, the argument goes, structurally incapable of the comprehension their fluency implies.

The argument transfers cleanly to music. A learned audio representation, trained on contrastive or generative objectives over a large corpus of waveforms, learns to predict plausible continuations or plausible neighbors. The fluency is real — the recommendations are acoustically appropriate, the generations sound like music. The comprehension is the question. To recommend a track in the wrong language at high cosine similarity is to display fluency without comprehension. To generate a track that sounds appropriate but is, on closer reading, a Christmas song offered for a summer playlist, is the same. The substrate has produced something. Whether it has understood anything is a different question.

The political stake, in Bender et al.'s account, is that fluency is routinely mistaken for comprehension when systems are placed in roles — search, recommendation, generation, decision support — that presume comprehension. The same confusion attends embedding-based music systems. Curators in the practice of Section 3 read acoustically-fluent-but-contextually-wrong recommendations as not-fits and supply the relational binding the encoder cannot. End listeners, without that mediation, have no equivalent vantage and receive the output as the system's understanding.

4.3 Musical Ontology: What Is Music That an Embedding Forgets It?

The third reading asks the prior question. To say that an embedding forgets something about music presupposes that music has a something to forget. Born (2010), and more recently in the introduction to her edited volume *Music and Digital Media* (Born, 2022), argues against a purely acoustic account: music is constituted relationally, across what she names four planes of social mediation — the immediate

sociality of performance and audition; the constitution of broader social identity formations through musical practice (genre, scene, generation); the institutional infrastructures within which music is produced and circulated; and the wider socio-economic and political conditions that shape these institutions.

Read against Born, embedding amnesia is not an engineering problem at all. It is an ontological mismatch. The audio embedding takes the waveform as if it were the music, when the waveform is one component of an object constituted across all four planes. Our four loci map onto Born's planes: cultural intelligibility belongs to the first and second; liturgical and seasonal function to the second and third; rights and clearance to the third and fourth; linguistic meaning cuts across all four. The mapping makes the heterogeneity of the loci visible as a feature: each is unfixable in a different way, because each lives on a different plane. Born has since pressed the point on MIR directly: the field's representations inherit ontological premises from a narrow slice of the world's musical practices, and diversifying its knowledge means revising those premises, not only enlarging datasets (Born, 2020). Amnesia names one mechanism of that narrowness — whatever a tradition binds extra-signally is precisely what a content-trained substrate will misread.

Born's framework also explains why curator agreement on retrieval fit is moderate rather than high in deployed practice: the relational object is not single-valued, and two trained curators bring different positions across the planes. Seaver (2022) arrives at a related point ethnographically — building recommenders is the construction of a "world of music" in which decisions about similarity are simultaneously decisions about which planes count. Drott (2018) makes the political-economy reading: a recommendation is not a guess about a pre-existing preference but an act of constructing a listener whose preference can be served. Embeddings absorb none of this. They treat the recommendation as if a single correct judgment lay somewhere in the audio; the relational ontology of music is the very thing they cannot encode and curators cannot help but encode.

5 THE GENERATIVE TURN

So far the framing has come from retrieval — the matching of an existing track to a seed track. The reader may reasonably ask whether the argument extends to generation, where music is not retrieved but *spawned*, in the conference's apt term. We argue that it does, and that the stakes are higher in the generative case.

Generative music systems — MusicLM (Agostinelli et al., 2023), MusicGen (Copet et al., 2023), and the rapid sequence of follow-ons that produce audio from text prompts, latent codes, or conditioning signals — depend on related learned audio and audio-text representations, even when their decoding architectures differ from retrieval systems. They embed audio (or text-paired audio) into a learned space; they sample from that space; they decode samples back into waveforms. Whatever the embedding cannot encode in the retrieval direction, it cannot encode in the generative direction either. A model that does not represent "this is a Christmas song" as a separate dimension cannot reliably avoid producing a Christmas song when prompted to write a summer pop track. A model that does not represent "this lyric refers to gun violence" cannot reliably refrain from generating one for a children's catalog. A model that does not represent "this sample is unlicensed" cannot refuse to interpolate it. The structural absence of an attribute from the substrate is a structural absence from everything the substrate produces.

This does not mean generative systems fail catastrophically. They mostly do not. They fail diffusely. They produce music that sounds plausible across an enormous range of conditioning signals and, on closer inspection, often fails along dimensions the human commissioner cared about but the substrate could not register. A platform that spawns millions of tracks daily has millions of opportunities to produce something acoustically appropriate but contextually wrong. The retrieval reading becomes, in the generative regime, we conjecture, not a per-recommendation curiosity but a continuous low-grade flattening at scale.

The conference theme proposes that "music has entered a state of hyper-reproduction." We agree, and we add: the substrate doing the reproducing is amnesiac. Hyper-reproduction without representational depth is a recipe for the flattening of *aesthetic signatures*, not because the models are biased toward a majority — though they often are — but because the dimensions along which difference might be preserved are dimensions the substrate has never registered. The flattening is not malicious. It is constitutive. Linguistic specificity collapses because language is not in the embedding. Liturgical specificity collapses because calendar is not in the embedding. Cultural specificity collapses because the relational network constituting the work is not in the embedding. What the substrate cannot tell apart, the generative system cannot keep apart — and this is the condition under which *new forms of musicianship* and *distributed creative agency* are now being asked to emerge.

Stated more sharply: hyper-reproduction does not *dilute* amnesia, it *multiplies* it. The sense of "multiplies" is narrow. Each generated track is a fresh draw from the same amnesiac substrate, so opportunities for contextually-wrong output scale linearly with throughput, while the mediating layer that catches such output — curatorial attention — does not scale with it. Nothing about any single draw worsens; what grows is the unmediated share. The error is not amplified; the exposure to the failure-mode is. This is a structural argument about scaling, not a measurement (we flag it as conjecture in Section 6.3), but its direction does not depend on the numbers: throughput is the quantity hyper-reproduction maximizes, and relational binding is the labor it economizes away.

A small composite illustration. Consider a brief that asks for warm, contemporary, female-led music for a Tuesday morning in November — explicitly no Christmas. An audio embedding can return tracks acoustically near a seed pulled from a prior November playlist; some of those returns will be unusable for reasons the embedding cannot register (a sleigh-bell timbre fine in July, impossible now; a gospel melisma read as devotional; a hit whose post-chorus has been adopted by a competing brand's recent campaign). A generative system asked to produce music for the same brief inherits the same blindnesses: it may write something passable while still producing, with non-trivial frequency, material the brief excludes by definition. This is the small unit out of which hyper-reproductive flattening is plausibly composed.

The standard responses — prompting, fine-tuning, retrieval-augmentation, content moderation — are forms of com-

pensation built around an amnesiac center. Prompts reintroduce textual specificity; metadata filters re-introduce calendrical and rights specificity; curatorial review reintroduces cultural intelligibility. They acknowledge through their existence what is not in the substrate. The interesting research question is whether the substrate itself can be reconceived to make fewer of them necessary — through paired multimodal training, through explicitly relational representations, or through hybrid architectures that bind the encoded waveform to a regularly-updated map of cultural and legal context. Until something of that order arrives, "hyper-reproduction" is what amnesiac scale produces, and "mediated musicianship" is what we propose to call the corresponding creative labor on the human side.

6 AFTER AMNESIA

A diagnostic paper without a constructive turn is incomplete. We do not propose to solve embedding amnesia — the argument of the preceding sections is that it is not solvable within the audio-representation paradigm — but we develop one creative-research orientation in detail — *the brief as a score* — and gesture briefly at three companions, in the spirit of the recent in-community programmatic statements on AI Music Studies (Sturm et al., 2024; Kaila and Sturm, 2024) and the relational framing of human–AI musicking offered by Vear et al. (2023).

6.1 The brief as a score: curators as composers of the listening occasion

The work that allows retrieval and generation systems to function in deployment is not a back-office layer of moderation. We propose a re-reading of curatorial labor that AIMC, more than any neighboring venue, has the vocabulary to receive: in deployed practice, the binding artifact is not the track. It is the *brief*. The brief specifies an occasion — mood, register, cultural fit, calendar, listener-in-room, rights territory — against which a track is selected as a performance. Read carefully, the brief is a kind of score. It is a notational artifact that defines a listening occasion; the track that satisfies it is a realization of that score; the curator who weighs candidates against the brief is doing the work of score-following in reverse. We submit this as a candidate for the *new forms of musicianship* the conference call invites: not authorship-as-prompting and not orchestration-as-curation, but composition-of-the-occasion, distributed across encoder, brief, and curator.

A worked enactment. Brief: *warm, female-led, contemporary, secular, no holiday cues, morning retail*. Read against Table 1, each clause does a different kind of work. *Warm, female-led, contemporary* are in-signal: the encoder is genuinely competent here, and the brief delegates them to cosine ranking. *Secular* and *no holiday cues* are extra-signal, relational: they bind the selection to a calendar and a community, and the brief carries them precisely because the embedding cannot. *Morning retail* names the listener-in-room — a relational fact about the occasion, not the track. (In fuller practice a rights-territory clause would add the stable extra-signal locus; we omit it here for compactness.) Now the selection: Candidate A is acoustically nearest a seed but quotes a holiday standard in its post-chorus; Candidate B is less near and honors every binding constraint. The curator chooses B. Note what has and has not happened to the amnesia. It is not transcended — no representation has been repaired. It has been *redistributed*: each extra-signal or relational constraint has been moved out of the substrate and into a notational artifact where a human can hold it, and the encoder has been confined to the loci where its ear is adequate. The brief, read as a score, is a division of representational labor — and the choice of B over A is not error-correction but the act of deciding which constraints are binding.

Three implications follow.

The labor is creative, not corrective. A curator selecting a track for a brief is not fixing a model's mistake; the model never had access to the brief. The curator authors the relational binding the embedding cannot supply, and the catalog assembled, pruned, and updated over time is a slowly-composed work. This is consistent with the values frame Cotton and Tatar (2023) develops for musical AI: care is a constitutive practice, not an afterthought.

The tools belong on the composer side. A composer's tool affords iteration, comparison, citation, revision over time; a moderator's tool affords approval and rejection. Most "human-in-the-loop" interfaces sit on the moderator side. Tools that edit the brief itself, annotate which relational facts a candidate honors or breaks, and hold a brief over weeks of catalog development are an under-occupied design space we name rather than propose to fill.

The reading folds back into evaluation. Reporting AUC against a curator panel mixes the encoder's acoustic competence with the panel's tacit application of constraints the encoder does not have. Constraint-conditional reporting — accuracy on items that satisfy the binding constraints, with violation rate per locus — separates the two; a locus-labelled diagnostic set, a *benchmark of amnesia*, would let future encoders be evaluated for what they cannot register, not only for what they can.

6.2 Three companion orientations, briefly

Architectural humility. Hybrid systems — audio encoders bound to rule-based filters for rights, calendar-aware filters for season, language tags as constraints — are not workarounds; they are the architecture. The external layers are load-bearing, and their tools deserve to be developed as carefully as the encoder.

Composition in the gaps. If the brief is the score and the embedding is amnesiac about it, there is design space for generative systems whose primary conditioning is the brief and whose audio is auxiliary — the inverse of the present architecture, in which audio is primary and text is a thin caption.

Anti-amnesiac instruments. A separate creative move is to use the gaps productively. Tools that generate not audio but *prompts to the curator*, designed to surface relational facts the encoder cannot reach (*is this a holiday track for your audience? a parodic cover? unsuited to a children's catalog?*), invert the generative paradigm: AI as prompt for human creative attention rather than as replacement for it.

6.3 What this paper does not claim

The contribution is theoretical. We do not validate the generative-flattening conjecture of Section 5 empirically; that claim awaits locus-specific failure-rate measurements of the kind a benchmark of amnesia would support. We do not propose specific instruments for the design space named in Section 6.1. We do not establish that working curators themselves prefer the brief-as-score self-description;

that is an empirical claim about practitioner identity. And we do not claim the four loci are exhaustive; they are diagnostic of a wider condition.

7 CONCLUSION

We have proposed embedding amnesia as a name for the structural underdetermination of certain features by learned audio representations. We have named the practitioner standpoint from which the loci are visible, illustrated by stylized composites rather than numbers, and read the concept across three theoretical traditions — Kittler and Flusser, Bender et al., Born — to argue that the engineering observation has historical and ontological depth. We have turned the same argument toward the generative regime, where the substrate's amnesia, we conjecture, becomes a continuous low-grade flattening of the specificities a hyper-reproductive medium might be expected to honor.

The constructive turn is the brief-as-score reading: in deployed practice the binding artifact is not the track but the brief, and the work of selecting a track against a brief is composition. The reading re-positions the curator as a composer of the listening occasion, and changes what tools, evaluations, and generative systems belong around the work.

The primary question of the generative era is not what AI can produce but what it can authoritatively understand — and on a substrate whose ear is constitutively partial, what it can produce will exceed what it can understand. Closing the gap is a matter of building representations that include what audio leaves outside, of building practices that compensate for what they leave outside, or, most ambitiously, of treating the gaps themselves as a creative substrate. Hyper-reproduction makes this a question about the politics of music-making: which specificities a generation of listeners will be invited to attend to, and which a substrate will quietly let go.

ACKNOWLEDGMENTS

The author thanks the curators at Feed Media Group — Eric "Stens" Stensvaag, Juan Hernandez-Cruz, Claire McConnell, and Mike Savage — whose daily judgments about what a brief binds and what an embedding misses are the situated ground of this paper's argument. Thanks to Michael Beaver, Eric Lambrecht, Brita Nordin, and Jay Caston for supporting and steering the work; to Jeff Yasuda and Feed Media Group for backing this line of research; and to Juan Sebastián Gómez-Cañón, whose research shaped an early version of the embedding system this paper reflects on. Thanks also to the two anonymous AIMC reviewers, whose comments sharpened the typology's examples, the generative-scaling claim, and the dialogue with Flusser.

ETHICS STATEMENT

This paper writes from a practitioner standpoint inside a deployed retrieval practice; the illustrative scenarios in Section 3 are stylized composites and do not report identifiable curator judgments, listener data, or proprietary catalog details. The argument is theoretical and does not rest on empirical claims drawn from any particular deployment.

The audio representation models referenced as illustrative examples (CLAP, MERT, and the broader class of joint audio-text and self-supervised audio encoders) are publicly released by their authors and trained on third-party data per their respective published terms; this paper engages their behavior conceptually and does not redistribute, fine-tune, or extend them.

The embedding-amnesia framing names cultural categories (linguistic, devotional, seasonal, children's, lyrical, popularity) whose acoustic markers, if treated asymmetrically by downstream systems, risk algorithmic harm to communities defined by non-dominant languages, traditions, or aesthetics. The paper's argument is precisely that these categories are not reducible to acoustic markers and therefore should not be inferred from audio alone; we recommend that any practitioner adopting the embedding-amnesia framework audit downstream systems for asymmetric treatment of non-dominant contexts.

AI USAGE STATEMENT

Per the AIMC 2026 policy on the use of artificial intelligence, we disclose limited use of large-language-model assistants during the preparation of this paper. AI tools were used for grammar, syntax, readability, and aesthetic copy-editing of author-drafted prose, and for technical assistance with the AIMC LaTeX template (formatting, bibliography hygiene). The research questions, theoretical framework, conceptual definitions, theoretical readings, choice of empirical material, and conclusions are the author's own. AI was not used to generate research content, ideas, analyses, or arguments. The author has read every line of the submission and is responsible for its content.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text, 2023.

Jean-Julien Aucouturier and François Pachet. Improving timbre similarity: How high's the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623, 2021. doi: 10.1145/3442188.3445922.

Georgina Born. For a relational musicology: Music and interdisciplinarity, beyond the practice turn. *Journal of the Royal Musical Association*, 135(2):205–243, 2010.

Georgina Born. Diversifying MIR: Knowledge and real-world challenges, and new interdisciplinary futures. *Transactions of the International Society for Music Information Retrieval*, 3(1):193–204, 2020. doi: 10.5334/tismir.58.

Georgina Born. Introduction — music and digital media: A planetary anthropology. In Georgina Born, editor, *Music and Digital Media: A Planetary Anthropology*. UCL Press, London, 2022.

Òscar Celma, Perfecto Herrera, and Xavier Serra. A multimodal approach to bridge the music semantic gap. In

Poster and Demo Proceedings of the 1st Int. Conf. on Semantic and Digital Media Technologies (SAMT), CEUR Vol. 233, Athens, 2006.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kelsey Cotton and Kıvanç Tatar. Caring trouble and musical AI: Considerations towards a feminist musical AI. In *Proceedings of the 4th Conference on AI Music Creativity (AIMC)*, 2023. URL https://aimc2023.pubpub.org/pub/zwjy371l/release/1.

Eric Drott. Why the next song matters: Streaming, recommendation, scarcity. *Twentieth-Century Music*, 15(3): 325–357, 2018.

Eric Drott. *Streaming Music, Streaming Capital*. Duke University Press, Durham, NC, 2024.

Vilém Flusser. *Towards a Philosophy of Photography*. Reaktion Books, London, 2000. Translated by Anthony Mathews. Originally published 1983.

Perfecto Herrera, Juan Bello, Gerhard Widmer, Mark Sandler, Òscar Celma, Fabio Vignoli, Elias Pampalk, Pedro Cano, Steffen Pauws, and Xavier Serra. SIMAC: Semantic interaction with music audio contents. In *Proc. 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT)*, London, 2005.

Anna-Kaisa Kaila and Bob L. T. Sturm. Agonistic dialogue on the value and impact of AI music applications. In *Proceedings of the 5th Conference on AI Music Creativity (AIMC)*, 2024. URL https://aimc2024.pubpub.org/pub/p3rww87r/release/1.

Friedrich A. Kittler. *Gramophone, Film, Typewriter*. Stanford University Press, Stanford, CA, 1999. Translated by Geoffrey Winthrop-Young and Michael Wutz.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: Acoustic music understanding model with large-scale self-supervised training. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2024.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

James E. K. Parker and Sean Dockray. 'All Possible Sounds': Speech, music, and the emergence of machine listening. *Sound Studies*, 9(2):253–281, 2023.

Markus Schedl, Emilia Gómez, and Julián Urbano. Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2–3):127–261, 2014. doi: 10.1561/1500000042.

Nick Seaver. *Computing Taste: Algorithms and the Makers of Music Recommendation*. University of Chicago Press, Chicago, 2022.

Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000. doi: 10.1109/34.895972.

Jonathan Sterne. Is machine listening listening? *communication +1*, 9(1), 2022.

Bob L. T. Sturm. Classification accuracy is not enough: On the evaluation of music genre recognition systems. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.

Bob L. T. Sturm. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

Bob L. T. Sturm, Ken Déguernel, Rujing Stacy Huang, Anna-Kaisa Kaila, Petra Jääskeläinen, Elin Kanhov, Laura Cros Vila, David Cabrera Dalmazzo, Luca Casini, Oliver Bown, Nick Collins, Eric Drott, Jonathan Sterne, Andre Holzapfel, and Oded Ben-Tal. AI music studies: Preparing for the coming flood. In *Proceedings of the 5th Conference on AI Music Creativity (AIMC)*, 2024. URL https://aimc2024.pubpub.org/pub/ej9b5mv1/release/2.

Craig Vear, Steve Benford, Juan Martinez Avila, and Solomiya Moroz. Human-AI musicking: A framework for designing AI for music co-creativity. In *Proceedings of the 4th Conference on AI Music Creativity (AIMC)*, 2023. URL https://aimc2023.pubpub.org/pub/zd46ltn3/release/2.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.