Beyond Engagement: A Rule-Based Approach to AI-Driven Music Recommendation for Emotional Wellbeing

Kieran Hertis Raymond White
Department of Engineering & Computing
University of East London
London, England
u2637328@uel.ac.uk

Dr Rahime Belen Saglam
Department of Engineering & Computing
University of East London
London, England
R.Belen-Saglam@uel.ac.uk

Abstract

Commercial music recommendation platforms are typically optimised for engagement rather than emotional wellbeing. This paper presents MoodFlow, a prototype AI-driven playlist generation system that prioritises mood regulation using music therapy principles. The system integrates a custom NLP ensemble, combining a GoEmotions-based transformer, with a domain-specific keyword lexicon, to classify free-text mood input; a three-tier rule engine, informed by a professional music therapist survey (N=15), public survey (N=203), and the MxMH dataset (N=736), maps emotional states to target audio features. The ISO-Principle is operationalised as a dynamic blend ratio between current and target mood states, weighted by user-reported intensity. Evaluation comprised functional testing (46 cases, 96% pass rate) and a pilot user study (32 sessions). Participants reported high perceived emotional safety (mean=4.68/5), and a moderate correlation (R^2=0.36) found between perceived playlist relevance and self-reported emotional change. These findings provide preliminary evidence for incorporating music therapy principles into recommender systems.

1 INTRODUCTION

As the advancements in AI continue, the prevailing assumption is that the future of AI in music is generative, automated, and oriented toward ever-higher output. Platforms like Spotify, Apple Music, and their generative successors optimise for a single metric: continued listening. Engagement is both the design objective and the measure of success. While engagement-based optimisation has proven commercially effective, it does not necessarily align with users' emotional wellbeing. Systems designed to maximise listening time may reinforce existing moods, including negative or maladaptive emotional states, rather than supporting recovery or regulation.

This paper reports on a system that deliberately runs counter to that notion. MoodFlow[1] is a retrieval-based, rule-driven playlist generator built not to maximise listening time, but to support emotional progression. The question it was designed to answer is straightforward: what changes when therapeutic outcomes replace engagement as the objective function of a music recommendation system?

The question is not merely theoretical. Shen et al. (2024) highlight the growing use of generative AI in music for mental health applications, while also identifying risks associated with emotionally misaligned outputs. Such systems may produce acoustically plausible content that does not reliably support users' emotional states, raising concerns for therapeutic use.

Sun et al. (2024), through co-design research with professional music therapists, report a clear preference for systems that are explainable and controllable, rather than fully automated. The demand is not for less AI, but for AI whose reasoning is auditable. No widely accessible consumer tool currently meets that standard.

Narayanan and Tarafdar (2025) further observe that existing commercial systems optimise for engagement signals rather than emotional resonance. While mood-aware recommendation is technically feasible, many approaches rely on opaque machine-learning models that remain difficult to validate in clinical contexts.

MoodFlow addresses these gaps through a different architectural choice: a rule-based system whose decision logic is fully inspectable, grounded in primary data from qualified professionals, and constrained by clinical safety as a non-negotiable first principle.

Beyond the architectural argument, this paper situates itself within the emerging discourse on curation and listening ecologies: specifically, what happens to platform infrastructure and its users when the objective function of a recommendation system is reoriented from engagement toward therapeutic outcome. In doing so, it engages directly with questions of curation infrastructure, platform governance, and the ethics of deploying AI in emotionally sensitive listening contexts.

[1] Accessible at: MoodFlow - AI Music Therapy (Section 6.5 Limitations Apply)

2 RELATED WORKS

2.1 Music Therapy Principles in Digital Systems

The ISO-Principle, as examined by Starcke, Mayr and von Georgi (2021), is one of the most widely cited concepts in clinical music therapy. Its operational definition, in a controlled experimental study, is precise: patients first listen to music that matches their current emotional state, then the music gradually shifts toward a desired target state. This confirmed the principle's efficacy for sadness induction and modulation, providing empirical grounding for its use in system design. Despite this, the ISO-Principle is almost entirely absent from computational music recommendation research.

To operationalise emotional progression computationally, models of affect are required. Han et al. (2022) does so by drawing on James A. Russell's Circumplex Model of Affect (Russell, 1980). This model offers a widely used framework, representing emotions along continuous dimensions of valence and arousal. This structure maps naturally to audio features and enables the classification of emotional states in a way that supports systematic mood transitions. It forms the theoretical basis for the mood categorisation used in MoodFlow, as well as its alignment with the GoEmotions taxonomy described in Section 3.

2.2 AI Music Recommendations: Capabilities and Limitations

While AI-driven music recommendation systems have become increasingly sophisticated in modelling user preferences, their design remains largely optimised for engagement rather than emotional understanding or therapeutic impact. Panda et al. (2021) benchmarked Spotify's commercial audio feature set against dedicated Music Emotion Recognition (MER) frameworks, showing that while Spotify's features correlate with perceived emotion, they oversimplify emotional nuance compared to specialised approaches. This limitation is significant for therapeutic contexts, where distinguishing between closely related emotional states requires greater granularity.

Sadhu et al. (2025) analysed the MxMH dataset across six machine-learning classifiers to assess the relationship between musical habits and mental health outcomes. Their methodology is comprehensive, but the resulting system is a diagnostic and predictive tool rather than a therapeutic one, with the inclusion of classifier models, ranging from K-Nearest Neighbours to Gradient Boosting, functioning as black boxes. Through this, a correlation between musical features and mental health scores is established, but the causal rules needed for therapeutic intervention are not surfaced. The system demonstrates that the relationship exists; it cannot explain why, or how to act on it safely.

Narayanan and Tarafdar (2025) take a closer step toward therapeutic design by integrating lyrical sentiment analysis alongside audio features and health reflections. Their demonstration that tonal dissonance - where a song's acoustic properties conflict with its lyrical content - affects therapeutic efficacy directly informed the vocal and instrumental filtering logic in MoodFlow.

2.3 The Transparency and Safety Gap

Sun et al. (2024) conducted co-design research with professional music therapists to understand the requirements for AI integration in clinical practice. The findings are unambiguous: therapists reject autonomous AI recommendation in favour of explainable, controllable systems. Two specific requirements emerged: users must be able to understand why a recommendation was made, and clinicians must be able to audit and override system decisions. The study also found that patients' trust in AI-driven tools is fragile and heavily dependent on perceived safety.

Shen et al. (2024) extended this concern to generative systems specifically, documenting the hallucination risk and concluding that human oversight remains essential when AI-generated audio is used in clinical contexts.

3 SYSTEM DESIGN AND ARCHITECTURE

3.1 Design Philosophy

Three constraints shaped the architecture from the outset. Clinical safety must override all other considerations: any parameter flagged as a risk by qualified professionals cannot be relaxed by user preference or algorithmic inference. Second, the system's reasoning must be fully inspectable: every playlist must be accompanied by a plain-language explanation of why it was generated. Third, the system should operationalise specific, well-understood music therapy principles rather than attempt to replicate clinical judgment. These constraints currently favour retrieval-based approaches over generative ones, which cannot currently satisfy the first and second requirements at present capability levels, and exclude opaque ML classifiers for the same reasons.

Table 1: Design principles derived from data analysis categorised by their sources (See Section 4.1 for results)

Data Source	Design Principle	Manifestation
Professional + Public	Gradual mood transitions	Audio features blend between current and target state, weighted by intensity
Professional	Neurodivergent accessibility	Dedicated filter capping energy, narrowing tempo range
Professional	User agency in crisis	Choice-based intervention with persistent support resources
Professional	Therapeutic vocabulary	Clinical terms replaced with alternatives
Professional + Public	Severity-based filtering	User toggle for vocals; instrumental bias at high distress

Literature Review	Granular emotion-to-mood mapping	28 AI-detected emotions mapped to seven therapeutic categories
Literature Review	Explainable recommendations	Every playlist includes a rationale card citing data sources

Each design principle in Table 1 derives directly from specific research finding. The gradual mood transition logic blends audio features between current and target states, weighted by user-reported intensity on both sliders, rather than jumping immediately to the target mood. The neurodivergent filter caps energy, narrows tempo range, and excludes acoustically unpredictable genres, reflecting 80% professional consensus on its necessity. The crisis intervention presents three user-controlled options rather than halting or ignoring the signal, implementing the majority professional position from a three-way split. Clinical mood labels are replaced with functional alternatives displayed to the user whilst the system processes the clinical term internally. The instrumental filter applies automatically at high distress intensity, reflecting both the public survey split and professional lyrical risk guidance. GoEmotions' 28 labels are mapped to seven therapeutic categories via Russell's Circumplex, and every playlist includes a rationale card implementing the transparency requirement of Sun et al. (2024).
MoodFlow connects directly to Spotify via the Spotify Web API and Web Playback SDK. Users authenticate with their Spotify account (must be added to the API and have Spotify Premium); the system retrieves tracks from their library or via mood-aware search queries, then plays them in-browser. The full system flowchart is found via the link to the supplementary materials in Appendix B and source code is accessible at: GitHub Repository

3.2 NLP Ensemble for Mood Classification

User mood input is free text. Classification uses an ensemble of two components: the *SamLowe/roberta-base-go_emotions* model (Lowe, 2023), derived from the original GoEmotions (Demszky et al., 2020), is a RoBERTa variant fine-tuned on 58,000 Reddit comments across 28 emotion labels, and a domain-specific keyword lexicon developed for this project. A consensus mechanism determines the final classification: where both engines agree above a 60% confidence threshold, the classification is accepted. The 60% confidence threshold was selected as an implementation choice, calibrated against the developer test register to balance false positive avoidance against over-reliance on the teach-the-system fallback Where they disagree or fall below threshold, the system either presents the user with a clarifying prompt or routes to the teach-the-system mechanism.
GoEmotions was selected over standard sentiment analysis because the therapeutic response to anger and sadness differs substantially, despite both registering as negative sentiment in binary models. The 28 GoEmotions labels were mapped to seven therapeutic mood categories (Anxious, Introspective, Angry, Restless, Content, Happy, and Energised) using Russell's Circumplex (Russell, 1980) as the theoretical framework.

The keyword lexicon addresses a known limitation of GoEmotions: its training corpus is American English Reddit, which systematically underperforms on British slang, sarcasm, and culturally specific expressions. Colloquialisms such as 'bricking it' (acute anxiety) or 'buzzing' (elation) are not in the GoEmotions vocabulary but are captured by the lexicon. The ensemble approach provides broader coverage than either component alone.
Display-label separation, informed by the therapist survey finding that 73% of professionals agreed clinical mood labels should be replaced with functional alternatives, means the user sees 'Sad' while the system processes 'Introspective.' This prevents negative label reinforcement without obscuring the system's behaviour.

3.3 Three-Tier Rule Engine and Data Hierarchy

The rule engine translates classified mood, user-specified target mood, and intensity inputs into target audio feature parameters. It operates across three data tiers arranged in strict precedence.
Tier 1 (Clinical Safety) draws from a survey of N=15 professional music therapists. Responses were analysed for consensus and thematically coded to extract safety constraints, the outputs of which are reported in Section 4.1.
Tier 2 (User Preferences) draws from N=203 public respondents using majority-vote thresholding: parameters where a single response exceeded 60% became hard-coded defaults; parameters below this threshold were exposed as user-adjustable preferences. This operationalises the distinction between what most users want and what individual users may need.
Tier 3 (Dataset Tie-breaking) uses the MxMH dataset (Rasgaitis, 2022; N=736), analysed in Python using correlation coefficients and grouped mean mental health scores.

3.4 ISO-Principle Dynamic Blend Formula

The ISO-Principle is operationalised as a blend ratio between the audio feature parameters of the detected current mood and the user-specified target mood. For distressed moods (Anxious, Introspective, Angry, Restless), the formula is:
*iso_blend = 0.3 + (intensity_factor * 0.5) [range: 0.30-0.80]*
where *intensity_factor* is the user-reported current mood intensity normalised to *[0, 1]*. At maximum reported intensity (10/10), *iso_blend* reaches 0.80, meaning 80% of the audio feature parameters reflect the current mood while 20% pull toward the target. For non-distressed moods, the formula inverts: high intensity means the user wants the target state strongly, so the blend moves directly toward it. A second slider (target intensity) scales the target feature parameters toward or away from a neutral midpoint, allowing the user to specify how strongly they want to reach the target state. The lower bound of 0.30 ensures a minimum target mood influence at all intensity levels, preventing the playlist from remaining entirely anchored to the current state.
The conservative cap at 0.80 was chosen to prevent jarring acoustic transitions, which the professional survey identified as a distress risk comparable to abrupt silence. Both blend values are surfaced to the user in the

transparency rationale card alongside a plain-language explanation of their effect.

3.5 Crisis Detection

Crisis detection operates on two tiers. Tier A fires on any match from a curated crisis keyword lexicon, regardless of reported intensity; the language itself is a sufficient signal. Tier B fires when the classified mood is within the distressed category and reported intensity is 7 or above. The threshold of 7/10 for Tier B was an implementation choice, reflecting the point at which self-reported intensity consistently mapped to distressed mood classifications during developer testing. Both tiers apply negation and past-tense filtering to reduce false positives from statements such as 'I used to feel hopeless' or 'I am not feeling anxious today.'

When either tier fires, the system surfaces helpline resources (Samaritans, Mind, NHS, Crisis Text Line) and presents three options: continue to playlist generation, take a break, or access support. This choice-based model directly implements the professional survey finding. A recovery state pathway, separate from the active crisis pathway, acknowledges that users reporting recent improvement from distress have different needs from those currently in distress.

3.6 Additional Design Features

The neurodivergent accessibility filter caps energy at 0.6, constrains tempo to a 15 BPM range around the target, removes genres associated with high acoustic unpredictability, and prefers studio recordings over live material. The cap of 0.6 represents an implementation choice within the range identified as comfortable by professional respondents, and the 15 BPM range is informed by Starcke et al. (2021), who identify tempo stability as a key factor in mood-matched listening. Amongst said respondents, there was an 80% consensus rating these features as 'Beneficial' or 'Critical'. An additional instrumental filter applies instrumentalness and speechiness thresholds alongside name-based vocal detection for cases where metadata is unreliable.

A cross-user teach-the-system mechanism allows users to correct low-confidence classifications (below 60%), with corrections stored in a shared SQLite database and applied to future sessions. Per-session music profiling is fully opt-in: returning users are shown a summary of their previous session and choose explicitly whether to apply that data to future recommendations.

The transparency rationale card, generated for every playlist, states in plain language: the detected current mood and target mood; the ISO-Principle blend percentage and its basis; the target intensity scaling applied; which filters were active; and the specific survey data points that influenced the audio feature parameters. This directly implements the transparency requirement identified by Sun et al. (2024).

4. METHODOLOGY

4.1 Data Collection

Primary data collection proceeded in three stages. A semi-structured interview with a qualified music therapist (Arianna Berardi) established foundational clinical constraints and introduced the ISO-Principle as a design requirement. This was followed by a professional survey of N=15 music therapists, distributed through Participant A's professional network, using a mix of Likert-scale and open-ended questions covering mood labelling, audio feature priorities, neurodivergent accessibility, and crisis response preferences. Key outputs included: 73% agreed with replacing clinical mood labels with functional alternatives; 80% rated neurodivergent accessibility features as Beneficial or Critical; and opinion on crisis response was split across three positions (halt: 4, continue: 3, user choice: 8). The split itself was an important finding. The professional reasoning - that abrupt silence can increase distress for someone in acute crisis - directly contradicted the engineering default of a kill switch and shaped the crisis intervention design.

A public survey of N=203 participants used majority-vote thresholding to establish user preference defaults across the same parameters. Key public survey findings included: 66% preferred mood-opposing music when feeling overwhelmed, becoming a hard-coded default; 50% preferred slow, steady tempo when anxious, falling below the 60% threshold and exposed as an adjustable preference; 71% indicated lyrics matter to their mood experience, informing the optional instrumental filter; and 78% reported using music to shift their emotional state rather than maintain it, validating the ISO-Principle transition model. Parameters where no single option exceeded the threshold were exposed as user-adjustable controls, operationalising the distinction between what most users want and what individual users may need. Both surveys were administered via SurveyMonkey.

Secondary data analysis used the MxMH dataset (Rasgaitis, 2022; N=736), accessed publicly via Kaggle. Python-based correlation analysis and grouped mean calculations were applied to identify genre-mood relationships and resolve conflicts between the primary data sources. The most significant finding was that genre alone is a weak predictor of emotional outcome ($r < 0.2$ for most genre-mood pairs), which confirmed the decision to prioritise audio features over genre labels. Counter-intuitive findings were used as cautionary tie-breakers: Lo-Fi listeners, for example, reported the highest mean depression score (6.6/10) despite the genre's popular therapeutic reputation.

4.2 Ethical Approval

This research was conducted under the University of East London's short-form ethics approval pathway, which applies to research involving adult professionals surveyed about their practice without collection of identifying personal data. The signed approval form was co-signed by the project supervisor on 20 November 2025. No names were taken from participants. No participant under 18 was involved. No clinical data was collected from any source. Spotify authentication uses the OAuth 2.0 Authorization Code flow; no user passwords are handled or stored by the system. Full details are provided in the Ethics Statement

4.3 Developer Testing

A structured test register of 46 cases across eight categories was developed prior to user testing. Categories covered: NLP classification accuracy, rule engine parameter calculation, ISO-Principle blend output at defined intensity values, both tiers of crisis detection, neurodivergent filter behaviour, music profile application, and conflict resolution edge cases. Each test case specified the exact input conditions, the expected system output, the actual output, and a pass/fail determination. Two cases required manual verification rather than automated assertion, given the Spotify API constraint described in Section 6.3. The full test register is available via the link to the supplementary materials in Appendix B.

4.4 User Testing

User testing used a structured protocol of seven tasks, designed to exercise the full feature set including mood input, target mood selection, intensity adjustment, filter usage, track rating, and session report review. Testers were recruited from willing repeat participants and informally through a known network; each tester was asked to complete at least one session and submit a post-use questionnaire immediately afterwards. Spotify's development mode restriction of five registered users at a time, with a 24-hour interval before adding further users, meant testing was conducted across 18 days. A total of 32 completed questionnaire responses were collected across this period, each representing a distinct usage session rather than a distinct individual.

The questionnaire comprised nine Likert-scale items (five-point, Strongly Agree to Strongly Disagree), two multiple-choice discovery questions, and three open-ended free-text questions. The full test register is available via the link to the supplementary materials in Appendix B.

5. RESULTS

5.1 Developer Testing

44 of 46 test cases passed, a rate of 96%. The two failures both required direct numerical verification of audio feature parameters on retrieved tracks, which was not possible following the deprecation of the Spotify Audio Features API in November 2024. In both cases, the rule engine produced the correct target parameter output; the failure was the inability to confirm that returned tracks matched those parameters under current API restrictions.

5.2 User Testing: Likert Scores

As seen in Figure 1 (and further supported by Figure 3) Emotional safety scored 4.68/5 with 100% positive responses across 31 valid answers. Several participants referenced the crisis detection and safeguarding features in open-ended responses without being prompted to do so. The rationale card scored 4.12/5 with no negative responses; qualitative criticism was presentational rather than substantive. User control scored 4.22/5, with specific praise for the intensity sliders and music profile personalisation among users who completed multiple sessions; one participant noted the playlist was "better tailored on the second run-through." Genre consistency returned the lowest score at 3.72/5.

Table 2: User testing results - post-use questionnaire (N=32 sessions, 5-point Likert scale unless otherwise noted)

Dimension	Mean (/5)	Notes
Emotional safety	4.68	100% positive responses; highest score across all dimensions
User control	4.22	Intensity sliders and song source mode cited
Transparency / rationale card	4.12	No disagreement responses; sole criticism was presentational density
Current mood detection accuracy	3.97	84% positive; lower scores linked to colloquialisms and non-English input
Emotional progression	3.94	Moderate positive correlation with playlist match ($R^2 = 0.36$)
Target mood understanding	3.88	38% neutral; suggests target mood interface merits redesign
Overall playlist quality	3.94 stars	Two outlier low ratings linked to instrumental filter failures
Playlist genre consistency	3.72	Lowest score; directly attributable to Audio Features API deprecation

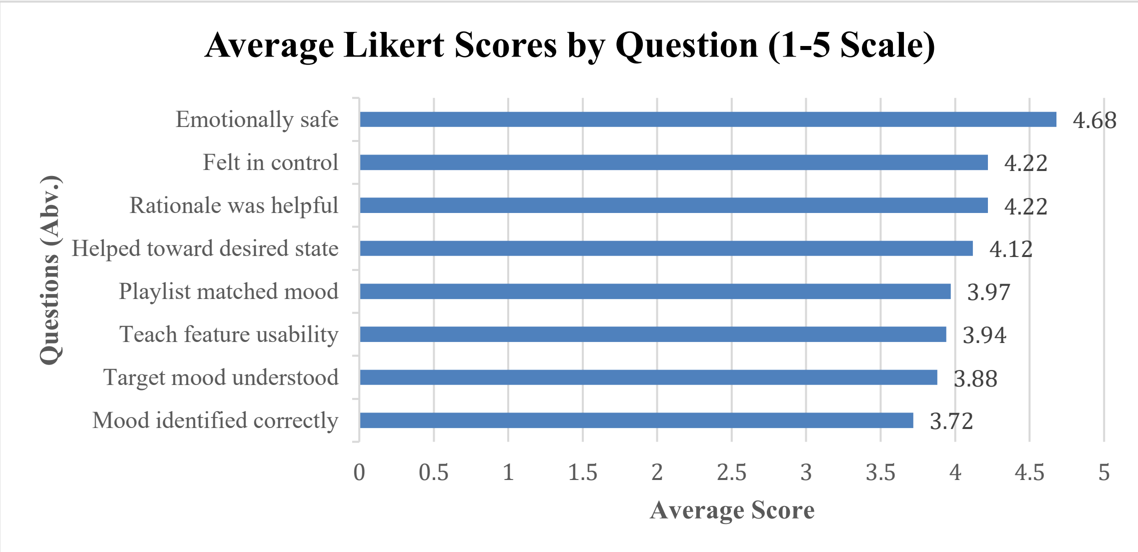

Figure 1: Mean Likert Scores (N=32 sessions).

5.3 ISO-Principle Correlation

In Figure 2, participants rated both playlist match and emotional progression on a five-point scale. Across 32 sessions, a moderate positive correlation was observed between the two dimensions (R^2=0.36, r=0.62, p<0.05). A jitter of ±0.2 was applied to separate overlapping Likert values for visualisation purposes; the raw distribution shows clustering at the upper end of both dimensions.

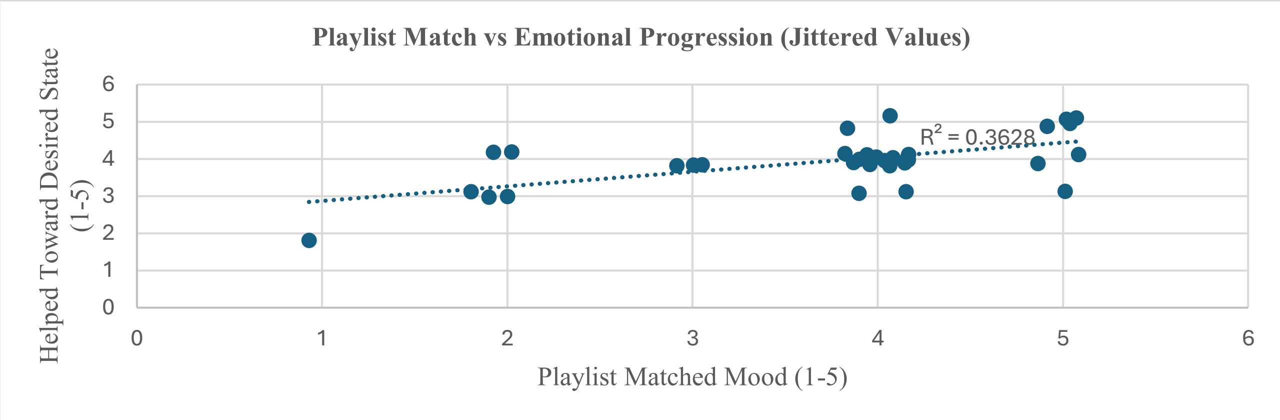

Figure 2: Playlist match vs emotional progression across 32 sessions (jitter of ±0.2 applied). R^2 = 0.3626, r = 0.602.

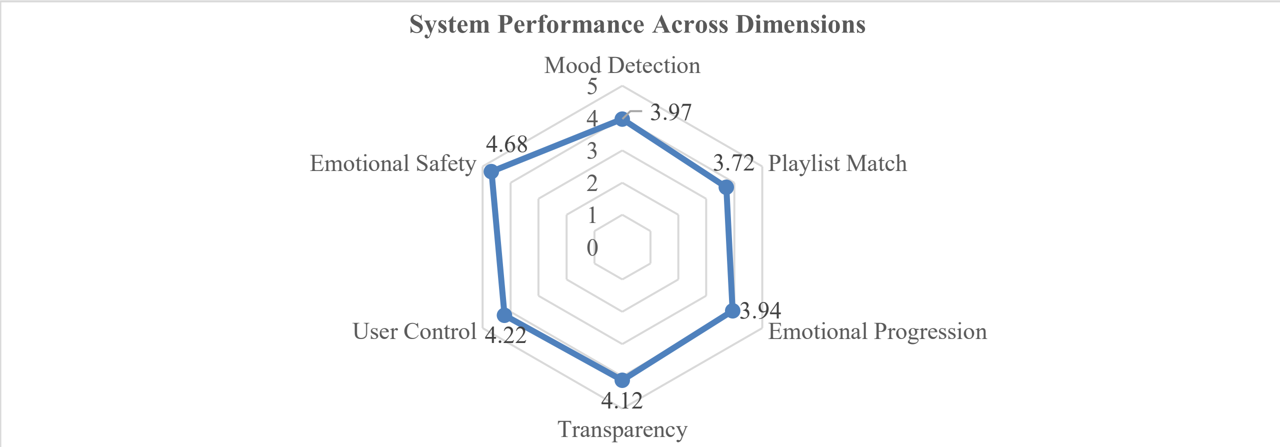

Figure 3: Mean performance scores across dimensions. This reflects the gap between emotional safety (strongest) and playlist genre consistency (weakest).

5.4 Overall Playlist Quality

Overall playlist quality ratings were collected as a five-star scale alongside the Likert questionnaire. The distribution is positively skewed: 69% of sessions were rated four or five stars, yielding a mean of 3.94 stars. The two outlier low ratings correspond to sessions where the instrumental filter underperformed, consistent with the developer testing failures documented in Section 4.3.

Table 3: Playlist Quality Ratings (Mean 3.94)

Rating	Frequency
1 ★	1
2 ★	1
3 ★	8
4 ★	11
5 ★	11

Regardless, of the 32 sessions, 24 resulted in participants discovering new songs, 7 used only pre-existing liked songs, and 1 reported no new songs found.

6. DISCUSSION

6.1 Retrieval, Generation, and Ethical Readiness

The question of whether to build a retrieval-based or generative system was, at its core, an ethical one. Generative AI in music is developing rapidly; Shen et al. (2024) document its growing application in mental health contexts whilst also noting the risks of emotionally misaligned outputs such as acoustically plausible content that a model cannot self-audit for therapeutic suitability. This is not a reason to dismiss generative approaches; it is a reason to ask what conditions would need to be in place before they could be deployed responsibly in emotionally sensitive contexts.

MoodFlow takes a position on that question by design. A retrieval-based system operates on music that already exists in the world: tracks a user can recognise, contextualise against personal memory, and verify against the system's stated reasoning. The transparency rationale card makes every curation decision legible, showing which mood was detected, what blend ratio was applied, and which clinical data points influenced the result. As a prerequisite for accountability, a system making recommendations to users in emotional distress should be able to explain itself and offer a tractable path.

The ISO-Principle arc itself, selecting music that matches the listener's current emotional state, then guiding them gradually toward a desired one through deliberate acoustic progression, demonstrates that meaningful creative and clinical intention can be encoded in AI-based curation logic.

The rapid advancement of generative systems gives reasonable grounds for optimism that they can meet the accountability and transparency demands that emotionally sensitive deployment requires. When that point arrives, frameworks such as the three-tier data hierarchy described in Section 6.2, where clinical safety constraints take strict precedence over user preferences and statistical defaults, represent the kind of structured, auditable foundation that responsible deployment will need.

6.2 The Three-Tier Hierarchy as a Design Argument

The precedence hierarchy in the rule engine encodes a specific position about whose priorities should dominate in an emotionally sensitive system. Clinical safety constraints override user preferences; user preferences override dataset statistics. This is a value judgment, not a technical default, and it produces observable consequences that a purely data-driven system would not produce.

The most striking example is the crisis response. The engineering default - a kill switch - was explicitly rejected because professional music therapists identified abrupt silence as a potential distress amplifier. A standard ML system trained on interaction data would not surface this finding because the data would not contain it; users in crisis rarely have their subsequent outcomes tracked against the intervention that preceded them. The professional survey, small as it was, provided exactly the kind of safety-first constraint that engagement-optimised systems structurally cannot accommodate.

The 100% safety rating is, in this context, a meaningful outcome for platform design more broadly. It is not a consequence of the system being cautious to the point of uselessness; the user control score (4.22/5) and transparency score (4.12/5) confirm that users found the system both usable and informative. Safety and utility were not in tension in this evaluation. Unprompted qualitative corroboration of a specific safety feature carries more evidential weight than a prompted Likert rating; these responses indicate that users noticed and valued the mechanism independently. The finding suggests that clinically informed design constraints do not necessarily degrade user experience; they can coexist with, and arguably reinforce, usability.

6.3 The Spotify API Constraint

The deprecation of Spotify's Audio Features API in November 2024 (Spotify, 2024) affected this prototype directly: the system cannot retrieve per-track energy, valence, or tempo to verify that returned tracks match the rule engine's calculated parameters. The Extended Quota Mode that would restore access is available only to registered organisations, not to individual or academic developers (Spotify, n.d.-a).

This represents a concrete instance of the infrastructural asymmetry that characterises platform-dependant research: the metadata most relevant for therapeutic curation sits behind access tiers calibrated for commercial scale, not clinical need. This is a governance question as much as a technical one, and it will not resolve itself through engineering alone.

Regardless, the practical implication for this paper's results is clear. This divergence between safety (4.68) and consistency (3.72) is not random; it reflects the Spotify API constraint precisely. The rule engine generates correct target parameters; without per-track audio feature access, retrieved tracks cannot be filtered to match them, and genre coherence suffers as a result.

Future work should explore open music datasets (such as Free Music Archive or OpenMusicLabs) that provide audio feature access without commercial gating, or engage with the legal and institutional pathways that might qualify a therapeutic application for extended API access.

6.4 The ISO-Principle Implementation Gap

The most significant gap between the current implementation and a fully realised therapeutic system is in playlist sequencing. The ISO-Principle, as defined by Starcke et al. (2021), implies a three-phase structure: matching (music mirrors the current emotional state), transition (acoustic properties shift gradually), and targeting (music reflects the desired state). The current system operates at the aggregate playlist level, blending feature parameters across the full playlist; it does not guarantee any particular ordering of tracks.

Spotify does not provide an API endpoint for crossfade or track-sequencing control, which means explicit track-by-track ISO-Principle construction is not possible within the current infrastructure. The $R^2 = 0.36$ correlation between playlist match and emotional progression suggests that aggregate-level blending produces some ISO-Principle effect; but this is not the same as demonstrating that a three-phase sequential structure was presented to the user. This moderate positive correlation holds across sessions where individual track quality ratings were mixed, which is consistent with the hypothesis that the transition logic - rather than any individual track - is driving the effect. This distinction matters for any future attempt at clinical validation: the correlation is encouraging but it does not isolate the sequencing mechanism as the causal factor.

6.5 Limitations

Several limitations constrain the conclusions that can be drawn from this evaluation. The user testing sample (N=32 sessions) is small; thus, the correlation data requires cautious interpretation and provides preliminary evidence rather than clinical validation. Participants were recruited through a known network and are unlikely to represent the full range of emotional states the system was designed to handle. One participant noted the playlist was 'better tailored on the second run-through,' which confirms the profile system's core intent while also highlighting the ancillary nature of repeated sessions. Notably, no participant was in acute crisis during testing, by design, for ethical reasons, meaning the crisis detection mechanism was not evaluated under the true conditions it was built for. Self-reported Likert ratings measure perceived emotional fit, not actual emotional change; whether the system produces genuine mood regulation outcomes, as opposed to a satisfying user experience, cannot be fully established from this data alone.

The system itself carries technical constraints that affects general applicability. The GoEmotions model was trained on American English Reddit comments and systematically underperforms on British slang and sarcasm; the keyword lexicon provides partial mitigation but cultural and linguistic coverage remains an open issue. The ISO-Principle implementation operates at aggregate playlist level rather than individual track sequencing, as discussed in Section 6.4. The Spotify Premium requirement and limited API access represent meaningful barriers to equitable use.

Critically, the emotional label mappings and audio feature parameter assignments derive from primary survey data and the GoEmotions taxonomy but have not been independently validated against psychometric instruments; whether these mappings reliably produce consistent emotional outcomes across users remains an open question, to be addressed in future work through pre- and post-session affect measurement using the PANAS.

7. CONCLUSION

This paper has described a prototype music recommendation system that operationalises music therapy principles within a consumer-facing application, and has argued that retrieval-based, rule-driven architectures offer a more tractable path to accountability and safety in emotionally sensitive domains at the current stage of AI development. The core contributions are: a three-tier hierarchical rule engine in which clinical safety constraints take precedence over user preferences and statistical defaults; a computational operationalisation of the ISO-Principle as a dynamic audio feature blend ratio; a two-tier crisis detection mechanism shaped directly by professional clinical input; and preliminary evaluation evidence that ISO-Principle sequencing supports self-reported emotional progression independently of individual track quality.

The system is a prototype. It makes no clinical claims and should not be used as a substitute for professional support. What it demonstrates is that the most consequential design questions in emotionally sensitive recommendation, whose priorities the system serves, how safety and preference conflicts are resolved, and how decisions are made legible, can be addressed within a rule-based retrieval architecture at prototype scale; and that doing so produces measurably different outcomes from engagement-optimised systems.

Three directions for future work are most pressing. Clinical validation using a pre- and post-session affect measure, such as the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988) with an appropriate control condition, is the most important next step. The current correlation data is suggestive, but it does not isolate the system's therapeutic contribution from confounding factors. Second, implementing true three-phase ISO-Principle sequencing requires either audio feature access at scale or migration to an open music dataset. Third, the access equity issue raised by Spotify's development-mode restrictions warrants engagement with clinical and legal pathways that could qualify a therapeutic application for appropriate data access.

ETHICS STATEMENT

Research was conducted under the University of East London's undergraduate short-form ethics approval pathway. Ethical approval was granted and co-signed by the project supervisor on 20 November 2025. All participants were willing adults. The professional music therapist survey (N=15) was distributed by a qualified therapist participant; no clinical patient data was collected at any stage. User testing participants were recruited through established channels and informally; no identifying data was collected. Mood input text, track ratings, and playback data are stored in a SQLite database on the project's hosting server and are not transmitted to third parties beyond the Spotify OAuth 2.0 authentication flow. Spotify login uses the Authorization Code flow; no user passwords are handled or stored by the system at any point. The music profiling system is explicitly opt-in: users are shown a summary of their previous session and must actively confirm before any profile data influences future recommendations. The system presents established crisis helpline resources and states explicitly that it is a wellbeing support tool, not a clinical service. AI-assisted tools were used to assist in the development of this system and in thematic preparation of this paper, in accordance with the University of East London's academic integrity policy.

ACKNOWLEDGEMENTS

This work was supported by Accenture UK&I, whose sponsorship of the degree apprenticeship programme made this research possible, and whose partnership with a music therapy charity provided access to the professional network from which survey participants were recruited. I am grateful to Dr Rahime Belen Saglam for her supervision and guidance throughout, and to Arianna Berardi, qualified music therapist, for her time, clinical expertise, and her efforts in distributing the professional survey. Thanks are also due to Lorenzo, Romeo, Rhea and Yana for their patience and support; and to Karen Panton-White, whose belief in this project, and whose willingness to distribute a survey to everyone she knew, meant more than she realises. This work is dedicated in part to the memory of Amy Yang, whose encouragement gave me the confidence to pursue it.

REFERENCES

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4040-4054). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.372

Han, D., Kong, Y., Han, J., & Wang, G. (2022). A survey of music emotion recognition. Frontiers of Computer Science, 16(6). https://doi.org/10.1007/s11704-021-0569-4

Lowe, S. (2023). *SamLowe/roberta-base-go_emotions* [Model]. Hugging Face. https://huggingface.co/SamLowe/roberta-base-go_emotions

Narayanan, V. S., & Tarafdar, A. (2025). Music therapy-driven mood-based music recommendation system integrating user emotion, song lyrics, and health reflections. In 2025 AI-Driven Smart Healthcare for Society 5.0. https://doi.org/10.1109/ieeeconf64992.2025.10963042

Panda, R., Redinho, H., Goncalves, C., Malheiro, R., & Paiva, R. P. (2021). How does the Spotify API compare to the music emotion recognition state-of-the-art? In Proceedings of the 18th Sound and Music Computing Conference. https://doi.org/10.5281/zenodo.5045100

Rasgaitis, C. (2022). Music and mental health survey results [Dataset]. Kaggle. https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Sadhu, B., Khasnobish, T., Dash, S. R., Barra, P., & Hoang, V. T. (2025). The inferential strength of music: Machine learning towards mental health estimation. In Proceedings of the 2025 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC) (pp. 1-6). IEEE. https://doi.org/10.1109/ASSIC64892.2025.11158696

Shen, L., Zhang, H., Zhu, C., Li, R., Qian, K., Meng, W., Tian, F., Hu, B., Schuller, B. W., & Yamamoto, Y. (2024). A first look at generative artificial intelligence-based music therapy for mental disorders. IEEE Transactions on Consumer Electronics, 71(3), 7439-7453. https://doi.org/10.1109/TCE.2024.3514633

Spotify. (2024). Introducing some changes to our web API. Spotify for Developers. https://developer.spotify.com/blog/2024-11-27-changes-to-the-web-api

Spotify. (n.d.-a). Quota modes. Spotify for Developers. https://developer.spotify.com/documentation/web-api/concepts/quota-modes

Starcke, K., Mayr, J., & von Georgi, R. (2021). Emotion modulation through music after sadness induction: The Iso Principle in a controlled experimental study. International Journal of Environmental Research and Public Health, 18(23), 12486. https://doi.org/10.3390/ijerph182312486

Sun, J., Yang, J., Zhou, G., Jin, Y., & Gong, J. (2024). Understanding human-AI collaboration in music therapy through co-design with therapists. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1-21). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642764

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54(6), 1063-1070. https://doi.org/10.1037/0022-3514.54.6.1063

APPENDIX A: DEVELOPER TEST REGISTER SUMMARY

Table A1: Developer test register results by category

Category	**Total Cases**	**Pass / Fail**	**Notes**
NLP Classification	8	8 / 0	Ensemble consensus logic and lexicon override cases
Rule Engine Parameters	7	7 / 0	All 13 audio features validated across mood pairs
ISO-Principle Blend	6	6 / 0	Blend formula verified at intensity 1, 5, 8, 10; both distressed and non-distressed paths
Crisis Detection	7	7 / 0	Tier A and Tier B; negation and past-tense filtering confirmed
Neurodivergent Filter	4	4 / 0	Energy cap, tempo range, genre exclusions all validated
Music Profile	5	5 / 0	Opt-in confirmed; liked artist boost and disliked track exclusion validated
Conflict Resolution	5	5 / 0	Tier precedence confirmed across five collision scenarios
Spotify Integration	4	2 / 2	2 failures: audio feature verification not possible without deprecated API endpoint
TOTAL	**46**	**44 / 2 (96%)**	

APPENDIX B: SUPPLEMENTARY MATERIALS

All supplementary materials referenced in this paper, including the professional and public survey questions, the semi-structured interview schedule, and the complete 46-case developer test register, can be found at this GitHub Folder: Supplementary Materials
These materials are intended to support methodological transparency and reproducibility.