Predicting Moral Values in Lyrics Through Audio

Charalampos Saitis*, Ben Heyderman*, Vjosa Preniqi*, Kyriaki Kalimeri† and Johan Pauwels*
*Centre for Digital Music, Queen Mary University of London, London, UK
†ISI Foundation, Turin, Italy
c.saitis@qmul.ac.uk

***Abstract*—This paper introduces the task of music morality recognition—predicting moral values in song lyrics using only audio features, which can be considered a form of music tagging with a set of new and well-defined tags grounded in social and cultural psychology research. Unlike previous research focused on lyrics analysis alone, this approach examines how musical elements correlate with moral content in associated lyrics. We used human-annotated lyrics and a set of experiments with XGBoost classifiers to establish a baseline recognition performance. Despite working with small and imbalanced data, we found audio features often outperformed a state of the art language model fine-tuned to detect moral content in lyrics, with some moral values being more reliably predicted than others in line with related work. SAGE and SHAP analyses revealed that specific timbral, harmonic, and melodic features play a prominent role in audio-lyrics moral associations. These findings advance our understanding of musical semantics and have potential applications in music and multimedia recommender systems and healthcare interventions. We provide a public repository containing all code and data used in this study.**

***Index Terms*—Audio features, moral foundations, music emotion recognition, music morality recognition, music tagging, timbre**

I. Introduction

The detection and classification of moral values in music is an emerging topic of research within the field of Music Information Retrieval (MIR), proposing a novel type of tag for automatic recognition in addition to the established genre, emotion, and instrument tags. Recent work has explored predicting moral values in song lyrics [1]–[3], on the premise that words are arguably more effective in conveying morality than non-verbal forms. However, the focus on lyrics alone limits the scope of understanding the full affective and ethical impact of music. Extending this analysis to audio signal data is a logical and necessary next step, opening up new possibilities for applications across several domains.

People often select music that aligns with their empathy levels and personality needs, and enables them to express their values [4], [5]. The possibility to extract morality rapidly from audio can thus improve how we experience and interact with music. Such tools can enhance music composition and creation processes by enabling artists and producers to align their music with specific moral themes. They can also lead to better, more personalised, and more diverse music recommender systems [6], [7], with applications in streaming and entertainment, but also in mental health support. Music listening interventions can be tailored to better align with a participant's values or ethical conflicts, potentially offering a more targeted approach to wellbeing through music [8], [9]. Additionally, as music often reflects cultural norms and societal values (and biases) [3], [10], audio morality detection tools can enable researchers in the social sciences and digital humanities to gain insights into the prevailing moral attitudes of different times or cultures.

In light of the above, this paper explores the novel task of automatic detection of moral values in song lyrics through audio signals, using an integrated approach. Our experiments centered on exploring the following research questions: To what extent can lyrics-based moral values be recognised from audio? Which features are most useful to align audio with lyrics-derived moral values? What insights does this bring?

We realise that this endeavour is impossible to fully achieve. Discrepancies between values in audio and lyrics are possible, and even deliberate as an artistic choice. For example, Rogers and Ogas [11, pp. 120–121] remark of the song "50 Ways to Say Goodbye" by the band Train: "*The lyrics tell us that the singer is heartbroken... But melodically, the record delivers a very different message. The fast tempo keeps the mood light and upbeat.* [And so does the official video.[1]] *The accompaniment features telenovela-style mariachi horns and acoustic guitar for tongue-in-cheek drama.*" Audio is the most common representation of music though, and first transcribing or retrieving its lyrics before applying text-based moral recognition is not practical nor always feasible. Furthermore, moral concepts may be expressed differently across different people and cultures. For instance, values of spirituality and self-discipline are not understood in the same way by religious and non-religious individuals [3], [12]. Nonetheless, it will be interesting to explore the extent of audio-lyrics moral associations and the most relevant audio features.

Challenges such as personal and cultural subjectivity are shared with the field of Music Emotion Recognition (MER), which is the most closely related topic. Work on what might be termed *music morality recognition* can take inspiration from MER, but one major difference is that moral values are strongly formalised in social and cultural psychological theory and uniquely characterized by well-developed term dictionaries. In contrast, categorical MER models differ in the number and vocabulary of their labels since it is unclear how many adjectives are sufficient to represent the wide scope of

VP is supported by a PhD studentship from Queen Mary University of London's Centre for Doctoral Training in Data-informed Audience-centric Media Engineering.

KK acknowledges support from the Lagrange Project of ISI Foundation funded by CRT Foundation.

Code: https://github.com/comma-lab/audio-mft

[1]https://youtu.be/GSBFehvLJDc?si=z4EbCgbNdCFZdg42

emotions. Continuous emotion models such as Russel's popular Valence-Arousal circumplex model suffer from the difficulty to numerically quantify emotions [13].

Specifically, we ground our approach on the Moral Foundations Theory (MFT) [14]. In its original conception, the MFT posits five foundational psychological systems that give rise to moral intuitions ("ethics") across cultures, characterized by polarised values (virtues/vices): *Care/Harm* is about kindness and empathy; *Fairness/Cheating* about equality and rights; *Loyalty/Betrayal* about ingroup solidarity and intergroup competition; *Authority/Subversion* about deference to superiors and respect for traditions; and *Purity/Degradation* about spirituality and "naturalness" (often present in religious narratives). A recent reformulation of the MFT [15] splits Fairness into *Equality* (focus on societal well-being) and *Proportionality* (focus on societal power). *Liberty* has been proposed as a potential additional foundation [16]. Cultures diverge in how they develop virtues (and vices), narratives, and institutions around these intuitive morals, which can be reflected in various forms of cultural expression, including music.

II. Related Work

The combination of music and lyrics creates what Davies [17] terms, "compositionally composite artworks," in which the effect created by the two mediums becomes a unified whole with greater meaning than its constituent parts. Alperson and Carrol [18] suggest that, because songs have propositional content, the role of the music is to clarify the meaning or more importantly the *significance* of the lyrical content, often providing emotional direction. MFT posits specific emotional responses to violation of moral values, such that harm (e.g., attacking a queer person) is linked to anger, and purity (e.g., committing adultery) to disgust [19]. Constructionist accounts also acknowledge a morality-emotion link, but suggest more complex, mixed emotions are produced by moral transgressions [20]. It is thus plausible to assume that music explicitly composed to match lyrics that reflect certain moral values would aim to clarify or amplify the intended emotional response to the moral message (cf. [21]).

Audio features, which comprise several abstraction levels from perceptual (e.g., melody, tempo) to more complex aggregated (e.g., timbre, valence), can predict musical emotions [22], virality [23], and gendering [24]. Some studies have used musical audio features to interpret genre-based preference dimensions [25], but direct links to psychological traits and values have only recently started to be explored [3], [5], [26]. For example, moral values of metal music fans, assessed via MFT, can explain a portion of the variance in their lyrics preferences [3]. Favouring lyrics about depression, hardships, love, and emotional turmoil was related to valuing virtues of care and harm. Whereas degradation was associated with liking songs about violence and Satanism. Czedik-Eysenberg et al. [27] found correlations between audio features that predict perceived *hardness* (rhythmic and spectral density, percussive energy) and *darkness* (spectral complexity, minor mode) and lyrical content, extracted with topic modelling, dealing with dystopia, occultism, Satanism, violence, love, and madness.

TABLE I
Audio Features Used to Predict Lyrics Moral Labels[a]

Category	Feature
Melody (17)	Pitch Height; Pitch Range; Direction; Step size (μ and σ of absolute inter-note pitch height difference); Melodic Intervals (normalised histogram of inter-note pitch height differences, modulo 12) [22], [28]
Harmony (64)	Chromagram (bespoke, 12 bins, μ and σ); Pitch Salience (μ and σ); Chord Histogram (24); Key (12); Scale; Key Strength [29]
Rhythm (9)	Tempo; Beats Loudness (2); 1st and 2nd peak from inter-onset interval histogram (value and weight); Danceability; Rhythm Density [29]
Dynamics (2)	Loudness (integrated); Dynamic Complexity [29]
Timbre (88)	MFCCs, Delta MFCCs (bespoke, 12 bins); Spectral Flatness; Spectral Skewness; Spectral Kurtosis; Spectral Spread; Spectral Flux; Spectral Centroid; Spectral Complexity; Spectral Contrast (6 peaks and valleys); Zero Crossing Rate (μ and σ for all) [29]

[a]Brackets indicate number of features.

Preniqi et al. [26] provided additional evidence that audio characteristics of songs can to some extent predict listeners' moral values, in some cases with higher accuracy than lyrical features. Audio features were found to be better than lyrical features in inferring values of empathy and equality, but less so for tradition and hierarchy. Specifically models built with either timbre or pitch features had good accuracy for predicting each of the five moral foundations. A limitation of that study is that audio features were used off the shelf from the Spotify API, hindering deeper interpretation. In their survey on audio features for MER, Pandas et al. [22] identify the need for bespoke musical features, tailored towards the task at hand rather than reusing features designed for other tasks.

III. Method

A. Human-Annotated Lyrics

We used a published dataset of 200 English language song lyrics tagged with 10 moral values (the polarities of the five original moral foundations are considered as separate labels) by two trained annotators [2]. To our knowledge, this is currently the only such dataset available. Titles were sampled from the WASABI Song Corpus [30] using a semi-random approach to match its distribution of music genres. The chosen songs represent a balanced mix of genres including Rock, Pop, Hip-Hop/Rap, R'n'B/Soul, and Country/Folk. Eighteen titles were from the 60s and 70s, 78 from the 80s and 90s, and 116 from the post-2000 era. Lyrics were extracted through the Genius.com API. For each moral value, annotators assigned presence or absence in the entire song lyrics (see also Sec. III-C).

B. Audio Features

For all songs except two, a corresponding YouTube video was found and full-length compressed audio tracks were downloaded. We extracted state of the art features that aim to capture relevant melodic, harmonic, rhythmic, timbral,

and dynamics content from musical audio signals, including many that have traditionally been associated with emotions in music [22]. In particular, we extracted a total of 180 features (Table I) using primarily the Essentia library [29]. In addition, librosa [31] was used to compute chromagrams [32] and Mel frequency cepstrum coefficients (MFCCs) [33], [34], and pitch contour characteristics were calculated with the MELODIA vamp plug-in [28].

C. MoralBERT

MoralBERT [35] is a collection of transformer-based language models (Bidirectional Encoder Representations from Transformers) fine-tuned on MFT human-annotated datasets sourced from social media to capture moral content in social discourse. It outperforms lexicon-based approaches, Word2Vec embeddings, and zero-shot classification with large language models such as GPT-4 (Generative Pre-trained Transformer).

Preniqi et al. [2] further fine-tuned MoralBERT on synthetic lyrics generated by GPT-4 (and called it MoralBERT SL) and evaluated the models on the 200-song dataset described in Sec. III-A, reporting higher performance than baselines. To our knowledge, this is the first language model explicitly fine-tuned to detect moral content in lyrics.

A potential limitation of MoralBERT is that it does not differentiate between structural elements of songs such as verses, bridges, and choruses. Instead a whole song is tagged with a single moral valence value, potentially missing within-song variations and nuances that might better reflect alignment of moral values between audio and lyrics. The focus also lies on English lyrics, and the model was built on English social media data, which limits their applicability to other languages.

D. Experiments and Evaluation

We trained a series of XGBoost binary classifiers to predict the presence or absence of a moral value in the human-annotated dataset with 10 labels. Previous work demonstrates that predicting one moral value at a time results in higher accuracy [2], [35], [36]. We see this paper as presenting proof of concept towards establishing a new MIR task of music morality recognition, going beyond lyrics only efforts. As such, and because of the limited data available, we consider XGBoost and standard audio features to be appropriate and sufficient, as opposed to, for example, neural networks, which might be prone to overfitting and are less interpretable.

The left panel in Fig. 1 illustrates the underrepresentation of all labels in the dataset. To gain some insight of how class imbalance influences predictions, we trained an additional five binary classifiers where labels that corresponded to polarities of the same moral foundation were merged. In MFT research this approach is generally as informative as considering virtues and vices as separate labels [15], [37]. As shown in the right panel of Fig. 1, Care/Harm thus became overrepresented while the other labels remain underrepresented albeit within-label the balance of absent or present is improved.

Given the small number of data points (198 songs) and large number of audio features (180), each model was trained on a reduced, optimal set of audio features via recursive feature elimination with 5-fold cross-validation (RFECV). Hyperparameters were tuned in an Optuna [38] study with 200 trials, including correcting for the minority class as well as preventing the model from overcorrecting for it. Model performance was assessed using a 5-fold cross-validation. Two evaluation metrics were used during both model training (with RFECV and Optuna) and testing: standard F1 score (hereafter F1 Binary) and weighted average F1 score (hereafter F1 Weighted). The latter accounts for both moral and non-moral (neutral) classes, binary scores only for the former.

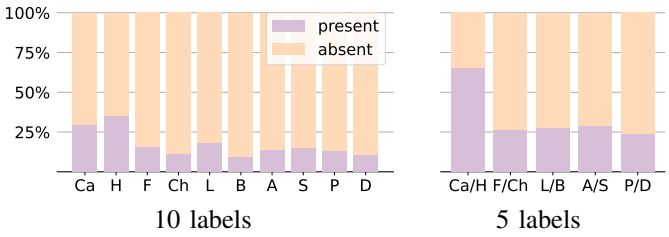

Fig. 1. Moral value class distribution (left panel) when considering moral foundation polarities as separate labels and (right panel) when grouping polarities of the same moral foundation.

Since this is a novel task, there is no established baseline. As a starting point, we considered constant classifiers which always predict the same label. We also compared the audio predictions with those of MoralBERT SL [2, Table 1] (hereafter just MoralBERT). For our 5-label experiments, we retrained MoralBERT accordingly. Our audio models, similar to MoralBERT, are binary classifiers that predict the presence of a moral value label (positive class) or its absence (neutral class). F1 Binary reports how well these models work for the former, not for the latter; as such, we compared F1 Binary scores between our models and constant classifiers which always predict the positive class. F1 Weighted accounts for both moral (positive) and non-moral (neutral) classes; as such, we compared F1 Weighted scores between our models and constant classifiers which always predict the majority class (effectively this is always the neutral class with one exception, see Fig. 1).

We used Shapley Additive Global importancE (SAGE) [39] and SHapley Additive exPlanations (SHAP) [40] to interpret our best models' general behaviour and the importance of different audio features, contributing to explainable machine learning [24], [41]. SAGE summarises a model's dependence on each feature across the whole dataset (global interpretability). SHAP assigns each feature a value that represents whether it pushes the prediction (of a single data point) higher or lower (local interpretability).

We make data, audio feature models, and associated code fully available via a GitHub repository.[2]

IV. Results and Discussion

Table II reports F1 scores averaged across folds. In the 10-label experiments, audio classifiers achieved F1 Binary scores

[2]https://github.com/comma-lab/audio-mft

TABLE II
F1 BINARY AND WEIGHTED AVERAGE SCORES[a]

Moral Value	F1 Binary			F1 Weighted		
	Constant	Audio Features[b]	MoralBERT[c]	Majority	Audio Features[b]	MoralBERT[c]
moral foundation polarities as separate labels						
Care	.46	.63 (.05) [22]	.75 (.04)	.58	.77 (.05) [22]	.83 (.03)
Harm	.52	**.69 (.10)** [09]	.69 (.04)	.51	**.79 (.05)** [09]	.70 (.03)
Fairness	.27	**.52 (.20)** [02]	.38 (.06)	.77	**.85 (.05)** [02]	.74 (.03)
Cheating	.20	.30 (.20) [01]	.32 (.06)	.84	**.85 (.05)** [37]	.69 (.03)
Loyalty	.31	**.56 (.15)** [07]	.27 (.09)	.74	.78 (.05) [53]	.79 (.04)
Betrayal	.18	**.43 (.30)** [05]	.37 (.08)	.86	**.90 (.05)** [05]	.84 (.02)
Authority	.25	**.43 (.10)** [01]	.39 (.09)	.79	**.86 (.05)** [13]	.84 (.03)
Subversion	.26	**.46 (.15)** [04]	.43 (.06)	.78	**.86 (.05)** [04]	.71 (.03)
Purity	.23	.31 (.15) [02]	.63 (.08)	.81	.84 (.05) [36]	.90 (.02)
Degradation	.19	**.60 (.35)** [21]	.32 (.10)	.84	**.94 (.05)** [21]	.86 (.03)
Average	.29	**.49 (.18)**	.46 (.07)	.75	**.84 (.05)**	.80 (.03)
grouping polarities of the same moral foundation						
Care/Harm	.79	**.79 (.01)** [66]	.78 (.03)	.51	**.66 (.05)** [73]	.56 (.04)
Fairness/Cheating	.42	**.50 (.15)** [11]	.48 (.05)	.62	**.77 (.05)** [11]	.62 (.03)
Loyalty/Betrayal	.44	**.56 (.10)** [04]	.50 (.06)	.61	.68 (.05) [53]	.75 (.03)
Authority/Subversion	.45	.54 (.10) [59]	.63 (.05)	.59	.75 (.10) [59]	.78 (.03)
Purity/Degradation	.38	.39 (.05) [06]	.50 (.06)	.66	**.72 (.01)** [06]	.72 (.03)
Average	.50	.56 (.08)	.58 (.05)	.60	**.72 (.05)**	.69 (.03)

[a]In bold are scores of audio models performing better than, or as good as, both the baseline and lyrics-only models.
[b]Mean and standard deviation (in brackets) over 5-fold cross-validation, and number of optimal features via RFECV (in square brackets).
[c]Standard deviation (in brackets) estimated via 1,000 bootstraps.

that are between 0.8% (Purity) and 41% (Degradation) higher than their constant counterparts. When looking at F1 Weighted, audio features still performed better than majority classifiers across the board, scoring up to 28% more (Harm). When comparing with MoralBERT, and insofar as we can compare predicting moral values of listeners [26] to predicting moral values in lyrics (the present study and [2])—two fundamentally different prediction tasks, we improved predictions of Fairness (by 14%), Loyalty (29%), Degradation (28%) and of three more labels (marginally) from audio versus from lyrics in terms of F1 Binary. When examining F1 Weighted, these improvements shrunk, but others increased from marginal to considerable (Cheating, Subversion). Care and Purity, on the other hand, appear to be better captured in lyrics than in the instrumental signal across both F1 scores.

The slightly improved present/absent balance in the 5-label experiments also resulted in comparable or higher performance than the constant and majority classifiers, although generally to a lesser extent than in the 10-label experiments. Better performance against MoralBERT was also demonstrated for some of the moral foundations but not for others. Overall, Care and Harm (and Care/Harm) are consistently the best predicted in terms of F1 Binary, with (Fairness/)Cheating and Purity(/Degradation) being the most challenging to infer. This result, especially when considering the 5-label experiments, corroborates previous findings [26] where audio features were found to be better predictors of empathy and equality than lyrical features, but less so for tradition, sanctity and hierarchy.

In both 10-label and 5-label experiments, the RFECV selected audio features are the same across F1 Binary and F1 Weighted models (e.g., Care, Fairness/Cheating), or a subset between the two evaluation metrics (e.g., Cheating, Care/Harm). In the latter case, F1 Weighted ends up with more predictors since it considers both the present/1 and absent/0 classes (i.e., which features push moral/1 *and* non-moral or neutral/0 predictions). Additional audio features are therefore picked up that help balance performance across both classes. When the optimal feature sets are the same for the two evaluation metrics, this indicates those audio features are stronger predictors for the corresponding moral values.

For some moral values, in F1 Binary experiments the number of optimal features is very low (< 3: Fairness, Cheating, Authority, Purity). This suggests that these values might be poorly represented in the investigated audio features—i.e., only a very small number of features carry a meaningful signal for these values while the rest contribute mostly noise, therefore the corresponding classifiers depend on a very limited set of predictors. This appears more likely for Fairness since, unlike Cheating, Authority and Purity, the respective F1 Weighted model did not pick up additional predictors.

As such, in the following we will focus our SAGE and SHAP analysis on the top nine most important audio features for Care and Harm when predicted separately, and when predicted as part of the same class. We use the best models under the F1 Binary score, and nine corresponds to the total number of optimal features for the Harm model. The SAGE plots in Fig. 2 (upper row) show that timbral, harmonic, and melodic features underlined predictions of Care and Harm, with Harm relying

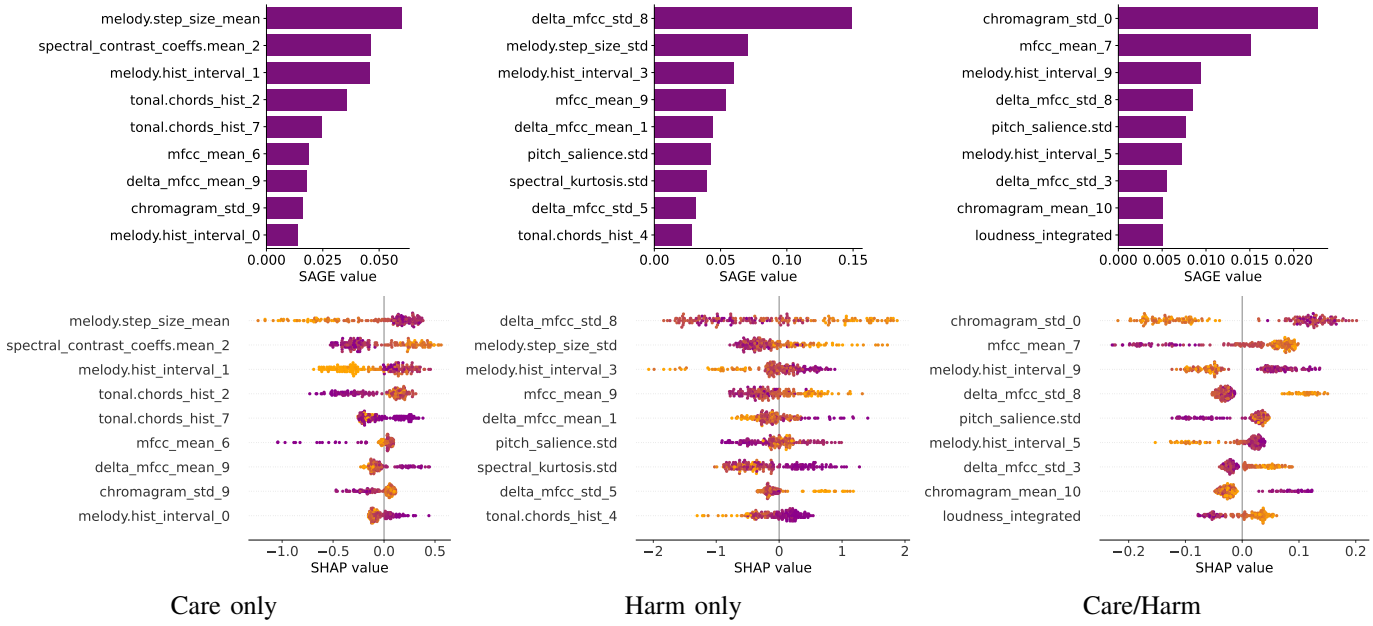

Fig. 2. SAGE and SHAP values of the top nine audio features for Care and Harm when predicted as separate labels and part of the same label, using the best models under F1 Binary (Table II). High feature values are depicted in orange and low values in dark magenta; hist = histogram, std = standard deviation.

more on timbre—as in the global timbre quality arising from instrumentation and arrangement [34]. Overall, dynamics and rhythm were not as relevant, except for Loudness (integrated) contributing positively to predictions of Care/Harm.

The presence of several low and higher-order MFCCs—four out of the nine RFECV selected predictors for Harm—suggests that timbre has a prominent role in audio-lyrics associations of kindness and empathy or the lack thereof. Low order MFCCs account for the slowly changing spectral envelope; higher-order ones describe the fast variations of the spectrum. We see that higher feature values of higher-order MFCCs have positive SHAP values (the points extending rightward are increasingly orange) for Harm and Care/Harm; lower values of these features have negative SHAP values (the points extending leftward are increasingly dark magenta). This indicates that, in the dataset we used, lyrics expressing Harm tend to be accompanied by music characterised by rapid spectral envelope variation.

Looking at the SHAP plot for Care, we see that lower values of "delta_mfcc_mean_9" contribute positively to predictions. We inspected SHAP values for the remaining 13 optimal features for Care (not reported here) and found similar patterns for two more higher-order MFCCs. This suggests that slow versus fast variations of the spectrum of a song can differentiate between Care versus Harm expressions in the song's lyrics, when predicting one moral value at a time. Higher Spectral Contrast values, indicating a clearer and more harmonically rich sound, also contribute positively to lyrical Care. Songs with lyrical Harm tend to have a flatter spectral distribution (lower Spectral Kurtosis values). However, when inferring the presence of either Care or Harm in the lyrics, the importance of these two timbral features diminishes.

Melodically, a feature which appears to discriminate between the two moral values is Step Size variation: songs with lyrical Harm tend to contain wider intervalic leaps than those with Care. Panda et al. [22] observe that some of the strongest relations between music and emotion are found between wider pitch ranges (e.g., wider intervalic leaps) and high arousal emotions such as fear. Music supporting lyrical Care is also less likely to contain intervals smaller than two semitones, which have been associated with lower arousal emotions of melancholy and sadness [22].

Considering harmonic features, in Care/Harm the SAGE and SHAP values for "chromagram_std_0" and "_mean_10" (as well as "_mean_0" and "_std_10" which are amongst the top 14 most important features for this model but not depicted here) suggest an association between lyrics expressing Care/Harm and music with prominent C pitch class content (bin 0) while avoiding strong B-flat (bin 10) presence. This points to Folk, Blues or Rock genres (pentatonic scales, guitar tunings, G-major/E-minor modes). Further, we see that less standard deviation of Pitch Saliency contributes negatively to Harm and Care/Harm predictions. This indicates that songs with lyrics expressing cruelty and neglect tend to involve less stable and clear melodic lines, where the focus instead shifts to timbral qualities as discussed above.

The connection between specific Chord Histogram bins and moral expression is more difficult to delineate, because the former need to be interpreted relative to the detected key. The SHAP plots indicate that the more present the dominant is, a chord a fifth (seven semitones) above the tonic (which in Essentia's key-offset chord histogram is represented by bin 2) that creates harmonic tension, the more likely it is for the lyrics to express Care values. Conversely, increased presence of a chord two fifths above the tonic (bin 7), which is some key

and scale contexts creates strong directional pull toward the dominant, tends to contribute negatively to Care predictions.

V. Conclusion

The present paper investigated the inference of moral values in music through audio features, rather than lyrics alone which has been the focus of recent efforts, with a view to formulating a new MIR task of music morality recognition. Part of what makes our approach interesting is that we try to predict morality present in lyrics from the music that was co-created with these lyrics but has not itself been used in the annotation process. From a more practical standpoint, as highlighted in our introduction, it is almost always easier to obtain a song's audio than to transcribe or retrieve its lyrics. Using an established, widely studied morality framework, the Moral Foundations Theory, makes the proposed task unique and better formulated compared to other types of music tagging where such a well-framed theory is currently lacking (e.g., emotion, genre).

Supporting our proof of concept approach, the reported experiments highlight how, despite the small size and high imbalance of the available human-annotated lyrics dataset, audio features offered on average the most accurate moral value inferences, often outperforming lyrics-based predictions. Overall, we found some moral values to be more reliable to infer than others, which echoes previous findings from predicting moral values in lyrics and of listeners. We intend to study the discrepancy in moral values between lyrics and audio, through further data creation and data-driven analysis and multimodal modelling methods. For example, the importance of each feature in the recursive feature elimination algorithm can be determined by its SHAP value [42].

SAGE and SHAP analyses revealed that timbral features play a prominent role in explaining predictions of Care and/or Harm in lyrics, and that music written for lyrics expressing Care tends to involve clearer melodic lines and be more harmonious. For certain genres (e.g., Hip-Hop/Rap) the human-annotated lyrics we used exhibit correlation with certain moral value labels (e.g., Degradation) [2, Fig. 2]. Future work could aim to examine genre mediated explanations for morality but also other tagging tasks (e.g., taking inspiration from [43]).

This research contributes to a deeper understanding of how audio features of music songs reflect and/or reinforce moral values expressed in the songs' lyrics, raising interesting new questions in the fields of music cognition, psychoacoustics, and information retrieval. The insights gained have a broad range of potential uses, including improving music and multimedia tagging and recommender systems, and music therapy. They also open up interesting avenues of cross-cultural research, given the cross-cultural premise of the MFT. The present work has a potential limitation in this respect, as we worked only with songs that have English lyrics and did not account for any cultural differences across songwriters and between lyricist and composer. Future work is encouraged to further examine moral expression through music in empirical research, ethnographic work, and intersectional studies.

References

[1] Vjosa Preniqi, Kyriaki Kalimeri, and Charalampos Saitis, ""More Than Words": Linking music preferences and moral values through lyrics," in *Proc. Int. Soc. Music Inf. Retriev. (ISMIR) Conf.*, 2022, pp. 797–805.

[2] Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Kyriaki Kalimeri, and Charalampos Saitis, "Automatic detection of moral values in music lyrics," in *Proc. Int. Soc. Music Inf. Retriev. Conf.*, 2024.

[3] Kyle J Messick and Blanca E Aranda, "The role of moral reasoning & personality in explaining lyrical preferences," *PLOS ONE*, vol. 15, no. 1, pp. e0228057, 2020.

[4] Antonis Gardikiotis and Alexandros Baltzis, "'rock music for myself and justice to the world!': Musical identity, values, and music preferences," *Psychol. Music*, vol. 40, no. 2, pp. 143–163, 2012.

[5] Ian Anderson, Santiago Gil, Clay Gibson, Scott Wolf, Will Shapiro, Oguz Semerci, and David M Greenberg, ""just the way you are": Linking music listening on spotify and personality," *Soc. Psychol. Personality Sci*, vol. 12, no. 4, pp. 561–572, 2021.

[6] Feng Lu and Nava Tintarev, "A diversity adjusting strategy with personality for music recommendation.," in *IntRS@ RecSys*, 2018, pp. 7–14.

[7] Sandy Manolios, Alan Hanjalic, and Cynthia CS Liem, "The influence of personal values on music taste: towards value-based music recommendations," in *Proc. ACM Conf. Rec. Syst.*, 2019, pp. 501–505.

[8] Dean McShane, "Using personalised music to enhance the well-being of people with dementia," *Mental Health Practice*, vol. 27, no. 1, 2024.

[9] Bleiz Macsen Del Sette, Dawn Carnes, and Charalampos Saitis, "Sound of care: Towards a co-operative ai digital pain companion to support people with chronic primary pain," in *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, New York, NY, USA, 2023, CSCW '23 Companion, p. 283–288, Association for Computing Machinery.

[10] Luca Marinelli, Petra Lucht, and Charalampos Saitis, "A multimodal understanding of the role of sound and music in gendered toy marketing," *PsyArXiv*, 2024.

[11] Susan Rogers and Ogi Ogas, *This is what it sounds like: What the music you love says about you*, Vintage, 2023.

[12] Don E Davis, Matthew T Dooley, Joshua N Hook, Elise Choe, and Stacey E McElroy, "The purity/sanctity subscale of the moral foundations questionnaire does not work similarly for religious versus non-religious individuals.," *Psychol. Relig. Spiritual.*, vol. 9, no. 1, pp. 124–130, 2017.

[13] Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang, "A survey of music emotion recognition," *Front. Comput. Sci.*, vol. 16, no. 6, pp. 166335, 2022.

[14] Jonathan Haidt and Jesse Graham, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," *Soc. Justice Res.*, vol. 20, no. 1, pp. 98–116, 2007.

[15] Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani, "Morality beyond the weird: How the nomological network of morality varies across cultures.," *J. Pers. Soc. Psychol.*, pp. 1157–1188, 2023.

[16] Ravi Iyer, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt, "Understanding libertarian morality: The psychological dispositions of self-identified libertarians," *PLOS ONE*, vol. 7, no. 8, pp. e42366, 2012.

[17] David Davies, "The dialogue between words and music in the composition and comprehension of song," *J. Aesthet. Art Critic.*, vol. 71, no. 1, pp. 13–22, 2013.

[18] Philip Alperson and Noël Carroll, "Music, mind, and morality: arousing the body politic," *J. Aesthet. Educ.*, vol. 42, no. 1, pp. 1–15, 2008.

[19] Helen Landmann and Ursula Hess, "Testing moral foundation theory: Are specific moral emotions elicited by specific moral transgressions?," *Journal of Moral Education*, vol. 47, no. 1, pp. 34–47, 2018.

[20] C Daryl Cameron, Kristen A Lindquist, and Kurt Gray, "A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions," *Pers. Soc. Psychol. Rev.*, vol. 19, no. 4, pp. 371–394, 2015.

[21] Halbert H Britan, "Music and morality," *Int. J. Ethics*, vol. 15, no. 1, pp. 48–63, 1904.

[22] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva, "Audio features for music emotion recognition: a survey," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 68–88, 2023.

[23] Gabriel P Oliveira, Ana Paula Couto da Silva, and Mirella M Moro, "What makes a viral song? unraveling music virality factors," in *Proceedings of the 16th ACM Web Science Conference*, 2024, pp. 181–190.

[24] Luca Marinelli and Charis Saitis, "Explainable modeling of gender-targeting practices in toy advertising sound and music," in *Int. Conf. Acoust. Speech Signal Process.* 2024, IEEE.

[25] Peter J Rentfrow and Samuel D Gosling, "The do re mi's of everyday life: the structure and personality correlates of music preferences.," *J. Pers. Soc. Psychol.*, vol. 84, no. 6, pp. 1236, 2003.

[26] Vjosa Preniqi, Kyriaki Kalimeri, and Charalampos Saitis, "Soundscapes of morality: Linking music preferences and moral values through lyrics and audio," *PLOS ONE*, 2023.

[27] Isabella Czedik-Eysenberg, Oliver Wieczorek, and Christoph Reuter, "'Warriors of the Word'–Deciphering Lyrical Topics in Music and Their Connection to Audio Feature Dimensions Based on a Corpus of Over 100,000 Metal Songs," *arXiv:1911.04952*, 2019.

[28] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 6, pp. 1759–1770, 2012.

[29] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Herrera Boyer, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra, "Essentia: An audio analysis library for music information retrieval," in *Proc. Int. Soc. Music Inf. Retriev. Conf.*, 2013, pp. 493–498.

[30] Michel Buffa, Elena Cabrio, Michael Fell, Fabien Gandon, Alain Giboin, Romain Hennequin, Franck Michel, Johan Pauwels, Guillaume Pellerin, Maroua Tikat, and Marco Winckler, "The WASABI Dataset: Cultural, Lyrics and Audio Analysis Metadata About 2 Million Popular Commercially Released Songs," in *The Semantic Web. ESWC 2021. Lecture Notes in Computer Science, vol 12731.*, May 2021, pp. 515–531.

[31] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python.," *SciPy*, vol. 2015, pp. 18–24, 2015.

[32] Mark A Bartsch and Gregory H Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[33] Marcelo Caetano, Charalampos Saitis, and Kai Siedenburg, "Audio content descriptors of timbre," in *Timbre: Acoustics, Perception, and Cognition*, Kai Siedenburg, Charalampos Saitis, Stephen McAdams, Arthur N. Popper, and Richard R. Fay, Eds., pp. 297–333. Springer, 2019.

[34] J-J Aucouturier, François Pachet, and Mark Sandler, ""The way it sounds": Timbre models for analysis and retrieval of music signals," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1028–1035, 2005.

[35] Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri, "MoralBERT: A Fine-Tuned Language Model for Capturing Moral Values in Social Discussions," in *ACM Int. Conf. Inf. Tech. Social Good*, 2024.

[36] Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazazian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al., "The moral foundations Reddit corpus," *arXiv preprint arXiv:2208.05545*, 2022.

[37] Kyriaki Kalimeri, Mariano G Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto, "Predicting demographics, moral foundations, and human values from digital behaviours," *Comput. Hum. Behav.*, vol. 92, pp. 428–445, 2019.

[38] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM Int. Conf. Knowl. Discov. Data Min.*, 2019, pp. 2623–2631.

[39] Ian Covert, Scott M Lundberg, and Su-In Lee, "Understanding global feature contributions with additive importance measures," in *Adv. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds., 2020, vol. 33, pp. 17212–17223.

[40] Scott M. Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Proc. Syst.*, Red Hook, NY, USA, 2017, NIPS'17, p. 4768–4777.

[41] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020.

[42] Jing Huang, Yang Peng, and Lin Hu, "A multilayer stacking method base on rfe-shap feature selection strategy for recognition of driver's mental load and emotional state," *Expert Syst. Appl.*, vol. 238, pp. 121729, 2024.

[43] Shreyan Chowdhury, Andreu Vall, Verena Haunschmid, and Gerhard Widmer, "Towards explainable music emotion recognition: The route via mid-level features," in *Proc. Int. Soc. Music Inf. Retriev. Conf.*, 2019.