Predicting Lyric Morality Through Handcrafted Audio Features

1st Benjamin Heyderman
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
benheyderman@live.co.uk

***Abstract*—This study investigated trends in how moral content in song lyrics is accompanied in the musical and sonic domains. Two datasets were analyzed: a collection of 3,100 pop songs from the Billboard hot 100 list for each year 1960-2023 and a sample of 34,000 songs from the WASABI dataset. Ten moral values were extracted from the lyrics using a model developed by Preniqi (2024), based on the five pillars of morality outlined in Moral Foundation Theory - care/harm, fairness/cheating, loyalty/betrayal, purity/degradation and authority/subversion. For each song, 181 handcrafted audio features were extracted, describing the timbre, rhythm, dynamics, harmony and melody of the tracks, and used to train 10 individual XGBoost classification models to predict the presence of a given moral value. Model performance varied across different moral values. Models predicting the presence of care and harm demonstrated reasonable accuracy in both datasets with maximum F1 scores of 0.68 and 0.62 respectively. Additional models with F1 scores of 0.4 or above were found for cheating, subversion, and degradation when using the WASABI dataset for training. The best performance was achieved by a model trained to distinguish between music that featured any of the positive moral values (care, fairness, etc.) versus the negative or immoral values (harm, cheating, etc.). SAGE and SHAP analyses were conducted to evaluate the contribution of individual audio features on the model's predictions.**

***Index Terms*—XAI, Morality, Emotion, MIR, MVR**

I. Introduction

Musical composition is a complex domain, drawing on cultural knowledge and understanding to convey mood and meaning. A composer has control of a number of essential musical building blocks — timbre, harmony, melody, dynamics, texture, form and rhythm — as well as the ability to augment these features with language in the form of lyrics. The aim of this paper is to explore the relationship between these building blocks and lyrical content.

This work builds on existing research by Preniqi (2024) into extraction of morality values from lyrics. These values are based on Haidt and Graham's (Haidt and Joseph, 2004; Haidt and Graham, 2007; Graham et al., 2011; Graham et al., 2013) Moral Foundations Theory (MFT), identifying 5 polar moral qualities: *Care and Harm, Fairness and Cheating, Loyalty and Betrayal, Authority and Subversion, Purity and Degradation*. Through the deployment of eXplainable AI (XAI) models to find links between audio and lyric content, this research hopes to extend this understanding of morality in music, asking — *how is morality expressed in the sonic domain?*

As well as expanding the understanding of morality in music for composers and musicologists, an aim of this research is to create a framework that can be used by music AI systems in a number of contexts. For music generation systems, this framework could guide the creation of compositions in which musical content and moral messages work in unison. The findings could also be applied to music recommendation systems, enabling song grouping and recommendations based on moral content and listener values.

The study uses two datasets: a dataset of the 100 top pop songs from each year since 1960 (3100 tracks available of 6400 total) and a random selection of 34000 tracks from the WASABI dataset (Buffa et al., 2021). The method begins by training XGBoost models to predict moral content of a track based on audio features, before using SHAP and SAGE analysis to assess the most important features used by each model to discern the moral content of the lyrics.

II. Related Work

A. Music, Meaning and Morality

The combination of music and lyrics creates what Davies (2013) 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 (2008) suggest that the role of the music is to "clarify the meaning" or more importantly the "significance" of the lyrical content, often providing emotional direction. On morality specifically, Britan (1904) states that music itself is not capable of conveying moral meaning, and is used to amplify the intended emotional response to the moral message.

Moral Value Theory posits standard emotional responses to violation of moral virtues. Violating care elicits compassion, fairness elicits anger, authority resentment, loyalty rage and purity disgust (Haidt and Joseph, 2004; Graham et al., 2011). The production of these emotions in response to moral violations was tested by a number of studies which supported the hypothesis (Yaşar and Akgün, 2023; Landman and Hess, 2017). However, this simple model is disputed by Cameron et al. (2015) who suggest a more complex, mixed emotion is produced by violating moral standards. Within the framework laid out in the previous paragraph, one could assume that the music used to match lyrics which display these moral values would aim to clarify or amplify these emotions.

Preniqi et al. (2023) explored the links between listener psychometric scores and lyric and audio features of their favourite songs. Their results show that there was a link with both lyric and audio features. This hints at some form of meaningful relationship between audio features and music morality.

This section has begun to lay out the complicated interplay between music and moral content, and raise the dual purpose of music: to convey meaning and emotion.

B. Predicting Musical Meaning with Audio Content

This study falls within the wider field of automatic morality detection but, as this research area is normally text based, the specific research area of finding links between lyric morality values and audio features is, to the author's knowledge, novel to this project. It will therefore be refered to as *Moral Value Recognition* (MVR).

Having shown the links between moral meaning and emotion in the previous section, MVR could be seen as proximate to the field of Music Emotion Recognition (MER). MER is the process of "using computers to extract and analyse music features, form the mapping relations between music features and emotion space, and recognize the emotion that music expresses" (Han and Jiayi, 2022).

MER is a challenging field of research, considered one of the topics within Music Information that resists a solution, something often attributed to the subjectivity of emotion labels and differing conventions between genres (Chowdhury et al., 2021). It is typically approached through machine learning techniques in which emotion labels are assigned to a collection of songs, and an algorithm is used to create a generalised model that links audio features to these labels. It is commonly either treated as a classification problem (Huang et al., 2016), in which specific emotion labels are assigned to a piece of music, or a regression problem (Delbouys et al., 2018) where the emotional content of a song is plotted on a continuous axis/axes such as Russel's circumplex model of mood (Posner et al., 2005), in which moods are plotted in two dimensions: valence and arousal.

There are generally two approaches to generating audio features: handcrafted features and learnt features. Handcrafted features are psychologically, acoustically or musically motivated features designed by a human or system with expert knowledge to describe specific qualities of the music. Learned features are automatically extracted from the waveform or time-frequency representation, often using deep machine learning algorithms.

The nature of ground truth labels is an area in which MVR and MER diverge. MER labels are assigned based on perception of the track as a whole, including music and lyrics, whereas the moral value labels used in this study are extracted from the lyrics only. This highlights the assumption that morality requires language to be expressed but emotion is based on the whole.

Despite their differences, at the core of both MVR and MER is the challenge of mapping quantifiable audio features to human values, and so it's hoped that methodologies and approaches in MER can offer valuable guidance for developing and refining MVR techniques.

In order to understand how features contribute to predicting lyric morality, model explainability is key to this project. Currently learnt feature approaches dominate the state of the art solutions in music information retrieval, however, they are generally more difficult for humans to interpret. Whilst work by Chowdhury et al. (2021) has shown that systems can be designed that employ learnt features in a explainable way, this study will focus on handcrafted features.

In their comprehensive survey on audio features for MER, Panda et al. (2023) identify the need for bespoke musical features, tailored towards MER rather than reusing features designed for other tasks. As a first step towards this they give an overview of the literature on music and emotion from music psychology. They then identify features that can target musical elements that have been proven to produce an emotive response. For example, a number of studies have shown that high pitch melodies are associated with surprise, anger, fear and happiness, whilst low pitch melodies are associated with sadness, boredom and pleasantness (Gabrielsson and Lindström, 2011; Juslin and Laukka, 2004). An audio feature that could target this musical element is average melody pitch height.

III. Methodology

A. Dataset

Two datasets were used in this study. The first, collated by Vjosa Preniqi, is a collection of the top 100 songs in the Billboard chart for each year between 1960 and 2023. From this set, a subset of songs with available lyrics and audio samples were chosen, leaving 3100 songs out of the original 6400. To assess whether this subset introduced bias, the average chart position, genre, and average year were calculated and revealed an insignificant difference in distribution.

The audio files used in the study are scraped from Spotify, who provide 30 second previews[1] of their tracks where allowed by copyright laws. Whilst the preview may not give an impression of the full song, Spotify states that "this segment is carefully selected to include the chorus or the most characteristic part of the music, providing a meaningful sample of the track" (Spotify, 2024).

The MFT values used as ground truths in the study were automatically extracted using Preniqi's (2024) MoralBert model. This gives 10 independent values between 0 and 1 — one for each pole of the 5 moral values.

Initial inspection of the moral distribution (Fig. 6, APPENDIX A) reveals a skewness across the data towards lower

[1]These previews are likely to include sung parts of the songs and so vocals are included in the audio being analysed. As the vocals contribute to the overall timbral, melodic, harmonic, dynamic and rhythmic picture of the track, it is hoped this will provide a better overall understanding of the song. However, it could also be interesting to repeat the study with the vocals removed from the audio to assess the relationship between the purely musical elements and lyrics.

Genre Key Word	Billboard %	WASABI %
Rock	34	36
Pop	55	17
Metal	2	12
Country	9	5
Punk	2	5
Folk	5	5
Hip hop	5	5
Blues	6	2
R&B	6	2
Jazz	4	1

TABLE I
GENRE DISTRIBUTION OF THE DATASETS.

values for each of the moral values, particularly in the case of loyalty, degradation and purity. This imbalance will have to be addressed in the model design and may undermine the performance of a number of the models.

Following the discovery of this imbalance in the data it was decided that a second larger dataset should also be included to see if this could improve performance. The second dataset is a randomly selected subset of the WASABI dataset, a collection of 2 million commercially released songs (Buffa et al., 2021).

The dataset was filtered for lyrics in the English language and a sample of 50,000 songs was selected. 50,000 is an arbitrary number, selected based on estimated storage space and feature extraction time. A maximum of 5 songs per artist were allowed to minimise the so-called 'album effect' (Vatolkin et al., 2015). Of these 50,000 tracks, 34,000 song previews were available and so this was the final number used.

The distribution of the morality data for the second dataset can be seen in Fig. 7 (APPENDIX A). Whilst this dataset still suffers from skew, it is generally better and the larger amount of data overall means that there are more examples of high values of each moral value.

Table 1 shows the genre distribution of the two datasets[2]. Genre data was searched for keywords to account for sub-genres — e.g. "Hard Rock" falls under the rock category. This shows the fairly significant difference in genre make up between the two datasets.

B. Audio Features

Panda et al. (2023) identify 8 categories of audio feature: melody, harmony, rhythm, dynamics, expressivity, timbre, musical texture and musical form. These will be used to organise the audio features in this study.

Of these 8 categories, form was removed due to the 30 second length of the audio samples, and expressivity and texture were removed due to computational restraints. APPENDIX B gives details of the features. The final feature sets contains 181 feature values representing each song.

C. Model

Following Han and Jiayi's (2022) review of the models most commonly used in MER, initial experimentation was undertaken comparing performance of XGBoost and Support Vector Machine models on the task. The XGBoost models displayed better performance and so were selected for the study.

Initial experimentation also compared treating the data both as a regression and classification problem - distinguishing between values above and below thresholds. Both model designs introduced a weighting to higher values, proportionate to the skew in the data, so that they had a greater effect on training. In general, the regression models were unable to perform above chance so the task was treated as a classification problem. Negative class values were classified as being less than 0.2, and positive more than 0.6.

For each of the ten moral values, an independent model was trained. In addition to this, a further model was trained on the WASABI dataset using a binary value generated for moral vs. immoral lyrics. Lyrics were deemed moral if they had a score of greater than 0.6 in any of the moral values — care, fairness, loyalty, authority and purity — and a score no greater than 0.2 in any of the negative valence moral values - harm, cheating, betrayal, subversion, degradation. Lyrics were deemed immoral if the opposite was true.[3]

Hyper parameters were optimised via a grid search, cross validated 3 times.

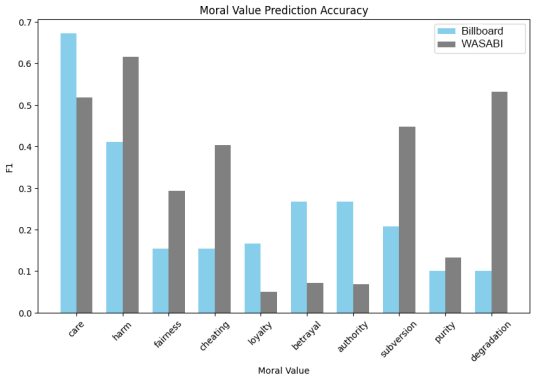

Fig. 1. F1 scores for each model. Colour indicates dataset.

IV. RESULTS

A. Model Performance

The performance of the models are illustrated in Fig. 1. This figure highlights the substantial variability in performance across different models and datasets. Whilst the best performing model overall is the care model trained on the Billboard dataset with an F1 score of 0.68, the mean F1 score of the WASABI trained models is 0.31 compared to 0.23 when trained on the Billboard data.

Fig. 2 gives the results of the cross-dataset performance testing. The performance was on average much lower than when using a test set from the same dataset. The two cases

[2]Note: only 50% of the WASABI dataset had genre information so this could be biased.

[3]Bi-polar models were trained for the individual moral values (e.g. Care vs Harm) but these were outperformed or similar to the independent models so these have not been included in this paper.

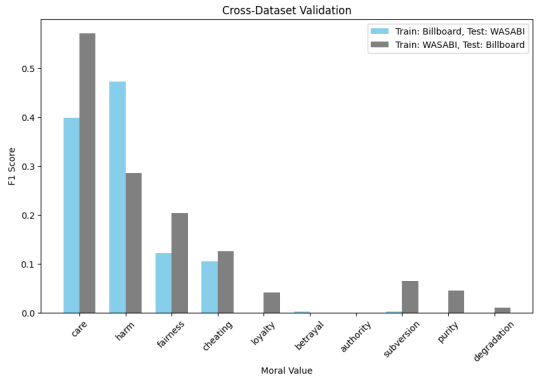

Fig. 2. f1 scores for each model when tested using the other dataset.

where performance was increased was the care model trained on the WASABI dataset and the harm model trained on the billboard dataset.

The bi-polar morality model greatly outperformed all of these with an overall F1 score of 0.81.

B. Global Feature Importance

SAGE is a method for analysing global feature importance whilst retaining complex feature relationships (Covert et al., 2020). Fig. 3 shows the top ten SAGE values for the care and harm models trained on the WASABI dataset. SAGE plots for the other models can be found in APPENDIX F. Models with an F1 score of less than 0.4 were not included as poorly performing models may provide false conclusions.

These graphs show a number of key trends. Firstly across the models we see the general reliance on timbral features to predict moral values. Spectral contrast is a particularly common feature. Spectral contrast is the difference between peaks and valleys in a given frequency band. The second bin, featured in all but one of the model's SAGE values, refers to the 800-1600 Hz frequency range. Later improved on by Akkermans and Herrera (2009), Jiang et al. (2002) who first described this statistic state that it "roughly reflect[s] the distribution of harmonic and non-harmonic components".

Other common features in the top SAGE values include integrated loudness, scale (major/minor), mean absolute inter-note melody pitch height difference, and the melody interval histogram bin 1. The melody interval histogram is an octave normalised histogram of pitch height intervals in the melody. Bin 1 represents a minor second.

SHAP analysis of features at the local level will help to uncover the nature of these relationships.

C. Feature Elimination

Following the primacy of the timbral features in the SAGE values and the absence of a number of feature categories, experimentation with subsets of features was undertaken. The first method for feature selection was categorical elimination. All 31 combinations of the 5 feature categories were used to train models. The results of these tests can be found in APPENDICES C and D. The results show that performance generally remains fairly high when using a reduced feature set and can even improve the model performance. Also in some cases the best performing model does not include the timbre category despite the prevalence of timbre features in the model trained on the full feature set.

Having shown that using a reduced feature set can increase accuracy, cross validated recursive feature elimination was applied to assess whether a better combination of features could be found by iteratively removing the least relevant features. Whilst this was able to improve the performance of some of the models by a small amount, this was not as effective overall when compared to the categorical elimination.

D. Local Feature Importance

SHAP analysis assesses the contribution of a feature to the decision made on an individual data point (Lundberg and Lee, 2017).

Given to the dominance of timbral features in the top SAGE values and the strong performance of models that exclude this feature category, SHAP values have been calculated for each model both with and without timbral features. This approach will provide better insights into feature importance within other feature categories.

Fig. 4 is beeswarm plots of the SHAP values for the care and harm models trained on the WASABI dataset. These show the top features used by the models and how they are used to affect the model's prediction. SHAP beeswarm plots for the other models can be found in APPENDIX G.

The results show that negative morality models use a common (to most if not all) set of features in the same way to make predictions. Color coding on the plots make these features clearer. These features are also used by the bi-polar morality model. These include low mean spectral contrast coefficient in the second bin; minor scale; high frequency of one-step intervals in the melody; high frequency of six-step intervals in the melody; high integrated loudness; large inter-note intervals steps overall; high dynamic complexity and overall higher pitch melodies. The importance of these features is different for different models. The remaining features used by the models seem to be more specific to each model individually.

The care model, the only positive morality model, appears to mirror some of these features. Preference for high mean spectral contrast coefficient in the second bin; low frequency of one-step intervals in the melody; low integrated loudness; and smaller inter-note intervals steps overall.

V. Discussion

A. Global Insights

Overall the performance of the models are variable. Despite this there are a number of models that show reasonable comprehension of the data form both datasets. This variability raises the difficult to answer question of whether this is something inherent in the data - that there is no unifying sonic imprint of purity in popular music for example - or a problem with model design.

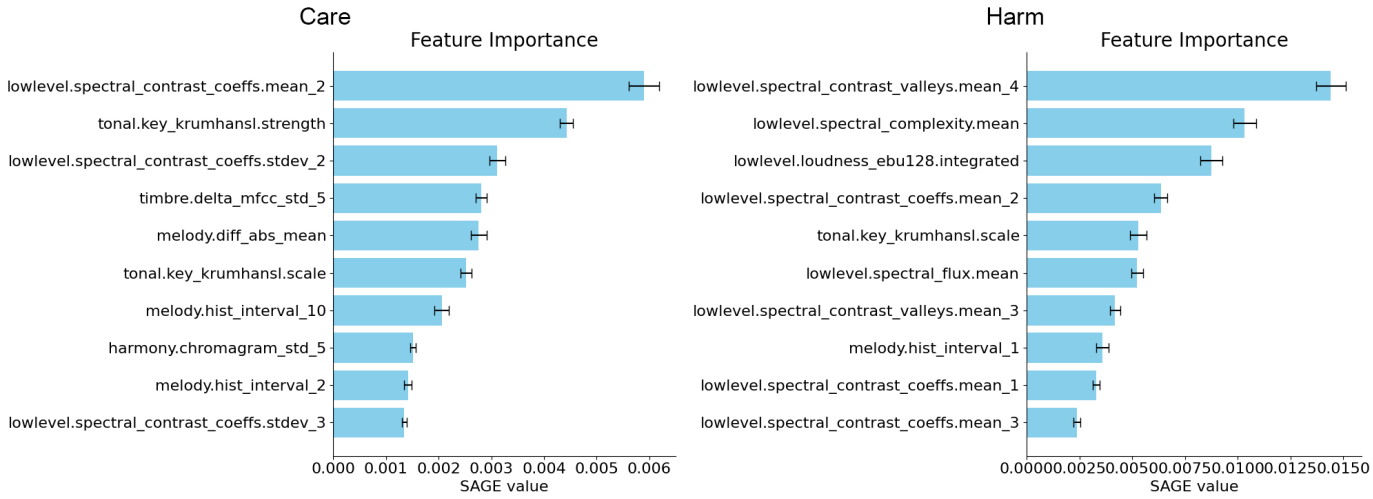

Fig. 3. Top 10 SAGE values for the care and harm models of the WASABI Dataset.

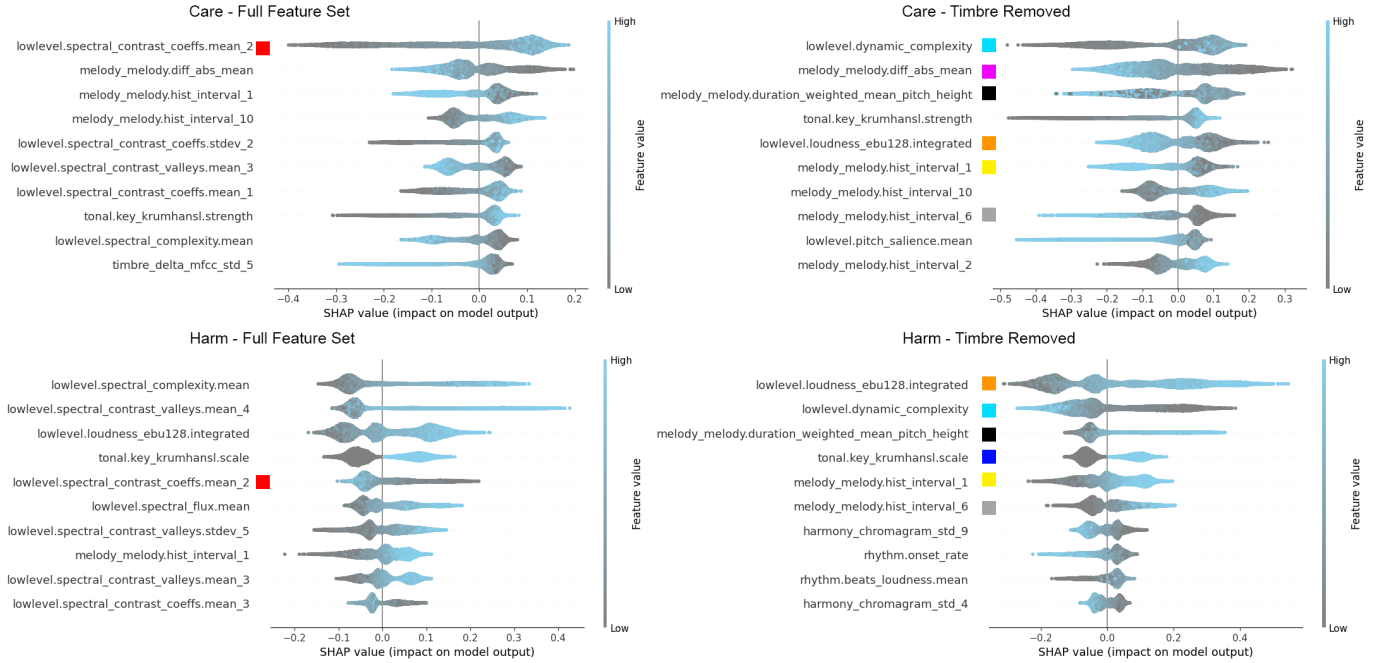

Fig. 4. Top 10 SHAP values for the care and harm models of the WASABI Dataset. Common features colour coded to reveal patterns.

Considering model design as the cause of this variability, one key trend emerges: model performance seems to be negatively correlated with skewness of the data. This can be seen in Fig. 9 and 10 (APPENDIX E).

One possible explanation for this negative correlation could therefore be that the method for addressing the class imbalance is inadequate. To examine this, two further techniques were tested: random oversampling (ROS) and Synthetic Minority Oversampling Technique (SMOTE). ROS is the process of randomly duplicating data points in the minority class until the classes are balanced. In general, ROS did not improve the model performance and was detrimental in a number of cases. It did however increase the subversion model's F1 performance from 0.22 to 0.32. While this is a fairly substantial improvement, the model still performs relatively poorly. SMOTE, the process of creating synthetic data points for the minority class using existing data, gave the same or decreased performance across the board apart from with the betrayal and authority models which rose from an F1 of 0.26 to 0.33 and subversion from 0.22 to 0.34, which again is substantial but not enough to make the models perform to a high standard generally. This variation in the effect of different class imbalance strategies, along with the feature selection findings showing improved performance with different combinations of feature subsets,

suggests that a grid search of a wider set of hyper parameters for each model could yield improved results. This however would require a very large amount of computing power as you would have to do a full grid search of all of the XGBoost parameters for each combination to find the best performing model.

If we instead treat this finding as something inherent in the data, studies on emotion in music may help to explain this. A number of studies have shown that lyrics interact with the mood of a piece of music in different ways in different pieces of music (Mori, 2022; Delbouys et al., 2018; Ali and Peynircioğlu, 2006; Brattico et al., 2011; Xu et al., 2011), broadly agreeing that sad songs rely on sad lyrics to convey emotion but happy music is unaffected by lyrical content. This means that instrumental music is equally effective at expressing happiness as lyrical music (Ali and Peynircioğlu, 2006) and that happy songs can feature sad lyrics and still be perceived as happy (Naseri et al., 2022). Returning to morality, this could mean that the emotional response created by the expression of certain moral values may exist purely in the lyrical domain and hence no link between lyrics and audio can be found.

Another interesting observation is the difference in performance of the models trained on the different datasets. The WASABI dataset gave a higher mean F1 score across the models. This is chiefly due to the large improvement in the cheating, subversion and degradation models. This could be attributed to the larger quantity of data and the more even distribution of the moral values.

It is interesting to note the tracks in the Billboard dataset are unified by the fact that they all achieved commercial success meaning that one can assume that the vast majority conform to the norms of commercially successful music: they were produced in professional recording studios, and conform to established songwriting structures and trends. The same cannot be said for the WASABI dataset which contains more varied and potentially noisier data. It is therefore unexpected that the WASABI dataset has better mean performance, suggesting that there are other, more important factors at play.

The two datasets display a difference in genre distribution (see Table 1). The WASABI dataset exhibits a swing away from pop primarily but also country, blues, R&B and jazz. It also exhibits a large increase in the representation of metal and punk music. Fig. 8 (APPENDIX A) plots the percentage of data points that are classified as having a high value (¿ 0.6) for each moral value, split by genre. Firstly this shows that moral values other than care and harm are generally lower across the board for the pop dataset. This could suggest that pop music contains a more simple set of moral messages. Parada-Cabaleiro et al. (2024) have found a trend of increasing simplicity in lyrics of pop songs which may be amplifying this.

We also see higher prevalence of negative moral values which could be due to the fact that music intended for chart success generally avoids explicit or dark themes. In the context of metal and punk being much more commonly represented in the WASABI dataset, it is notable that harm and degradation are particularly common values in these genres and care particularly uncommon, matching the trends in model performance.

This highlights the importance of considering the effect of dataset choice on the models. It would be interesting to expand the net further, looking at other collections of music with different genre distributions or that are grouped by other cultural metrics other than chart success. If MoralBert could be retrained on other languages, it would also be interesting to look at the ways sound and lyric morality interact in other musical cultures globally.

The cross-dataset validation results also provide some interesting insights. Firstly they show a large drop in performance overall suggesting that moral values are encoded in fairly different ways in the two datasets, perhaps for the reasons highlighted above. The two values that retain some degree of comprehension of the data are care and harm which tracks with the fact that they were the only two moral values that had reasonable performance on the Billboard set.

The feature elimination showed that, even with the timbre features removed which dominated the SAGE and SHAP values when trained on the full set, performance was maintained to a reasonable degree. This suggests that a large number of the features give helpful insights in predicting moral values, beyond those highlighted by the top SAGE and SHAP values. Similar results were found by Grekow (2017) in their study into categorical feature elimination in MER.

A further issue, identified by Han and Jiayi (2022) in relation to MER, is the use of global labels for moral content that can change over time. Drawing an example from the Billboard dataset, *Bohemian Rhapsody by Queen* (1976) is musically very diverse. Split into 5 distinct sections spanning over 6 minutes, it's difficult to see how a 30 second audio sample could support meaningful analysis of this work. Lu et al. (2006) make the case for dynamic MER where emotion tags are windowed and can therefore fluctuate throughout a piece of music. In the future it could be interesting to produce a dataset of dynamically morally annotated music and see if this improves performance and could provide new insights.

The high performance of the moral vs. immoral model seems to suggest that there is a universality in the sonic and musical rendering of moral and immoral lyrics, regardless of the specific moral value exhibited. This is also supported by the common SHAP values between the negative moral values. Finding the nuances that separate these moral values of the same valence is potentially much more difficult.

A possible other reason for the increased performance of the morality vs. immorality model is that, by only including songs that are moral *or* immoral, ambiguous lyrics were not included in the training set for this model, thus artificially increasing performance. Whilst this means the model would not generalise to unseen unfiltered data, it can still provide insights into what separates music that exhibits either positive or negative valance moral values.

Care and degradation are not diametrically opposed and can exist in tandem in the lyrical domain. This is true for

171 songs in the WASABI subset, including *Living on an Island* by Status Quo (1979). The lyrics of Living on an Island are momentarily joyous, discussing waiting for a friend and picturing their smile, but also explore themes of drug use and suicide. Musically the song also feels joyous and is in a major key, something the SHAP values associate with care, and so whilst the care model is able to identify the high care exhibited in the lyrics, the degradation model does not predict high degradation. In addition to morally ambiguous music, some songs feature counterpoint between music and lyric meaning. The aesthetic impact of this is explored by Sizer and Dadlez (2023) who state that it creates two forms of dissonance: affective and cognitive. Affective dissonance is where the music and lyrics make you perceive opposing emotions. Cognitive dissonance occurs when tension between the music and lyrics changes how listeners perceive the meaning of the song overall. These are complex devices in a composer's toolkit and, given the fact that they are outliers in the data and so there is not a large corpus of examples, may be hard to model.

To better understand this problem it could be worth conducting some more qualitative research exploring, not whether morality can be linked to sonic qualities as this study has shown, but why composers choose to match lyrics and musical content in this way. Panda et al. (2023) highlighted the need for emotionally motivated features and models for MER, perhaps it's through this line of questioning and interdisciplinary collaboration with composers that more targeted, morally motivated experimental designs can be found for MVR.

B. Towards Compositional Frameworks

A key motivation for this project is to start to build a compositional framework for how music can be matched to morally coded lyrics. There are two main applications for this framework – analysis and generation – and certain features lend themselves to providing helpful insights in one over the other. It's generally easier to see how higher level musical features can be applied to generation whilst composing in a way that satisfies low level features is conceptually quite difficult.

For higher level insights: high loudness, minor key, high mean pitch height, high mean absolute inter-note melody pitch height difference and frequent use of the sixth and minor second interval in the melody all appear to correlate with high values in the immoral values, both for the independent and bi-polar models. These are all things that could be directly applied to generation systems.

However, whilst this study has shown that moral and immoral music can be differentiated and that the presence of independent moral values can be identified, it has not conclusively shown that the nuances of individual moral or immoral values can be differentiated. It would be interesting to try to train models to differentiate between high care and high purity music, or high degradation and high cheating music. This could start to inform more fine grained compositional frameworks that target specific moral values.

These general insights feel quite intuitive, however the association between high mean pitch height and immoral music is in contrast with the correlation between height - both pitch height and non musical height - and morality in existing research. Meier et al. (2007) showed that moral meaning is more quickly identified in moral words - "caring," "charity," "trustworthy" - when written at the top of a screen and immoral words - "adultery," "corrupt", "evil" - when at the bottom, and Huang and Labroo (2019) demonstrate that higher pitch music can encourage listeners to make moral decisions such as choosing healthy food. Even the phrase "the moral high ground" hints at cultural linguistic understanding of this relationship. It is important to note however that these studies are not assessing morality through the lens of MFT and different understandings of morality may be a factor here. It could be interesting to build alternative models for quantifying morality in lyrics and assess the ways in which these interact with audio content.

Another possible reason for the inverse finding in this study could be the fact that other factors that affect pitch height, such as instrumentation, could be be a involved and that this overrides the association between height and morality. Instrumentation is fundamental to music composition and inclusion of this as an audio feature could be very valuable to future MVR and MER studies.

Despite being interpretable to some extent and shown to be an effective predictor for moral values, information from the melody interval histogram can be non conclusive. Given the octave normalisation, frequent use of the minor second interval could actually mean the occurrence of +1, +13, -11 intervals which to a composer or listener could mean very different things. This is one example of how, now this study has demonstrated the possibility of MVR, more work can be done to refine the feature set to maximise interpretability and performance. Furthermore, having shown that greatly reduced feature sets can provide equivalent results, one approach to creating truly explainable models could be to start with a small set of high level compositional questions and build a feature set around these. These could then be used to construct mid-level learnt features similar to Chowdhury et al. (2021).

When training the models with the full feature set, the SHAP and SAGE values are dominated by low level timbral features, demonstrating their ability to aid MVR. For example the mean spectral contrast in bin two correlates with care and is negatively correlated with all of the immoral values. This means that moral values are typically associated with more harmonic content in the 800-1600 Hz frequency range. How this could be used to inform generative systems is less immediately obvious as the higher level statistics but its prevalence as the top feature in a number of models suggests that it is an important quality to understand. Metrics like this do have their place in a system that is designed to analyse goodness of fit of music and lyric content.

What this study is not able to discern is correlation vs causation, and this is something that is worth keeping in mind when imagining compositional frameworks. For example, the

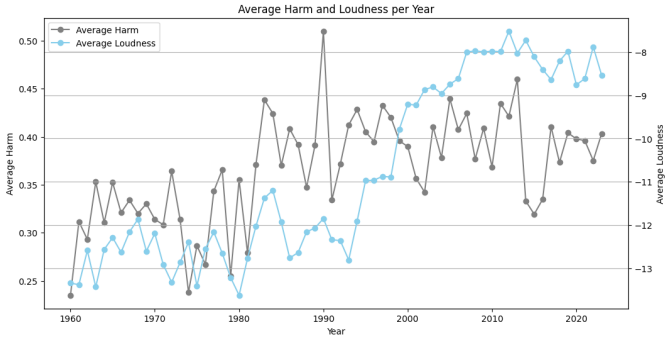

Fig. 5. Average harm moral value and integrated loudness plotted against release date.

results have shown that integrated loudness is a a key feature used by the model to predict harm, however, plotting average harm and integrated loudness against the release date (Fig. 5) reveals that these features exhibit similar increases — harm in the early 1980s and loudness in the early 1990s. The former of these rises was observed by Brand et al. (2019) who describe a trend of music lyrics becoming increasingly negative over the last 50 years. The latter is due to the so-called *loudness wars* that happened in the 1990s in which the dynamic range of music recordings is increasingly compressed to make tracks sound louder. These two events can't necessarily be causally linked and so may lead to false conclusions on compositional practices. It would be interesting to split the dataset into smaller time frames and see if this affects the ways in which the models make predictions.

VI. Conclusion

This study has demonstrated the potential for predicting the moral content of lyrics using audio features. Models for care, harm, cheating, subversion, and degradation achieved reasonable accuracy, with F1 scores of 0.4 or above, indicating that certain moral values have a specific set of musical norms for matching lyric and audio content. The bipolar model distinguishing between moral and immoral content was the most successful, with an F1 score of 0.81.

Global and local analyses of these models using SAGE and SHAP statistics identified the most influential features in predicting moral content, revealing a common set of features that appear to differentiate between moral and immoral music.

This thesis also highlighted several challenges and areas for future research, including expanding the dataset and diversifying data sources, refining the audio features in collaboration with music composers, and exploring whether moral values with the same valence — positive or negative — can be effectively distinguished.

References

[1] Akkermans, V., Serrà, J. and Herrera, P. (2009) "Shape-based spectral contrast descriptor", *Sound and Music Computing Conference (SMC'09)*, Porto, Portugal: July 23-25.

[2] Ali, S., and Peynircioğlu, Z. (2006) "Songs and emotions: are lyrics and melodies equal partners?", *Psychology of Music*, 34(4), pp 511-534.

[3] Alperson, P and Carroll, N. (2008) "Music, Mind, and Morality: Arousing the Body Politic", *The Journal of Aesthetic Education*, 42(1), pp. 1-15.

[4] Brand, C., Acerbi, A, and Mesoudi, A. (2019) "Cultural evolution of emotional expression in 50 years of song lyrics", *Evolutionary Human Sciences*, 1(e11).

[5] Brattico, E., Alluri, V., Bogert, B., Jacobsen, T., Vartiainen, N., Nieminen, S., and Tervaniemi, M. (2011) "A functional MRI study of happy and sad emotions in music with and without lyrics", *Frontiers in Psychology*, 2: 308.

[6] Britan, H. (1904) "Music and Morality", *International Journal of Ethics*, 15(1), pp. 48-63.

[7] Buffa, M., Cabrio, E., Fell, M.J., Fabien Gandon, Alain Giboin, Romain Hennequin, Michel, F., Pauwels, J., Pellerin, G., Maroua Tikat and Winckler, M. (2021) "The WASABI Dataset: Cultural, Lyrics and Audio Analysis Metadata About 2 Million Popular Commercially Released Songs", *Lecture Notes in Computer Science*, 12731, pp.515–531.

[8] Cameron, C.D., Lindquist, K. and Gray, K. (2015) "A Constructionist Review of Morality and Emotions", *Personality and Social Psychology Review*, 19(4), pp.371–394.

[9] Chowdhury, H., Praher, V., Widmer, G. (2021) "Tracing back music emotion predictions to sound sources and intuitive perceptual qualities", *Proceedings of the 18th Sound and Music Computing Conference (SMC2021)*. Virtual Conference: June 29-July 01.

[10] Covert, I., Lundberg, S. and Lee, S. (2020) "Understanding Global Feature Contributions With Additive Importance Measures", *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 33, pp.17212–17223.

[11] Davies, D (2013) "The Dialogue between Words and Music in the Composition and Comprehension of Song", *The Journal of Aesthetics and Art Criticism*, 71(1), pp. 13-22.

[12] Delbouys, R., Hennequin, R., and Piccoli, F. (2018) "Music Mood Detection Based On Audio And Lyrics With Deep Neural Net", *19th ISMIR Conference*. Paris, France: September 23-27.

[13] Gabrielsson, A., and Lindström, E. (2011) "The role of structure in the musical expression of emotions," in Juslin, P. and Sloboda, J. (eds.) *Handbook of Music and Emotion: Theory, Research, Applications*. London, UK: Oxford University Press, pp. 367–400.

[14] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., and Ditto, P. (2013) "Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism", *Advances in Experimental Social Psychology*, 47, pp 55-130.

[15] Graham, J., Nosek, B.A., Haidt, J., Iyer, R., Koleva, S. and Ditto, P.H. (2011) "Mapping the Moral Domain", *Journal of Personality and Social Psychology*, 101(2), pp.366–385.

[16] Grekow, J. (2017) "Audio features dedicated to the detection of arousal and valence in music recordings", *INISTA Conference*. Gdynia, Poland: July 3-5.

[17] Haidt, J. and Graham, J. (2007) "When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize", *Social Justice Research*, 20(1), pp.98–116.

[18] Haidt, J. and Joseph, C. (2004) "Intuitive ethics: how innately prepared intuitions generate culturally variable virtues" *Daedalus*, 133(4), pp.55–66.

[19] Han, D., Kong, Y., Han, J. and Wang, G. (2022) "A survey of music emotion recognition", *Frontiers of Computer Science*, 16(6).

[20] Huang, X., and Labroo, A. (2019) "Cueing Morality: The Effect of High-Pitched Music on Healthy Choice ", *Journal of Marketing*, 84(6), pp. 130-143.

[21] Huang, M., Rong, W., Arjannikov, T., Jiang, N., Xiong, Z. (2016) "Bi-Modal Deep Boltzmann Machine Based Musical Emotion Classification" in:

[22] Villa, A., Masulli, P., Pons Rivero, A. (eds) *Artificial Neural Networks and Machine Learning*, 9887, pp 199-207.

[23] Jiang, D., Lu, L., Zhang, H., Tao, J. and Cai, L. (2002) "Music type classification by spectral contrast feature", *IEEE International Conference on Multimedia and Expo (ICME'02)*, 1, pp. 113–116.

[24] Juslin, P. and Laukka, P. (2004) "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal New Music Research*, 33(3), pp. 217–238,

[25] Landmann, H. and Hess, U. (2017) "Testing moral foundation theory: Are specific moral emotions elicited by specific moral transgressions?", *Journal of Moral Education*, 47(1), pp.34-47.

[26] Lindberg, S., and Lee, S. (2017) “A Unified Approach to Interpreting Model Predictions”, *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, USA: 25 November.

[27] Lu, L., Liu, D. and Zhang, H. (2006) ”Automatic mood detection and tracking of music audio signals”, *IEEE Transactions on Audio, Speech and Language Processing*, 14(1), pp.5–18.

[28] Meier, B., Seldom, M., and Wigan, D. (2007) “Failing to take the moral high ground: Psychopathy and the vertical representation of morality ”, *Personality and Individual Differences*, 43, pp. 757–767 .

[29] Mori, K. (2022) “Decoding peak emotional responses to music from computational acoustic and lyrical features”, *Cognition*, 222.

[30] Naseri, S., Reddy, S., Correia, J., Karlgren, J., and Jones, R., (2022) “The Contribution of Lyrics and Acoustics to Collaborative Understanding of Mood”, *AAAI Conference on Web and Social Media*. Atlanta, United States: June 6-9.

[31] Panda, R., Malheiro, R., Paiva, R. (2023) “Audio Features for Music Emotion Recognition: A Survey”, *IEEE Transactions On Affective Computing*, 14(1), pp. 68-88.

[32] Parada-Cabaleiro, E., Mayerl , M., Brandl , S., Skowron, M., Schedl, M., Lex, E., and Zangerle, E. (2024) “Song lyrics have become simpler and more repetitive over the last five decades”, *Nature: Scientific Reports*, 14:5531.

[33] Posner, J., Russell, J. And Peterson, B. (2005) ”The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology”, *Development and Psychopathology*, 17(03), pp 715–734.

[34] Preniqi, V. (2024) *Automatic Detection of Morality Values in Music Lyrics* [Unpublished manuscript]. School of Electrical Engineering and Computer Science: Queen mary University of London.

[35] Preniqi, V., Kalimeri, K, and Saitis, C. (2022) “Soundscapes of morality: Linking music preferences and moral values through lyrics and audio”, *PLoS ONE*, 18(11).

[36] Salamon, J., Gomez, E., Ellis, D.P.W. and Richard, G. (2014) Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 31(2), pp.118–134.

[37] Sizer, L., and Dadlez, E. (2023) ”Why, Delilah? When music and lyrics move us in different directions”, *Philosophical Studies*, 181, pp. 1789–1811.

[38] Spotify (2024). *Spotify QuickPreview*. Available at: https://community.spotify.com/t5/Live-Ideas/Spotify-QuickPreview/idi-p/5929740 (Accessed: 13 August 2024).

[39] Spotify (2024) *Spotify QuickPreview*. Available at: https://community.spotify.com/t5/Live-Ideas/Spotify-QuickPreview/idi-p/5929740 (Accessed: 13 August 2024).

[40] Streich, S. and Herrera, P. (2005) ”Detrended Fluctuation Analysis of Music Signals: Danceability Estimation and further Semantic Characterization”, *AES 118th Convention*, Barcelona, Spain: May 28-31.

[41] Vatolkin, I., Rudolph, G., Weihs C. (2015) “Evaluation of Album Effect for Feature Selection in Music Genre Recognition”, *16th International Society for Music Information Retrieval Conference*, Malaga, Spain: October 26-30.

[42] Xu, L., Sun, Z., Wen, X., Huang, Z., Chao, C. and Xu, L. (2021) “Using machine learning analysis to interpret the relationship between music emotion and lyric features”, *PeerJ Computer Science*, 7:e785.

[43] Yaşar, M., and Akgün, S. (2023) “The Relationship Between Moral Foundations and Emotions”, *Studies in Psychology Cilt*, 43(3), pp. 429-466.

APPENDIX A
DATA DISTRIBUTION

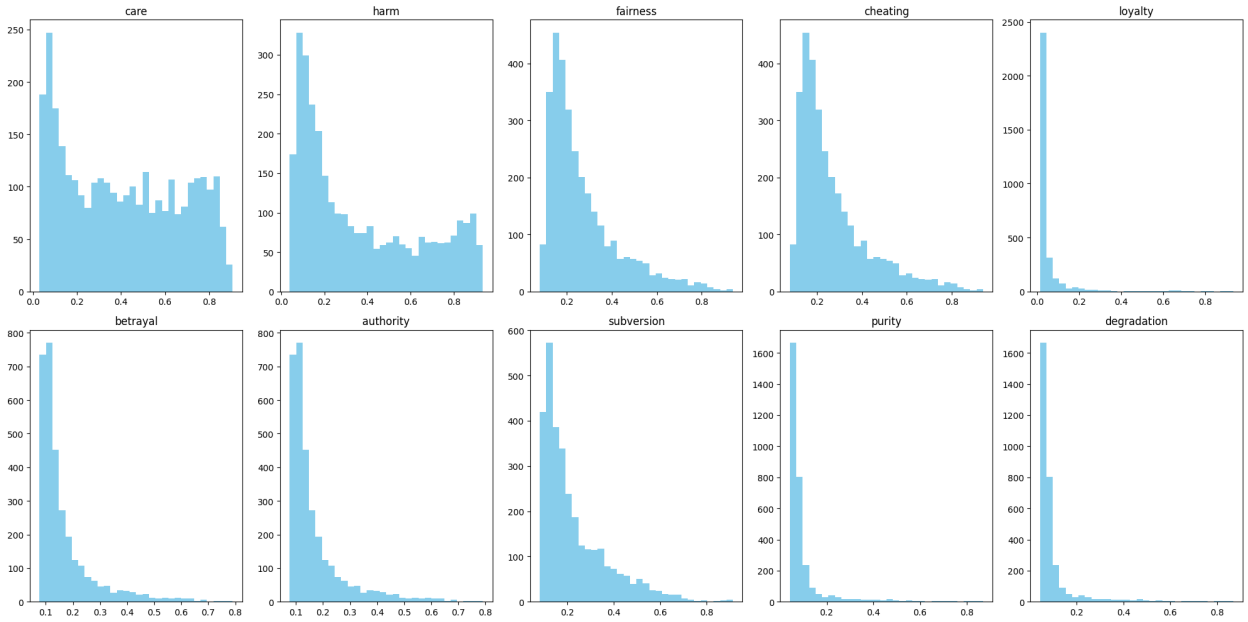

Fig. 6. Moral value distributions for Billboard dataset.

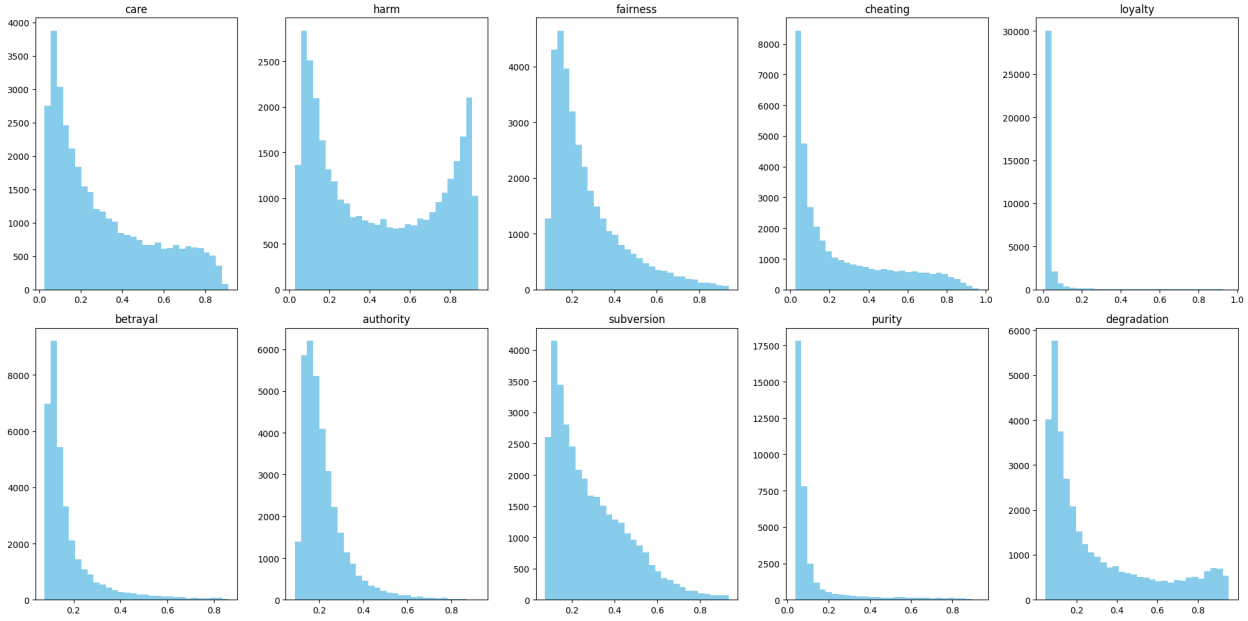

Fig. 7. Moral value distributions for WASABI dataset.

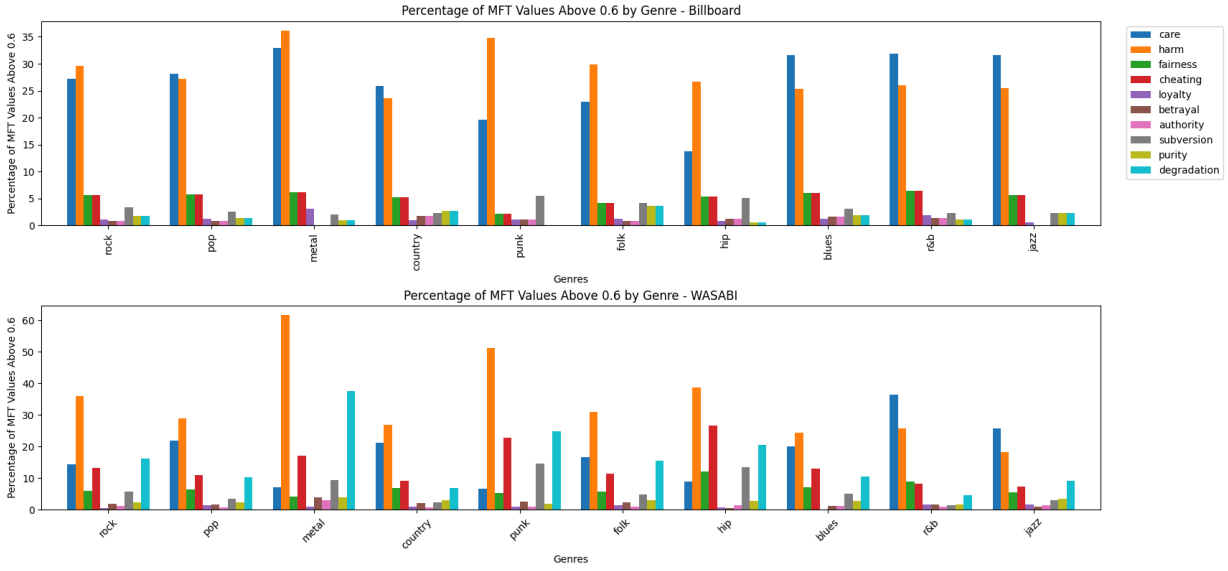

Fig. 8. Percentage of data points with a moral value above 0.6 for each genre.

APPENDIX B
FEATURE DESCRIPTIONS

The melody of the music is extracted using the Melodia (Salamon et al. 2014) vamp plugin. Analysis is then performed on this data.

FEATURE	**SOURCE**	**DESCRIPTION**
Pitch height	Bespoke	Duration weighted mean pitch height
Pitch Range	Bespoke	Standard deviation of pitch height
Direction	Bespoke	Normalised sum of pitch direction. +1 for ascending, -1 for descending.
Step size	Bespoke	Mean and standard deviation of absolute inter-note height difference.
Melodic intervals	Bespoke	Normalised histogram of inter-note pitch height differences, modulo 12, to fit into the chromatic scale (range 0 to +11).

TABLE II
Melody features.

FEATURE	**SOURCE**	**DESCRIPTION**
Pitch salience	Essentia	Measure of tone sensation. Ratio of the highest auto correlation value of the spectrum to the non-shifted auto correlation value.
Chord Histogram	Essentia	Histogram of chords present
Chromagram	Bespoke	Normalised histogram of pitches present in the piece on 12 note chromatic scale.
Key and Scale	Essentia	Key (e.g. Ab) and the mode (major/minor)
Key strength	Essentia	How strongly the harmonic content matches the key.

TABLE III
Harmony features.

FEATURE	**SOURCE**	**DESCRIPTION**
Tempo	Essentia	Estimated beats per minute.
Most common interval length	Essentia	Peak bin from inter-onset interval histogram. Weight and time.
Second most common interval length	Essentia	Second peak bin from inter-onset interval histogram. Weight and time.
Danceability	Essentia	Metric for how danceable a song is based on Detrended Fluctuation Analysis from Streich and Herrera (2005).
Rhythm density (onsets per minute)	Essentia	

TABLE IV
Rhythm features.

FEATURE	SOURCE	DESCRIPTION
Loudness	Essentia	Integrated loudness
Dynamic complexity	Essentia	Deviation from global loudness of 2 second segments.

TABLE V
Dynamic features.

FEATURE	SOURCE	DESCRIPTION
Spectral Content	Bespoke	Mean MFCCs with 13 bins.
Spectral Content Fluctuations	Bespoke	Mean First Differential of MFCCs.
Spectral Shape	Essentia	Mel flatness, skew, kurtosis, spread.
Spectral Flux	Essentia	L2-norm of the difference between two consecutive frames of the magnitude spectrum.
Spectral Centroid	Essentia	Related to the brightness of a signal. Weighted mean of the frequencies present in the signal, with their magnitudes as the weights.
Spectral Complexity	Essentia	The number of peaks in the input spectrum
Spectral Contrast	Essentia	Octave Based Spectral Contrast based on Jiang et al. (2009) and Akkermans et al. (2009).
Zero crossing rate	Essentia	Number of times the sign of the signal value changes, divided by the number of samples.

TABLE VI.
Dynamic features.

APPENDIX C
CATEGORICAL FEATURE ELIMINATION MODEL
PERFORMANCE (BILLBOARD)

Feature Set	**Calm F1**	**Harm F1**	**Num Features**
timbre	0.654	0.370	98
dynamics	0.553	0.420	12
rhythm	0.668	0.385	19
harmony	0.654	0.415	74
melody	0.610	0.407	27
timbre, dynamics	0.664	0.422	110
timbre, rhythm	0.678	0.431	117
timbre, harmony	0.672	0.433	172
timbre, melody	0.671	0.455	125
dynamics, rhythm	0.649	0.395	31
dynamics, harmony	0.672	0.389	86
dynamics, melody	0.591	0.392	39
rhythm, harmony	0.661	0.355	93
rhythm, melody	0.661	0.435	46
harmony, melody	0.669	0.358	101
timbre, dynamics, rhythm	0.692	0.422	129
timbre, dynamics, harmony	0.674	0.419	184
timbre, dynamics, melody	0.665	0.415	137
timbre, rhythm, harmony	0.679	0.406	191
timbre, rhythm, melody	0.670	0.398	144
timbre, harmony, melody	0.664	0.423	199
dynamics, rhythm, harmony	0.681	0.391	105
dynamics, rhythm, melody	0.651	0.424	58
dynamics, harmony, melody	0.682	0.336	113
rhythm, harmony, melody	0.682	0.373	120
timbre, dynamics, rhythm, harmony	0.686	0.383	203
timbre, dynamics, rhythm, melody	0.657	0.432	156
timbre, dynamics, harmony, melody	0.692	0.409	211
timbre, rhythm, harmony, melody	0.668	0.398	218
dynamics, rhythm, harmony, melody	0.695	0.356	132
timbre, dynamics, rhythm, harmony, melody	0.682	0.419	230

TABLE VII
PERFORMANCE ON FEATURE SUBSETS (BILLBOARD). BEST PERFORMING MODELS FOR EACH VALUE UNDERLINED.

APPENDIX D
CATEGORICAL FEATURE ELIMINATION MODEL
PERFORMANCE (WASABI)

Feature Subset	**Care**	**Harm**	**Cheating**	**Subversion**	**Degradation**
timbre	0.508	0.594	0.404	0.442	0.502
dynamics	0.450	0.589	0.331	0.335	0.438
rhythm	0.431	0.576	0.336	0.308	0.411
harmony	0.481	0.586	0.379	0.387	0.492
melody	0.480	0.552	0.371	0.391	0.480
timbre, dynamics	0.510	0.599	0.406	0.437	0.518
timbre, rhythm	0.506	0.595	0.411	0.432	0.510
timbre, harmony	0.508	0.616	0.409	0.434	0.525
timbre, melody	0.515	0.595	0.403	0.435	0.515
dynamics, rhythm	0.456	0.598	0.351	0.329	0.451
dynamics, harmony	0.488	0.597	0.381	0.384	0.494
dynamics, melody	0.497	0.589	0.380	0.417	0.496
rhythm, harmony	0.485	0.598	0.382	0.397	0.501
rhythm, melody	0.483	0.576	0.378	0.408	0.491
harmony, melody	0.502	0.586	0.390	0.413	0.514
timbre, dynamics, rhythm	0.510	0.604	0.410	0.431	0.517
timbre, dynamics, harmony	0.508	0.603	0.407	0.434	0.524
timbre, dynamics, melody	0.518	0.599	0.409	0.445	0.521
timbre, rhythm, harmony	0.514	0.603	0.408	0.451	0.526
timbre, rhythm, melody	0.515	0.595	0.402	0.442	0.519
timbre, harmony, melody	0.516	0.598	0.408	0.456	0.532
dynamics, rhythm, harmony	0.492	0.600	0.393	0.398	0.498
dynamics, rhythm, melody	0.499	0.590	0.390	0.410	0.497
dynamics, harmony, melody	0.509	0.593	0.398	0.422	0.515
rhythm, harmony, melody	0.502	0.593	0.399	0.415	0.514
timbre, dynamics, rhythm, harmony	0.511	0.605	0.410	0.432	0.535
timbre, dynamics, rhythm, melody	0.519	0.602	0.409	0.436	0.523
timbre, dynamics, harmony, melody	0.523	0.599	0.400	0.451	0.535
timbre, rhythm, harmony, melody	0.517	0.599	0.405	0.440	0.530
dynamics, rhythm, harmony, melody	0.508	0.599	0.399	0.431	0.517
timbre, dynamics, rhythm, harmony, melody	0.515	0.601	0.414	0.454	0.532

TABLE VIII
PERFORMANCE ON FEATURE SUBSETS (WASABI). BEST PERFORMING MODELS FOR EACH VALUE UNDERLINED.

APPENDIX E
EVENNESS OF DISTRIBUTION VS. PERFORMANCE

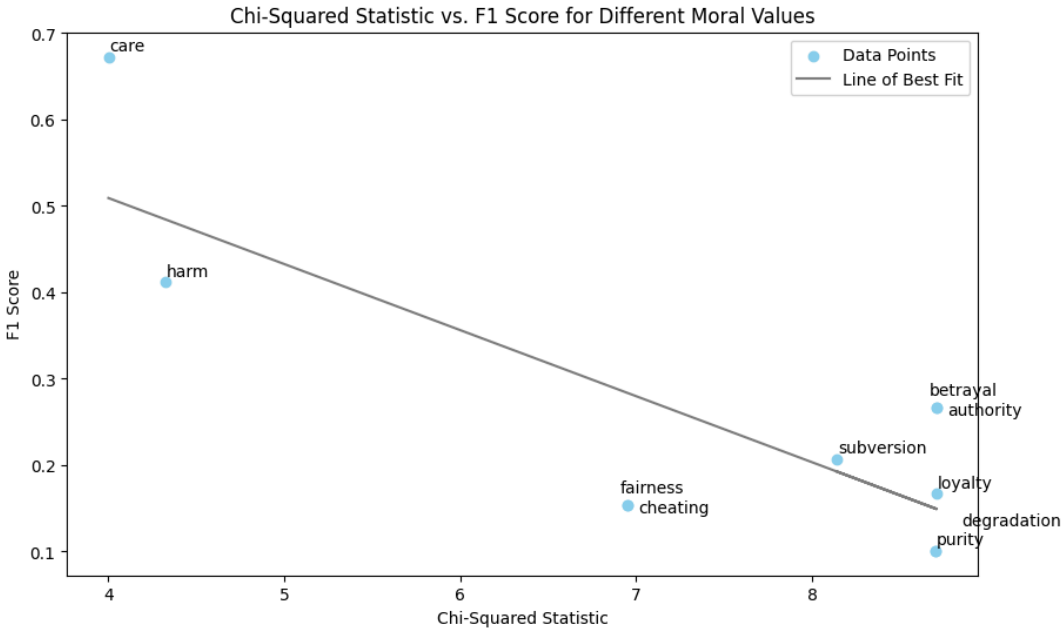

Fig. 9. Chi-squared statistic showing distance from even distribution plotted against performance (f1) for the Billboard dataset.

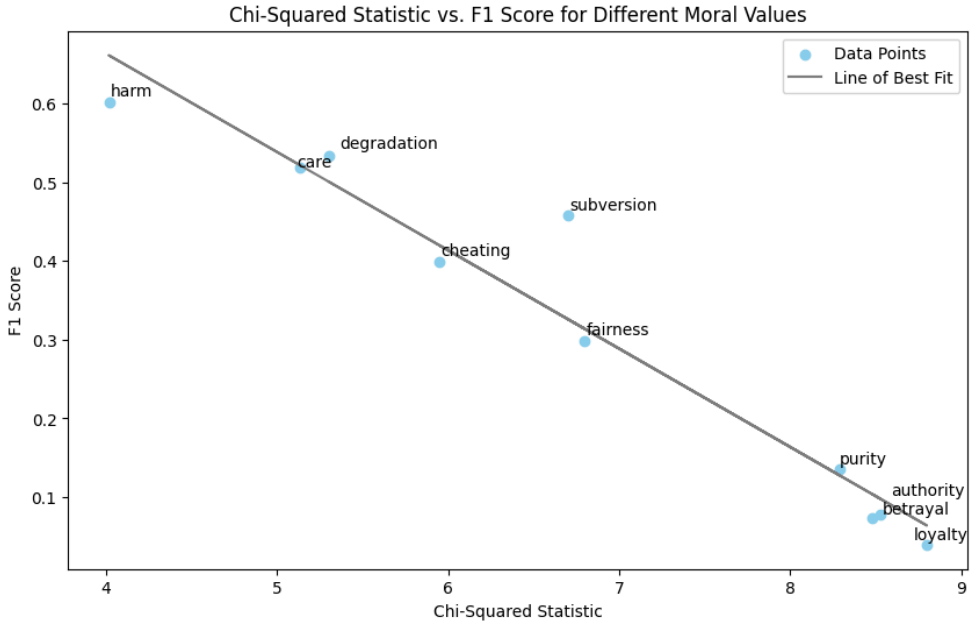

Fig. 10. Chi-squared statistic showing distance from even distribution plotted against performance (f1) for the WASABI dataset.

APPENDIX F
SAGE VALUES

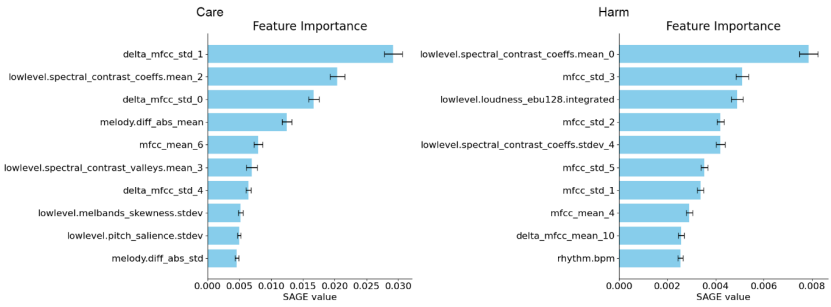

Fig. 11. Top 10 SAGE values for Billboard Dataset.

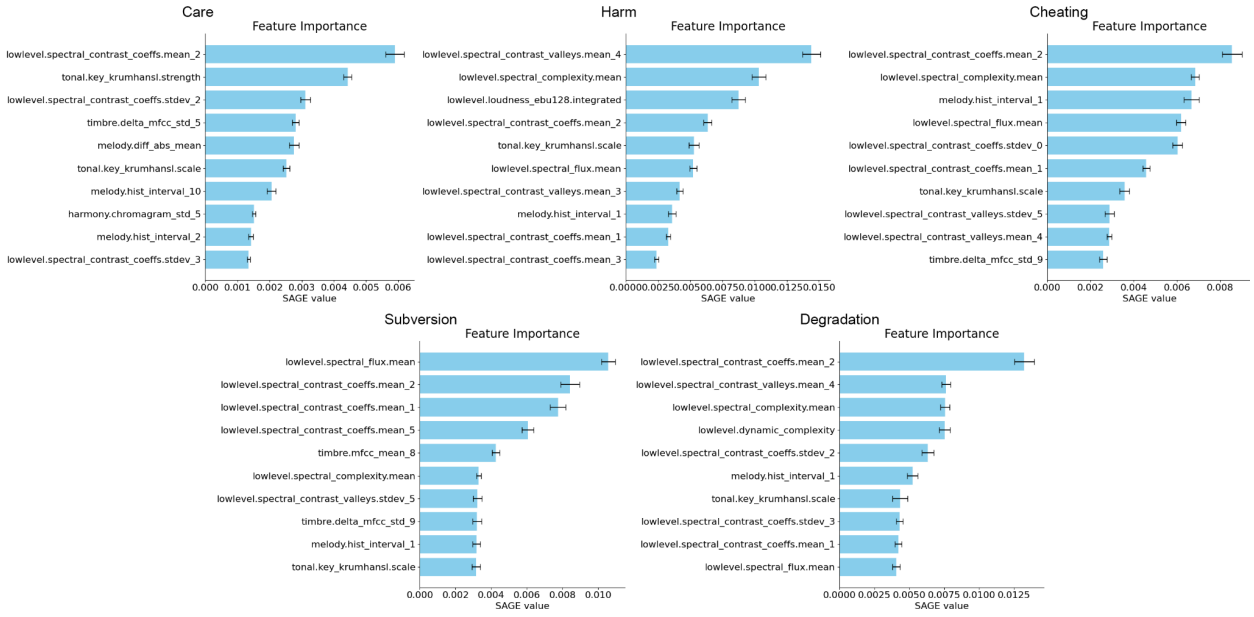

Fig. 12. Top 10 SAGE values for WASABI Dataset.

APPENDIX G
SHAP VALUES

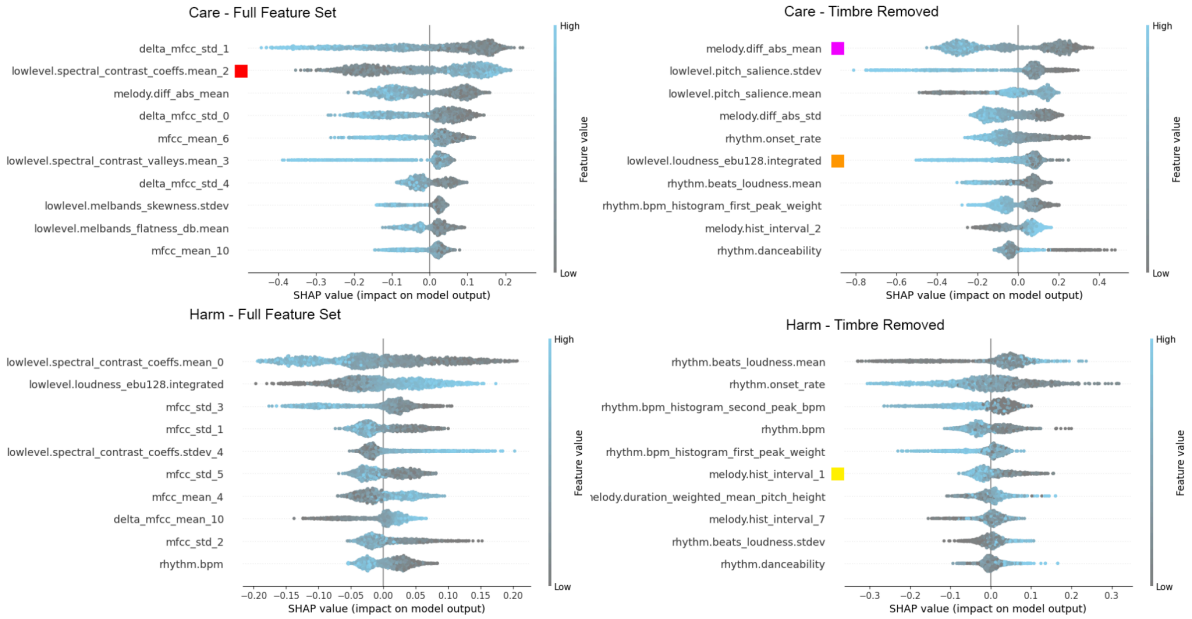

Fig. 13. Top 10 SHAP values for Billboard Dataset. Left: Full feature set. Right: Timbre removed from feature set. Common features colour coded to reveal patterns.

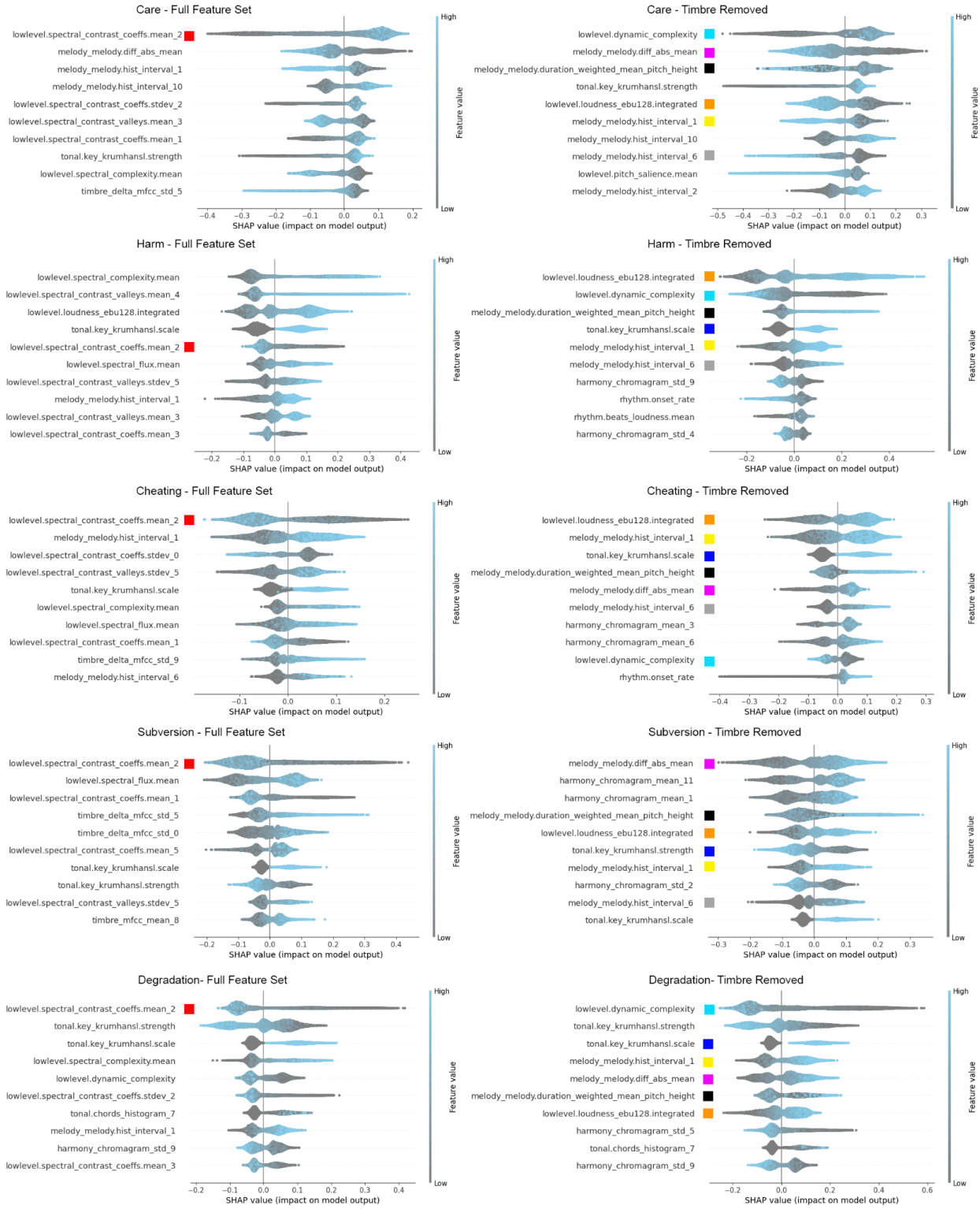

Fig. 12. Top 10 SHAP values for WASABI Dataset. Left: Full feature set. Right: Timbre removed from feature set. Common features colour coded to reveal patterns.

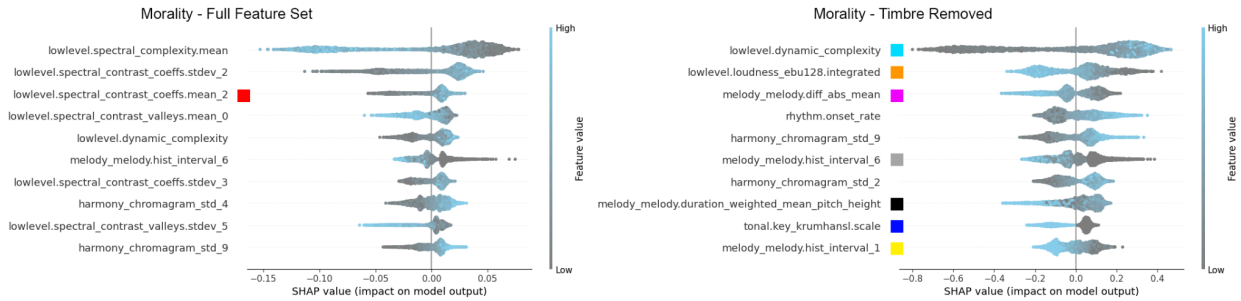

Fig. 13. Top 10 SHAP values for bi-polar moral vs. immoral model. Left: Full feature set. Right: Timbre removed from feature set. Common features colour coded to reveal patterns.