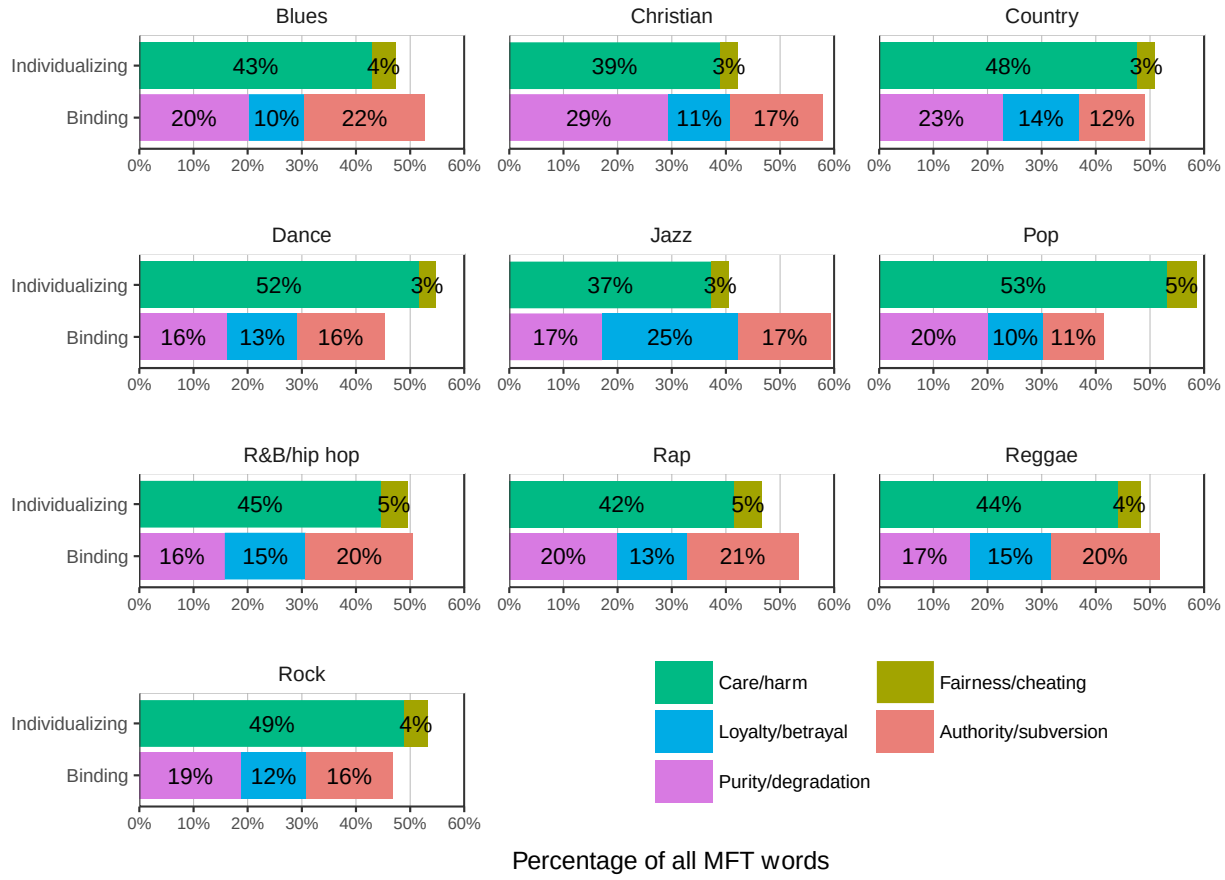
Appendix Table 1. Most-used words from Moral Foundations Dictionary by foundation.

Category	**Most-used Examples**
Individualizing	
Care/harm	fight*, care, hurt*, kill, war
Fairness/cheating	honest*, fair, justice, constant, equal*
Binding	
Loyalty/betrayal	together, family, enem*, nation*, cliqu*
Authority/subversion	control, father*, mother, respect, class
Purity/degradation	sick*, dirt*, holy, clean*, sin

Words are listed from most frequent to least frequent within each foundation, but encompass only the 5 most-used words out of lists that contain dozens of words in each. An asterisk signifies that any word that begins with the letters prior to the asterisk are counted as a match. The full dictionary can be downloaded from MoralFoundations.org.



Appendix Figure 1. Breakdown of moral word usage by foundation for each of ten genres. The five foundations are grouped into individualizing and bind categories to reflect the operationalization used in analyses.

Appendix: Description of Music Content Analysis Procedure

Automated Coding Procedure

To collect data for the content analysis, the first author wrote a software tool[1] to scrape weekly chart data from Billboard.com for each genre that was included in the survey. Complete listings of the top songs and albums for each week from January 2010 to September 2015 were stored in a database, organized by genre. The timeframe for analysis was selected to allow for a sufficiently large sample as well as one unlikely to be prone to random or seasonal variations. This was balanced against the desire to ensure that the songs analyzed were representative of the genres as they were currently understood by the typical survey respondent. Billboard maintains only a singles chart for pop music and only album charts for blues, classical, jazz, new age, and reggae; for all others, both single and album charts were scraped. The software then retrieved track listings for albums using an application programming interface (API) for Spotify, a music streaming service. When no match was found, the application would attempt to find the information via an API for Discogs, an online music database. The success rate in retrieving the lists of songs for charted albums ranged from 69.5% (reggae) to 89% (rock) among the genres included in the final analysis.

Once all songs were identified and categorized, the application searched for the accompanying lyrics using MetroLyrics.com and Lyrics.Wikia.com. Both report having licensed the lyrics from the copyright holders, which is taken as a rough heuristic for accuracy. When data were found at both sites, the lyrics from MetroLyrics were used in the analysis. Overall, success

[1] The command line program along with documentation can be found at https://github.com/jacob-long/Song-and-Lyric-Data-Scraper. Users can collect chart positions and lyrics for any publicly available Billboard chart and use the Spotify API to collect further metadata. The application is permanently archived with DOI: 10.5281/zenodo.1203368.

rates in lyric retrieval showed a great deal of variance. Fewer than 30% of relevant tracks were matched with lyrics for blues, jazz, and reggae. Country (78.5%) had the highest retrieval rate. While not ideal, the reasons for low availability of the lyrics for some genres are not likely related to the concepts of interest in this study. Moreover, some genres more than others—dance/electronic and jazz chief among them—make wide use of tracks that include no vocals.

Following other research linking media and MFT (e.g., Clifford & Jerit, 2013), the software program Linguistic Inquiry and Word Count (LIWC; see Tausczik & Pennebaker, 2010) was used to conduct the text analysis of lyrics. To assess the moral content of the corpus, the *Moral Foundations Dictionary* (Graham et al., 2009) was added to LIWC to generate quantities, via word counts, of the relevant words for each of the five moral foundations as well as general moral words like "good" and "evil" (see examples in Table 2). Although the dictionary's creators made a distinction between "vice" and "virtue" words (roughly corresponding to positive and negative valence) within each genre, no analytic distinction between them was made in the study that introduced the dictionary because the valence of the word may not be very useful in distinguishing endorsement (an endorsement of Purity can equivalently take the form of describing something as "pure" or "disgusting," for example). We follow that lead in focusing on the foundation categories rather than analyzing the "vice" and "virtue" subcategories separately.

Three music genres were dropped from consideration before final analyses were conducted. Latin music was dropped due to the unsurprising preponderance of Spanish-language lyrics, meaning the English-only LIWC dictionary used would systematically undercount the moral content significantly. Classical music was not included due to very poor retrieval rate of information both in terms of track lists and lyrics along with the fact that such a large portion of

the music is purely instrumental, although the exact portion is unknown. New age music was excluded for many of the same reasons, but also for the amount of non-musical (i.e., spoken word) albums that charted, meaning some of the "lyrics" included in the analysis would not have been from music at all.

Human Coding Procedure and Results

To investigate the appropriateness of using the automated procedure to measure the moral content of music lyrics, two human coders analyzed a subset of the lyrics. In particular, given that the dictionary was devised for a rather different corpus, the use of human coders was designed to detect irrelevant or otherwise non-moral uses of the dictionary words. If the dictionary approach seems to be too often wrong to human eyes, especially in ways that differ greatly from one genre to another, then proceeding further with data from the automated analysis would not be justified.

Past work based on the Moral Foundations Dictionary has used human coders for similar purposes in myriad ways, from rating word usage on a continuous scale (Graham, Haidt, & Nosek, 2009), to creating rules for dropping words from the dictionary which too frequently are not used to evoke the foundation (Lipsitz, 2017), to manually checking non-matches to see if they, too, are synonyms to dictionary words (Clifford, 2014). Our approach differs somewhat from other users of the dictionary, partly to suit the particular goals of this paper and otherwise to ensure methodological quality.

Given the goal of human coding in this case is to evaluate the word counting approach, some consideration of the unit of analysis is necessary. LIWC and other dictionary-based methods look at each word and decide whether it fits into a category or not. In the case of a small set of categories like the Moral Foundations Dictionary, with few exceptions a word is either a

word that taps one of the foundations, or it taps none of them. At this level of analysis, there are three main types of errors the automated analysis may make. First, the dictionary may omit words that tap a foundation, causing errors of omission (words are treated as fitting into no categories when they do fit into one). Second, the dictionary may cause words to be treated as if they evoke a foundation when, in some or all contexts, they do not (they should have been coded as having no foundation relevance). Lastly, the dictionary could be right to code a word as moral, but in some or all contexts the choice of a particular foundation is wrong. With this in mind, to check the work of the automated method, human coders need to investigate words coded as belonging to each of the foundations as well as words that were not coded as belonging to any of them.

Two coders not otherwise involved in the project were given focused training on Moral Foundations Theory after having been introduced to it in an undergraduate course. The coders were told the general purpose of the coding, namely how word counting methods for content analysis work and that the research team expected the computer to sometimes be wrong in its categorizations, hence the need for human checking. Coders, like LIWC, used single words as the unit of analysis. However, unlike LIWC, coders were shown the word along with the preceding and succeeding 15 words as context (except in cases in which the song did not have that many words prior to or subsequent to the sampled word). But, coders were specifically instructed to focus on the word's meaning, using the context only to disambiguate the meaning of the word and not to choose a category based only on context (since each word within the context could also be judged as being related to a moral foundation). Coders could choose any of the five foundations or indicate that the word did not evoke any of the foundations.

Coders were advised that dictionary words may have multiple meanings or atypical usages and told not to make judgments just on the basis of their assumption or knowledge that a word was included in a particular LIWC dictionary category. Coders were not told the song or genre the lyrics originated from, nor were they shown the category chosen by LIWC for any passage. To construct a sample that covered all relevant subsets of lyrics, passages were stratified in a 10 genres x 6 categories (5 foundations plus "none") scheme, with the latter dimension defined by the LIWC's categorizations. Units were randomly sampled from the full corpus within each stratum. For the "none" category, a standard list of "stop words" ("the," "and," etc.)[2] were omitted to ensure the sampled units would have some substantive meaning. The coders were reliable between themselves in an overlapping sample of 120 passages, reaching a Krippendorff's α (Hayes & Krippendorff, 2007) of .76. Each then was given an additional 240 passages to code independently, yielding a total 588 distinct coded passages.[3] As an initial basis of evaluation, Krippendorff's α was calculated for the entire sample of passages, treating LIWC as a third rater. The coefficient ($\alpha = .75$) in this case was virtually unchanged in comparison to using just the two human coders, suggesting our human coders in aggregate agreed with LIWC at a level that would be deemed acceptable in research that utilizes only human coders. On a genre-by-genre basis, the genre with the lowest reliability was rock, with $\alpha = .69$, while the highest reliability was Christian, with $\alpha = .84$. This suggests there is no particular genre that is so mis-

[2] The list originates with the SMART information retrieval system (described in Lewis, Yang, Rose, & Li, 2004) and obtained from the R package "quanteda" (Benoit et al., 2017).
[3] The total does not sum to 600 due to chance overlap in the coders' independent samples — these overlapping passages were used in the calculation of inter-coder reliability as well, but the inclusion/exclusion of these passages does not alter the value enough to change the rounded estimate.

coded by LIWC as to introduce high levels of systematic (on a between-genres basis) or random error into the combined analyses based on the automated coding.

References

Benoit, K., Watanabe, K., Nulty, P., Obeng, A., Wang, H., Lauderdale, B., & Lowe, W. (2017). quanteda: Quantitative Analysis of Textual Data (Version 0.99). Retrieved from https://cran.r-project.org/web/packages/quanteda/index.html

Clifford, S. (2014). Linking issue stances and trait inferences: A theory of moral exemplification. *The Journal of Politics*, *76*, 698–710. https://doi.org/10.1017/S0022381614000176

Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral Foundations Theory and the debate over stem cell research. *The Journal of Politics*, *75*, 659–671. https://doi.org/10.1017/S0022381613000492

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046. https://doi.org/10.1037/a0015141

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89. https://doi.org/10.1080/19312450709336664

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, *5*, 361–397.

Lipsitz, K. (2017). Playing with emotions: The effect of moral appeals in elite rhetoric. *Political Behavior*. https://doi.org/10.1007/s11109-017-9394-8

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*, 24–54. https://doi.org/10.1177/0261927X09351676