

## Extracting Latent Moral Information from Text Narratives: Relevance, Challenges, and Solutions

René Weber, J. Michael Mangus, Richard Huskey, Frederic R Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn & Ron Tamborini

To cite this article: René Weber, J. Michael Mangus, Richard Huskey, Frederic R Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn & Ron Tamborini (2018): Extracting Latent Moral Information from Text Narratives: Relevance, Challenges, and Solutions, Communication Methods and Measures, DOI: [10.1080/19312458.2018.1447656](https://doi.org/10.1080/19312458.2018.1447656)

To link to this article: <https://doi.org/10.1080/19312458.2018.1447656>



Published online: 15 Mar 2018.



Submit your article to this journal [↗](#)






View related articles [↗](#)



View Crossmark data [↗](#)



# Extracting Latent Moral Information from Text Narratives: Relevance, Challenges, and Solutions

René Weber <sup>a,b</sup>, J. Michael Mangus<sup>a,b</sup>, Richard Huskey <sup>c</sup>, Frederic R Hopp<sup>a</sup>, Ori Amir<sup>a,b</sup>, Reid Swanson<sup>d</sup>, Andrew Gordon<sup>d</sup>, Peter Khooshabeh<sup>e</sup>, Lindsay Hahn <sup>f</sup>, and Ron Tamborini<sup>f</sup>



<sup>a</sup>Department of Communication, UC Santa Barbara Media Neuroscience Lab, Santa Barbara, CA, USA; <sup>b</sup>Institute for Collaborative Biotechnologies, UC Santa Barbara, Santa Barbara, CA, USA; <sup>c</sup>School of Communication, The Ohio State University Cognitive Communication Science Lab, Columbus, OH, USA; <sup>d</sup>Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA; <sup>e</sup>Human Research and Engineering Directorate, US Army Research Laboratory, Adelphi, MD, USA; <sup>f</sup>Department of Communication, Michigan State University, East Lansing, MI, USA

## ABSTRACT

Moral Foundations Theory (MFT) and the Model of Intuitive Morality and Exemplars (MIME) contend that moral judgments are built on a universal set of basic moral intuitions. A large body of research has supported many of MFT's and the MIME's central hypotheses. Yet, an important prerequisite of this research—the ability to extract latent moral content represented in media stimuli with a reliable procedure—has not been systematically studied. In this article, we subject different extraction procedures to rigorous tests, underscore challenges by identifying a range of reliabilities, develop new reliability test and coding procedures employing computational methods, and provide solutions that maximize the reliability and validity of moral intuition extraction. In six content analytical studies, including a large crowd-based study, we demonstrate that: (1) traditional content analytical approaches lead to rather low reliabilities; (2) variation in coding reliabilities can be predicted by both text features and characteristics of the human coders; and (3) reliability is largely unaffected by the detail of coder training. We show that a coding task with simplified training and a coding technique that treats moral foundations as fast, spontaneous intuitions leads to acceptable inter-rater agreement, and potentially to more valid moral intuition extractions. While this study was motivated by issues related to MFT and MIME research, the methods and findings in this study have implications for extracting latent content from text narratives that go beyond moral information. Accordingly, we provide a tool for researchers interested in applying this new approach in their own work.

## Introduction

Moral intuitions frequently motivate individuals' personal and political choices. There is mounting evidence that humans possess innate moral sensibilities which enable the understanding and enforcement of norms regarding what is best for society as a whole. A well-known conceptual framework supporting this view is Moral Foundations Theory (MFT; Graham et al., 2012; Haidt & Joseph, 2007), which contends that moral judgment and decision-making are built on a universal set of basic, intuitive moral foundations.<sup>1</sup> Advocates of MFT propose at least five moral foundations: (1) care/harm (an intuitive concern for the suffering of others); (2) fairness/cheating (an intuitive

**CONTACT** René Weber  [renew@comm.ucsb.edu](mailto:renew@comm.ucsb.edu)  Department of Communication, University of California, Santa Barbara, CA 93106-4020

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/hcms](http://www.tandfonline.com/hcms).

© 2018 Taylor & Francis Group, LLC

preference for reciprocity and justice); (3) loyalty/betrayal (an intuitive concern for the common good and bias against outsiders); (4) authority/subversion (an intuitive deference to dominance hierarchies); and (5) sanctity/desecration (an intuitive concern for purity, broadly defined, including pathogen avoidance). A sixth foundation—liberty/oppression (an intuition about the feelings of reactance and resentment people feel toward those who dominate them and restrict their liberty)—is currently under consideration (see <http://moralfoundations.org>). The relative salience of each foundation varies both across and within cultures, and the variation in individuals' moral intuition salience can be used to explain differences in attitudes and behaviors. Research has shown, for instance, that political conservatives tend to emphasize sanctity, loyalty, and authority (the binding foundations) more than liberals; conversely, liberals tend to place the greatest emphasis on care and fairness (the individualizing foundations; Graham, Haidt, & Nosek, 2009; Haidt & Graham, 2007).

Moral themes are latent in a wide range of media content, and a theoretical frame for understanding the impact of moral content embedded in mediated narratives is the Model of Intuitive Morality and Exemplars (MIME; Tamborini, 2013). The MIME suggests that, over time, consistent exposure to messages emphasizing the superiority of one moral foundation over another will increase the salience of that foundation among audiences and maintain its salience in the face of other influences (Tamborini, 2013). Furthermore, the MIME holds that insulation from value-inconsistent messages will foster polarized values within ideological groups and reduce openness to divergent views (Leidner & Castano, 2012; Moscovici, 1985). For example, both American Protestant fundamentalist (Ammerman, 1991) and Islamic fundamentalist groups (Armstrong, 2000) have isolated networks of interpersonal and mass-media communication in which individuals are exposed exclusively to messages consistent with group values.

The MIME's predictions regarding both differences in content produced for different sub-groups, as well as the effects of exposure to that content, have found substantial empirical support in recent years. For example, several studies have supported the predicted differences in media produced for sub-groups that differ by age (Lewis & Mitchell, 2014), political interest (Bowman, Lewis, & Tamborini, 2014), and culture (Mastro, Enriquez, Bowman, Prabhu, & Tamborini, 2012). Feinberg and Willer (2013, 2015) have also shown that political messages are more persuasive when they are framed in terms of moral intuitions that align with the intuitions of the target population. Other studies have provided evidence in support of the MIME's predictions about the effect on intuition salience of both long-term exposure (e.g., Grizzard et al., 2016; Tamborini, Weber, Eden, Bowman, & Grizzard, 2010) and short-term exposure to moral intuitions embedded in narratives (e.g., Lewis, Grizzard, Mangus, Rashidian, & Weber, 2016).

## **MFT and MIME: previous moral intuition extraction from text**

Many MFT- and MIME-related studies use latent moral information in narratives as an important variable. For example, researchers have coded for the presence of MFT foundations in content analyses of participants' text summaries about their moral acts throughout the day (Hofmann, Wisneski, Brandt, & Skitka, 2014), religious interviews (McAdams et al., 2008), tweets (Sagi & Dehghani, 2014), television programs for children (Lewis & Mitchell, 2014), and political YouTube videos and newspaper articles (Bowman et al., 2014; Feinberg & Willer, 2013). Researchers have also made use of the Moral Foundations Dictionary (MFD) provided by Graham and colleagues (2009) to code newspaper articles mentioning stem cell research (Clifford & Jerit, 2013) or religious sermons (Graham et al., 2009). Yet despite their common goal of extracting moral foundations, the coding procedures used in these studies vary considerably. Differences in the procedures and coder characteristics are summarized in Table 1. Notably, details of the coder training procedure are poorly documented in most cases, leaving open questions for researchers who might hope to replicate those procedures.

Compared to general MFT studies, research utilizing MIME-based coding schemes provides relatively more systematic coder training and uniform coding procedures for extracting moral

**Table 1.** Summary of coding procedures and interrater reliabilities for content analyses using an MFT and MIME rationale.

Study	Procedure	Coder Details	Range of Intercode Reliability for Moral Foundations
<b>MFT-Based Content Analyses</b>			
Graham et al. (2009)	Word count software analyzed (n = 103) religious sermons for MFD words. Following this, human coders assessed the context surrounding each word in all sermons.	Linguistic Word Count Program; four coders blind to the study's hypotheses	N/A for the word count program; Krippendorff's alpha = .79, collapsed for all intuitions.
Hofmann et al. (2014)	Coders categorized each moral response (n = 3823) as it fit into one of the MFT foundations.	One rater (an author, Hofmann) coded participant responses (n = 3823) and one rater (a second author, Wisneski) coded 50% of this content.	Kappa = .85, collapsed for all intuitions.
Feinberg and Willer (2013)	In each study, coders indicated the extent to which each video (n = 51 videos; study 2a) or newspaper article (n = 232 articles; study 2b) was grounded in the 5 moral foundations on a 7-point scale ranging from 0 (not at all) to 6 (extremely).	Five coders blind to the study hypotheses (study 2a), and seven coders blind to the study hypotheses (study 2b).	Krippendorff's alpha = .73, collapsed for all intuitions for study 2a, and Krippendorff's alpha = .73 collapsed for all intuitions in study 2b.
McAdams et al. (2008)	Each interview (n = 128) being coded on a 5-point scale (1 = no concern; 5 = high concern) for each intuition.	Two coders—one of whom was blind to the hypotheses, and one of whom worked closely with an author to develop the rating system for each of the five intuitions	Alpha for care = .80; fairness = .76; loyalty = .82; authority = .82; purity = .86.
Clifford and Jerit (2013)	Coders identified: (1) MFD words associated with care, purity, and general moral words, (2) the contextual valence of the word, and (3) whether context associated with the moral word was being endorsed or rejected by the overall article.	One rater coded n = 3192 words from articles and one rater coded "a randomly selected subset of stories" (p. 664).	Krippendorff's alpha = .76, collapsed across care, purity, and general moral words.
<b>MIME-Based Content Analyses</b>			
Tamborini et al. (2017)	Coders identified (1) presence/valence of intuitions, (2) the extent to which these intuitions were rewarded or punished, and (3) whether the associated character was good or bad.	Four undergraduate coders coded n = 27 children's television episodes; a fifth rater acted as a referee to address disagreements.	Percent agreement was assessed for two sets of coders for care (74%, 75%), fairness (80%, 89%), loyalty (93%, 90%), authority (90%, 90%), and purity (95%, 94%).
Hahn et al. (2017)	Same procedure as Tamborini et al. (2017) above, except that coders coded (1) only scenes which contained intuitions in conflict, and (2) the characters' choice in this conflict	Three undergraduate coders coded n = 40 conflict scenes identified in Tamborini et al.'s (2017) sample.	Krippendorff's alpha for care = .74, fairness = .94, loyalty = .73, authority = .93, purity = 1.
Lewis and Mitchell (2014)	Coders identified (1) scenes which contained intuitions in conflict and (2) what intuitions were in conflict.	Two undergraduate coders coded n = 30 popular children's television programs (this sample was used in Tamborini et al. (2017) and Hahn et al. (2017).	Krippendorff's alpha for care = .93, fairness = 1.00, loyalty = 1.00, authority = 0, purity = 0 (note: Alphas that are zero denote categories in which at least one of the coders marked "absent" for all units of analysis).

(Continued)

**Table 1.** (Continued).

Study	Procedure	Coder Details	Range of Intercode Reliability for Moral Foundations
Bowman et al. (2014)	Coders identified (1) presence/absence (2) valence, and (3) intensity of intuitions.	Two female coders coded $n = 401$ headlines and $n = 352$ subheads of newspaper articles from sources based in U.S. counties	Krippendorff's alpha for presence/absence of loyalty = .98, fairness = 1, purity = .67; valence for loyalty = .89, fairness = .99; intensity for care = .71, fairness = .99.
Experimental MIME Studies that Include Content Analyses			
Tamborini et al. (2016)	Coders identified the extent to which experimental stimuli featured exemplars of any moral intuitions.	Three coders blind to the study's hypotheses rated a 40-min TV episode.	Krippendorff's alpha for care = 0.84, fairness = 0.91, loyalty = 0.90, authority = 0.85, purity = 0.92.
Tamborini, Lewis et al., (2016)	Coders identified the extent to which (1) experimental stimuli and (2) participant thought listings featured exemplars of any moral intuitions.	Three coders blind to the study's hypotheses rated (1) a 40-min TV episode and (2) participant thought listings ( $n = 173$ ).	Stimuli: Krippendorff's alpha for care = 0.84, fairness = 0.91, loyalty = 0.90, authority = 0.85, purity = 0.92. Thought listings: Krippendorff's alpha for care = .78, fairness = .89, loyalty = .88, authority = .82, purity = .82.
Grizzard et al. (2016)	Coders identified the extent to which experimental stimuli featured exemplars of any moral intuitions.	Two coders (who were authors but blind to the hypotheses at the time of coding) coded ( $n = 10$ ) movie plot summaries.	Krippendorff's alpha = .68 (81% agreement) for all coding categories combined.

intuitions from content. A typical procedure for MIME studies involves training coders for two to three months on definitions and examples of MFT foundations. In training sessions, coders code examples together, discuss the coding protocol aloud, and complete weekly assignments where they determine whether moral foundations are present in text and, if present, whether they are upheld or violated (e.g., Tamborini, Hahn, Prabhu, Klebig, & Grall, 2017). For instance, this procedure has been used in studies examining popular children's television programming. Tamborini et al. (2017) coded for the presence/absence of each moral foundation within a given episode. If a foundation was present, they then evaluated if it was in conflict with other foundations (e.g., should I choose to uphold care or fairness?; see also Hahn et al., 2017; Lewis & Mitchell, 2014). Although not content analytic research per se, three recent experimental studies have utilized a MIME-based coding scheme and procedure to assess the extent to which their stimuli feature moral foundations (Grizzard et al., 2016; Tamborini et al., 2016; Tamborini, Prabhu, Lewis, Grizzard, & Eden, 2016).

Although the extent of coder training for MIME-based studies may be relatively more uniform than other MFT content analyses, the procedures employed by MIME studies still varied in the examples used for coder training, the amount of in-person training coders received, and the characteristics of the coders themselves. Furthermore, a key difference in these studies is that they ask coders not to simply code explicit content, but instead to consider and classify how *latent* moral content activates their own subjective moral intuitions.

### **The current studies**

Despite considerable heterogeneity in the procedures described in Table 1, reported reliabilities nonetheless vary from a low of 0.73 (Feinberg & Willer, 2013) to a high of 1.00 (Hahn et al., 2017). This range is surprisingly high considering the subjective nature of moral intuitions and, as discussed in detail below, we believe reliability may be artificially inflated at the expense of validity. And, while this article focuses only on MFT- and MIME-based coding procedures, it is possible that this

concern may generalize to other studies focused on extracting latent content. Accordingly, the procedure described in this article offers a roadmap for researchers interested in evaluating other topics.

In this article, we subject different content extraction procedures to rigorous tests, underscore challenges by identifying a range of reliabilities each procedure is capable of producing, and provide solutions that maximize the reliability and validity of moral intuition extraction. In six content analytical studies we demonstrate that: (1) traditional content analytical approaches lead to rather low reliabilities when extracting moral content from news articles; (2) variation in coding reliabilities can be predicted by both text features and characteristics of the coders; (3) variation in coding reliabilities and coder agreement are largely unaffected by the intensity and detail of coder training—relying on a small group of highly trained and involved coders does not lead to substantially higher reliabilities than relying on a large group of coders with little training and involvement; and (4) a coding task with simplified training and a coding technique that treats moral foundations as the products of fast, spontaneous intuitions leads to plausible and acceptable inter-rater agreement. We discuss implications of these findings for future MFT and MIME research and suggest that the application of simplified coding techniques in a large crowd of coders leads to more valid extraction of latent moral information in text, and perhaps of latent information in general.

In this article, we focus our attention to a specific domain of latent information—moral intuitions in news narratives, which we consider an important and most difficult test case in communication studies. Beyond the extraction of latent moral information in news narratives, however, our methods and findings provide further recommendations and evidence for the usability of crowdsourcing for coding latent constructs in other domains such as in general political texts (Benoit, Conway, Lauderdale, Laver, & Mikhaylov, 2016; Lind, Gruber, & Boomgaarden, 2017). Lastly, our conclusions may also provide valuable methodological insights for coding and extracting more general frames in news using (human) supervised machine learning techniques (Burscher, Odiijk, Vliegthart, De Rijke, & De Vreese, 2014). Our discussion section considers these issues in greater detail.

## Content analyses 1–4: setting the baseline

### Coders

We conducted four separate content analyses using diverse human coder groups that differed in involvement and training time. The first coder group ( $n_1 = 3$ ) consisted of undergraduate research assistants who participated for a total of two academic quarters at the University of California Santa Barbara. Using a small group of trained coders is a common procedure in traditional content analyses. This first group of coders received an initial training using a Web-based platform (see the “Procedures - Online Platform” section below) which lasted for about one hour. Subsequently, these coders attended weekly one-hour research meetings where issues were discussed and questions clarified. Our second and third coder groups  $n_2 = 5$  and  $n_3 = 14$  consisted of undergraduate students participating in separate year-long honors seminars at Michigan State University. These students were highly involved as the outcome of their coding was relevant for a presentation of their work at a university-wide undergraduate conference. At the same time, these students also received a high level of training on MFT and the MIME (3 semester-units of course credit) in addition to a training using the Web-based platform, weekly training meetings, and example items to code for weekly homework (taken together,  $\approx 2.5$  hr/wk) that were specific to the content analysis procedure. Finally, a sample from the undergraduate research pool ( $n_4 = 223$ ) at UC Santa Barbara completed the 1 hr-long online platform training and coded articles for course credit. No additional training was provided and no additional incentives were issued to this last coder group.

Given the above, we understand these coder groups as follows:  $n_1$  = high-involvement, medium training;  $n_2$  = high-involvement, high-training;  $n_3$  = high-involvement, high-training; and  $n_4$  = low-

**Table 2.** Coder and coding statistics for content analyses 1–4.

Group	Coders	Pairs	Codings	Documents
High Involvement - Maximum Training	19	171	3,511	413
High Involvement - Medium Training	3	3	909	374
Low Involvement - Low Training	225	25,200	1,837	40

involvement, low-training. Table 2 summarizes the number of coders per group together with the number of coded articles.

### Text materials

We collected articles published between 2013 and 2015 from four major news outlets: *The New York Times*, *Reuters*, *CBS News*, and *The Washington Post*. Once per day, the politics section of each source was automatically crawled using a Scrapy spider (<http://scrapy.org>) and the full text of each article was stored, along with relevant metadata, in a relational database. Additionally, named entities were automatically extracted using the Stanford Named Entity Recognition engine (<http://nlp.stanford.edu/> Finkel, Grenager, & Manning, 2005), which provides a list of the people, locations, and organizations referenced in each article.

The Python Natural Language Toolkit (<http://www.nltk.org/> Bird, Loper, & Klein, 2009) was used to tokenize and stem the text of each article, which was then subjected to a simple word-frequency analysis: using the MFD created by Graham and colleagues (2009), which associates certain words with particular moral foundations, we counted the number of words for each moral foundation. These word-count measures were used to help select articles for human coding, ensuring that articles contain some moral information and prioritizing articles with high variance in moral content. Within this pre-filtered set, assignments of documents to coder groups were made randomly. Each coder group coded a common set of at least ten articles, although, because some coders completed more codings than others, certain pairs of coders have many more articles in common.

### Measures

#### Coders' moral and political views

We pre-tested our coders' political knowledge using a five-item index created by Delli Carpini and Keeter (1993). Coders' moral intuition salience was measured with the Moral Foundations Questionnaire (MFQ; Graham et al., 2009). To measure political views, we used the Society Works Best Index (SWB; Smith, Oxley, Hibbing, Alford, & Hibbing, 2011), which produces an additive index of liberalism/conservatism from subscales that reflect preferences for a society that takes care of its neediest members, has a tolerant approach to outgroups, promotes forgiveness of rule breakers, and favors egalitarian leadership practices with a flexible approach to moral codes of behavior. Participants also self-reported their political affiliation on an 11-point rating scale ("extremely liberal" to "extremely conservative").

#### Other coder characteristics

In addition to the measures above, we collected self-reported gender and age. System usage information was collected using a combination of client- and server-side logging in order to filter out low-quality participants, such as those who spent only a few seconds on training or coding pages.

## Text difficulty

We computed three measures of text difficulty. First, because *ceteris paribus*, longer articles require more sustained attention and cognitive engagement to understand, the total word-count of an article served as a simple proxy for its difficulty. Second, we computed lexical diversity using the uncorrected type-token ratio (TTR). Articles with a higher TTR—i.e., a greater proportion of unique words to total words—exhibit greater lexical diversity and thus may be more difficult to read (however, although TTR has been widely-used for many decades, this relationship is not uncontroversial; see Vermeer, 2000). Finally, based on the notion that articles that reference many different actors may have more complex latent moral narratives, we treat the number of entities identified by the automated entity-recognition system as an indicator of text difficulty.

## Procedures

### Online coding platform and coder training

An online training platform, the Moral Narrative Analyzer (MoNA; <https://mnl.ucsb.edu/mona>), was developed to assist in coding moral content in news articles. We deliberately chose an online platform that manages both coder training and the coding procedure so as to minimize inconsistencies that might be introduced by subtle differences in face-to-face interactions. This choice allowed for the rigorous testing of different training and coding procedures. It also allows for easy sharing and modification which should allow other researchers to implement these procedures in their own work. Interested readers should email the corresponding author for access.

Upon registering with the system, coders completed basic demographic questions followed by the political knowledge, SWB, and MFQ scales. Coders were then required to complete an online training procedure before they were qualified to code articles. This procedure included reading detailed descriptions of each moral foundation, step-by-step guidelines for article coding (with examples), and practice articles that had already been coded. Coding instructions and the conceptual definitions of moral foundations followed established protocols in MIME content analyses that have been used to code a diverse set of narratives in fictional (e.g., movies) and non-fictional (e.g., news) media content (see Tamborini et al., 2017). Coders were then required to complete several comprehension checks designed to assess their understanding of each moral foundation and the coding procedure. Automated feedback was given when coders did not pass a comprehension check, and coders were unable to advance until they correctly passed all comprehension checks.

### Coding procedure and units of analysis

After completing the training, coders began the coding process. Coders were presented with one article at a time to read and code. For each article, coders were required to specify which moral foundation was most salient overall (e.g., care/harm). Coders were also given the option to indicate that an article did not contain any moral content, which advanced the coder to the next article in their queue. For articles where an overall primary foundation was identified, coders were asked to identify the valence on a 5-point scale (e.g., completely care, mostly care, both care and harm, mostly harm, completely harm). This procedure was then repeated for the second-most-salient foundation within the article.

In addition, a given article may contain several moral actors who uphold or violate different moral foundations, thereby confounding attempts at article-level moral codings. Furthermore, long-standing traditions favoring balanced journalism tend to produce articles which avoid explicit moralizing despite describing morally relevant actions taken by the entities (e.g., people or organizations) discussed within an article. It is possible, then, that entities discussed within an article represent a more accurate unit of analysis for the assessment of moral content. If true, this suggests that reliabilities for codings at the article-level could end up being quite low, even if reliabilities for codings at the entity level are high. Accordingly, to test this premise, coders also coded morally-

**Table 3.** Coder-pairwise Krippendorff  $\alpha$  by group, content analyses 1–4.

Group	Primary Only			Primary or Secondary		
	Mean	SD	Range	Mean	SD	Range
High-max	0.14	0.13	(−0.04, 0.17)	0.18	0.31	(−0.32, 0.48)
High-med	0.10	0.08	(−0.01, 0.19)	0.14	0.24	(−0.09, 0.39)
Low-low	0.11	0.25	(−0.55, 0.99)	0.17	0.36	(−0.95, 0.99)

**Table 4.** Coder-pairwise Cohen  $\kappa$  by group, content analyses 1–4.

Group	Document Primary Foundation			Entity-Specific Foundations		
	Mean	SD	Range	Mean	SD	Range
High-max	0.20	0.10	(−0.10, 0.37)	0.21	0.13	(−0.13, 0.55)
High-med	0.15	0.05	(0.09, 0.22)	0.14	0.09	(0.07, 0.26)
Low-low	0.13	0.24	(−1.00, 1.00)	0.09	0.24	(−1.00, 1.00)

relevant entities within the article. Entities detected by the Stanford NER algorithm were presented as a list from which coders could select up to four entities. Each selected entity was coded according to their most salient moral foundation and the valence of that foundation.

Finally, coders were asked to rate their overall confidence in their coding on an 11-point scale. Coders who rated their confidence below 7 were required to select at least one reason for their lack of confidence from a predefined list (e.g., “the article was too long”).

## Results

### Reliabilities

We calculated corrected hit rates via Cohen’s Kappa (Cohen, 1968) and Krippendorff’s Alpha (Krippendorff, 2013) for all available coder pairs across three variables: article-wide primary moral foundation alone, article-wide primary and secondary moral foundations combined (liberally considering the pair to agree if either foundation matched), and the moral foundation assigned to any entities that were selected by both coders. Tables 3 and 4 summarize the results. Overall, reliabilities were quite low, ranging from 0.09–0.21, which is below generally accepted standards, and does not correspond to those reported in the MFT/MIME literature.

In addition to our general reliability analyses, we also reviewed the confusion matrices of many coder-pairs to assess trends in inter-rater agreement. Table 5 provides an example confusion matrix for one coder pair.

Overall, we found that the liberty and sanctity foundations are rarely used and often subject to substantial confusion when they do occur. We further noted that when the liberty foundation was selected as the primary foundation, coders were more likely to choose the midpoint of the moral valence scale (i.e., they were morally ambivalent). This result is consistent with the generally weak evidence for liberty as a distinct MFT foundation (e.g., Clifford, Iyengar, Cabeza, & Sinnott-Armstrong, 2015).

**Table 5.** Representative primary foundation confusion matrix for a pair in the high-max group ( $\kappa = 0.192$ ).

Foundation	Authority	Fairness	Care	Liberty	Loyalty	Sanctity
Authority	4	9	7	1	4	0
Fairness	1	23	3	2	2	1
Care	2	10	20	1	4	0
Liberty	2	9	3	1	2	0
Loyalty	1	1	2	0	5	0
Sanctity	0	4	0	1	1	0

**Table 6.** Significant predictors of inter-coder reliabilities for content analyses 1–4.

Predictors	$\beta$	t	Sig.
Lexical diversity	–.273	–3.26	.001
Coding confidence	.134	12.85	.000
Society works best index	.118	10.93	.000
Gender	.043	4.37	.000
MFQ fairness	.031	2.90	.004
Political affiliation	.031	3.04	.002
Age	.025	2.46	.014

R = .237; R<sup>2</sup> = .056 (5.6%); R<sup>2</sup>adj = .054 (5.4%); F(23, 9846) = 25.38; p < .001

### Predicting reliabilities

Next, we modeled pairwise reliability measures in a linear regression model with a number of coder-pair-specific qualities. Put differently, we analyzed whether pairwise reliabilities can be predicted by variables such as a coder pair’s similarity (euclidean distance) in political views, the text difficulty of a coder pair’s common article set, etc. (see the section “Measures” above). The analysis included n = 9869 coder pairs.

We found that text difficulty (as measured by lexical diversity) and—not surprisingly—coding confidence are the two strongest predictors of pairwise reliabilities. The more difficult the text material and the less confident coders are in their codings, the lower are their reliabilities (see Table 6). Furthermore, the analysis revealed that the more similar coders are in their SWB (Society Works Best attitudes) and self-reported political affiliation, the higher their reliabilities. Age and gender were also important predictors of pairwise coder reliabilities, with older coders and gender homogeneous coder pairs showing slightly higher reliabilities. A coder pair’s similarity in terms of moral foundation salience was only a significant predictor in the fairness foundation; similarity in other foundations did not significantly predict coders’ reliabilities. Likewise, all other measures, such as number of entities within a text, did not produce significant results. Notably, the number of care/harm, fairness/cheating, loyalty/betrayal, and authority/subversion words in a pair’s common article set (as captured by the MFD) did not make a difference. Only a higher number of sanctity/desecration words predicted significantly higher reliabilities, which can be explained by the rather low number of articles of this type among our news articles.

### Discussion

Our analyses show that our trained coder groups were—on average—not able to replicate the high levels of reliability and inter-coder agreement reported in the literature (see Table 1). In fact, even when evaluating the most highly trained coders using our most liberal metric, reliabilities do not meet the typical  $\alpha > 0.8$  threshold. Notably, while reliabilities increase slightly from the low-involvement, low-training to the high-involvement, maximum-training groups, the reliabilities do not differ substantially. Our pairwise reliability prediction model revealed that even when coders are extensively trained, text difficulty measures, coding confidence, political attitudes and affiliation, and even gender play an important role in explaining low reliabilities.

Overall, our results indicate that our human coder groups performed rather poorly on this type of (widely-used) moral foundation extraction procedure. There are several possible explanations for these findings. For instance, while consistent with previous content analytical paradigms, the decision to code moral information first at the article-level, and subsequently on entity level, makes several assumptions. Specifically, it assumes that an article contains just one or two overall moral foundations that are adhered to (e.g., a coder rates an article “completely/mostly care”) or violated (e.g., a coder rates an article as “completely/mostly harm”), that coders are sensitive to these adherences/violations at both the article-level and entity level, and that coders can be trained in such a way that they interpret these adherences or violations in a systematic and

reliable way. Despite the successes of previous content analyses, it is possible that these are untenable assumptions for news articles. However, some of the MFT content analyses which did code non-fictional news content (e.g., Feinberg & Willer, 2013) also report reliabilities above 0.7 (occasionally the human coding procedures for moral foundations were reduced to a simple “newspaper headline keyword find task” which also can explain a surprisingly high inter-coder reliability; see Bowman et al., 2014).

Drawing from our experiences with a number of pilot studies over a period of three years plus the four content analyses presented here, we think it is also possible that more fundamental assumptions about moral intuition extractions specifically, and about extracting latent information from text generally, may be flawed. In the following section, we explore the possibility that largely unchallenged assumptions made by traditional content analyses do not hold when applied to subjective, intuition-driven tasks like identifying latent moral information in text.

## Myths of trained human codings?

Generations of social scientists have used traditional quantitative content analysis as a tool to collect intersubjective, reliable, and valid data that allow inferences about messages (e.g., Holsti, 1969). Those messages can be provided in different modalities, but are usually represented via text. In the early years of content analyses researchers focused largely on the manifest content of messages (e.g., Berelson, 1952), which all coders can be reasonably expected to understand in the same way. In contrast, contemporary content analyses include the measurement of latent information in messages (for an overview, see Riffe, Lacy, & Fico, 2005; Vlieger & Leydesdorff, 2011), which requires some form of subjective inference from coders during the coding task (for instance, inferring a character’s intention within a narrative). Nevertheless, the quantitative content analyses used in social science research today predominantly emphasize: (1) a sound conceptual basis for all coding dimensions (both manifest and latent), (2) a methodical strategy for sampling and unitizing content, and (3) a detailed procedure for the selection and training of expert coders (Krippendorff, 2013). While there is little controversy regarding emphasis (1), the results of our four studies, as well as methodological innovations in the area of “big data social science” (see Lazer et al., 2009), challenge emphases (2) and (3).

Recent experimental research has shown that, despite the sophistication of the machine learning algorithms being applied to make sense of “big data,” human codings must still be considered an essential benchmark for the extraction of latent information from text data. However, analytical techniques for making sense of those codings are largely based on the outdated ideal of a single correct ground-truth (Hsueh, Melville, & Sindhvani, 2009). More specifically, supported by evidence from a series of experiments, Aroyo and Welty (2015) set out to debunk a number of myths in traditional content analyses. Four of the myths they identify are of particular interest here: (1) there is one correct interpretation and coding of every coding unit (ground truth); (2) disagreement of coders (low inter-rater reliability) is inherently bad and ideally should be eliminated; (3) coder training reduces disagreement by constraining possible interpretations; and (4) expert coders with conceptual knowledge of the coding categories always provide more reliable and valid data. To refute these myths, Aroyo and Welty (2015) suggest a new theory of *crowd* truth which assumes that human codings are inherently subjective (despite any training attempts), and that “measuring annotations on the same objects of interpretation [...] across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations” (p. 15).

Rejecting myth (1) in the context of moral intuition codings seems almost obvious. If moral intuition salience varies between individuals as MFT suggests, and intuitions represent a fast, mostly unconscious cognitive process that is largely unaffected by slow, conscious deliberations, then we should expect inter- and even intra-coder variation in evaluations of moral information in text. Furthermore, coder training, which focuses on conscious deliberation, should not be able to override intuitions substantially. Evidence for rejecting a ground-truth logic in other domains

is plentiful (see, e.g., Sheng, Provost, & Ipeirotis, 2008). Myth (2) can be rejected on the basis of our reliability prediction analyses above. The fact that we were able to identify coder (e.g., political attitude) and text (e.g., lexical diversity) characteristics that explained a significant amount of variance in inter-coder reliability is a testament that disagreement does not exclusively represent noise, but signal. With this information it becomes possible, for instance, to identify text with high and low moral ambiguity (i.e., high or low inter-coder agreement) or to identify a group of coders with a specific political attitude profile that is the best group to code texts of different complexity (best in terms of agreement or disagreement). Similarly, our results in studies 1–4—surprisingly we must admit—suggest that myth (3) and (4) can be dismissed in moral intuition coding procedures. Our coder groups clearly differed in the amount of knowledge, training, and involvement in the coding task: from a group of undergraduate students who read only a few pages of instructions and received little credit, to a group of undergraduate students who were highly trained over a period of 10 weeks, attended an honors seminar on MFT and the MIME, and had a personal interest in best practice, highly reliable codings for their research projects. Our results in studies 1–4 (and in previous pilot studies not reported here in which we tested different versions of our coder trainings) have shown that coder training and expert knowledge do not make a substantial difference in our moral intuition coding procedure.

We might conclude that, when it comes to coding latent moral foundations, tasks that follow the guidelines of traditional content analysis are unlikely to meet common standards for inter-coder reliability (e.g., Krippendorff  $\alpha > 0.8$ ), yet the published literature seems to demonstrate just the opposite. In light of our findings in content analyses 1–4, one possible explanation is that the reported reliabilities might be inflated by methodological practices that reduce the independence of coders. Additionally, although interesting predictions have been made about latent moral frames in news content, we believe those frames to be far more difficult to reliably identify than the more explicit moral content found in fictional narratives. More broadly, we question the ground-truth coding logic that undergirds the bulk of prior work when extracting moral foundations represented in text; low reliabilities should not be mistaken for noise. With this in mind, we now turn to content analyses five and six, in which we test a moral intuition extraction procedure that is not constrained by traditional content-analytical methodology in that it: (1) accounts for the inherent subjective nature of the moral intuition concept; (2) applies new metrics for inter-rater agreement; and (3) allows the procedure to be implemented on crowdsourcing platforms using a large number of human coders.

## Crowd content analyses 5–6: highlighting intuitions

For our fifth and sixth content analysis, we sought to radically redesign our coding procedure to capitalize on the crowd-truth paradigm discussed above. In doing so, we looked to other projects that developed simplified procedures for an otherwise-complex coding task. While a number of successful projects proved quite interesting (in fact, the Amazon Mechanical Turk platform was originally designed for exactly these sort of tasks), the EyeWire project (Kim et al., 2014) was most inspirational.

EyeWire is a large-scale coding project that simplifies the otherwise-complex task of tracing neurons in the retinae of mice. Historically, such a task required slow and painstaking work by highly trained specialists. The EyeWire project convincingly demonstrated that it is possible to break a complex project down into a series of small tasks that can be quickly and easily accomplished by a large number of minimally trained coders. The success of this project relied on some rather counterintuitive methods (at least according to traditional content analytical approaches). First, any single coding is not particularly useful. Codings were only useful in aggregate. Relatedly, codings for a given piece of content only provided useful data after a considerable number of coders had coded the same content. Contrary to the assumption behind myths 3 and 4 (presented above),

EyeWire analyses showed that an individual coder's coding quality was positively correlated with the number of codings completed; although the authors noted that such an outcome is uncommon in other crowd truth approaches. Accordingly, we set out to redesign the MoNA platform to allow for rapid training, highly modular coding tasks, the ability for coders to quickly code a multitude of articles, and scalability that allows for a large number of coders to code a substantial amount of news articles. We describe this revised procedure below.

## Coders

A new and fifth coder group ( $n_5 = 227$ ) was comprised of low-involvement/low-training students from the undergraduate research pool at UC Santa Barbara. These students received a simplified training procedure (see "Procedures" section below) and completed their article codings for course credit. No other training or incentives were provided.

In order to replicate findings in a larger, more heterogeneous crowd of human coders drawn from the general United States population, we used the Prolific Academic (<https://www.prolific.ac/>) platform and recruited 854 human coders, of which  $n_6 = 557$  fully completed all assigned tasks. In contrast to other crowd platforms (e.g., Amazon's Mturk), the Prolific Academic platform offers higher levels of "workers' quality control" and provides a more heterogeneous and more motivated group of human coders, in part due to a significantly increased pay rate requirements (we paid approximately four times more than compared to Amazon's Mturk) and stronger pre-screening of participants; (see Necka, Cacioppo, Norman, & Cacioppo, 2016). We attempted to match our sample of coders to the US population in terms of political affiliation and gender as best as possible within the constraints of Prolific Academic sampling frame, which includes more Democrats than Republicans. The final sample consisted of 195 female democrats (35%), 187 male democrats (34%), 40 female republicans (7%), 84 male republicans (15%), 24 unaffiliated females (4%), and 27 unaffiliated males (5%). The reported mean age was 32.59 years ( $SD = 11.45$ ). Political leaning was also assessed by using a single-item 11-point Likert scale ("Think about your personal political views. Where would you place yourself on a continuum ranging from very liberal to very conservative?," 0 = very liberal, 10 = very conservative), which had a mean of 3.26 ( $SD = 2.87$ ), further indicating that our sample leans somewhat toward the political left (across all student samples for which we have collected this measurement,  $n = 656$ , mean = 3.39,  $SD = 2.25$ ).

## Text material and measures

For coder group five, the news articles were drawn from the database described for studies 1–4 above. We selected a subset of 20 articles which had relatively high levels of inter-rater agreement on the earlier coding task, with an equal number of articles for each moral foundation (as labeled by a plurality of coders in studies 1–4). Coder group six read articles that were more recent (published in 2016 or later) and from more politically diverse sources than previous groups. Articles were drawn from *The Washington Post*, *Reuters*, *The Huffington Post*, *The New York Times*, *Fox News*, *The Washington Times*, *CNN*, *Breitbart*, *USA Today*, and *Time*. We utilized metadata provided by the Global Database of Events, Language, and Tone (GDELT; Leetaru & Schrodtt, 2013) to gather Uniform Resource Locators (URLs) of articles with at least 500 words. GDELT includes word-frequency scores for each moral foundation in the MFD. To make sure that our human coders received articles that included at least some moral content, we only selected articles which contained some MFD words. Using a purpose-built Python script, we attempted to scrape headlines and article text from those URLs, yielding a total of 8,276 articles. After applying a combination of text-quality heuristics and random sampling, 3,980 articles were selected for coding, of which 1,010 were coded by at least one participant. Compared with study 5, study 6 decreased the number of coders per article and increased the number of total articles. By capturing highlighted words (see below) from a

wider variety of articles, more training data is available to develop a successor to the MFD (which tends to produce relatively low variance in news articles; Graham et al., 2009).

All coders for study 5 and 6 provided informed consent before completing the same MFQ, SWB, political knowledge, and demographics questions as coders in content analyses 1–4 did. Likewise, the conceptual definitions of moral foundations for coders followed the same protocols as in study 1–4 (see above).

## Procedures

### Online coding platform and coder training

We developed a fast, crowd-truth-driven coding task: for each article, coders were instructed to simply highlight portions of the text which they understood to be related to an assigned moral foundation. This new coding model was designed to be much simpler for users, thereby minimizing training time and time-per-coding while emphasizing the intuitive nature of moral judgments.

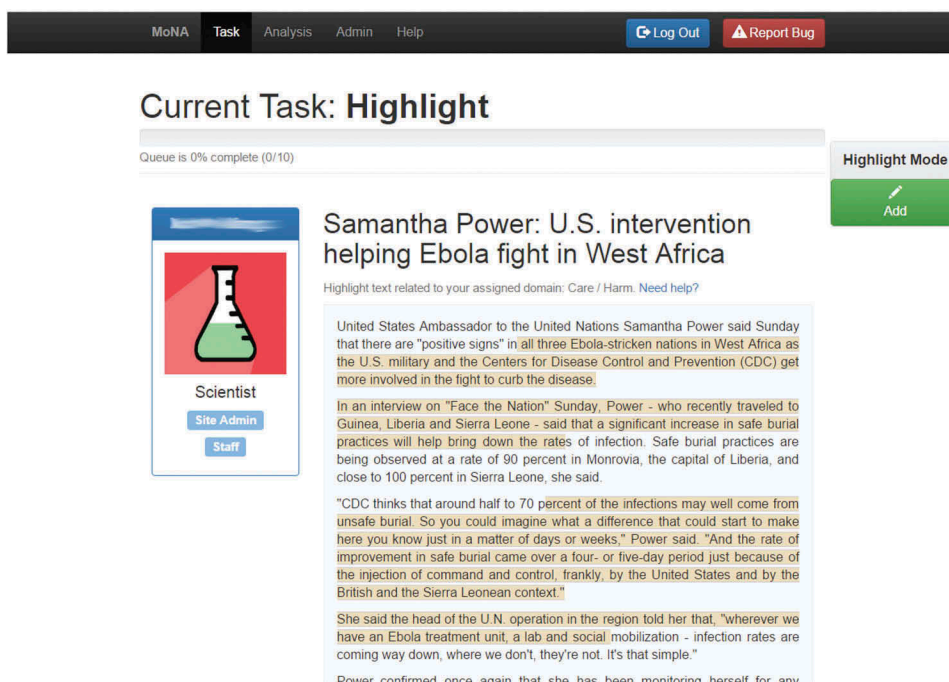
All coders received information about the background and purpose of the MoNA project, as well as text and a 7-min video explaining the general ideas behind MFT and each moral foundation. From there, coder training diverged with coders split into two groups: single-foundation coders and multi-foundation coders.<sup>2</sup> Single-foundation coders were tasked with learning about just *one* moral foundation (e.g., care/harm). This training included example images (e.g., a mother nursing a child, refugees in a war-torn country), a text-based description of the foundation, and detailed examples where the foundation was upheld or violated that were adapted from training materials used in previous content analyses (e.g., Tamborini et al., 2016, 2016). Multi-foundation coders were presented with the same materials, but for *all* (not just one) moral foundations. Subsequently, all coders (both single- and multi-foundation) were presented with text- and video-based training materials instructing them on how to complete the highlighting task (described below). Importantly, single-foundation coders were instructed to only code content pertaining to the specific moral foundation they were trained on. Multi-foundation coders were *also* tasked with coding an article according to just one moral foundation, however, the selected moral foundation differed for each article. This single- and multi-foundation coder strategy was adopted to empirically address an ongoing debate within the research team about whether a single-foundation coding strategy imposes too great of a restriction on user choice and potentially leads to lower coding validity (the “law of the instrument” argument—“give a small boy a hammer, and he will find that everything he encounters needs pounding”; Kaplan, 1964, p. 28).

### Text highlighting procedure

Upon completion of the training procedure, coders were directed to the coding interface where they began to highlight articles one at a time. A “cheat-sheet” was provided with five simple rules for effective highlighting. Items included: “only highlight your specific foundation”, “only highlight relevant content”, “moral content often relates to an entity”, “how much you highlight will change with each article”, and “when in doubt, don’t highlight”. The interface was designed such that coders were provided with a toggle button that allowed for adding or removing highlights (see Figure 1). Single-foundation coder highlights were always in yellow. A color-coding scheme was adopted for multi-foundation coders where foundation-specific highlight colors were applied. In total, student coders in group five generated 12,653 text highlights; general U.S. population coders in group six generated 68,983 highlights.

### Revised measurement of inter-coder agreement

Novel methods are required to evaluate the quality of our new coding procedure. Whereas many content analyses aim for categorical classification of discrete coding units pre-selected by the researchers, our highlight-based codings pre-assign a particular moral category and then allow coders to freely demarcate relevant units of information in the text. We adapted techniques from



**Figure 1.** A screen capture of the MoNA platform showing the document highlighting task.

natural language processing to assess inter-rater agreement as measured by the similarity of highlighted text, then pitched our empirical data against simulated random coding data to evaluate the effectiveness of our procedure for identifying moral content.

In order to evaluate inter-rater agreement, we need a measure of how similar the text highlighted by any given coder is to the text highlighted by other coders. We consider text highlighted by a coder as a judgment of that coder. Highlights for each article were preprocessed by tokenizing them into a list of words, filtering out the English stop words (e.g., “is” and “the”) provided by the NLTK stopwords corpus, and applying the Porter (1980) algorithm to reduce words to their stems. We then evaluated shared information between highlights using a vector space model. This space has as many dimensions as there are unique word-stems in the collection of all highlights for a given article. Each highlight can be represented as a vector, which will contain non-zero values for all the words that occur in that highlight. As is common practice in text summarization procedures, the vector space was transformed using term frequency-inverse document frequency weighting (TF-IDF; see Leskovec, Rajaraman, & Ullman, 2014) to account for the fact that more frequently used words provide comparatively less information about the semantic differences between two selections of text.

The cosine similarity was measured between all possible pairs of highlight vectors for a given article, yielding a two-dimensional matrix with 1’s on the diagonal such that the cell *similarity\_matrix[i][j]* contains the cosine similarity between highlight *i* and highlight *j*. The mean value of row *i* in the matrix therefore represents the mean cosine similarity of highlight *i* to all other highlights. Row masks were generated to filter a row’s values by assigned moral foundation; each highlight therefore has five mean-similarity scores, one for each moral foundation. When a coder is assigned to code content related to care, for example, each highlight should be more similar to other highlights for the care foundation than to highlights for any of the other moral foundations.

This technique was used to generate a data structure that is conceptually similar to a traditional confusion matrix used for categorical content analysis. The procedure is as follows.

- (1) Start with a  $5 \times 5$  matrix of 0's, one cell for each possible combination of moral foundations.
- (2) For each highlight, find the moral foundation that has the highest mean-similarity score. For instance, if the assigned foundation is care, and care is also the foundation with the highest similarity score, then count this as a match and increment the care/care cell. If the assigned foundation is care but the foundation with the highest similarity score is fairness, then count this as a miss and increment the care/fairness cell.
- (3) Once all highlights have been processed, divide each cell by the sum of its column to get a proportion. This allows us to know the proportion of highlights for which the assigned foundation was also the maximum-scoring foundation. If coders' codings are able to consistently distinguish between moral foundations, then the final matrix will have high values on the diagonal and low values off the diagonal.

## Results

### *Inter-coder agreement - study 5 (student coders)*

To compare foundation-assignment techniques, this procedure was run two separate times, once for single-foundation coders and again for multi-foundation coders. This data is summarized in Tables 7a and 7b below.

As expected, values are highest on the diagonal and exceed a naive baseline proportion of 0.2 for the within-foundation comparison. Furthermore, it seems that the multi-foundation group does a slightly better job in distinguishing between most foundations.

In order to evaluate whether these results are likely to have occurred by chance, we developed a simulated coding system to provide random highlights. The “robocoder” simulation was built such that the number of highlights per article and the number of words per highlight match the empirical distributions from our human coders. However, unlike our human coders, “robocoders” are naive to the semantic content of the text. Instead, each simulated highlight begins at a randomly selected word in the article and is associated with a randomly selected moral foundation. The result is a set of simulated highlights that match the formal elements of our empirical data (highlights per article and words per highlight) but should not distinguish moral foundations.

The “robocoder” procedure was used to generate 100 simulated datasets, each containing approximately the same number of highlights as the empirical sample, which were analyzed with the same procedure described above. The mean and standard deviation of the simulated results were used to standardize the values from Tables 7a and 7b. These standardized scores (z-values) are

**Table 7a.** Foundation-score proportions for the single-foundation coders in study 5. Columns are the assigned foundations; values are rounded.

Foundation	Authority	Care	Fairness	Loyalty	Sanctity
Authority	0.308	0.132	0.153	0.151	0.143
Care	0.130	0.285	0.195	0.202	0.222
Fairness	0.223	0.231	0.302	0.236	0.185
Loyalty	0.179	0.195	0.191	0.277	0.174
Sanctity	0.159	0.158	0.160	0.136	0.276

**Table 7b.** Foundation-score proportions for the multi-foundation coders in study 5. Columns are the assigned foundations; values are rounded.

Foundation	Authority	Care	Fairness	Loyalty	Sanctity
Authority	0.308	0.109	0.156	0.171	0.124
Care	0.129	0.317	0.169	0.190	0.235
Fairness	0.229	0.202	0.353	0.207	0.230
Loyalty	0.219	0.168	0.181	0.285	0.180
Sanctity	0.116	0.204	0.141	0.146	0.231

**Table 8a.** Z-scores for foundation-score proportions, single-foundation coders, study 5. Standardized based on a simulated random baseline. columns are the assigned foundations; values are rounded.

Foundation	Authority	Care	Fairness	Loyalty	Sanctity
Authority	7.10	-4.48	-3.11	-3.26	-3.73
Care	-4.58	5.57	-0.36	0.11	1.45
Fairness	1.50	2.01	6.73	2.34	-0.99
Loyalty	-1.36	-0.34	-0.61	5.04	-1.74
Sanctity	-2.67	-2.76	-2.65	-4.24	5.01

**Table 8b.** Z-scores for foundation-score proportions, multi-foundation coders, study 5. Standardized based on a simulated random baseline. columns are the assigned foundations; values are rounded.

Foundation	Authority	Care	Fairness	Loyalty	Sanctity
Authority	7.08	-5.96	-2.91	-1.88	-4.97
Care	-4.68	7.70	-2.03	-0.65	2.28
Fairness	1.89	0.12	10.09	0.47	1.97
Loyalty	1.25	-2.11	-1.26	5.61	-1.35
Sanctity	-5.53	0.25	-3.90	-3.55	2.07

presented in [Tables 8a](#) and [8b](#). The strong positive z-scores along the diagonal indicate that the mean pairwise similarity of highlights within each moral foundation are significantly higher than would be expected if highlights were made at random (all  $z$ 's  $> 1.65$ ,  $p < 0.05$ ). Conversely, strongly negative z-scores indicate mean pairwise similarities that are significantly lower than would be expected if highlights were made at random, while scores close to zero indicate similarity that is roughly equivalent to what would be expected from a sample of random highlights. For instance, [Table 8a](#) shows that highlights from single-foundation coders assigned to authority are significantly similar to each other ( $z = 7.10$ ,  $p < 0.0001$ ), and significantly dissimilar from care coders' highlights ( $z = -4.58$ ,  $p < 0.0001$ ), when compared to the baseline similarity level of random highlights.

### *Inter-coder agreement - study 6 (general U.S. population coders)*

Our results in study 5 were largely replicated in our sample of coders from the general U.S. population in study/coder group 6. We applied the same analytical procedure described for study 5 above; results, which exhibit the same general pattern as study 5, are summarized in [Tables 9](#) and [10](#). Note that the primary motivation for study 6 was to collect data across a wide variety of articles, so each article was coded by at most 15 coders. Consequently, the raw data presented in [Table 9](#)

**Table 9.** Foundation-score proportions for study 6. Columns are the assigned foundation; rows are the maximally similar foundation; values are rounded.

Foundation	Authority	Care	Fairness	Loyalty	Sanctity
Authority	0.241	0.151	0.182	0.192	0.165
Care	0.165	0.251	0.195	0.177	0.203
Fairness	0.180	0.181	0.236	0.180	0.163
Loyalty	0.200	0.165	0.171	0.222	0.161
Sanctity	0.214	0.252	0.216	0.229	0.307

**Table 10.** Z-scores for foundation-score proportions, study 6. Standardized based on a simulated random baseline. columns are the assigned foundation; rows are the maximally-similar foundation; values are rounded.

Foundation	Authority	Care	Fairness	Loyalty	Sanctity
Authority	6.97	-5.14	-1.01	0.47	-2.91
Care	-2.55	6.66	0.86	-1.00	1.89
Fairness	-0.89	-0.95	5.24	-0.85	-2.89
Loyalty	1.30	-2.84	-2.01	3.27	-3.34
Sanctity	-4.87	0.38	-3.89	-2.91	6.34

must be interpreted carefully and the simulation-standardized results in Table 10 should be preferred. This is most noticeable in the inflated similarities for the sanctity foundation seen in Table 9. That trend is apparent in both the simulated and empirical data, suggesting it is an artifact of the analytical procedure. The simulation therefore naturally accounts for this inflation, since it applies an identical analysis procedure on equally sparse but randomly generated highlights, leading to z-values for sanctity in Table 10 which are relatively low compared to the raw proportions reported in Table 9.

## Discussion

The results in our revised coding task demonstrate that our highlighting procedure greatly outperforms a random baseline in both a homogenous student group and more heterogeneous general population group of human coders. Furthermore, it seems that the multi-foundation coders are generally better able to distinguish between foundations compared to the single-foundation coders, although these differences are relatively small. We interpret these findings and their replication in a large independent group of human coders as good evidence that a highlighting procedure for moral intuition extraction from text does indeed produce consistent, non-random results while better accounting for the inherent latent and subjective nature of moral intuitions.

## Overall discussion, limitations, and outlook

In our content analyses we found that traditional content-analytical approaches lead to moral intuition extraction from text narratives with highly variable but generally low reliabilities which can be predicted by both text and coder characteristics. We also found that this variation is largely unaffected by coder selection and coder training. In our section “Myths of Trained Human Codings?” above, we discuss the divide between current practices in traditional content analyses and evidence that refutes fundamental assumptions of trained human codings, which provided a theoretical basis for our reformed content analytical approach in studies 5 and 6. We understand the following *Myths of Human Annotations* (Aroyo & Welty, 2015) as especially noteworthy: the notion that there is one correct interpretation and coding of every coding unit; that disagreement of coders is inherently undesired; that coder training reduces coder disagreement; and that expert coders with conceptual knowledge of the coding categories provide more reliable and valid data. We interpret our findings as further evidence against these myths which seem especially prevalent in content analyses that focus on the extraction of latent information from text. Moral foundations, as represented in text, can be considered as latent information because human coders’ perception and interpretation of moral information crucially depends on the salience of coders’ individual moral intuitions. Furthermore, if moral intuitions follow largely a fast, spontaneous, subconscious cognitive process, then it is not surprising that deliberations (i.e., coder trainings) are mostly ineffective.

However, this does not necessarily mean that in traditional content-analytical approaches we should generally see much lower inter-coder reliabilities than reported in the literature. We believe that due to the typical setup of content analytical studies in communication research, in which coders are either part of the research team (frequently including the investigators as expert coders; see, e.g., Grizzard et al., 2016) or at least are able to communicate among each other during coder trainings and even during the actual coding to “clarify confusion” (see, e.g., Goranson, Ritter, Waytz, Norton, & Gray, 2017, a study with serious implications), those coders inadvertently adapt to each other’s coding and potentially outcome expectations. This may be especially true when traditional content analysts report

“spending months in training sessions with coders, during which time they refined categories, altered instructions, and revised data sheets until the coders felt comfortable with what was expected of them and the analysts

were convinced they were getting the data they needed. It is typical for analysts to perform reliability tests during the development of coding instructions until the reliability requirement is met as well” (Krippendorff 20, p. 130).

As communication science often aims for an understanding of how a broader, more diverse public, and not extensively trained human coders respond to media messages, a crowdsourcing content analytical approach may actually reflect a more ecologically valid procedure for the assessment of latent constructs embedded in media content (Lind et al., 2017).

The real issue here is then: Does the common logic that validity never trumps reliability still apply under these circumstances? (Potter & Levine-Donnerstein, 1999). Could it be that reliability in traditional moral foundation extraction procedures is inflated to the detriment of validity? These are difficult questions to address, as a thorough answer would require meta-analysis of a large number of published studies which employ diverse methods and multiple (theoretically) predicted outcomes in statistical models. Unfortunately, the field of MFT and MIME research is not developed enough yet to warrant such an investigation. Nevertheless, in our research we deliberately chose to develop MoNA as an online platform which standardizes and manages both coder training and the coding task itself. This decision made truly independent trainings and codings possible, which we believe is a major reason for the substantially lower reliabilities we have observed in our analyses compared to previous studies.

It is possible, of course, that compared to previous research, all 1,028 coders who were involved in our six content analyses and completed the task, from small, highly-trained, highly-involved groups of 3 coders, to a large, less-trained, less-involved group of 557 coders from the general U.S. population, were just poorly trained and produced by and large random codings. This is unlikely, however, for two main reasons: (1) If a generally inferior coding procedure is indeed responsible for the findings, then we should not be able to find coder groups with systematically higher reliabilities in our reliability prediction analyses. We did find, however, groups of coders with high inter-coder reliabilities when coders align in moral intuition salience and other characteristics. (2) Our research team has developed and tested numerous iterations of coder trainings and coding procedures with care and over a period of three years. In addition, at least five of the authors (RW, MM, RH, LH, and RT) have extensive experience in MFT and MIME related research and are well versed in the development of coder trainings. There is no plausible explanation why a generally inferior coding procedure has found its way into all six content analyses presented here but not into previous content analyses.

Possibly our most important finding for future MFT and MIME research, and perhaps for the extraction of latent content in general, is that a simplified, intuitive coding procedure using a large heterogeneous crowd of mildly trained coders leads to acceptable inter-coder agreement. Considering the increasing availability of crowdsourcing platforms such as Mechanical Turk (<https://www.mturk.com>), Prolific Academic (<https://www.prolific.ac>), and CrowdFlower (<https://www.crowdflower.com>), as well as intensifying research that studies the weighting and selection of high-quality coders in crowdsourcing tasks (e.g., Raykar & Yu, 2012; Sheng et al., 2008), we suggest a crowd truth approach in combination with computational methods for text preparation and selection, entity extraction, and reliability tests as presented in this article as a general and promising solution for future moral intuition extractions from text. This conclusion confirms and specifies recent findings in studies testing the usability of crowdsourcing for coding latent constructs in political texts (Benoit et al., 2016; Lind et al., 2017).

## Limitations

As in all research, the studies reported in this manuscript are not without their limitations. A major limitation of our studies is that our content analyses presented here only include non-fictional, news narratives as text material. In line with early theorizing within moral foundation theory (Graham et al., 2009), we believe that analyzing news narratives with respect to moral information can be considered as “worst-case-scenario”, because news narratives’ primary goal is to deliver unbiased information rather than produce a dramatic narrative structure. It is plausible that moral intuition extraction procedures that use fictional, dramatic narratives, which are more likely to maximize the

prevalence of moral information and moral conflict (e.g., Booker, 2004), do not suffer from the same limitations. In this sense, we believe that our findings represent a lower baseline of reliabilities and coder agreement for moral foundation extraction.

## Outlook

The freely available, open source MoNA platform (<http://mnl.ucsb.edu>) which manages text selection, coder training, reliability tests, and moral intuition extraction based on a highlighting task, can be easily combined with crowdsourcing platforms. Furthermore, this new procedure has the additional benefit that the text highlight data can be processed with natural language processing algorithms and with the goal of creating new, crowd-sourced Moral Foundation Dictionaries (MFDs; i.e., extensions of Graham & Haidt, 2012) which are “less subject to the bias and oversight from dictionaries made by a small number of experts” (Schwartz & Unger, 2015, p. 81), but are instead based on methodical content analyses, are empirically tested, and can subsequently be used to improve the analysis of moral information in large amounts of text data (e.g. global online news, see <http://gdeltproject.org>). A promising approach for extending MFDs is to identify words and phrases that are highly discriminative of particular MFT categories based on our text highlights in content analyses 5 and 6. For instance, pointwise mutual information (PMI) is a generalized measure of correlation that is often used in natural language processing applications to identify word collocations and automatically extract dictionaries from textual documents (Manning & Schuetz, 1999). The creation of the extended MFD-E is currently underway.

## Notes

1. The terms “moral foundations” and “moral intuitions” are sometimes used interchangeably in the literature. We use the term “foundations” to refer to the conceptual dimensions of MFT, i.e., the *universal dimensions* that categorize moral judgments. We use the term “intuitions” to refer to the *experiential, subjective processes* of moral judgment.
2. We only used multi-foundation coders in coder group six.


## Acknowledgments

This research would not have been possible with the help of our research assistants at the University of California Santa Barbara (Mitch Grimes, Brandon Mims, Rachel Glikes, Douglas Keith, Sierra Scott, Cathy Chen, and Dane Asto) and at Michigan State University (Brandon Walling, Kathryn Hollemans, Erica Lydey, Maryssa Mitchell, Anna Young, Erika Lentz, Ellen Grimes, Kristin Barndt, Tyler Lawrence, Allison Aigner, Riley Hoffman, Elizabeth Paulson, Sierra Richards, Savannah Jenuwine, Pooja Dandamundi, Will Marchetti, and Abigail Johnson).

## Funding

Contract grant sponsors: US Army Research Laboratory (to R.W.), contract grant number: W911NF-15-2-0115.

## ORCID

René Weber  <http://orcid.org/0000-0002-8247-7341>  
 Richard Huskey  <http://orcid.org/0000-0002-4559-2439>  
 Lindsay Hahn  <http://orcid.org/0000-0002-0039-9782>

## References

- Ammerman, N. T. (1991). North American fundamentalism. In M. E. Marty, & R. S. Appleby (Eds.), *Fundamentalisms observed* (pp. 1–65). Chicago, IL: University of Chicago.
- Armstrong, K. (2000). *The battle for God: Fundamentalism in Judaism, Christianity, and Islam*. New York, NY: Knopf.

- Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1), 15–24. doi:10.1609/aimag.v36i1.2564
- Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, 110(2), 278–295. doi:10.1017/S0003055416000058
- Berelson, B. R. (1952). *Content analysis in communication research*. New York: Free Press.
- Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with python*. Sebastopol, CA: O'Reilly.
- Booker, C. (2004). *The seven basic plots*. New York: Continuum.
- Bowman, N. D., Lewis, R. J., & Tamborini, R. (2014). The morality of May 2, 2011: A content analysis of US headlines regarding the death of Osama bin Laden. *Mass Communication and Society*, 17, 639–664. doi:10.1080/15205436.2013.822518
- Burscher, B., Odijk, D., Vliegthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, 8(3), 190–206. doi:10.1080/19312458.2014.937527
- Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, 47(4), 1178–1198. doi:10.3758/s13428-014-0551-2
- 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. doi:10.1017/S0022381613000492
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. doi:10.1037/h0026256
- Delli Carpini, M. X., & Keeter, S. (1993). Measuring political knowledge: Putting first things first. *American Journal of Political Science*, 37(4), 1179–1206. doi:10.2307/2111549
- Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science*, 24(1), 56–62. doi:10.1177/0956797612449177
- Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence. *Personality and Social Psychology Bulletin*, 41(12), 1665–1681.
- Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363–370.
- Goranson, A., Ritter, R. S., Waytz, A., Norton, M. I., & Gray, K. (2017). Dying is unexpectedly positive. *Psychological Science*, 1–12. doi:10.1177/0956797617701186
- Graham, J., & Haidt, J. (2012). The moral foundations dictionary. Retrieved from <http://moralfoundations.org/othermaterials>
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., & Ditto, P. H. (2012). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 55–130.
- 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. doi:10.1037/a0015141
- Grizzard, M., Shaw, A. Z., Dolan, E. A., Anderson, K. A., Hahn, L., & Prabhu, S. (2016). Does repeated exposure to popular media strengthen moral intuitions?: Exploratory evidence regarding consistent and conflicted moral content. *Media Psychology*, 1–27. doi:10.1080/15213269.2016.1227266
- Hahn, L., Tamborini, R., Prabhu, S., Klebig, B., Grall, C., & Pei, D. (2017). The importance of altruistic versus egoistic motivations: A content analysis of conflicted motivations in children's television programming. *Communication Reports*, 1–13. doi:10.1080/08934215.2016.1251602
- Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1), 98–116. doi:10.1007/s11211-007-0034-z
- Haidt, J., & Joseph, C. (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 3, pp. 367–391). New York: Oxford University Press.
- Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, 345, 1340–1343. doi:10.1126/science.1251560
- Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.
- Hsueh, P., Melville, P., & Sindhwani, V. (2009). *Data quality from crowdsourcing: A study of annotation selection criteria*. Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing (pp. 27–35). Boulder, CO: Association for Computational Linguistics.
- Kaplan, A. (1964). *The conduct of inquiry. Methodology for behavioral science*. San Francisco, CA: Chandler Publishing Company.
- Kim, J. S., Greene, M. J., Zlateski, A., Lee, K., Richardson, M., Turaga, S. C., ... Seung, H. S. (2014). Space-time wiring specificity supports direction selectivity in the retina. *Nature*, 509(7500), 331–336. doi:10.1038/nature13240
- Krippendorff, K. (2013). *Content analysis. An introduction to its methodology*. Thousand Oaks, CA: Sage.
- Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... Jebara, T. (2009). Life in the network: The coming age of computational social science. *Science*, 323(5915), 721–723. doi:10.1126/science.1167742

- Leetaru, K., & Schrodt, P. A. (2013). GDELT: Global Data on Events, Location and Tone, 1979-2012. Paper presented at the International Studies Association Meeting, San Francisco, CA, April 2013. Retrieved from <http://data.gdelproject.org/documentation/ISA.2013.GDELT.pdf>
- Leidner, B., & Castano, E. (2012). Morality shifting in the context of intergroup violence. *European Journal of Social Psychology*, 42(1), 82–91. doi:10.1002/ejsp.v42.1
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of Massive Datasets*. Cambridge, UK: Cambridge University Press.
- Lewis, R. J., Grizzard, M., Mangus, J. M., Rashidian, P., & Weber, R. (2016). Moral clarity in narratives elicits greater cooperation than moral ambiguity. *Media Psychology*, 1–24. doi:10.1080/15213269.2016.1212714
- Lewis, R. J., & Mitchell, N. (2014). Egoism versus altruism in television content for young audiences. *Mass Communication and Society*, 17, 597–613. doi:10.1080/15205436.2013.816747
- Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd. Assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods and Measures*, 11(3), 191–209. doi:10.1080/19312458.2017.1317338
- Manning, C. D., & Schuetze, H. (1999). Collocations. In *Foundations of statistical natural language processing* (pp. 178–183). Cambridge, MA: MIT Press.
- Mastro, D., Enriquez, M., Bowman, N. D., Prabhu, S., & Tamborini, R. (2012). Morality subcultures and media production: How Hollywood minds the morals of its audience. In R. Tamborini (Ed.), *Media and the moral mind* (pp. 75–92). London, UK: Routledge.
- McAdams, D. P., Albaugh, M., Farber, E., Daniels, J., Logan, R. L., & Olson, B. (2008). Family metaphors and moral intuitions: How conservatives and liberals narrate their lives. *Journal of Personality and Social Psychology*, 95(4), 978–990. doi:10.1037/a0012650
- Moscovici, S. (1985). Innovation and minority influence. In S. Moscovici, G. Mugny, & E. Van Avermaet (Eds.), *Perspectives on minority influence* (pp. 9–51). Cambridge, UK: Cambridge University Press.
- Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PloS One*, 11(6), 1–19. doi:10.1371/journal.pone.0157732
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137. doi:10.1108/eb046814
- Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, 27(3), 258–284. doi:10.1080/00909889909365539
- Raykar, V. C., & Yu, S. (2012). Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13, 491–518.
- Riffe, D., Lacy, S., & Fico, F. (2005). *Analyzing media messages: Using quantitative content analysis in research*. Mahwah, NJ: Lawrence Erlbaum.
- Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social Science Computer Review*, 32(2), 132–144. doi:10.1177/0894439313506837
- Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 78–94. doi:10.1177/0002716215569197
- Sheng, V. S., Provost, F., & Ipeirotis, P. H. (2008). *Get another label? Improving data quality and data mining using multiple, noisy labels*. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 614–622). New York, NY: Association for Computing Machinery.
- Smith, K. B., Oxley, D. R., Hibbing, M. V., Alford, J. R., & Hibbing, J. R. (2011). Linking genetics and political attitudes: Reconceptualizing political ideology. *Political Psychology*, 32(3), 369–397. doi:10.1111/pops.2011.32.issue-3
- Tamborini, R. (2013). Model of intuitive morality and exemplars. In R. Tamborini (Ed.), *Media and the moral mind* (pp. 43–74). London, UK: Routledge.
- Tamborini, R., Hahn, L., Prabhu, S., Klebig, B., & Grall, C. (2017). The representation of altruistic and egoistic motivations in children's television programming. *Communication Research Reports*, 34(1), 58–67. doi:10.1080/08824096.2016.1227312
- Tamborini, R., Lewis, R. J., Prabhu, S., Grizzard, M., Hahn, L., & Wang, L. (2016). Media's influence on the accessibility of altruistic and egoistic motivations. *Communication Research Reports*, 33(3), 177–187. doi:10.1080/08824096.2016.1186627
- Tamborini, R., Prabhu, S., Lewis, R. L., Grizzard, M., & Eden, A. (2016). The influence of media exposure on the accessibility of moral intuitions. *Journal of Media Psychology*, 1–12. doi:10.1027/1864-1105/a000183
- Tamborini, R., Weber, R., Eden, A., Bowman, N. D., & Grizzard, M. (2010). Repeated exposure to daytime soap opera and shifts in moral judgment toward social convention. *Journal of Broadcasting & Electronic Media*, 54(4), 621–640. doi:10.1080/08838151.2010.519806
- Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17(1), 65–83. doi:10.1177/026553220001700103
- Vlieger, E., & Leydesdorff, L. (2011). Content analysis and the measurement of meaning: The visualization of frames in collections of messages. *The Public Journal of Semiotics*, 3(1), 28–50.