Supplemental Materials

Method Validation

Prior to the primary Study 1 analysis, we conducted a preliminary study testing the validity of using DDR to measure the presence of moral values in tweets (Study 1a). Garten et al. (2017) conducted a similar validation using 3000 tweets annotated by 3 coders in overlapping sets of 2000 and we extend their validation using a set of 6000 tweets fully annotated by 3 coders. As in Garten et al. (2017), we compare the *moral loadings* calculated via DDR to the coders' annotations. We then use DDR to calculate moral loadings for the entire corpus and we use these to investigate differences in moral sentiment between tweets containing donation sentiment and tweets not containing donation sentiment (Study 1b).

Method. We Garten et al. (2017) extended this method validation by testing DDR on a larger collection of annotated tweets. Due to the relatively low base-rates of expressions of moral values, tweets were selected for annotation by taking the n tweets with the highest loading for each moral domain. The same approach was used to select non-moral tweets by sampling from the set of tweets with loadings between 0 and one standard deviation from 0 on all moral domains. In this case $n = 545$ (Total $N = 5{,}995$), which yielded a set that includes those analyzed in Garten et al. (2017) as well as an additional 2,995 tweets.

However, note that while 545 tweets within each dimension were preselected, duplicates were not controlled. Thus, a tweet selected to represent care could also be selected to represent fairness if the loadings for care and fairness were among the 545 for those dimensions. After selecting the full set of 5,995 tweets, duplicates were removed, yielding a final set of 4,939 unique tweets representing the 545 highest loading tweets in each of the 11 target dimensions. Three undergraduate research assistants were then trained by the researcher to detect expressions of moral values via repeated training sessions and reference to a training guide written by the researcher (See Appendix B). Specifically, annotators were instructed to label the moral content of each tweet, allowing both overlapping dimensions (e.g. care and fairness) and the absence of

moral sentiment, which was coded as Non-moral. Inter-coder reliability was then assessed using prevalence-and-bias adjusted Kappa (PABAK) (Sim & Wright, 2005b), due to the low base-rates of moral values expressions. For each set of annotated tweets, DDR loadings were then used to predict the binary presence of each moral foundation using univariate logistic regression with 10 repetitions of 10-fold cross-validation. In these models, a given coders annotations for a specific moral domain were regressed on the DDR loadings for that domain.

Results. Across the 11 categories, there was moderate (PABAK $>$ 0.40 to 0.60) to near perfect (PABAK $>$ 0.80) agreement, with mean agreement of $PABAK = 0.74$, indicating that, on average, there was substantial agreement regarding the presence/absence moral sentiment among the coders (See Figure 5 in the Supplementary Materials for a graphical representation of PABAK for each coder pair's agreement within each domain. These results indicate that expressions of moral sentiment on twitter are sufficiently robust for reliable inter-annotator signal detection.

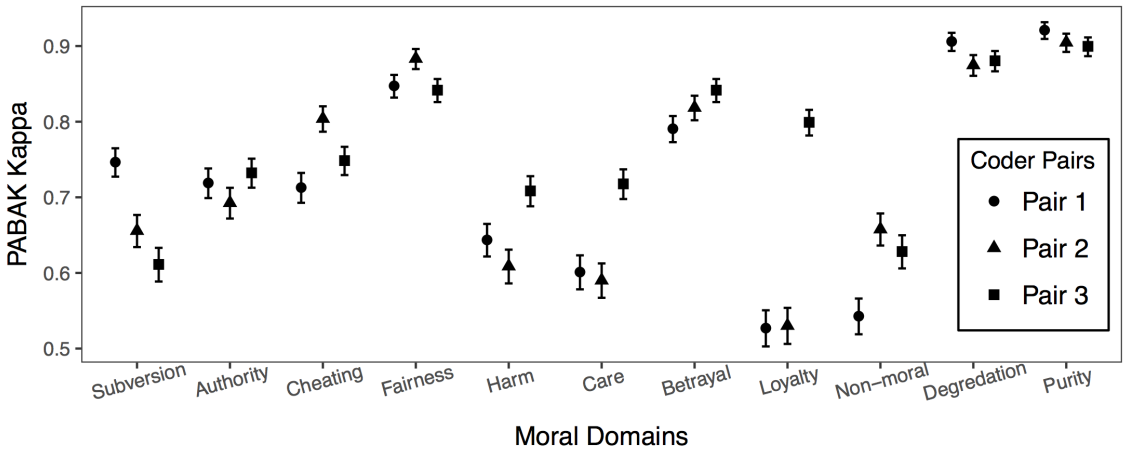

Figure 5. PABAK for annotated tweets

Further, classification results from the 10 logistic regression models indicate that the moral loadings estimated via DDR are reliably associated with their target domains (see Table 5). For all domains sensitivity (the proportion of true positives correctly identified as such) was $>$ 0.70 and specificity (the proportion of true negatives correctly identified as such) was greater than 0.67. However, the precision (ratio of identified true positives to the total number of identified positives) was markedly lower, ($M = 0.23$,

$SD = 0.02$), indicating that while the models were largely able to identify 70% of the tweets containing or not containing a given moral domain, the model also somewhat indiscriminately labeled tweets as containing moral sentiment. However, it is worth noting that model precision appears to be at least partially a function of the number available training cases. Indeed, precision is nearly perfectly correlated with the average number of cases available in a given domain ($r(df = 8) = 0.99, p < 0.5$), such that precision was high in domains for which there were more cases.

Table 5

DDR classification performance. M N coded indicates the number of tweets averaged across all three coders that were assigned a given class.

Class	Specificity	Sensitivity	F1	Precision	N_coded
Authority	0.71	0.77	0.36	0.25	559
Betrayal	0.7	0.76	0.23	0.13	288
Care	0.77	0.77	0.57	0.46	1007
Cheating	0.7	0.79	0.39	0.26	599
Degredation	0.7	0.7	0.14	0.08	177
Fairness	0.7	0.73	0.22	0.13	280
Harm	0.71	0.73	0.51	0.39	988
Loyalty	0.72	0.72	0.38	0.29	683
Purity	0.73	0.77	0.14	0.08	140
Subversion	0.67	0.74	0.35	0.25	614

These results are largely congruent with previous research (Garten et al., 2017) and indicate that DDR is able to reliably detect moral sentiment in twitter messages. Clearly, the detected signal does not correspond perfectly, or even close to perfectly, to human annotations. However, even a noisy measurement should be sufficient for examining associations between donation sentiment and moral sentiment, given effects

of sufficient magnitude and reliability. Accordingly, it is essential that the potential unreliability of our measurement of moral sentiment is counterbalanced by subsequent controlled experimentation.

Study 2

Diffusion Motivation. To test the hypotheses that care and loyalty rhetoric increases perceived diffusion motivation, t-tests were conducted, comparing the distribution of diffusion motivation reports from the care and loyalty conditions to the non-moral condition. The observed effects for diffusion motivation were comparable to those observed for donation motivation (See Table 6).

Table 6

Study 2: T-tests for donation and diffusion motivation

Test	$\hat{\mu}_{moral}$	$\hat{\mu}_{non-moral}$	Δ	Δ%95CI	d	t	p
Care Retweet	4.19	3.43	0.76	[0.4, 1.11]	0.42	4.22	« 0.001
Loyalty Retweet	4.25	3.43	0.82	[0.46, 1.19]	0.50	4.42	« 0.001

Donation and diffusion motivation controlling for 'sense'

In addition to asking participants to rate the degree to which solicitation tweets would motivate them to donate and retweet, participants were asked to report 'how much sense' each tweet made. To estimate (1) the degree to which 'sense' accounted for variation in donation and retweet motivation and (2) to determine whether moral framing effects had effects above and beyond sense, four regression models were estimated. Specifically, donation or diffusion motivation was set as the dependent variable, condition (care vs non-moral or loyalty vs non-moral) as the independent variable, and fluency as the covariate. In each model, the non-moral condition was set

as the intercept, thus the estimated effects of condition indicate the estimated differences between conditions means.

Results of the four regressions indicated that evaluations of the degree to which tweets made sense was associated both with perceived donation motivation and diffusion motivation. However, except for the effect of care rhetoric on donation motivation (Model 1), the between condition effects remained significant and diminished only moderately (See Table 7).

Table 7

Study 2: Regressions controlling for sense.

DV	*Model*	*IV*	*b*	*SE*	*t*	*p*
Donation	1	Care	0.26	0.16	1.57	0.12
		Sense	0.55	0.08	6.69	< 0.001
	2	Loyalty	0.36	0.16	2.26	0.02
		Sense	0.64	0.08	7.93	< 0.001
Diffusion	3	Care	0.42	0.18	2.29	0.02
		Sense	0.52	0.09	5.68	< 0.001
	4	Loyalty	0.47	0.18	2.60	0.01
		Sense	0.61	0.09	6.72	< 0.001

Further, both care and loyalty tweets received higher sense ratings ($Ms = 5.87$ and 5.78, $SDs = 1.03$ and 1.09) compared to the non-moral tweets, ($M = 5.13$, $SD = 1.09$), $t(310.51) = 6.20$ and $t(310.77) = 5.29$ $ps < 0.001$. That is, (1) participants thought the solicitations with moral frames simply made more sense and (2) the more sense a solicitation was perceived to make, the higher a donation or diffusion motivation rating it received, on average. However, except for the effect of care on donation motivation, which became non-significant at $p = 0.12$, the moral frame effects remained significant after controlling for the effect of sense.

One possibility of what 'sense' might represent in this study is is fluency (Reber & Schwarz, 1999; Reber et al., 1998). That is, it may be that popular cultural prototypes of donation solicitations using care and loyalty frames are more cognitively accessible and thus may be processed with a sense of cognitive ease. In other words, frames that use care and loyalty rhetoric might just seem *right*. Such an effect would not be a direct effect of moral framing, per se, but rather an effect of the interaction between cognitive processing and the process cultural sedimentation that led to certain moral frames being more prototypical than others. However, these analyses also indicate that moral framing has effects above and beyond those of sense or fluency. Accordingly, it may be that the effects of moral frames work through multiple pathways: one, through fluency effects induced by matching popular solicitation prototypes and two through moral concern activation.

Donation and diffusion motivation controlling for individual connections to Hurricane Sandy

To investigate whether there is meaningful variation in the effects of moral frames between groups of people who were affected by Hurricane Sandy, we estimated regression models that included an interaction between condition and either whether participants knew someone affected by the storm or whether, during the storm, they lived in an area affected by the storm.

Twenty-three percent of participants reported knowing at least one person who was affected by Hurricane Sandy and 13% reported living in an area affected by the storm. Regression estimates of the interaction effect of condition with knowing someone or living in an area affected by the storm yielded no evidence for an interaction, all *p*s >= 0.20 (See Tables 8 and 9. However, it should be noted that the estimates in these models were relatively imprecise due to the low number of participants who were

associated with Hurricane Sandy.

Table 8

Interaction effects of association with Hurricane Sandy and condition on donation motivation

Model	*IV*	*b*	*SE*	*t*	*p*
1	Care	−0.75	0.47	−1.59	0.11
	Area[a]	0.19	0.31	0.60	0.55
	Care∗Area	0.13	0.50	0.25	0.80
2	Loyalty	-1.03	0.48	−2.14	0.03
	Area	−0.02	0.32	−0.06	0.95
	Loyalty∗Area	0.33	0.52	0.64	0.52
3	Care	−0.23	0.34	−0.69	0.49
	Know[b]	0.55	0.28	1.96	0.05
	Care∗Know	−0.49	0.39	−1.27	0.20
4	Loyalty	−0.57	0.35	−1.62	0.11
	Know	0.26	0.29	0.91	0.36
	Loyalty∗Know	−0.21	0.40	−0.53	0.60

[a] Area is a binary indicator representing whether a participant lived somewhere affected by the storm (1) or not (0).

[b] Know is a binary indicator representing whether a participant knew someone affected by the storm (1) or not (0).

Study 3

To estimate the degree to which the effects of care and loyalty, compared to the other conditions, were driven by sense, we estimated a follow-up ANCOVA calculating

the same comparisons while controlling for ratings of sense. Finally, to maintain consistency with Study 2, we also estimated the same tests using motivation to retweet as the dependent variable.

Table 9

Interaction effects of association with Hurricane Sandy and condition on diffusion motivation

Model	*IV*	*b*	*SE*	*t*	*p*
5	Care	−1.29	0.51	−2.52	0.01
	Area[a]	−0.34	0.34	−1.01	0.31
	Care∗Area	0.62	0.55	1.13	0.26
6	Loyalty	−1.27	0.53	−2.39	0.02
	Area	−0.24	0.35	−0.68	0.50
	Loyalty∗Area	0.51	0.57	0.91	0.37
7	Care	−0.51	0.37	−1.38	0.17
	Know[b]	0.32	0.31	1.03	0.30
	Care∗Know	−0.32	0.43	−0.74	0.46
8	Loyalty	−0.76	0.39	−1.97	0.05
	Know	0.08	0.32	0.24	0.81
	Loyalty∗Know	−0.08	0.44	−0.18	0.86

[a] Area is a binary indicator representing whether a participant lived somewhere affected by the storm (1) or not (0).

[b] Know is a binary indicator representing whether a participant knew someone affected by the storm (1) or not (0).

As in Study 2, results from the ANCOVA model testing our target hypotheses while controlling for sense indicated that the degree to which participants thought a

solicitation made sense was associated with the degree to which they thought it would motivate them to donate, $F(1, 811) = 301.1, p < 0.001$. However, substantial between condition variance remained, indicating again that moral framing effects perceived donation motivation above and beyond the degree to which a solicitation makes sense, $F(6, 811) = 1071.7, p < 0.001$. Further, while post hoc tests found that after accounting for the effect of sense, the difference between the effects of care ($M = 2.07$, $SE = 0.16$) and cheating ($M = 1.82$, $SE = 0.12$) and loyalty ($M = 1.91$, $SE = 0.15$) and cheating were not significant, the differences between care and loyalty and the fairness ($M = 0.22$, $SE = 0.16$) and non-moral ($M = 0.57$, $SE = 0.17$) remained robust (See Table 10).

Table 10

Comparison	Δ	SE	t	p
care vs. cheating	0.25	0.15	1.62	0.47
care vs. fairness	1.85	0.14	12.81	0
care vs. harm	-0.08	0.14	-0.56	0.99
care vs. nonmoral	1.50	0.15	10.26	0
loyalty vs. cheating	0.08	0.15	0.57	0.99
loyalty vs. fairness	1.68	0.15	11.60	0
loyalty vs. harm	-0.24	0.14	-1.71	0.41
loyalty vs. nonmoral	1.34	0.14	9.27	0

In contrast, results of the analyses of retweet motivation diverged somewhat from those observed for tweet motivation. Specifically, the effects of care ($M = 4.03$, SE=0.12) and loyalty ($M = 4.10$, SE=12) were significantly higher than those of cheating ($M = 2.42$, SE=0.12) and harm ($M = 2.72$, SE=0.12); but they were not significantly different from fairness ($M = 3.66$, SE=0.12) or non-moral ($M = 4.17$, SE=0.12; see Table 11 for paired tests). Further, when sense was added to the model,

all target comparisons became significant in the expected direction except for those comparing care and loyalty with non-moral.

Table 11

Study 3 Retweet motivation analyses

ANOVA	care vs. cheating	1.61	0.17	9.67	< 0.001
	care vs. fairness	0.37	0.17	2.21	0.16
	care vs. harm	1.31	0.17	7.86	< 0.001
	care vs. nonmoral	-0.13	0.17	-0.8	0.94
	loyalty vs. cheating	1.68	0.17	10.06	< 0.001
	loyalty vs. fairness	0.43	0.17	2.61	0.06
	loyalty vs. harm	1.38	0.17	8.25	< 0.001
	loyalty vs. nonmoral	-0.07	0.17	-0.41	1
ANCOVA	care vs. cheating	0.69	0.15	4.6	< 0.001
	care vs. fairness	0.59	0.14	4.15	< 0.001
	care vs. harm	1.18	0.14	8.33	< 0.001
	care vs. nonmoral	-0.57	0.14	-3.96	< 0.001
	loyalty vs. cheating	0.94	0.15	6.37	< 0.001
	loyalty vs. fairness	0.84	0.14	5.82	< 0.001
	loyalty vs. harm	1.43	0.14	10.07	< 0.001
	loyalty vs. nonmoral	-0.33	0.14	-2.28	0.14

Overall, these analyses provide further evidence that the degree to which a solicitation is seen as making sense – which may be a proxy for fluency – is associated with its perceived donation efficacy. However, as in study 2, results indicated that the effects of care and moral frames cannot reduced to a function of sense, as significant differences in effects remained after controlling for sense.

Further, while the current research is focused primarily on charitable donation,

the current study suggests that moral framing may play a role in solicitation diffusion. However, based on the results of this study and their contradiction with those of study 2, it is not necessarily clear what the association between care and loyalty framing and retweeting is.

Study 4 Plot

As noted in the main text, none of the observed effects fit our hypotheses. The full distribution of the data is shown in Figure 6

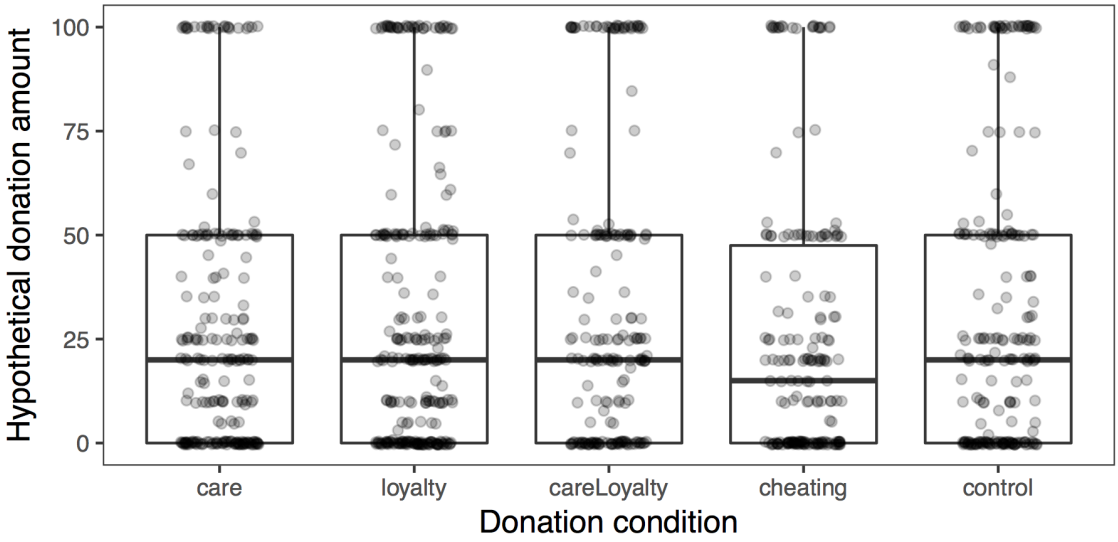

Figure 6. Hypothetical donations by moral frame condition

Reference

Sim, J., & **Wright, C. C.** (2005b). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*. DOI: https://doi.org/10.1093/ptj/85.3.257