Temporal variability in moral value judgement

Alexandra Surdina (a.surdina@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry CV4 7AL, United Kingdom

Adam Sanborn (a.n.sanborn@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry CV4 7AL, United Kingdom

Abstract

Moral judgments are known to change in response to changes in external conditions. But how variable are moral judgments over time in the absence of environmental variation? The moral domain has been described in terms of five moral foundations, categories that appear to capture moral judgment across cultures. We examined the temporal consistency of repeated responses to the moral foundations questionnaire over short time periods, fitted a set of mixed effects models to the data and compared them. We found correlations between changes in participant responses for different foundations over time, suggesting a structure with at least two underlying stochastic processes: one for moral judgments involving harm and fairness, and another for moral judgments related to loyalty, authority, and purity.

Keywords: morality, moral foundations theory, consistency, variability

Introduction

Morality is a vital part of who we are. A person's moral beliefs are tied into their identity (Aquino & Reed II, 2002; Aquino et al., 2009) – humans believe that if their moral values changed, they would change (Heiphetz et al., 2016). Are people's intuitions about this correct? Are our moral values consistent over time?

Since moral beliefs tend to be associated with a person's sense of identity, we should expect people's underlying moral values to largely endure over short time periods. Yet, there have been many recent explorations of moral inconsistency. These have included manipulations of two types – manipulations of response timing, or manipulations by exposure to new information or decisions. In terms of timing, we now know that time-limited decisions appear to be more altruistic (Rand et al., 2012) and that choices can be influenced by forcing decisions at a specific point in time (Pärnamets et al., 2015), indicating that the actual decision outcome is time-sensitive. Regarding information or decisions, dishonest behaviour increases future dishonesty (Garrett et al., 2016; Engelmann & Fehr, 2016). A morally good action makes a subsequent morally bad action more appealing and vice versa, effects known as moral cleansing and moral licensing (Merritt et al., 2010; Sachdeva et al., 2009). Exposure to a moral dilemma leads to belief revision in moral decisions that persists for multiple hours (Horne et al., 2015).

The fact that changes in external circumstances can influence the outcomes of moral decisions is hardly surprising assuming morality evolved as an adaptive strategy (Machery & Mallon, 2010). Likewise, viewing moral judgment as a decision process, we would expect the effects of changed response timing on general decision-making (McClelland, 1979; Usher & McClelland, 2001) to transfer into the moral domain. But in the absense of such manipulations, are our moral judgments fundamentally noisy? Outside of the moral domain, there is evidence in decision making research that people's decisions vary stochastically even in cases where external conditions remain constant (Mosteller & Nogee, 1951). We are interested in exploring whether there is a corresponding moral variability beyond the actual decision process: are our moral values different from moment to moment, even in the absence of new information or manipulations of response timing?

Moral Foundations Theory (MFT) provides a way to look at this. It is based on a dominant model of morality, the social intuitionist model, according to which moral choices are made primarily intuitively and then justified post hoc (Haidt, 2001). MFT maps out the moral domain in terms of six fundamental hidden parameters that appear to capture an individual's moral judgment (Graham et al., 2009), enabling us to distinguish between conservative and liberal political profiles on the basis of an agent's foundation weights. This idea that there are foundational categories that guide intuitive moral judgement has the potential to explain people's tendency to disagree on moral issues, and predict future moral judgement based on the individual scores. If we can find a systematic structure in the stochastic changes of different foundation scores beyond merely a layer of noise, this would point towards moral variability, rather than just motor variability or variability in how the response scale is used.

In line with the aforementioned results indicating temporal consistency, moral foundation scores appear stable over longer time periods; Graham et al. (2011) tested participants again after approximately a month and found that their moral foundation scores exhibited test-retest reliability. Yet, effects such as moral licensing and moral cleansing – where the outcome of an individual's moral decision influences subsequent moral de-

cisions, even decisions made by others in their ingroup (Kouchaki, 2011), over the course of single experimental sessions and thus shorter timescales – suggest the possibility of an interaction between moral foundations. Moreover, the list of known moral foundations is likely incomplete – a view shared by moral foundations theorists (Haidt & Joseph, 2011).

Viewing moral decisions as a sampling process from a distribution that represents an agent's moral values, we can use the framework provided by MFT to investigate hidden parameters which predict an individual's moral variability. Conversely, observing within-subject variability over time can help us understand to which extent individual moral variability reflects between-individual variability that has been used to support the existence of MFT (Graham et al., 2011). Are we all sometimes a little bit more conservative and sometimes a little bit more liberal in our moral judgments and values?

In this paper, we aim to discuss the extent to which randomness plays a role in moral judgment over time by collecting responses to the moral foundations questionnaire delivered repeatedly. We subsequently fit a set of models to the data and compare them. If the variability we observe stems merely from randomness in the decision process, we expect variation in individual responses to be explained by a single noise-generating process. We find evidence for at least two separate stochastic processes associated with different sets of moral foundations, indicating the existence of inherent variability in moral values.

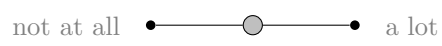
Method

Participants

The participant pool consisted of 80 psychology undergraduate students (mean age 19 years, 90% female). 14 participants were excluded from the analysis due to wrong responses on the two 'catch' trials, as done by Graham et al. (2011).

Materials

The original moral foundations questionnaire (MFQ30) asks participants to respond using a 1–6 scale; to enhance precision and avoid subjects simply recalling previous answers, the participants in our task had to use a slider bar to indicate their responses instead:



In addition, our version of the questionnaire contained four further questions (see Figure 1). Those were chosen so as not to correspond in any obvious way to the five foundations measured in the MFQ30, nor to the recent addition of the liberty foundation (Graham et al., 2012; Haidt, 2012). We added these questions because we wanted the same number of presumably neutral trials as the number of foundation-related questions – the MFQ30 includes six question for each foundation but only two neutral 'catch' items.

When you decide whether something is right or wrong, to what extent is the following consideration relevant to your thinking?

Please read the following sentence and indicate your agreement or disagreement:

HARM:

1. Whether or not someone suffered emotionally
2. Whether or not someone cared for someone weak or vulnerable
3. Whether or not someone was cruel
4. Compassion for those who are suffering is the most crucial virtue.
5. One of the worst things a person could do is hurt a defenseless animal.
6. It can never be right to kill a human being.

FAIRNESS:

1. Whether or not some people were treated differently than others
2. Whether or not someone acted unfairly
3. Whether or not someone was denied his or her rights
4. When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.
5. Justice is the most important requirement for a society.
6. I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.

LOYALTY:

1. Whether or not someone's action showed love for his or her country
2. Whether or not someone did something to betray his or her group
3. Whether or not someone showed a lack of loyalty
4. I am proud of my country's history.
5. People should be loyal to their family members, even when they have done something wrong.
6. It is more important to be a team player than to express oneself.

AUTHORITY:

1. Whether or not someone showed a lack of respect for authority
2. Whether or not someone conformed to the traditions of society
3. Whether or not an action caused chaos or disorder
4. Respect for authority is something all children need to learn.
5. Men and women each have different roles to play in society.
6. If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.

PURITY:

1. Whether or not someone violated standards of purity and decency
2. Whether or not someone did something disgusting
3. Whether or not someone acted in a way that God would approve of
4. People should not do things that are disgusting, even if no one is harmed.
5. I would call some acts wrong on the grounds that they are unnatural.
6. Chastity is an important and valuable virtue.

NEUTRAL:

1. Whether or not someone was good at math
2. Whether or not someone told the truth (*)
3. Whether or not someone made a smart decision (*)
4. It is better to do good than to do bad.
5. If one's children live a happy life, it is better to have children than not to have children. (*)
6. Destroying beautiful things that took long to create is worse than destroying things that took less time. (*)

Figure 1: Moral foundations questionnaire. Questions added by us are marked with (*).

Procedure

The questionnaire was presented six times in randomised order, with a word search task before the last two trials. In each trial, one of the two question types was displayed (see Figure 1, left and right side, respectively), along with one of the statements for that question type. Randomisation was implemented so that each statement was shown to the participant exactly once in each block: The set of questions within each block was shuffled, and presented within the block in randomised order, so no regular pattern in the order of foundations would occur.

After four blocks, a word search task[1] was shown for 6 minutes to provide a timed break[2]: Participants had to find and mark words in a 18x18 letter square filled with a selection of words and random letters, based on the WordFind.js library (Scheidel, 2012). With the exception of the timed word search task, participants provided responses at their own pace. The experiment took approximately 20-25 minutes to complete.

Results

Since participant responses are indicated using slider bars, foundation scores change between the blocks (participants will be unable to recall the exact position of the slider for previous trials). But beyond the expected variation resulting from differences in participant's slider operation accuracy, is there a relationship between these changes in different moral foundation scores?

Means

As found by Graham et al. (2011), we anticipated and found our psychology undergraduate subject pool in the UK to remain largely at the liberal end of the U.S. political spectrum. Welch's t-test shows that the differences between the means for harm and fairness (p=.16) and for loyalty and authority (p=.44) are not significant. All other pairs of means indeed differ significantly (p<.001). In particular, the first two foundation means differ significantly from the last three, with higher subject scores for harm (M = 72.9, SD = 24.6) and fairness (M = 70.3, SD = 23.6) and lower scores for loyalty (M = 51.1, SD = 27.1), authority (M = 48.8, SD = 26.1) and purity (M = 42.1, SD = 28.8). The between-subject standard deviation is notably larger than the within-subject standard deviation (see Figure 3), supporting the MFT framework for examining between-subject differences.

A within-subjects ANOVA[3] showed a main effect of both foundation (F(5,325) = 72.67, p<.001) and block (F(5,325)=6.26, p<.001) on average slider bar values, as well as an interaction between foundation and block (F(25,1625)=1.764, p=0.011). But we are mainly interested in changes in the absence of new information, and Figure 5 suggests that the very first block in which the whole questionnaire was new to the participant qualitatively differs from the others. Excluding the first

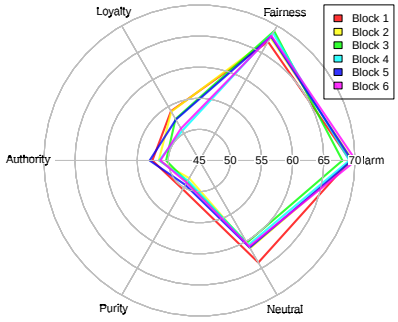


Figure 2: Spider plot of means for each foundation and block.

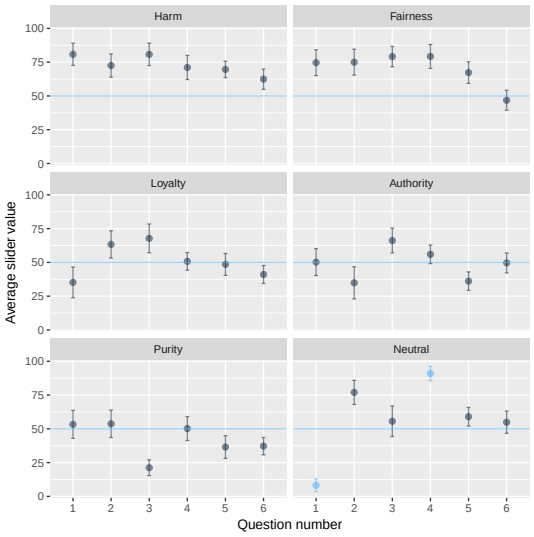


Figure 3: Average slider value for each response, and the average of within-subject standard deviations. The catch trials and the baseline level are marked in blue.

block from the analysis indeed makes the effects of block (F(4,260)=6.26, p=.47) and the interaction effect between foundation and block (F(20,1300) = 1.201, p=.24) in the ANOVA above no longer significant: While moral foundation scores differ between Block 1 and the other blocks, for the later blocks alone, this is no longer true.

Variability over time

We also expected that within-foundation variance, i.e. the variance between participant responses to the sets of questions for each respective foundation, would decrease over time: As time passes, people's certainty which choice they will make will increase as they get more familiar with the questionnaire. Moreover, we thought we might be able to observe a shift towards more extreme values for each question over time – as people become increasingly familiar with the set of questions they will

[1]We removed words such as 'excellent' from the task to reduce the likelihood that word valence in this task would influence future participant responses. This quiz block was followed by three more blocks.

[2]Due to an off-by-one error in our code, the first statement from the block after the word search task was erroneously displayed before the word search task. We excluded this error trial from the analysis.

[3]It should be noted that in this ANOVA, we are treating the block number as a factor variable rather than a numeric variable due to the non-linear relationship between block number and participant response; including the block number as a numeric variable yields qualitatively the same results.

encounter, there would be less need to for caution about new options which are more or less morally upsetting than the previous maximum or minimum, respectively.

We computed residual slider values by subtracting the within-subject mean for each foundation from the slider values for each trial. The two hypotheses above can be rephrased as: The slider residual variance for each participant and block decreases as a reflection of the increase in certainty; and the average absolute residual value increases over time as a result of the decision drifting towards the extremes.

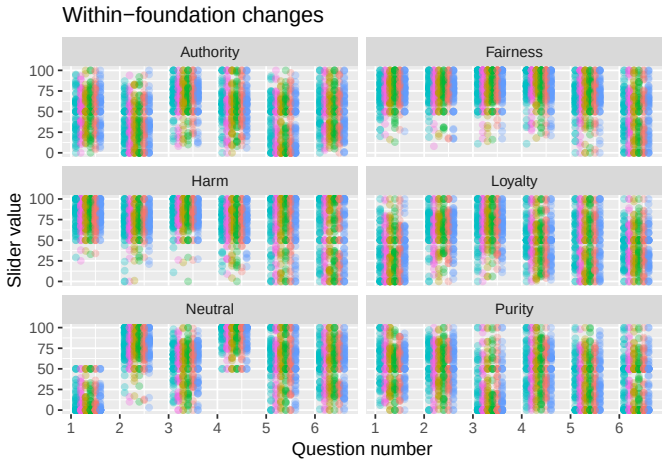


Figure 4: Changes over time, by foundation. The colours represent the different blocks.

In fact, we found no significant effect of block number on foundation variance: Again, an ANOVA only yields significant results for the variance hypothesis ($F(5,325)=18.71$, $p<.001$) and the absolute residual hypothesis ($F(5,325)=47.4$, $p<.001$) if we are taking the very first block into account – here, a slight decrease after the first block can be spotted (see Figure 5). If we are looking at only the other blocks, we do not find any significant change in the variance ($F(4,260)=1.90$, $p=.11$), nor in the absolute slider residual ($F(4,260)=1.22$, $p=.30$).

Between-foundation variability

One hypothesis is that changes in moral foundations that are opposed with respect to their representation on the political spectrum, such as harm and purity, will balance each other out – that is, they are negatively correlated (Fig. 6a). Each person may have a constant morality 'budget', and thus an increase in a moral foundation score will inevitably be accompanied by a decrease in others. This would imply that people's position on the liberal-conservative spectrum might not be fixed. Another hypothesis is that changes in opposing moral foundations are positively correlated (Fig. 6b). This would for instance be the case if people's moral profile was

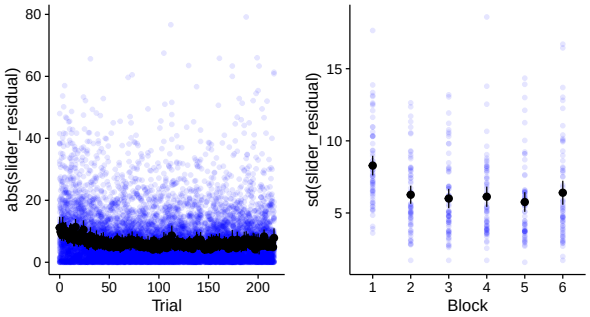


Figure 5: Absolute value and within-subject standard deviation of slider residual over time

indeed fixed, and the sampled moral foundation scores are scaled by a time-dependent factor. Alternatively, changes in different moral foundations may not be correlated at all.

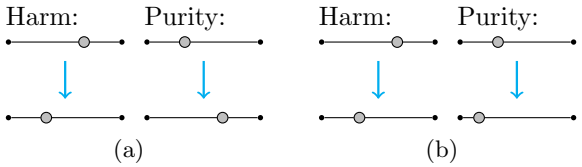


Figure 6: Relative changes in foundation scores

We did not find evidence for any of these relationships between participant scores for different foundations. On the contrary, the changes in foundation scores over time were not particularly large.

To test for interactions between changes in different foundation scores, we modelled the data using mixed effects models with a full covariance matrix and a diagonal covariance matrix, respectively. We created dummy coded variables for each foundation. Since we did not detect any notable change in the means after the first block, we now focused on the variability and removed the influence of the means entirely by modelling slider residuals: we calculated the mean slider value for each question for each participant, and subtracted it from the raw slider values. We used residuals for each question rather than for each foundation score because of the differences in responses to the different questions within each foundation (see Figure 4). Furthermore, we excluded the first block in which all information had been newly introduced from the analysis. We fitted two models to the data: First, a model including a full covariance matrix and thus allowing for interactions between the different foundations, and second, a model with a diagonal covariance matrix reflecting the assumption that sampling occurs for each foundation individually.

As a baseline model, we used a model assuming a random slider residual for each participant and block, sam-

pled from the same distribution for each foundation (random noise model). The models for the slider residual y_{ijkl} of Participant i in Block j for a question or statement l relating to Foundation k are:

$$y_{ijkl} = u_{ij} + u_{ijk} + \varepsilon_{ijkl}, \quad \text{(M1-M3)}$$

with $u_{ij} \sim \mathcal{N}(0, \sigma)$, and

$$u_{ijk} = 0 \quad \text{(M1)}$$

$$\begin{pmatrix} u_{ij1} \\ u_{ij2} \\ u_{ij3} \\ u_{ij4} \\ u_{ij5} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{bmatrix} \right) \quad \text{(M2)}$$

$$\begin{pmatrix} u_{ij1} \\ u_{ij2} \\ u_{ij3} \\ u_{ij4} \\ u_{ij5} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{22} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{33} & 0 & 0 \\ 0 & 0 & 0 & \sigma_{44} & 0 \\ 0 & 0 & 0 & 0 & \sigma_{55} \end{bmatrix} \right) \quad \text{(M3)}$$

The model M2 ($\chi^2(10) = 26.98$, $p = .003$) differs significantly from the baseline model M1. Comparing the models M1-M3 to each other suggests that M2 (BIC=71058) has a lower BIC value than M1 (BIC=70960) and M3 (BIC=70993). M2, which has a full covariance matrix, shows an interesting pattern of dependencies between the different foundation types:

Foundation	Harm	Fair	Loya	Auth
Fair	1			
Loya	-0.86	-0.86		
Auth	-0.95	-0.95	0.95	
Puri	-0.7	-0.7	0.64	0.80

Responses for harm and fairness appear to be positively correlated with each other and negatively correlated with responses for the other foundations, and vice versa. This would be less surprising if it was merely capturing a between-participant relationship between foundation scores. Note however that this model describes the slider residuals which add up to zero for each foundation and participant – yet, this result suggests that participants who drag the slider bar a bit further to the right for harm-related questions than in the last block will do a similar thing with the fairness-question slider, but the opposite with sliders on loyalty, authority and purity trials.

Is there some overlap between which property of morality harm and fairness on the one hand and loyalty, authority and purity on the other hand are measuring? Since the mean foundation scores for harm and fairness, and the scores for loyalty, authority, and purity seem similar to each other (see Figure 2), we introduced alternative models that only distinguish between these two groups instead of the individual foundations.

To find out if we could confirm the five-dimensional moral foundations structure, we fitted a set of linear mixed effects models to the data. As an alternative, we dummy-coded two foundation *types* (the 'individualising' foundations harm and fairness and the 'binding' foundations loyalty, authority, and purity (Graham et al., 2009)). Again, we fitted a full covariance model and a diagonal covariance model to the data, adding the two models below to our list of candidate models. They are describing the slider residual y_{ijml} of Participant i in Block j for a question l of Foundation type m:

$$y_{ijml} = u_{ij} + u_{ijm} + \varepsilon_{ijml}, \quad \text{(M4-M5)}$$

with

$$\begin{pmatrix} u_{ij1} \\ u_{ij2} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \right) \quad \text{(M4)}$$

$$\begin{pmatrix} u_{ij1} \\ u_{ij2} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix} \right) \quad \text{(M5)}$$

We find that out of these, the model M5 differs significantly from the baseline model ($\chi^2(2) = 27.85$, $p < .001$). Comparing M4 (BIC=70960) and M5 (BIC=70951) to the models above suggests that M5 is preferable to M2 and M4. Thus, it appears that from a model comparison perspective, the main distinction in the moral foundation framework lies in the two different foundation types rather than the individual foundations, and that at this level of description, between-foundation correlations do not play a prominent role.

Discussion

We found that people showed moral variability even in the absence of new information or time pressure. This moral variability is distinguishable from response variability because we found two random processes that were associated with different sets of moral foundations. The evidence for MFT is based on an analysis of between-individual responses to the MFQ (Graham et al., 2011), and much of this may actually be due to the within-individual variability that we have found. This within-individual variability may also be what allows timing interventions to have an effect (Pärnamets et al., 2015), and might potentially even allow to influence the outcomes of value-related decisions (such as election results).

While for our dataset a simpler two-type model was preferable to the more complex model including five moral foundations, we hesitate to draw general conclusions about the number of moral foundations due to the small size and relative cultural homogeneity of our subject pool. Yet, our brief glimpse at candidates for additional foundations suggests the possibility of a wider underlying structure of which MFT has captured but a part.

A common criticism of MFT is that the known moral foundations are unlikely to capture moral judgment in its entirety (Suhler & Churchland, 2011). We had expected

our added questions to be rated similarly irrelevant to morality as the more conservative moral foundations in our liberal subject pool. Somewhat surprisingly, the responses to our added, 'neutral' foundation appear to be less neutral overall. We chose the four additional statements in the neutral foundation because we suspected that they might turn out to be morally relevant. Figure 3 suggests that questions 2 and 5 in particular (see Figures 1 and 4) indeed resonate with our participants' values. While the act of lying may arguably be related to the purity scale, it is remarkably more morally relevant than any of the purity questions. This particularly utilitarian view on having children also appears to lie outside of the given scales.

Interesting open questions remain that reach beyond refining and expanding MFT. While we observe a range of scores for different moral foundations, we do not yet understand the actual decision process: How are different moral values integrated in a decision between options that are morally relevant for more than one moral foundation, or options that are uncertain? Which impact does moral variability have on the kinds of moral decisions we face every day?

Acknowledgments

This research was supported by a Leverhulme Trust Doctoral Scholarship within the Bridges programme at the University of Warwick.

References

Aquino, K., Freeman, D., Reed II, A., Lim, V. K., & Felps, W. (2009). Testing a social-cognitive model of moral behavior: the interactive influence of situations and moral identity centrality. *Journal of personality and social psychology*, *97*(1), 123.

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of personality and social psychology*, *83*(6), 1423.

Engelmann, J. B., & Fehr, E. (2016). The slippery slope of dishonesty. *Nature Neuroscience*, *19*(12), 1543–1544.

Garrett, N., Lazzaro, S. C., Ariely, D., & Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience*.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2012). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, Forthcoming*.

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*(5), 1029.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York: Pantheon.

Haidt, J., & Joseph, C. (2011). How moral foundations theory succeeded in building on sand: A response to Suhler and Churchland. *Journal of Cognitive Neuroscience*, *23*(9), 2117–2122.

Heiphetz, L., Strohminger, N., & Young, L. L. (2016). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive science*.

Horne, Z., Powell, D., & Hummel, J. (2015). A single counterexample leads to moral belief revision. *Cognitive science*, *39*(8), 1950–1964.

Kouchaki, M. (2011). Vicarious moral licensing: the influence of others' past moral actions on moral behavior. *Journal of personality and social psychology*, *101*(4), 702.

Machery, E., & Mallon, R. (2010). Evolution of morality. *The moral psychology handbook*, 3–46.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological review*, *86*(4), 287.

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and personality psychology compass*, *4*(5), 344–357.

Mosteller, F., & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, *59*(5), 371–404.

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, *112*(13), 4170–4175.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430.

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners the paradox of moral self-regulation. *Psychological science*, *20*(4), 523–528.

Scheidel, B. (2012). *Wordfind*. `https://github.com/bunkat/wordfind`.

Suhler, C. L., & Churchland, P. (2011). Can innate, modular "foundations" explain morality? challenges for haidt's moral foundations theory. *Journal of cognitive neuroscience*, *23*(9), 2103–2116.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550.