The relation between Honesty-Humility and moral concerns as expressed in language

Karolina A. Ścigała[1], Ioanna Arkoudi[2], Christoph Schild[3], Stefan Pfattheicher[1], Ingo Zettler[4,5]

[1]Department of Psychology and Behavioural Sciences, Aarhus University, Denmark

[2]Department of Technology, Management and Economics, Technical University of Denmark, Denmark

[3]Department of Psychology, University of Siegen, Germany

[4]Department of Psychology, University of Copenhagen, Denmark

[5]Copenhagen Center for Social Data Science, University of Copenhagen, Denmark

Please note that this manuscript has not been peer-reviewed.

Author Note

This investigation was funded by grants from the Carlsberg Foundation (CF16-0444) and the Independent Research Fund Denmark (7024-00057B) to the last author.

Correspondence concerning this manuscript should be addressed to Karolina Aleksandra Ścigała, Department of Psychology and Behavioural Sciences, Bartholins Allé 11, Building 1350, Aarhus University, 8000 Aarhus, Denmark; phone number: +45 52 68 62 48; e-mail: karolina.scigala@psy.au.dk

Abstract

Does the basic trait Honesty-Humility predict the type of moral concerns people express in language? We explore whether Honesty-Humility relates to the expression of five moral concerns in language—namely, care/harm, justice/fairness, loyalty/betrayal, authority/subversion, and sanctity/degradation—as conceptualized by the Moral Foundations Theory. Using Natural Language Processing, we screened 16,497 (un)ethical justifications—i.e., reasons for behaving (un)ethically—for the presence of the five moral concerns (N = 901). We found that Honesty-Humility related positively to justice/fairness concerns, but it did not relate to care/harm, loyalty/betrayal, authority/subversion, and sanctity/degradation concerns. Our findings thus suggest that justice/fairness concerns might serve as one of the mechanisms relating Honesty-Humility to anti- and prosocial behavior.

Keywords: Honesty-Humility, HEXACO, Moral Foundations, ethical justifications, unethical justifications, self-serving justifications, Natural Language Processing, NLP

Introduction

Pro- and antisocial behaviors—such as charity donations, bribery, and corruption—have far-reaching consequences for individuals and societies at large (e.g., United Nations, 2018). In the light of this relevance, researchers have identified a range of individual and situational factors that predict such behaviors (e.g., Ścigała et al., 2020a; Thielmann et al., 2021). For instance, the basic personality trait Honesty-Humility from the HEXACO Model of Personality—defined as "the tendency to be fair and genuine when dealing with others" (Ashton & Lee, 2007 p. 156)—has been found to consistently predict a range of pro- and antisocial behaviors, such as cooperation, dishonesty, and generosity (for meta-analytical evidence, see e.g., Zettler et al., 2020). But while the relation between Honesty-Humility and pro- and antisocial behavior is well-established, the question of *why* Honesty-Humility predicts such behavior has been investigated to a far lesser extent (e.g., Hilbig et al., 2015; Pfattheicher & Böhm, 2018; Ścigała et al., 2020b; Thielmann et al., 2021).

Recently, Ścigała and colleagues (2020b) hypothesized that people with higher levels of Honesty-Humility engage in more prosocial and less antisocial behavior because they generate better/more ethical justifications as well as worse/less unethical justifications (i.e., reasons for behaving ethically and unethically, respectively). To test these hypotheses, they asked participants to generate ethical and unethical justifications of several morally ambiguous behaviors (e.g., not returning excess change to a cafe). Then, they rated the justifications in terms of (1) quality, defined as the effectiveness of the ethical justifications in discouraging, and the unethical justifications in encouraging committing the morally ambiguous actions, as judged by raters, and (2) quantity, defined as the number of generated justifications. Results show that Honesty-Humility was positively related to the quality of ethical justifications, and negatively related to the quality and quantity of unethical justifications. Notably, this research focused exclusively on the quality and quantity of said

justifications, without exploring the actual content of the (moral) language participants used in the justifications. Importantly, though, investigating the relation between Honesty-Humility and moral language used in (un)ethical justifications might further broaden the understanding of the potential mechanisms relating Honesty-Humility to pro- and antisocial behavior. More precisely, exploring the moral concerns that individuals with high (vs. low) levels of Honesty-Humility rely on when considering (un)justifiability of actions could provide insights into the kinds of concerns that might motivate them to engage in pro- and antisocial behavior. Furthermore, such an investigation could shed light on the type of moral language that might be used to encourage prosocial, and discourage antisocial behavior among individuals with different levels of Honesty-Humility (e.g., moral nudges and frames; e.g., Capraro et al., 2021). Consequently, we herein explore the relation between Honesty-Humility and moral concerns expressed in language.

In our analyses of moral language, we employ the Moral Foundations Theory (MFT), an influential theory of moral judgment, which proposes that people's concerns about morality can be captured by five basic moral foundations: (1) care/harm, defined as minimizing harm to others; (2) fairness/justice, defined as a motivation to maintain justice within a group and sensitivity to inequality; (3) loyalty/betrayal, defined as a motivation to favor and protect one's own group; (4) authority/subversion, defined as respect for authorities and a motivation to maintain a social hierarchy within a group; and (5) sanctity/degradation, defined as a motivation to be pure both in a spiritual and a physical sense (e.g., Graham et al., 2011, 2013). Specifically, we analyzed the content of the (un)ethical justifications from Ścigała and colleagues (2020b) using Natural Language Processing (NLP) techniques. This approach allows us to identify whether the language used in each of the generated justifications included words referring to one (or more) of the moral foundations (Frimer et al., 2019). Notably, to the best of our knowledge, the MFT constitutes the only theoretical

account which both (1) distinguishes between several moral concerns and (2) is accompanied by a NLP tool for extracting said moral concerns from text (i.e., Frimer et al., 2019). Hence, the MFT is particularly suited for the current investigation (for criticism against this theory, see, e.g., Schein and Gray 2018).

With regard to the relation between Honesty-Humility and Moral Foundations, note that Honesty-Humility is conceptualized with a strong focus on fairness and non-exploitation (e.g., Ashton & Lee, 2007; Ashton & Lee, 2009), suggesting that people with higher (vs. lower) levels of Honesty-Humility might be more likely to rely on concerns related to care/harm and fairness/justice foundations—i.e., "individualizing values" that refer to the rights and welfare of individuals (e.g., Zeigler-Hill et al., 2015). On the other hand, Honesty-Humility might not relate to concerns pertaining to loyalty/betrayal, authority/subversion, and sanctity/degradation—i.e., "binding values" that refer to the maintenance of social order and group cohesion (e.g., Zeigler-Hill et al., 2015)—as these values do not specifically refer to fairness and non-exploitation.

Previous research already linked Honesty-Humility to Moral Foundations, using self-report scales for both (e.g., Ashton & Lee, 2009; Graham et al., 2011). Whereas some findings indeed indicate that Honesty-Humility relates to individualizing, but not to binding values (Zeigler-Hill et al., 2015), others observed rather mixed results (e.g., Međedović & Petrovic, 2016; Webster et al., 2021). Specifically, while Honesty-Humility consistently related positively to care/harm (e.g., Međedović & Petrovic, 2016; Webster et al., 2021), the positive relation with fairness/justice emerged in some (e.g., Webster et al., 2021), but not in other studies (e.g., Međedović & Petrovic, 2016). Furthermore, the relations between Honesty-Humility and binding values ranged from being positive, through non-significant, to negative (e.g., Međedović & Petrovic, 2016; Webster et al., 2021; Zeigler-Hill et al., 2015). Consequently, previous, self-report-based research provided rather mixed findings overall.

Herein, we extend previous research by measuring the moral foundations-based concerns using an NLP approach, rather than self-report scales, bypassing some of the typical limitations of self-report measures (e.g., Paulhus & Vazire, 2009). Furthermore, with analyzing participants' (un)ethical justifications using NLP, we capture the expression of moral concerns in a relatively realistic setting, i.e., where participants actually consider if and why a given action is right or wrong; that is, we let them actually generate (un)ethical justifications. On the other hand, self-report measures of Moral Foundations are rather hypothetical, that is, they typically require participants to imagine what they *would feel* when deciding if something is right or wrong (Graham et al., 2011).

Overall, we contribute to the previous inconsistent literature relating Honesty-Humility to Moral Foundations, the latter measured using NLP techniques. In doing so, we re-analysed the (un)ethical justifications from Ścigała and colleagues (2020b) using a moral dictionary based on the MFT which allowed us to annotate the (un)ethical justifications as containing moral concerns relating to one or more of the five moral foundations (Frimer et al., 2019). The study procedure was pre-registered (i.e., the original study was published as a Registered Report; Ścigała and colleagues, 2020b). The following re-analyses of the data from Ścigała and colleagues (2020b) were not pre-registered and are therefore exploratory. The detailed study procedure, analysis code, and data are available on the Open Science Framework (OSF; https://osf.io/ru39p/).

Methods

Procedure and participants

The study was conducted via the participant panel Prolific Academic (https://www.prolific.co/) across two measurement occasions. On the first measurement occasion, we asked participants for their gender and age, as well as to complete the HEXACO-60 (Ashton & Lee, 2009) and a divergent thinking task (which is not relevant for

the current investigation). One week later, on the second measurement occasion, participants took part in the (un)ethical justification generation task. At both measurement occasions, participants provided consent to participate, and were presented with basic information about the study. 1,024 participants completed the study on the first measurement occasion and 907 on the second measurement occasion. We excluded six participants from the following data analyses because they did not provide any justifications. The final sample size consists of 901 participants (589 females, 307 males, and 5 who indicated the response option "other"), aged from 18 to 78 (M = 37.25, SD = 12.81) years. A simulation based sensitivity power analysis for the five mixed logistic regression models used below indicated that the sample size of N = 901 was sufficient to detect effects between OR = 1.08 and OR = 1.13 with power between 80% and 90%, and a = .05 (for details, see Table S1 in Supplemental Material). The study procedure can be accessed at: https://wave1r.formr.org/ (measurement occasion 1) and https://wave2r.formr.org/ (measurement occasion 2; links masked for review).

Measures

Honesty-Humility

Honesty-Humility was measured via the HEXACO-60 (Ashton & Lee, 2009), which measures the six HEXACO dimensions via ten items each. Specifically, participants were presented with 60 statements in a random order, describing themselves and other people, and were asked to rate the extent to which they (dis)agree with said statements on a five-point Likert scale. We included two attention check items among the HEXACO-60 (specifically, 'This is an attention check. Please choose 1'; 'This is an attention check. Please choose 4'). Participants who failed at least one of the attention check items were excluded from further participation in the study. Each of the HEXACO subscales exhibited at least good internal consistency estimates (see Table S2 in the Supplemental Material).

(Un)Ethical justification generation task

Participants were presented with three scenarios, which described everyday morally ambiguous situations (see Table 1; adapted from Ścigała et al., 2020b). We asked participants to generate reasons why the behavior presented in each scenario could be (1) ethically unacceptable (i.e., ethical justifications) and (2) ethically acceptable (i.e., unethical justifications). Both the order of the scenarios and the order of ethical and unethical justification generation tasks were in a random order. In each scenario, participants had 20 seconds to read the instructions and then they had two minutes to generate the justifications to the scenario (while the instructions and the scenario were visible on the screen). Before completing the tasks, we presented participants with an exemplary scenario and (un)ethical justifications (see Table 1). Additionally, we included a control question: 'Which of the following is the aim of the task? 1) To generate ethical and unethical justifications of the presented scenarios, 2) To create follow-up stories of the presented scenarios, 3) To generate random associations with the presented scenarios.' If participants answered incorrectly, they were excluded from the study. Three members of the research team excluded non-understandable, nonsensical, and (for a given participant) repetitive responses, responses that did not contain an (un)ethical justification, as well as divided responses from participants who provided multiple justifications in one field.

Table 1.

Scenarios and (un)ethical justifications from the (un)ethical justification task.

Scenario	Exemplary ethical justification	Exemplary unethical justification
Scenario and justifications presented in the instructions		
'Your accounting course requires you to purchase a software package that sells for £50. Your friend, who is also in the class, has already bought the software and offers	'People who wrote the software have families to support.'	'These large software companies have loads of money anyway.'

to lend it to you. You now have a chance to load it onto your computer.'		

Scenarios presented in the actual task

'You've waited in line for 10 min to buy a coffee and muffin at a cafe. When you're a couple of blocks away, you realize that the clerk gave you change for £20 rather than for the £10 you gave him. You now have a chance to keep the £10.'	'Because the clerk could get into trouble.'	'I could donate the money to a good cause.'
'You work in a fast-food restaurant in downtown. It's against policy to eat food without paying for it. You came straight from another job and are therefore hungry. Your supervisor isn't around, so you now have a chance to make something for yourself and eat it without paying.'	'It's stealing.'	'I work hard at my job, so I deserve a little bonus at least sometimes.'
'You work as an office assistant for a department at a company. You're alone in the office making copies and realize you're out of copy paper at home. You therefore have a chance to slip a ream of paper into your backpack.'	'What's next? Pens? Notebooks? Staplers? Where's the limit to what one can steal from the workplace?'	'The company will never notice the loss.'

Note. Please note that the exemplary justifications for the scenarios presented in the task are actual examples generated by participants.

Annotating moral concerns in (un)ethical justifications

We used NLP techniques to trace the moral concerns in the language usage of the participants. Specifically, we followed a dictionary-based approach, allowing us to automatically identify and extract key terms found in the justification data that signify the presence of concerns for each of the five moral foundations according to the MFT. For this purpose, we used the Moral Foundations Dictionary for Linguistic Analyses 2.0 (MFD 2.0;

Frimer et al., 2019), originally containing 2,048 words conceptually relevant to the five Moral Foundations. This word list was enriched to include highly semantically similar words to the original ones, based on a pre-trained word embeddings model (Mikolov et al., 2013), resulting in an extended moral dictionary containing 5,254 words. By tracing the presence of moral words in the justifications generated by the participants, each justification was automatically annotated with *1 = yes* or *0 = no* across the five moral foundations, denoting whether the participant expressed a moral concern for a given foundation. The criterion for annotating a justification with 1 for a given foundation was the presence of at least one word associated with the respective moral foundation according to the extended dictionary.

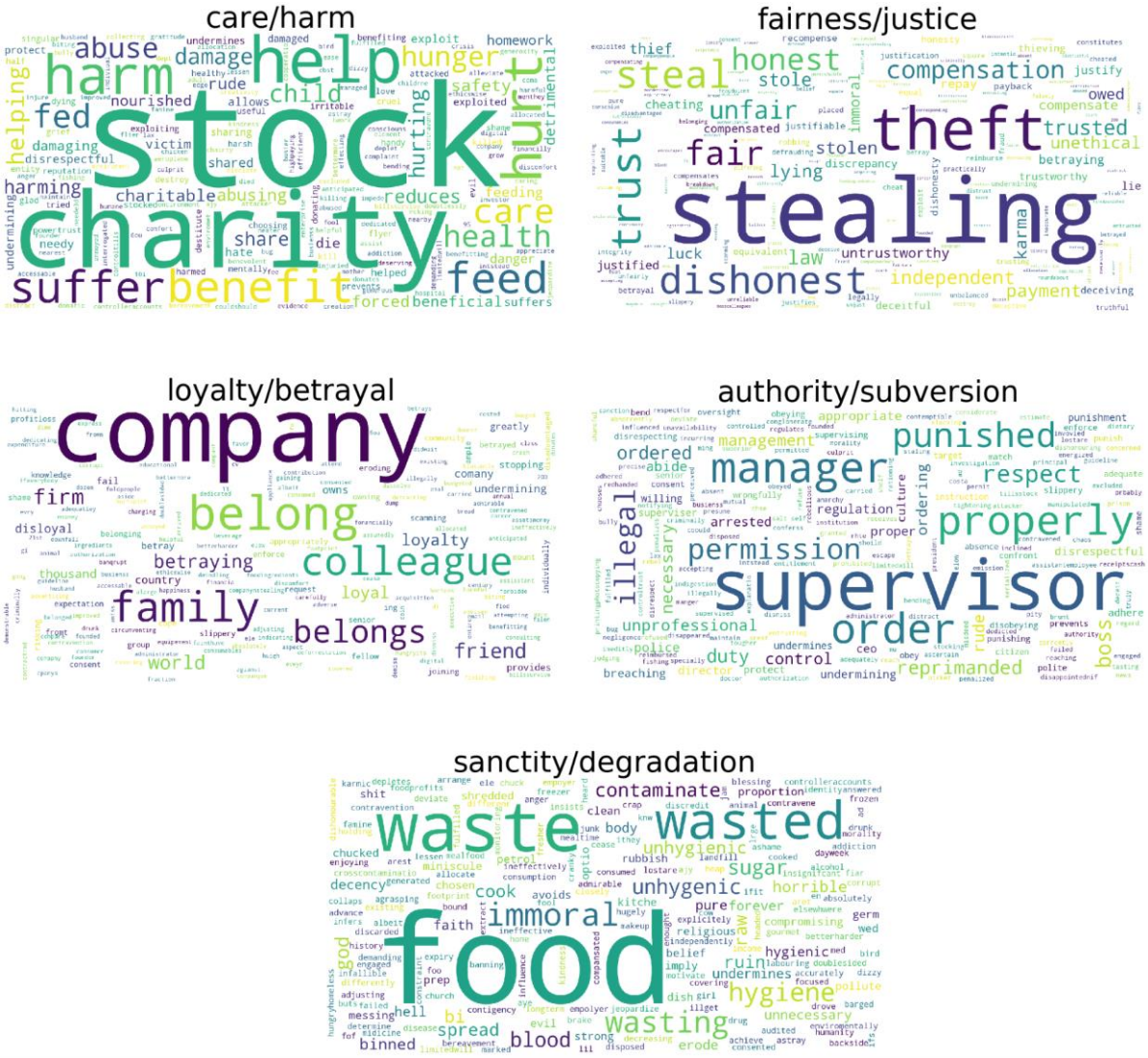
Results

Usage of moral language in (un)ethical justifications

We annotated 1,118 justifications as relating to care/harm, 3,084 as relating to fairness/justice, 2,946 as relating to loyalty/betrayal, 769 as relating to authority/subversion, and 1,605 as relating to sanctity/degradation. Additionally, we generated word clouds illustrating the most prominent terms in each annotated set of justifications (see Figure 1).

Figure 1.

Word clouds illustrating the most prominent terms in the justifications expressing concerns for each moral foundation.



Note. Some of the words presented in the word clouds might not relate to the moral foundations directly but co-occur with words relating to them in the justifications. For instance, the word “unnecessary” in the sanctity/degradation word cloud is not a part of the moral dictionary, but it co-occurs with the words “waste/wasted”, which belong to the moral dictionary.

Honesty-Humility and the use of moral language in (un)ethical justifications

In the following, we use five logistic mixed models with four predictors: Honesty-Humility as a fixed factor, as well as participant id, scenario (three levels: cafe, food, and office), and justification (two levels: ethical and unethical) as random intercept factors, and one dependent variable per model: the presence of at least one word related to a moral

foundation in a given justification (*1 = yes*, *0 = no*). Please note that we decided to analyze ethical and unethical justifications together because we found no significant interactions between Honesty-Humility and ethical vs. unethical justifications when predicting each of the moral concerns/foundations (see Table S3 in the Supplemental Material).

We found that people with higher (vs. lower) levels of Honesty-Humility were more likely to use fairness/justice-related words in their justifications (*OR* = 1.07, 95% CI = [1.02, 1.12], *p* = .004). On the other hand, Honesty-Humility did not significantly relate to the use of words relating to care/harm (*OR* = 1.03, 95% CI = [0.96, 1.11], *p* = .446), loyalty/betrayal (*OR* = 1.01, 95% CI = [0.95, 1.08], *p* = .715), authority/subversion (*OR* = 1.01, 95% CI = [0.94, 1.09], *p* = .756), and sanctity/degradation (*OR* = 0.96, 95% CI = [0.90, 1.03], *p* = .296). The results remained conceptually identical when controlling for the remaining five HEXACO traits (see Table S4 in the Supplemental Material). Furthermore, we report bivariate relations between all included variables (see Table S5 in the Supplemental Material).

Discussion

Does Honesty-Humility relate to the type of moral concerns people express in language? Herein, we examined whether Honesty-Humility predicts the expression of five moral concerns in language—namely, care/harm, justice/fairness, loyalty/betrayal, authority/subversion, and sanctity/degradation concerns—as conceptualized by the Moral Foundations Theory (Frimer et al., 2019; Graham et al., 2011, 2013). We found that participants with higher levels of Honesty-Humility were more likely to generate (un)ethical justifications including words related to fairness/justice concerns. On the other hand, we found that Honesty-Humility did not significantly relate to usage of words associated with concerns of care/harm, loyalty/betrayal, authority/subversion, and sanctity/degradation.

Taken together, our findings suggest that one of the reasons why individuals with

higher levels of Honesty-Humility engage in more prosocial and less antisocial behavior might be fairness/justice concerns, rather than care/harm, loyalty/betrayal, authority/subversion, and sanctity/degradation concerns. Indeed, such a possibility is in line with the theoretical understanding of Honesty-Humility as a trait primarily focused on fairness (e.g., Ashton & Lee, 2007; Ashton & Lee, 2009; Hilbig et al., 2015), as well as with research showing that in contexts when fairness is at odds with other moral concerns such as care and loyalty, high-Honesty-Humility individuals tend to act in line with the former (e.g., Ścigała et al., 2019; Thielmann et al., 2021). Other research, however, points out that under certain circumstances, individuals high in Honesty-Humility rather prefer loyalty over fairness/honesty (Ścigała et al., 2020a). Given these inconsistent findings, future research should systematically examine under which conditions individuals high in Honesty-Humility behave in line with fairness/honesty vs. care/loyalty values.

From a practical perspective, our findings suggest that because individuals high in Honesty-Humility are more concerned with fairness/justice when reasoning about (un)justifiability of actions, it might be beneficial to use such fairness/justice themes in nudges and frames attempting to increase prosocial, and decrease antisocial behavior among high- rather than low-Honesty-Humility individuals. Notably, these potential implications are based on one exploratory study only, and future research is needed before it can be more safely recommended to implement them in practice.

Our study has a few limitations and points towards several directions for future research. First, our findings are exploratory and hence future studies should examine the relations we investigated using confirmatory tests. Second, because we did not have access to behavioral data, we did not investigate whether generation of fairness/justice-related (un)ethical justifications actually mediated the relation between Honesty-Humility and pro- and antisocial behavior. Future research could examine such a mediation effect. Third, our

study was conducted on UK residents and therefore future studies could test whether the observed relations generalize to other samples. Lastly, the current study is delimited to investigate high-level aspects of language, i.e. the expression of concerns regarding different moral foundations. Future work could focus on a more fine-grained analysis of the language data, considering the polarity and magnitude of the expressed moral concerns (rather than their mere presence in the justifications). This would allow to examine in more detail the potential relationships between Honesty-Humility and moral language usage.

Summarizing, we examined the relation between Honesty-Humility and moral concerns expressed in language. We found that Honesty-Humility related positively to justice/fairness concerns, but did not relate to care/harm, loyalty/betrayal, authority/subversion, and sanctity/degradation concerns. Hence, our findings suggest that justice/fairness concerns might serve as one of the mechanisms relating Honesty-Humility to prosocial and antisocial behavior.

References

Ashton, M. C., & Lee, K. (2007). Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. *Personality and Social Psychology Review*, *11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Ashton, M., & Lee, K. (2009). The HEXACO-60: A Short Measure of the Major Dimensions of Personality. *Journal of Personality Assessment*, *91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Capraro, V., Vanzo, A., & Cabrales, A. (2021). *Playing with words: Do people exploit loaded language to affect others' decisions for their own benefit?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/yswxe

Frimer, J. A., Boghrati, R., Haidt, J., Graham, J., & Dehgani, M. (2019). Moral Foundations Dictionary for Linguistic Analyses 2.0. *Unpublished Manuscript*.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral Foundations Theory. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Elsevier. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

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), 366–385. https://doi.org/10.1037/a0021847

Hilbig, B. E., Thielmann, I., Wührl, J., & Zettler, I. (2015). From Honesty–Humility to fair behavior – Benevolence or a (blind) fairness norm? *Personality and Individual Differences*, *80*, 91–95. https://doi.org/10.1016/j.paid.2015.02.017

Međedović, J., & Petrovic, B. (2016). Can there be an immoral morality? Dark personality traits as predictors of moral foundations. *Psihologija*, *49*(2), 185–197. https://doi.org/10.2298/PSI1602185M

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. https://doi.org/10.48550/ARXIV.1301.3781

Paulhus, D., & Vazire, S. (2009). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology*.

Pfattheicher, S., & Böhm, R. (2018). Honesty-humility under threat: Self-uncertainty destroys trust among the nice guys. *Journal of Personality and Social Psychology, 114*(1),

179–194. https://doi.org/10.1037/pspp0000144

Schein, C., & Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Ścigała, K. A., Schild, C., Heck, D. W., & Zettler, I. (2019). Who deals with the devil? Interdependence, personality, and corrupted collaboration. *Social Psychological and Personality Science*, *10*(8), 1019–1027. https://doi.org/10.1177/1948550618813419

Ścigała, K. A., Schild, C., & Zettler, I. (2020a). Dishonesty as a signal of trustworthiness: Honesty-Humility and trustworthy dishonesty. *Royal Society Open Science*, *7*(10), 200685. https://doi.org/10.1098/rsos.200685

Ścigała, K. A., Schild, C., & Zettler, I. (2020b). Doing justice to creative justifications: Creativity, Honesty-Humility, and (un)ethical justifications. *Journal of Research in Personality*, *89*, 104033. https://doi.org/10.1016/j.jrp.2020.104033

Thielmann, I., Böhm, R., & Hilbig, B. E. (2021). Buying Unethical Loyalty: A Behavioral Paradigm and Empirical Test. *Social Psychological and Personality Science*, *12*(3), 363–370. https://doi.org/10.1177/1948550620905218

United Nations. (2018). *Global cost of corruption at least 5 percent of world gross domestic product*. https://www.un.org/press/en/2018/sc13493.doc.htm

Webster, R. J., Morrone, N., Motyl, M., & Iyer, R. (2021). Using trait and moral theories to understand belief in pure evil and belief in pure good. *Personality and Individual Differences*, *173*, 110584. https://doi.org/10.1016/j.paid.2020.110584

Zeigler-Hill, V., Noser, A. E., Roof, C., Vonk, J., & Marcus, D. K. (2015). Spitefulness and moral values. *Personality and Individual Differences*, *77*, 86–90. https://doi.org/10.1016/j.paid.2014.12.050

Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The Nomological Net of the HEXACO Model of Personality: A Large-Scale Meta-Analytic Investigation. *Perspectives on Psychological Science*, *15*(3), 723–760. https://doi.org/10.1177/1745691619895036