Predicting Bias in the Evaluation of Unlabeled Political Arguments

Nicholas Diana (ndiana@cmu.edu)
Human-Computer Interaction Institute
Carnegie Mellon University, Pittsburgh, PA, USA

John Stamper (john@stamper.org)
Human-Computer Interaction Institute
Carnegie Mellon University, Pittsburgh, PA, USA

Kenneth Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

While many solutions to the apparent civic online reasoning deficit have been put forth, few consider how reasoning is often moderated by the dynamic relationship between the user's values and the values latent in the online content they are consuming. The current experiment leverages Moral Foundations Theory and Distributed Dictionary Representations to develop a method for measuring the alignment between an individual's values and the values latent in text content. This new measure of alignment was predictive of bias in an argument evaluation task, such that higher alignment was associated with higher ratings of argument strength. Finally, we discuss how these results support the development of adaptive interventions that could provide real-time feedback when an individual may be most susceptible to bias.

Keywords: myside bias; moral foundations theory; distributed dictionary representations; civic reasoning

Introduction

The rise of social media has been accompanied by a rise in smaller, decentralized media sources. One clear negative consequence of this democratization of media has been an increase in access to unreliable or misleading news stories. Despite their lack of credibility and veracity, these stories are persuasive and appealing. Some estimates suggest that Americans fall for fake news headlines approximately 75% of the time (Silverman & Singer-Vine, 2016), and that these stories are generally more engaging than stories produced by traditional news outlets (Silverman, 2016).

The proposed solutions to these problems generally fall into two categories. The first category leverages various machine learning methods (Conroy, Rubin, & Chen, 2015) to create "fake news detectors." While some of these classifiers are quite sophisticated (Wang et al., 2018), these detectors tend to limit their scope to the detection of stories that are patently false. More nuanced instances of stories that are merely misleading are generally beyond the purview (and perhaps ability) of these systems (McGrew, Ortega, Breakstone, & Wineburg, 2017). Moreover, even if accurate classification was possible, one might question whether it is in our best interest to delegate this task to machines, potentially allowing our own ability to critically evaluate media sources to languish in the process.

In contrast to the content-driven detectors, other solutions focus on improving the critical thinking skills of the media consumers themselves. There is certainly evidence of a deficit in this regard. A recent study of students in middle school, high school, and college summarized the student's "civic online reasoning" (e.g. evaluating arguments, recognizing spin) as simply, "bleak" (Wineburg, McGrew, Breakstone, & Ortega, 2016). Non-detector solutions tend to focus on strengthening these kinds civic reasoning skills. For example, *Factitious* is a game created by the American University Game Lab (Hone, Rice, Brown, & Farley, 2018) that is marketed as a way to test the player's ability to distinguish fake and real news stories, but along the way teaches the player to identify features like reliable sources and neutral language.

While the detectors focus on the media content itself (hoping to fill the role of editor in the new democratized news space), the civic education solutions focus instead on the media consumers, with the hope that better critical thinkers might be more or less immune to the appeal of misleading content. Both of these approaches unfortunately tend to neglect the dynamic relationship between the media content and the media consumer. Consider the following actual fake news headline:

> "Pope Francis Shocks World, Endorses Donald Trump for President"

If you happen to be a religious Trump supporter, this story may seem plausible. After all, if you, a person of faith, have found reason to endorse him, why shouldn't another person of faith. This headline confirms what you already believe to be true. However, if you are not a Trump supporter, this headline might raise several red flags. It is in that wave of skepticism that you may dart your eyes to the URL in order to check the credibility of the source. Because this headline runs counter to your beliefs, you go searching for evidence to disprove it. In either case, the degree of critical thought that is brought to bear on the content is, at least to some extent, dependent on the values and beliefs of the reader.

This tendency to evaluate arguments more favorably when they align with your own views or beliefs (and conversely, more critically when they do not) is formally known as *myside bias* (Baron, 2000). Numerous studies have shown the effects of myside bias on reasoning to be reliable and

strong (Klaczynski & Robinson, 2000; Stanovich & West, 1997), irrespective of intelligence (Stanovich & West, 2007; Stanovich, West, & Toplak, 2013). Haidt's Social Intuitionist Model of moral reasoning (Haidt, 2001) suggests that the power of myside bias is likely due to the fact that the primary drivers of our moral judgments are intuitions and heuristics. According to the Social Intuitionist Model, when we encounter a new piece of information, we have an immediate and powerful intuition about whether we agree or disagree with the information. Haidt argues that the vast majority of these judgments are made automatically, using Kahneman's (Kahneman & Egan, 2011) System 1 thinking. Rational (or System 2) thinking always comes after an intuitive judgment has already been made, and only comes online if we are asked to justify our position. In short, we are not, by default, the rational thinkers we think we are. Moreover, when we do make use of our capacity for rational thought, it's generally to justify a decision we have already made, not to search for the truth.

Misleading and false news stories can exploit this vulnerability by designing stories that strongly align with the prior-held beliefs of the target audience. Because the reader wants to believe the story is true (to affirm their reality), System 2's critical reasoning skills are never engaged. The bias literature suggests that overcoming the strength of this intuitive appeal may require more than detecting falsehoods or training consumers to be more critical. Solutions that ignore the dynamic relationship between the user's beliefs and the beliefs latent in the misleading media content are perhaps ignoring the very feature that makes the target content so powerful.

Accurately capturing user beliefs is a daunting challenge. Each user likely possesses countless individual beliefs, and new beliefs are constantly being created in response to their current political context. One solution is to instead measure the foundational values that inform our beliefs. Moral Foundations Theory (Haidt & Graham, 2007) argues that moral decision making can be traced to a small set of foundational values (Care, Fairness, Loyalty, Authority, and Sanctity). These moral foundations have been empirically shown to be highly predictive of both general voting behavior (Franks & Scherr, 2015) as well as more specific political beliefs (e.g., "Climate change is real") (Koleva, Graham, Iyer, Ditto, & Haidt, 2012; Rottman, Kelemen, & Young, 2014). Moral Foundations Theory allows us to approximate beliefs in a theory-driven, context-general way. This is crucial for any solution intended for deployment on the internet, where the number of unique-contexts is virtually infinite.

After deriving a measure of user values (as a proxy for beliefs), the system must also be able to estimate the values latent in the text they are reading. Recently, Garten et al. (2018) developed a methodology for estimating the values latent in short pieces of text (tweets), and demonstrated that their methodology can accurately classify a tweet's most salient moral foundation (as measured by human raters). What remains to be seen is if value classification methods (like Garten's) can be combined with measures of user values to estimate the degree of alignment between the values of the media consumer and the content they are consuming.

We would expect that any measure that accurately captures this relationship between consumer and content should also be able to predict the presence of myside bias. That is, when alignment between user and content values is high, we expect that the user will be biased to evaluate the content more favorably. In this paper, we propose a method for measuring this dynamic relationship between consumer and content values, and demonstrate that the resulting metric can be used to predict bias in argument evaluation. Specifically, we test whether the alignment between participants' values and the values latent in an argument predicts participants' ratings of argument strength. We hypothesized that higher alignment between participant and content values will be associated with higher ratings of argument strength, and that this relationship will be present even in arguments specifically designed to confuse our natural language processing method.

Practically, this methodology lays the groundwork for future hybrid solutions that leverage technology alongside human critical thought to mitigate the impact of content designed to confirm our values rather than disseminate true information. Perhaps more importantly, this methodology presents a novel, context-independent way to estimate the impact of myside bias, a known, powerful moderator of everyday reasoning.

Methods

Participants

Eighty (80) participants were recruited using the participant recruitment platform Prolific. Participants were required to be between 18-65 years of age, U.S. citizens, and not have participated in any of our research group's prior studies (due to content similarity). Five participants exited the study before completing any significant portion of the main experiment and were excluded from analyses. The estimated completion time (based on prior pilot studies) was approximately 15 minutes, and participants were paid $2.50 ($10/hour) for participating.

Data Quality We mitigated the impact of potential gamers in several ways. First, the post-test questionnaire included a reading-check question. Participants who failed the reading-check (n=7) were excluded from analyses. We also used timing data to identify participants who were likely clicking through the problems without reading the prompts. Specifically, participants who selected an answer less than two seconds after a prompt loads (roughly the time needed to select an answer after the page loads), at least 10 times (for half of the problems), are labeled as gamers. We chose a threshold of 10 problems for two reasons: 1) it is reasonable to assume that participants who begin the experiment with the intention of gaming the system will exhibit this behavior for at least half of the problems, and 2) if we set this threshold too low, we risk excluding participants who begin the experiment with

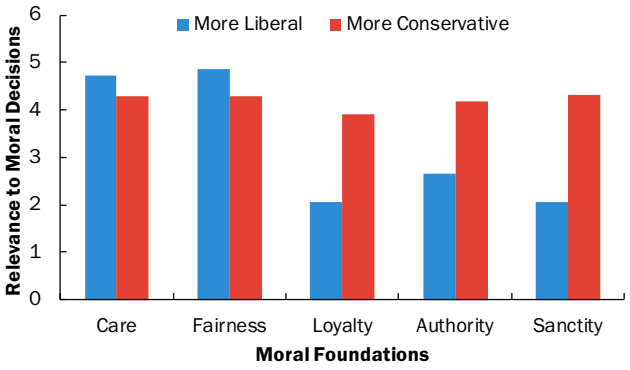

Figure 1: Relevance to Moral Decisions by Moral Foundation for more conservative and more liberal participants. These values closely match previously observed values for liberals and conservatives (Haidt & Graham, 2007), suggesting that our sample was politically diverse.

good intentions, but get fatigued towards the end. Eight participants met this criteria for gaming, and were excluded from analyses.

Demographics Of the remaining 60 participants, 38 identified as male, 20 as female, and 2 as other. Participants ranged in age from 18-62 years old (M=31.10). With respect to race and ethnicity, 50 participants identified as Caucasian, 6 as Hispanic, 3 as Black, and 1 as Asian. The majority (59%) of participants reported having completed a college level education or higher, and a high number of participants reported completing a master's degree (n=19).

Political Diversity One of the key benefits of recruiting participants from Prolific is that participants are drawn from all over the country. If instead, we were to recruit participants from our local community, we would likely get an unbalanced distribution of political beliefs (as our city has a history of voting overwhelmingly in favor of one party). Recruiting from across the country gives our sample a degree of political diversity that would be impossible to achieve otherwise.

To evaluate if our sample was indeed politically diverse, we used the composite measure of the Moral Foundations Theory Questionnaire (described below), called *progressivism*, to divide our sample into two groups along the mean score. Then, for each of the two groups we graphed the mean scores of each foundation and compared them to known averages. Figure 1 shows the mean scores for more liberal and more conservative participants across the five moral foundations. These values closely match previously observed values for liberals and conservatives (Haidt & Graham, 2007), suggesting that our sample was politically diverse.

Experiment Environment and Procedure

Participants completed the experiment online by navigating through a web-based application. After completing a consent form, participants were informed that the study consisted of two sections. In the first section, they were asked to complete a questionnaire (described below), and in the second section, they were asked to rate a series of arguments (presented in random order). After completing the two sections, the participants were directed to a post-test questionnaire where demographics information was collected, and then finally, to the debriefing page, which clarified that any facts and figures used in the study were entirely fictitious. The experiment was estimated to take approximately 15 minutes to complete (actual median completion time was 12 minutes).

Moral Foundations Theory Questionnaire In the first section of the experiment, each participant was required complete the Moral Foundations Questionnaire (MFQ) (Haidt & Graham, 2007). This 30-item questionnaire is designed measure how relevant each one of the five moral foundations (Care, Fairness, Authority, Loyalty, Sanctity) is to one's moral decision making. For example, participants are asked to indicate the degree to which they agree with the following statement: "Respect for authority is something all children need to learn." The final output of the questionnaire is a vector of five scores that indicate the relative importance of the five moral foundations to the participant's moral decision making. Ultimately, we are interested in constructing a model that relates the values latent in text to the values and beliefs of an individual person. This vector of five scores represents the human side of that relationship.

It is worth noting that having the participants take the MFQ before answering arguments designed to evoke moral decision making is not ideal. The questionnaire may cause participants to be more conscious of their beliefs than they might normally be if encountering these arguments in the real world. However, this ordering is unfortunately necessary for later stages of this project, where adaptive interventions designed to promote analytic thinking use an individual's scores on the MFQ to decide when targeted feedback is needed most. These future directions are explored in more detail in the *Discussion* section.

Rating the Strength of Arguments Participants were asked to read and rate the strength of 20 arguments on a nine-point Likert scale (1=Very Weak; 9=Very Strong). Each argument had three key features. First, each argument was designed to evoke a specific moral foundation. For example, the following argument was designed to evoke the *Authority* foundation:

> Greenville School District requires students to address all adults as "Sir" or "Ma'am" and their students always score higher on state tests than ours. Instilling a strong respect for authority for their teachers helps students learn.

Regardless of the argument's actual strength, we would expect that if a participant believes that respecting authority is important, this argument will resonate with them. Each of the five foundations is the focus of an argument four times, for a

total of 20 arguments.

The second key feature is the relative quality of an argument. This is a categorical feature with two levels, *high quality* and *low quality*. The above argument is an example of a *low quality* argument. In contrast, consider the following argument:

> The number of suspensions at Redbridge School District has been slowly increasing for the past 5 years. Last year they added three police officers to their staff and saw a 10% decline in suspensions. The presence of a strong authority figure reduces bad behavior.

While this argument is certainly not airtight, it has several attributes that make it a relatively higher quality argument. First, it shows the reversal of a long-term trend, in contrast to the *low quality* argument where no temporal context is established. Second, it uses concrete figures that are relative to the norm, as opposed to the *low quality* argument which uses vague terms like "higher" to quantify changes. In general, high quality arguments include information that can be used to rule out some alternative explanations. Low quality arguments leave open the possibility of alternative explanations. Of the 20 arguments, half are *high quality* and half are *low quality*.

The third key feature is *congruence* with the target foundation. A potential limitation of the distributed dictionary representation methodology (described below) is that statement representations are formed using the representations of single words. This means that, while this methodology should have no problem knowing that the word "son" in the context of the word "king" likely refers to the concept "prince," it will likely have more difficulty identifying the cultural nuances between statements like "God is good" and "God is dead." The *congruence* feature is designed to test the robustness of this methodology's ability to adapt to these kinds of unfavorable circumstances. Consider again the two previous example arguments. Both arguments 1) use language that evokes the authority foundation, and 2) are supportive of that foundation. In contrast, consider the following argument:

> Woodford School District doesn't allow teachers to reprimand students, and last year they had fewer detentions than our district. Students behave better when they're treated like equals instead of children

While this argument also evokes the Authority foundation, this example argues against an increased respect for authority. We would expect that participants that value authority will be more skeptical of the claims in this argument, because they violate their intuitions. Whether the model's representation of the values latent in the argument is nuanced enough to make the distinction between *incongruent* and *congruent* arguments is an open question. Again, half (10) of the arguments are *congruent*, half *incongruent*.

Analysis

Distributed Dictionary Representations The broad goal of this experiment is to evaluate a method for comparing an individual's values with the values latent in the media they are consuming. Using the MFQ, we are able to generate a theory-driven estimation of the participant's values. The next step is identifying the values latent in a particular piece of text. While historically this has been done using word-frequency methods (i.e., counting the number of times terms in a concept dictionary appear in the target text), these methods are much less effective for analyzing small bodies of text (e.g., news headlines, tweets), which may not contain any of the dictionary terms.

In contrast to word-frequency methods, distributed representations (Mikolov, Chen, Corrado, & Dean, 2013) estimate the meaning of words by comparing the numerous, varied contexts that the word appears in within a large text corpus. These models are rooted in the distributional hypothesis, which states that words that appear in similar contexts likely share some semantic features. The distributed representation of a word is simply that word's location in a low-dimensional (10-10,000 dimensions) space. This location can be represented as a vector, which allows us to compute the semantic distance between two concepts using cosine similarity.

Garten et al. (2018) extends this work in distributed representations to incorporate concept dictionaries. A distributed dictionary representation is computed by simply averaging the distributed representations of all the words in the dictionary. The result is a point in the semantic space that amplifies the shared, core features of each of the component dictionary terms. Because we are ultimately using an abstract representation of a concept, our dictionaries can be highly focused, including only the most relevant terms. The current study uses five such dictionaries (one per moral foundation), and each dictionary contains four positive words (e.g., fairness, equality) and four negative words (e.g., unfair, injustice) related to the moral foundation (e.g., fairness). Using distributed representations allows for the effective analysis of small bodies of text (such as the short arguments used in the current study), because the method does not require any of the dictionary terms to be present in the text. We used gensim (a Python implementation of Word2Vec (Mikolov et al., 2013)) and the pre-trained Google News corpus (approximately 100 billion words) Word2Vec model[1].

Alignment The output of the distributed dictionary representation analysis is a vector of five values, indicating the average semantic distance between the words in the argument and the words in each of the five moral foundation concept dictionaries. To compute *alignment*, we compute the cosine similarity between this vector and the vector of moral foundation relevance scores outputted by the Moral Foundations Questionnaire (i.e., the participant's values). We then used a

[1]The pre-trained Google News model can be found here: https://code.google.com/p/word2vec/

normalized log-transformation to correct for skew. *Alignment* is computed for each participant and argument combination.

Linear Mixed Effects Models Because it is impossible to determine an objective rating of argument strength for the arguments used in this study, we are less interested in the individual rating of each argument and more interested in how a participant rates arguments relative to one another (i.e., high alignment vs. low alignment or high quality vs. low quality). To make use of all the data while accounting for differences in ratings across participants, we use a series of linear mixed effects models, with *participant* as a random effect. Similarly, to control for unintended variations in argument strength, we include *Argument ID* as an additional random effect. We compare models to one another using the Akaike information criteria (AIC), which estimates the fit of a model (lower scores are better). All reported coefficients are standardized.

Results

We used a series of mixed effects models to examine the relationship between *alignment* (between participant and argument values) and ratings of argument strength. To reduce the possibility that any effect of *alignment* on ratings could be attributed to differences in demographics, we tested for collinearity between *alignment* and each demographic variable (*age*, *gender*[2], *race*, and *education level*). A series of one-way analyses of variance (ANOVA) between *alignment* and each of the three categorical variables (*gender*, *race*, and *education level*) showed no evidence of collinearity. Similarly, there was no significant correlation between *alignment* and *age*.

A mixed effects model with *participant* and *argument ID* as random effects, ratings of strength as the outcome variable, and *gender*, *race*, *educational level*, and *alignment* as fixed effects was generated[3]. *Alignment* was a significant predictor of ratings when included alongside these demographics variables ($\beta = 2.48$, $p < 0.01$), providing further evidence that any effect of *alignment* on ratings was not due to differences in demographics.

Impact of Alignment when Controlling for Quality To test if *alignment*'s impact persists when controlling for argument quality, we built a mixed effects model with *participant* and *argument ID* as random effects, ratings of strength as the outcome variable, and *alignment* and *argument quality* as fixed effects. We found that *alignment* was predictive of ratings despite the presence of *argument quality*. It should be noted that although participants on average rated high quality arguments as significantly stronger ($t(59) = 8.07$, $p < .001$) than low quality arguments ($M = 5.06$, $SD = 1.72$) (suggesting some categorical validity), the labels "high" and "low" are very much subjective labels. As such, we cannot objectively compare the impact of *alignment* to the impact of quality. Still, we can make a meaningful, subjective comparison between the impact of *alignment* and "quality" (as operationally defined in this context). In this context, the impact of *alignment* on ratings of strength ($\beta = 3.06$, $p < 0.001$) was greater than the impact of *argument quality* ($\beta = 1.33$, $p < 0.01$).

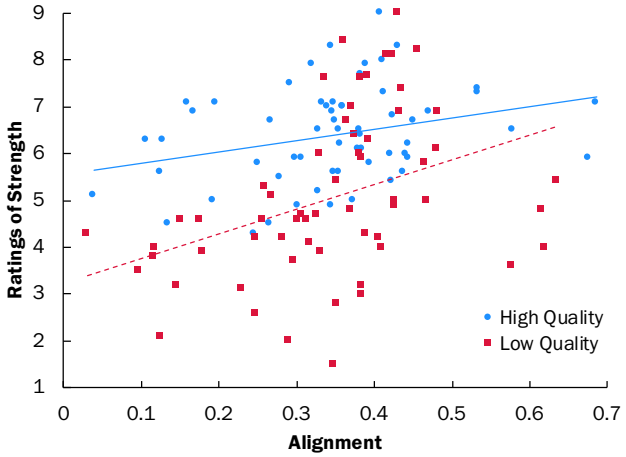

Figure 2: Relative impact of alignment on the ratings of high and low quality arguments. Each data point represents the average rating and alignment for all arguments within a category (high or low quality) for one participant. On average, participants rated high quality arguments as stronger than low quality arguments. The ratings of both types of arguments were associated with alignment scores.

While on average, participants rated congruent problems ($M = 5.89$, $SD = 1.35$) as significantly stronger ($df(59) = 2.27$, $p = 0.02$) than incongruent problems ($M = 5.57$, $SD = 1.40$), *congruence* was not a significant predictor of ratings when added to this model.

Interaction between Age and Alignment Previous research suggests that, because reliance on heuristic reasoning increases with age, older adults may be more likely to exhibit biases in everyday reasoning (Klaczynski & Robinson, 2000). To test whether this was true of our sample, we built a mixed effects model with *participant* and *argument ID* as random effects, ratings of strength as the outcome variable, and *argument quality* and *alignment*age* as fixed effects (where *alignment*age* is an interaction term). We found that there was a significant interaction between *alignment* and *age* ($\beta = 15.01$, $p < 0.001$), such that *alignment*'s impact increases as *age* increases. This finding aligns with previous research. Additionally, this *alignment*age* interaction model had a better fit (AIC=5033.05) than the previous model built without the interaction term (AIC=5058.63).

[2]Participants identifying as "other" were excluded from this analysis because the sample was very small (n=2).

[3]Note that age was excluded from the model because a one-way ANOVA between *age* and *education level* indicated a significant relationship between *age* and *education level*. A Tukey's HSD test showed that participants at the graduate level ($M = 35.33$, $SD = 5.38$) were significantly older than those below the college level ($M = 27.96$, $SD = 9.05$). A likelihood ratio test showed that *education level* was more explanatory than *age*, so *age* was excluded in favor of *education level*.

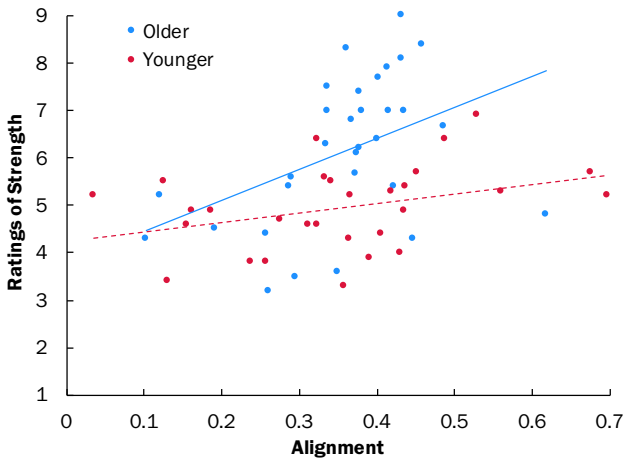

Figure 3: The interaction between age and alignment. Each data point represents one participant's average rating and alignment scores. Alignment had a much larger impact on ratings of strength for older participants (participants above the median age) than younger participants. This conforms with previous findings examining the relationship between bias in argument evaluation and age.

Performance on Incongruent Problems A potential limitation of this particular NLP method is its reliance on the semantic relationships between isolated words. A robust methodology should be able to accurately determine the valence of an argument that may contain several words related to a foundation, but nonetheless is incongruent with the beliefs of someone who values that foundation. To test the robustness of our method, we built another iteration of the above, best performing mixed effects model (including the *alignment*age* interaction), but selected only *incongruent* arguments (previously both *congruent* and *incongruent* problems were used). The impact of *alignment* on ratings of *incongruent* arguments also appears to be dependent on *age*, as the interaction term *alignment*age* was again a significant predictor of ratings of argument strength ($\beta = 15.01$, $p < 0.001$). To examine this relationship further, we divided the sample into two groups (older and younger) along the mean age, and then calculated the correlation between participants' mean *ratings* and mean *alignment* for each group. While we found a significant correlation between *ratings* and *alignment* in the older group ($r = 0.26$, $p < 0.001$), we found no such correlation in the younger group (see Figure 3).

Discussion

Our results demonstrate that distributed dictionary representations (DDR) combined with a measure of user values may provide a reliable method for identifying when users may be prone to biased reasoning. Because our method does not require labeled text data, it can be easily applied to real-world data (such as social media posts). The only limitation on this front is the identification of the user's values. We do this formally with the Moral Foundations Questionnaire, but research has demonstrated that political orientations can be predicted with a high degree of accuracy purely based off of social media activity (Colleoni, Rozza, & Arvidsson, 2014). Whether these predictions are as nuanced as those generated by the theoretically grounded Moral Foundations remains to be seen, but the potential for a fully automated method for measuring a user's susceptibility to myside bias exists.

We used *incongruent* problems to test the robustness of our methodology. These problems were intentionally designed to confuse the DDR method, and produced some interesting results. While alignment was predictive of ratings on incongruent problems, this was only true for older participants. One potential explanation for this difference is a lack of clarity about what low scores on the moral foundations questionnaire mean (specifically in this context as a proxy for beliefs). High scores indicate a value is relevant, but do low scores indicate indifference or impassioned opposition? Future work will require a qualitative exploration of these nuances.

Toward Adaptive Interventions

Our results suggest that we can leverage the dynamic relationship between user and content values to predict when the user may be prone to biased reasoning. This work is the first step toward providing adaptive, targeted interventions when high alignment between user and content values is detected (i.e., when the user is most prone to biased reasoning). It is in these cases of high alignment that we are least likely to move from System 1 (intuitive) to System 2 (rational) thinking, and engage the reasoning processes that may mitigate bias. Adaptive interventions may be able to facilitate the engagement of System 2 thinking in exactly these critical moments, making users less vulnerable to content designed to exploit natural biases. This kind of hybrid solution leverages the strength of sophisticated machine learning methods, while still preserving the need for and power of human reasoning.

Conclusion

In this paper, we demonstrated that a measure of alignment between a participant's values and the values latent in short arguments was a significant predictor of ratings of argument strength. This was true even for nuanced arguments, designed to confuse our methodology. These results underscore the impact of values on the evaluation of everyday arguments, and lay the groundwork for adaptive interventions designed to mitigate everyday reasoning biases.

Acknowledgements

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Baron, J. (2000). *Thinking and deciding*. Cambridge University Press.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, *64*(2), 317–332.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th asis&t annual meeting: Information science with impact: Research in and for the community* (p. 82).

Franks, A. S., & Scherr, K. C. (2015). Using moral foundations to predict voting behavior: Regression models from the 2012 us presidential election. *Analyses of Social Issues and Public Policy*, *15*(1), 213–232.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, *50*(1), 344–361.

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

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

Hone, B., Rice, J., Brown, C., & Farley, M. (2018). Factitious. Retrieved from `factitious.augamestudio.com`

Kahneman, D., & Egan, P. (2011). *Thinking, fast and slow* (Vol. 1). Farrar, Straus and Giroux New York.

Klaczynski, P. A., & Robinson, B. (2000). Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning biases. *Psychology and Aging*, *15*(3), 400.

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality*, *46*(2), 184–194.

McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (2017). The challenge that's bigger than fake news: Civic reasoning in a social media environment. *American Educator*, *41*(3), 4.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Rottman, J., Kelemen, D., & Young, L. (2014). Tainting the soul: Purity concerns predict moral judgments of suicide. *Cognition*, *130*(2), 217–226.

Silverman, C. (2016, Nov). *This analysis shows how viral fake election news stories outperformed real news on facebook.* Retrieved from `https://www.buzzfeednews.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook`

Silverman, C., & Singer-Vine, J. (2016, Dec). *Most americans who see fake news believe it, new survey says.* Retrieved from `http://www.buzzfeed.com/craigsilverman/fake-news-survey`

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*(2), 342.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225–247.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., ... Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857).

Wineburg, S., McGrew, S., Breakstone, J., & Ortega, T. (2016). Evaluating information: The cornerstone of civic online reasoning. *Stanford Digital Repository.*.