Debiasing Politically Motivated Reasoning with Value-Adaptive Instruction

Nicholas Diana[1][0000−0002−8187−3692], John Stamper[2][0000−0002−2291−1468], Ken Koedinger[2][0000−0002−5850−4768], and Jessica Hammer[2]

[1] Colgate University ndiana@colgate.edu
nickdiana.com
[2] Carnegie Mellon University

Abstract. While there is a substantial appetite in the United States for improving media consumption skills, little work has focused on the biases that can make inaccurate or misleading claims feel true. This skill is particularly difficult to teach, as effective instruction requires the instructor to adapt course content to the specific beliefs of individual students, a process that is unscalable in most classrooms. Here we examine the impact of a novel method of user-centered personalized instruction that uses value-adaptivity to highlight and address user bias in the context of a civics education game. This intervention uses estimates of player and content values to predict when players may be most susceptible to biased reasoning and then intervene in those instances. We found that the intervention successfully reduced bias among high bias-regulators with practice. These results suggest that value-adaptive systems may be able to support debiasing instruction in an effective, scalable way.

Keywords: Myside Bias · Confirmation Bias · Personalization · Educational Games · Civic Technology · Civics Education

1 Introduction

The co-opting of social media platforms in large-scale disinformation campaigns has spurred the development of novel tools and methods for more responsible media consumption. Much of this work focuses on the media content itself, with researchers developing sophisticated machine learning models for classifying patently false information [20, 15]. Other work focuses on the opposing side of the media equation: the user. This work examines methods for improving media literacy (i.e., their ability to evaluate the credibility of the information they are consuming) [10].

Less work, however, has focused on the dynamic relationship between the media consumer and the content they are consuming, and in particular: how that relationship can be a powerful source of bias and how best to mitigate those biases. Recognizing and reducing bias (sometimes called "debiasing") [14] is a critical component of any comprehensive 21st century civics education. Civics teachers (along with those in the English Department) are often tasked with

equipping students with the media literacy skills they need to navigate an increasingly fraught media landscape. One common approach to debiasing in civics class is to ask a student to defend a political position that they themselves do not hold (or actively oppose). This act of *perspective taking* can be powerful [11], but efficient perspective taking requires that the teacher: 1) knows the positions of each student with respect to each topic, and 2) takes the time to match students to positions individually. In a class of 30 students that might discuss a topic a week, a systematic adherence to this approach is likely unscalable.

In the current experiment, we used Moral Foundations Theory [9] in conjunction with natural language processing methods to model student and content values. We used the relationship between those two sets of values to power a value-adaptive debiasing intervention. This debiasing intervention was integrated into an educational game designed to help students practice engaging in productive civil discourse. We tested the efficacy of this novel approach to debiasing by examining students' bias regulation, or their ability to ignore an intuitively correct option (biased response) in favor of the actual correct option. Specifically, we hypothesize that the bias regulation of students who saw the debiasing intervention will improve over time relative to their peers who did not see the intervention.

The primary contribution of this work is the demonstration of a scalable approach to debiasing instruction in civics education that is powered by a novel method of personalized instruction: value-adaptive instruction.

2 Related Work

The fallibility of human rationality has long been established as an important and consequential area of study [1, 4]. In many cases, human cognition fails in regular and predictable ways. As such, it is not unreasonable to attempt to identify the circumstances under which we may be most susceptible to these cognitive biases and to develop training programs designed to mitigate the impact of the most common or most critical biases. Despite the large body of work pertaining to the identification and measurement of cognitive biases, the body of work pertaining to the development and testing of so-called *debiasing* training programs is relatively small [14]. This may be due to the fact that many cognitive biases are quite robust, persisting even in the face of explicit debiasing instruction [5].

With respect to debiasing instruction, tasks that require participants to consider the opposing viewpoint may mitigate the impact of biased reasoning. This strategy shares many features with a skill called *perspective taking*, a common instructional goal in civics curricula that can be complicated by the personal nature of political beliefs. For example, a civics instructor might ask a student who is anti-immigration to defend a pro-immigration stance. The serious consideration of opposing perspectives may reduce the impact of bias. This kind of individualized debiasing instruction is an example of what we call *value-adaptive instruction* (i.e., instruction that is adapted to the specific values of the learner).

Unfortunately, the traditional approach to value-adaptive instruction described above is simply unscalable in classrooms of 20-30 students and in courses that might cover 40 topics over the span of a school year. One potential solution is to use technology to support the estimation of student and content values. These estimations will never be as accurate as those from an expert human instructor, but reasonably accurate estimations may allow us to provide numerous individualized practice opportunities to students in a scalable way. Moreover, it would allow us to use educational technologies to:

1. Estimate the impact of bias on informal reasoning tasks,
2. Predict when students may be most susceptible to biased reasoning, and
3. Provide targeted debiasing interventions precisely in those moments of vulnerability.

Myside Bias in Civics Education In this study, we explore a particular type of bias, termed *Myside Bias*, that is often found in civil discourse. In brief, *Myside Bias* refers to one's tendency to evaluate claims or evidence more favorably if the claim or evidence supports one's own beliefs or worldview [19]. Myside bias has been characterized as both a more accurate term for *Confirmation Bias* [16] and a subclass of *Confirmation Bias* [19] in various works. In the context of civil discourse, myside bias can manifest as one's inability to speak across ideological lines to the values that motivate the beliefs of those they disagree with. It is our tendency to reach for the argument that seems strongest to us, rather than the argument that would appeal most strongly to whomever we are trying to persuade.

Effectively choosing arguments that will be most persuasive to those with a differing ideology requires two skills:

1. The ability to identify the values that underpin the beliefs of your interlocutor
2. The ability to choose an argument that best aligns with those values

Inherent in this second skill is the challenge of overcoming myside bias (in this context, our tendency to choose an argument that aligns with our own values instead of more persuasive options). In the current study, we examine a value-adaptive intervention designed to mitigate the impact of myside bias when choosing effective arguments in civil discourse. This intervention was integrated into an educational game designed to give students opportunities to practice these key discourse skills.

3 Methods

A total of 87 students from high schools located in the Northeastern Region of the United States participated in the study. Note that all demographics questions were optional, and a small number of students chose not to answer some

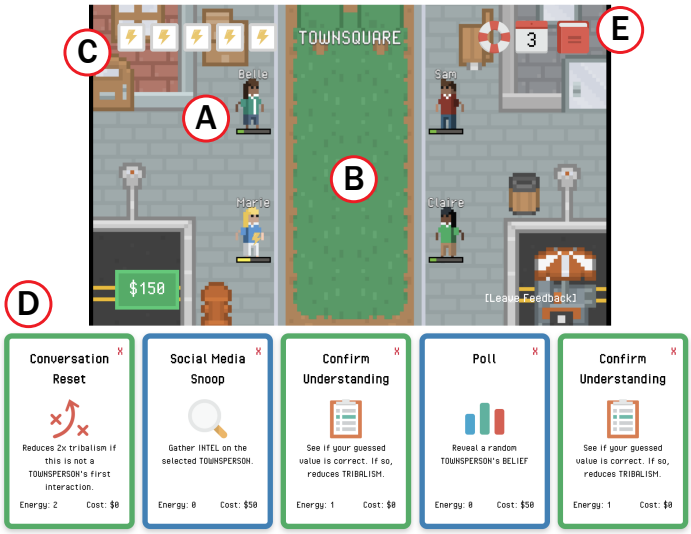

Fig. 1. An annotated screenshot of a scenario. In each scenario, players must persuade NPCs like Belle (A) to move into the TOWNSQUARE (B). To do this players must identify which argument from the opposing side appeals to what Belle values. Persuading NPCs costs Energy (C). The bar positioned below each NPC represents their political tribalism. Players must reduce an NPC's tribalism before attempting to persuade them. They can do this by playing Discourse Cards (D) like *Conversation Reset*. Finally, the action menu (E) allows the player to request a hint, end the day/turn, or reference information in their notebook.

questions. The students where evenly split with respect to sex (41 females, 43 males, 1 other), and reflective of the racial demographics of the area (9 black or African American students, 72 white or Caucasian students, and 4 students identifying as other/more than one race). Students from six classes (three English classes and two Social Studies classes) participated in the study.

3.1 Leveraging NLP Alongside Theories of Moral Judgments

Any given student will, by definition, only exhibit myside bias when presented with information that aligns with their beliefs. As such, crafting instructional events that give students the opportunity to wrestle with their biases requires three critical pieces of information:

1. An **estimate of a particular student's values**. That is, if we were to cover a new topic in class, can we be relatively confident that we could use our understanding of their values to predict the kind of belief this student will espouse?
2. An **estimate of the political values latent in the content** we are presenting to this student. In the case of commonly debated topics, these latent

political values may be obvious. However, in the case of uncommon or novel topics, it may be substantially more difficult to predict which aspects of the content will resonate with a particular student.
3. A **means of understanding the *relationship*** between the prior two sets of values. For example, to what extent do a student's values align with the values latent in a news article? Will one aspect of a problem be more important to this student than another aspect?

Moral Foundations Theory We estimate the student's values using the Moral Foundations Theory Questionnaire [9]. Moral Foundations Theory [8, 9] argues that the moral judgements we make are directly related to the importance we ascribe to a small set of *moral foundations* (e.g., care, fairness, authority, loyalty, sanctity). For example, if someone values the authority foundation (i.e., they generally respect laws, traditions, and authority figures), we might expect them to side with the police and the military on controversial matters. These moral foundations have been empirically shown to be highly predictive of both general voting behavior [6] as well as more specific political beliefs (e.g., "Climate change is real") [13, 18]. The output of the Moral Foundations Questionnaire is a vector of five scores, representing the degree to which the student values each of the five foundations when making moral judgments.

It is worth reiterating that these are *estimates* of user values. The reasons humans hold beliefs are numerous and personal. As such, we can say with relative certainty that the model of human beliefs (based on values) employed in this study is incomplete and flawed. What remains to be seen is whether the model provides a good enough estimate of user beliefs to be useful in the context of debiasing instruction.

Distributed Dictionary Analysis We estimate the values latent in text content using natural language processing, specifically distributed dictionary representations (DDR) [7]. DDR builds off of Word2Vec [17], which invoves modeling a large corpus of text data in a low-dimensional space, where each word can be represented as a point in that semantic space. DDR was created to model psychological constructs (such as the foundations in Moral Foundations Theory) using this semantic space. That is, the foundation referred to as *Fairness* actually encompasses more than just the concept of fairness; it includes equality, injustice, rights, and fraud. To find the point in the semantic space that matches this more nuanced concept that we label *Fairness*, we first generate a concept dictionary (i.e., a list of terms that approximate the meaning of the concept). Because each word in the concept dictionary can be represented as a vector, we can simply average across all word vectors in the dictionary to find the vector that corresponds to our operational definition of the concept *Fairness*.

In this work, we follow the original procedure outlined in [7] to generate a vector for each of the five moral foundations. Next, to estimate the values latent in a piece of text, we compute the cosine distance between the representations of each of the five moral foundations and the average representation of all words in

the text. Like the Moral Foundations Questionnaire, the output of this process is a vector of five scores, representing the degree to which the text was semantically similar to each of the five foundations (see [2] for a more complete discussion of this process). In our analysis, we used the pre-trained Google News corpus (approximately 100 billion words) Word2Vec model[3], and a Python implementation of Word2Vec [17] called *gensim*.

Computing Alignment Because the outputs of the Moral Foundations Questionnaire and the DDR analysis are two vectors of equal length, measuring the relationship between the two vectors can be done simply by computing the cosine similarity between them. The extent to which the student's values are similar to the values latent in the content is termed *Alignment*. Previous work has shown that *Alignment* is predictive of bias in argument evaluation tasks [3]. In the current study, we use *Alignment* to predict where students might be most susceptible to biased reasoning during gameplay and to adapt the debiasing intervention accordingly. It is worth clarifying that *Alignment* is essentially measuring the presence of foundational concepts and their relationship to the user's estimated values. As such, we expect it to fail in the face of sufficiently nuanced language. What the current experiment aims to test is whether or not the resulting model of bias is good enough to be useful.

3.2 A Value-Adaptive Debiasing Intervention

The debiasing intervention was incorporated into *Persuasion Invasion*, an educational game called designed to help students practice productive civil discourse skills. The goal of each level in *Persuasion Invasion* is to persuade ideologically entrenched townspeople to engage with those they disagree with. Successfully persuading a townsperson requires that the student 1) identify which of the five moral foundations the townsperson values most, and 2) identify which of three arguments from the opposing side appeals most to someone who values that foundation. We expect that players may be biased to choose the argument that aligns most to their own values rather than the argument that aligns with the values of the townsperson they are attempting to persuade.

To mitigate the impact of bias, we integrated a value-adaptive debiasing intervention into this *Persuasion* interaction. All students were randomly assigned to one of two conditions: an adaptive condition or a control condition. The two conditions were identical in every respect with one exception: When players in the adaptive condition were asked to choose which of the three listed arguments would be most persuasive to a townsperson, they saw one of the options presented in orange-colored text with an additional piece of instruction that read:

> **Caution: Orange-colored options might seem more persuasive to you (based on your values). Remember to choose the best response for [NPC NAME].**

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

The orange-colored option corresponds to the option with the highest *Alignment* score (i.e., the option that, based on this methodology, aligns most closely to this specific player's values). The color orange was chosen because it is attention-grabbing, but isn't traditionally associated with correctness (as colors like red and green are in the United States). Importantly, the orange-colored option was not any more or less likely to be the correct answer; this intervention is simply designed to elicit a more critical analysis of the options.

3.3 A Composite Measure of Bias Regulation

In previous work, *Alignment*-based estimates of potential bias have represented the extent to which the user's values align with the values of the correct response. This made a direct comparison between *Alignment* and other baseline measures possible. However, this *Alignment*-based estimate of bias is limited in that it fails to account for the alignment between the user and the other potential options. Imagine, for example, a case in which the correct option happens to also be the option with the highest alignment. We would expect that, in this case, the choice is easy, as there is no conflict between the intuitive choice and the correct choice. This case also tells us nothing about the user's ability to regulate their own bias. Contrast this with a scenario in which the correct option happens to be the option with the lowest alignment (i.e., least congruent with the user's values). Choosing the correct option in this case, may require the user to overcome their own bias.

We used the alignment scores of all options presented to the user to generate a more nuanced estimate of the amount of potential bias a student may be overcoming at each opportunity. This novel composite metric, which we call the *Bias Regulation Index* (BRI), is computed as follows:

$$BRI = (A_{highest} - A_{chosen}) + (A_{correct} - A_{chosen}) \tag{1}$$

Here $A_{highest}$ represents the alignment score of the option with the highest alignment score (i.e., the option we would expect a completely biased player to pick). Similarly, A_{chosen} represents the alignment score of the option the player chose, and $A_{correct}$ represents the alignment score of the correct option. The first set of parentheses in this equation essentially gives the player credit for choosing an option that isn't the option with the highest alignment, and gives them more credit the farther away their choice's score is from that highest score. This first set of parentheses cannot penalize players, as they cannot chose an option with a score higher than the highest score.

The second set of parentheses penalizes the player if they chose an option with a higher alignment score than the correct option. If $A_{correct} < A_{chosen}$, then the result of this second set of parenthesis is set equal to 0 to keep the metric from crediting the players for choosing an incorrect option with lower alignment than the correct option. Importantly, players are neither penalized nor credited in this metric for choosing the correct option. The resulting sum of these two sets of parentheses represents a student's ability to overcome bias to choose the correct answer. Positive scores on this metric capture those instances

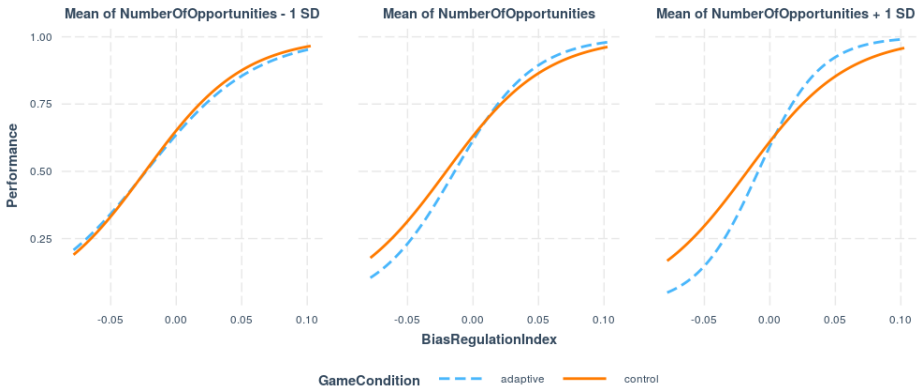

Fig. 2. The three-way interaction between condition, opportunities and *Bias Regulation Index* (BRI). We see that, while the relationship between BRI and performance remains relatively constant with additional practice in the control condition, the relationship seems to change in the adaptive condition. As the number of practice opportunities increase, students in the adaptive condition with low bias-regulation appear to do worse than their peers in the control condition, whereas students in the adaptive condition with high bias-regulation appear to benefit from the intervention compared to their peers in the control condition.

in which a student chooses the low-aligned correct score over the high-aligned incorrect one. Negative scores capture those instances in which the player chooses a high-aligned incorrect option over a lower-aligned correct one.

This metric is more nuanced than simply including the *Alignment* score of the correct option, as it mitigates the impact of the option's correctness on choice. That is, did the player choose this because it is the correct option, or because it aligned with their values. When the correct option is also the option with the highest alignment score, the choice is easy and uninteresting. In contrast, this metric focuses on instances in which the choice is difficult. We expect that bias regulation will improve over time for students in the adaptive intervention condition.

4 Results

We examined the impact of an intervention (designed to reduce bias) on in-game performance. Recall that students in the adaptive condition had an in-game experience identical to those in the control condition with one exception: during *Persuade* actions, students in the adaptive condition saw an additional piece of instruction that highlighted the option that most aligned with their values (i.e., had the highest computed *Alignment*) alongside a message warning the player that they may be biased to select the highlighted option.

4.1 Interaction Between Condition, BRI, and Number of Opportunities

We expected that the relationship between bias regulation and performance will be impacted by both experimental condition (i.e., the presence or absence of the intervention) and practice. To test for interactions between experimental condition, practice, and students' ability to regulate bias with respect to performance, we incorporated this new *Bias Regulation Index* (BRI) into the following hierarchical mixed effects model:

$$Outcome \sim BRI * PO * condition + (1|AP/Student) \tag{2}$$

Where *Outcome* represents the binary correctness score (0 = incorrect, 1 = correct) for the multiple choice problem, and *PO* (Prior Opportunities) represents the number of times, prior to the current opportunity, that the player has attempted a *Persuade* action. The model also includes the nested random effects of *AP* status[4] and the *Student* identifier. Table 4.1 shows the model results.

	Estimate	SE	P-val	Sig
PO	-0.015	0.009	0.070	.
BRI	23.339	4.954	0.000	***
Condition(control)	0.067	0.145	0.646	
PO:BRI	1.558	0.611	0.011	*
PO:Conditioncontrol	0.000	0.013	0.973	
BRI:Condition(control)	3.163	6.603	0.632	
PO:BRI:Condition(control)	-1.583	0.806	0.049	*

Table 1. Results from the hierarchical mixed effects model. There was a significant three way interaction between Prior Opportunities, Bias Regulation and Condition.

As expected, we found a significant three-way interaction between *Bias Regulation Index*, the number of prior practice opportunities, and experimental condition ($\beta = -1.583, p < .05$). We used the R library *interactions* to explore and visualize this interaction. Figure 2 shows the relationship between BRI and Performance at three different opportunity counts. We see that, while the relationship between performance and BRI remains relatively stable across practice opportunities in the control condition, the relationship between these variables changes with practice in the adaptive condition. Recall that BRI scores below zero indicate opportunities in which the student chose a high-aligned incorrect option over a low-aligned correct one, and positive scores indicate opportunities in which the student chose a low-aligned correct option over a high-aligned incorrect one. This graph suggests that the intervention may have caused students with low bias-regulation to perform worse (potentially choosing the visually salient orange-colored option more). However, students with high bias-regulation seemed to benefit from seeing the intervention, outperforming their peers in the control condition.

[4] AP Status was shown to be predictive of performance in previous work.

5 Discussion

We found that, by comparing estimates of student values to estimates of the values latent in text content, we could provide an adaptive intervention that appears to have reduced the impact of bias on task performance (for high bias-regulators). In the context the educational game, this effect appears to be gradual, increasing with additional practice opportunities. This suggests that regulating bias in this context is likely a skill that can be learned, but that it may also require many practice opportunities to hone.

To measure the intervention's impact, we developed a novel metric, the *Bias Regulation Index* (BRI). BRI more accurately captures the difficulty of bias-prone tasks, allowing us to measure a student's capacity to overcome (or regulate) their own biases. In the future, value-adaptive systems may use BRI to provide additional practice opportunities or individualized feedback to students exhibiting low bias-regulation.

Why the adaptive debiasing intervention had a differential impact on low and high bias-regulators remains an open question. While it may be tempting write this off as another example of the "rich get richer" effect that can occur in educational technology work, this would not explain the discrepancy in the performance of low bias-regulators across conditions. That is, the intervention seems to not only have made the rich richer, but the poor poorer as well. One potential explanation for this effect is a simple misunderstanding about the nature of the intervention. Low bias-regulators may have incorrectly interpreted the orange color as an indicator of an option's correctness (e.g., assumed it was a hint), when in fact, there was no such relationship between correctness and color. This may explain why low bias-regulators in the adaptive condition displayed worse performance than their peers in the control condition.

Such confusion may have been avoidable with additional instruction about the nature of the intervention. However, because students in the same class were randomly assigned to either the control or adaptive condition, drawing attention to the debiasing intervention (seen by those in the adaptive condition) may have tainted the independence of the control condition.

5.1 Limitations

Perhaps the largest limitation of this debiasing intervention is the absence of powerful social influences. It was unfortunately necessary to test the intervention at the individual level, separate from peer-influence simply because half the students within a classroom were assigned to the control group (no intervention). Thus, the instruction pertaining to bias was given to each student in the adaptive condition individually (via the interface). Future experiments might instead provide the bias instruction to the class as a whole, which might add social pressure to make unbiased decisions.

Other limitations pertain to our participant population and experiment structure. While we believe our sample was representative of late high school-aged students, there are known interactions between bias and age [12] that leave us

unable to confidently generalize these results to a population that includes older players. Similarly, an important part of debiasing research is the longevity of the effects [14]. As part of this work, we had originally planned to return to our participants' classrooms both one week and three weeks later to examine potential effects of the game on real-world classroom discussions. However, the onset of the COVID-19 pandemic cut our original data collection plan short. Both of these limitations are important areas of future work.

5.2 Potential Applications

This work has several potential applications. First and foremost, we believe that educational technologies that implement value-adaptive debiasing interventions allow instructors to provide students with opportunities to recognize and overcome their biases. These technology-based interventions will never be as nuanced as an intervention from an expert human instructor, but unlike traditional instruction, technology-based interventions like the one described in the current study are scalable to any number of students. Because of these tradeoffs, we see value-adaptive debiasing systems ultimately as a tool for supporting the critical, real-world classroom discussions.

Beyond the classroom, value-adaptive debiasing systems might be embedded into our interactions with media content. Here, such systems could make the content consumer aware of the degree to which the content they are consuming aligns with their own values. Alternatively, the system could alert the user to engage their critical thinking faculties when the alignment between the content's values and their own is above a certain threshold. What remains to be seen is how users will react to these kinds of interventions absent the affordances of game environments.

6 Conclusion

In this study, we examined the impact of a value-adaptive debiasing intervention on myside bias in the context of an educational game designed to teach productive civil discourse skills. We found a significant three-way interaction between the number of prior practice opportunities, our measure of bias (BRI), and condition (adaptive vs. control). Further investigation revealed that students in the adaptive condition (i.e., who saw the adaptive intervention) got better at mitigating the impact of bias with practice relative to their peers in the control condition. However, this was only true for high bias-regulators. While further improvements are necessary to ensure that the impact of debiasing interventions is equitable, this encouraging result demonstrates that value-adaptivity, this novel method of personalized learning, may be a useful tool for *scalable* debiasing instruction. Value-adaptivity allows us to craft instruction that recognizes and reacts to the dynamic relationship between the media content and the media consumer. With it, we can provide the rich, user-centered practice necessary for any comprehensive media literacy education.

References

1. Baron, J.: Thinking and deciding. Cambridge University Press (2000)
2. Diana, N., Stamper, J., Koedinger, K.: Towards value-adaptive instruction: A data-driven method for addressing bias in argument evaluation tasks. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–11 (2020)
3. Diana, N., Stamper, J.C., Koedinger, K.: Predicting bias in the evaluation of unlabeled political arguments. In: CogSci. pp. 1640–1646 (2019)
4. Evans, J.S.B., Barston, J.L., Pollard, P.: On the conflict between logic and belief in syllogistic reasoning. Memory & cognition **11**(3), 295–306 (1983)
5. Evans, J.S.B., Newstead, S., Allen, J., Pollard, P.: Debiasing by instruction: The case of belief bias. European Journal of Cognitive Psychology **6**(3), 263–285 (1994)
6. Franks, A.S., Scherr, K.C.: 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 (2015)
7. Garten, J., Hoover, J., Johnson, K.M., Boghrati, R., Iskiwitch, C., Dehghani, M.: Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. Behavior research methods **50**(1), 344–361 (2018)
8. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H.: Moral foundations theory: The pragmatic validity of moral pluralism. In: Advances in experimental social psychology, vol. 47, pp. 55–130. Elsevier (2013)
9. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological review **108**(4), 814 (2001)
10. Hone, B., Rice, J., Brown, C., Farley, M.: Factitious (2018), factitious.augamestudio.com
11. Johnson, D.W., Johnson, R.T., Tjosvold, D.: Constructive controversy: The value of intellectual opposition. (2000)
12. Klaczynski, P.A., Robinson, B.: Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning biases. Psychology and Aging **15**(3), 400 (2000)
13. Koleva, S.P., Graham, J., Iyer, R., Ditto, P.H., Haidt, J.: Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. Journal of Research in Personality **46**(2), 184–194 (2012)
14. Lilienfeld, S.O., Ammirati, R., Landfield, K.: Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? Perspectives on psychological science **4**(4), 390–398 (2009)
15. McGrew, S., Ortega, T., Breakstone, J., Wineburg, S.: The challenge that's bigger than fake news: Civic reasoning in a social media environment. American Educator **41**(3), 4 (2017)
16. Mercier, H.: Confirmation biasmyside bias. (2017)
17. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
18. Rottman, J., Kelemen, D., Young, L.: Tainting the soul: Purity concerns predict moral judgments of suicide. Cognition **130**(2), 217–226 (2014)
19. Stanovich, K.E., West, R.F., Toplak, M.E.: Myside bias, rational thinking, and intelligence. Current Directions in Psychological Science **22**(4), 259–264 (2013)
20. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J.: 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. ACM (2018)