

Of a Different Persuasion: Perception of Minority Status and Persuasive Impact

Logan Fields and Zaid Marji and John Licato

Department of Computer Science and Engineering
Advancing Machine and Human Reasoning (AMHR) Lab
University of South Florida

Abstract

Racial and gender bias, from advertisement to political rhetoric, is ubiquitous in persuasion. However, the impact of bias on persuasive discourse is often muddled by intent and framing. Reasoners practicing anti-racism may be more likely to scrutinize racially-specific arguments, while arguments made by women may only be diminished when they are emotionally charged. We sought to study how humans evaluate interpretive arguments, what makes certain arguments persuasive, and the impact of bias and emotionality on persuasiveness. We found that shallow heuristics such as argument length and readability are poor indicators for persuasive impact, but reasoners are more likely to be persuaded by arguments made by White people, particularly White women. Further, no difference was observed based on a reasoner's ability to see the arguer's face, implying that judgments are made solely by name recognition. Our focus on written arguments has broad implications for information literacy and racial justice.

Keywords: interpretive arguments, persuasion, racial bias, gender bias, emotionality

Introduction

As decision-making autonomy is increasingly delegated to artificially intelligent systems, we also require reliable means of ensuring such systems reason and act appropriately. In humans, we encourage moral rectitude by assigning laws and rules that use open-textured language, thereby allowing their exact interpretations to be delegated to the rational discretion of collective human reasoners. However, this relies on such discretion being free of bias and on the general agreement of the collective as to what is a rational interpretation. For example, a rule instructing a driver to keep to the right side of the road "as far as safely possible" may invite disagreement. Disagreements about the proper interpretation of open-textured terms in rules are often settled through interpretive arguments (Sartor, Walton, Macagno, & Rotolo, 2014). Unstructured argumentation, however, often invites the deleterious effects of cognitive bias (Kahneman, 2011; Stanovich & West, 2007). The impact of such biases on argument persuasiveness can be particularly difficult to disentangle and may vary from one reasoner to another. Thus, the exact role that bias plays on interpretive arguments and their persuasiveness is not fully understood.

As such, to fully determine the quality of an interpretive argument as a means of persuasion, we must consider two variables: (1) the judgment of the argument's persuasive impact relative to an alternative argument; and (2) the inter-rater reliability of said judgment, i.e., the agreement between

multiple reasoners about an argument's persuasiveness. A judgment indicating a lower average persuasiveness with high inter-rater reliability would suggest an objectively poorer argument. However, a judgment indicating a lower average persuasiveness with low inter-rater reliability may suggest polarizing impacts of bias. Therefore, in this paper, we aim to determine what factors affect the persuasiveness and inter-rater reliability of written interpretive arguments. Specifically, we set about to answer:

- RQ1** Do any shallow linguistic features influence interpretive argument persuasiveness?
- RQ2** Is there a significant difference in persuasiveness or inter-rater reliability based on an arguer's race or gender?
- RQ3** Is there a significant difference in persuasiveness or inter-rater reliability based on the ability to see an arguer's face?
- RQ4** Is there a significant difference in inter-rater reliability based on the incidence of emotional reactions to an argument?

Before we discuss each of these research questions in-depth, we will present the dialogue environment we use throughout our experiments.

Aporia: A Platform to Study Interpretive Reasoning

Aporia (Marji & Licato, 2021) is an argumentation dialogue environment designed to create structured datasets of opposing interpretive arguments. Players compete by arguing for or against certain interpretations of open-textured rules. The game is designed to be fun in order to incentivize willing participation and obtain useful datasets. *Aporia* is played in rounds by any group of three or more people. At the beginning of each round, two players are randomly chosen to argue against each other, and a third player is designated as a judge. The players are provided with an ethical rule for a given professional association and a scenario. For example, the rule "teachers need to act professionally with students" could be paired with the scenario "some teacher exchanges some light-hearted jokes with a student during recess." Would the teacher's action be considered "professional" in the sense meant by the rule?

Profession: professional economic developer

Description: A professional economic developer is responsible for planning, designing, and implementing economic development strategies, as well as acting as a key liaison between public and private sectors and the community.

Rule: Professional economic developers shall carry out their responsibilities in a manner to bring respect to the profession, the economic developer and the economic developer's constituencies.

Scenario: An economic developer decided to have a residential community rezoned to include commercial businesses, basing the decision on a survey given five years ago to residents.

The economic developer needs to take their responsibility seriously. Basing a decision on an outdated survey where no strong evidence exists that it is still relevant is a case of not going the extra mile to ensure the quality of their decision-making process.

John

Carrying out a survey is a costly endeavor. Claiming that an economic developer is ignoring their responsibilities to their constituents is an unsubstantiated claim, since they need to consider cost and time constraints, and factor in those elements in their decision-making. The economic developer in this situation has made a reasonable and justified judgement that basing their decision on a five years old survey is likely a reliable measure as demographics typically do not change rapidly.

Zaid

I believe that Lindsay's argument does not fully account all of the economic developer's considerations as explicated by John. I will judge in favor of John.

Show Scenario

Welcome Zaid,

The Judge is Zaid
 PLAYER 1 = Lindsay
 PLAYER 2 = John

Judging Phase

TIMER = 67 / 180

Figure 1: Screenshot of Aporia during the judge's turn.

After the players read the rule and the scenario, Player 1 decides if they wish to argue for or against the scenario complying with the rule, and they submit an argument for their interpretation. Player 2 then presents a counterargument by casting doubt on the validity of the first argument. Player 2 does not necessarily need to argue for the opposing view, only that Player 1's argument is invalid or incomplete. This caveat is designed to counter-balance Player 1's advantage in choosing which side they wish to argue, as it can be assumed that they will take the most easily defensible position. Finally, the judge decides if Player 1 has convincingly made their argument or if Player 2 has successfully invalidated it and announces the winner. The judge must also provide a brief justification for their judgment. Once a winner is announced, the round ends, and a new round begins. All actions, including reading the scenario and announcing the winner, are time-constrained to ensure the game proceeds in an orderly fashion. Figure 1 shows a partial screenshot of the Aporia interface.

The scenarios presented to the players were collected from a previous experiment (Licato, Marji, & Abraham, 2019) where participants were asked to write scenarios which were ambiguous with respect to specified open-textured terms within a given rule. While the use of interpretive arguments was not enforced for this experiment, the scenarios were chosen to encourage interpretive argumentation, as studying interpretive arguments is the primary motivation behind Aporia's design.

Can linguistic features predict persuasion?

Several studies have attempted to quantify the impact of linguistic features in isolation or in combination with other factors, and simple features have been shown to have measurable predictive power for persuasion (El Baff, Wachsmuth, Al Khatib, & Stein, 2020; Longpre, Durmus, & Cardie, 2019; Durmus & Cardie, 2018). Linguistic features may impact trustworthiness and competence, the main components of arguer credibility (McGinnies & Ward, 1980; McCroskey & Young, 1981), which in turn affects persuasion (Pennebaker, Boyd, Jordan, & Blackburn, 2015). For example, word count has been shown to be a reliable measure of trustworthiness (Larrimore, Jiang, Larrimore, Markowitz, & Gorski, 2011; Lucas, Stratou, Lieblich, & Gratch, 2016). Sophisticated language, as measured by average number of letters and argument readability, impacts engagement in conjunction with other features (Xu, Ellis, & Umphrey, 2019). The use of uncommon words might suggest technical competency but may also decrease readability.

Thus, we first set out to understand whether these surface-level linguistic and textual features have an effect on interpretive argument persuasiveness. If so, this might suggest that the factors that make certain arguments more persuasive than others can be reduced to simple patterns, shallow heuristics, or surface-level features in the arguments themselves. Furthermore, as this was the first time that Aporia was used in an experimental setting, it is worthwhile to establish its value as a platform for studying interpretive reasoning.

Method

Participants A total of 19 participants were recruited from Amazon Mechanical Turk (mTurk) to play Aporia using a web interface we developed. Each participant received \$30 for full participation. Participants were asked to provide their gender, race, occupation, and annual income. All demographic questions were optional.

Procedure All game sessions were announced 3 hours in advance on mTurk, and players were provided the scheduled session time. After signing up for the announced game session, players watched a 15-minute training video. They then answered qualification questions that we reviewed to ensure they understood the instructions clearly and were willing to make the expected effort. Players were informed once their qualifications were accepted and were reminded of the scheduled session time. Players were instructed to allocate 3 hours to play 12 rounds in a game session. In practice, most game sessions lasted less than 2 hours.

The game consisted of four phases. The first phase was the reading phase, where players were given 90 seconds to read the rule and the scenario for that round. In the second phase, Player 1 was given 120 seconds to announce the side they wished to argue for and provide their argument. In the third phase, Player 2 was given 30 seconds to read Player 1’s argument, and 120 seconds to deliver their counterargument. Finally, the fourth phase was the judging phase where the judge had 180 seconds to decide the winner and provide an explanation for their decision. If the time ran out in a player’s turn, whatever they had written was automatically submitted.

Results

We collected 48 complete game rounds. Each record includes the profession, rule, scenario, side chosen by Player 1, the first and second arguments, the winner, and the explanation provided by the judge. Table 1 shows some statistics of the collected dataset. For more statistics, including average words per sentence and average word length of the arguments, refer to (Marji & Licato, 2021).

Table 1: Some statistics of the collected dataset

	<i>Player 1</i>	<i>Player 2</i>	<i>Total</i>
<i>Argues For</i>	30	18	48
<i>Argues For and Wins</i>	22	9	31
<i>Argues Against and Wins</i>	9	8	17
<i>Total</i>	31	17	48

We analyzed the data for statistical features which may predict the winning argument. We used simple algorithms that do not require much data, such as support-vector machines, Naive Bayes, and decision trees. We calculated features such as the number of words, the average number of letters in each word, the Flesch reading score (Flesch, 1948) (among other

readability scores), Term Frequency - Inverse Document Frequency (TF-IDF) (K. S. Jones, 1972), and similar features. The results indicate no simple heuristics or obvious biases that may predict the winning argument aside from a preference for the first argument over the counterargument.

Can arguer’s race / gender impact persuasion?

Interpersonal biases may also influence persuasive discourse by impacting an arguer’s perceived credibility. However, such bias may not be readily apparent or inherently negative. Reasoners have been found to rate arguments as less persuasive when made by Black arguers, but only when the argument was also deemed to be extremely emotional or racial (Apfelbaum, Sommers, & Norton, 2008; Schultz & Maddox, 2013). Moreover, the persuasiveness of Black arguers is routinely diminished by poor argument quality (Schultz & Maddox, 2013) and improved when the arguments are counter-indicative of Black self-interests (Petty, Fleming, Priester, & Feinstein, 2001). However, some studies also suggest that non-biased reasoners are more likely to scrutinize arguments made by Black arguers than biased reasoners in an effort to reduce prejudice (Devine, Monteith, Zuwerink, & Elliot, 1991; Petty, Fleming, & White, 1999).

The effect of gender bias on persuasion is equally nuanced. Little gender gap is reported in persuasiveness or perceived competence of expert recommendations (Greve-Poulsen, Larsen, Pedersen, & Albæk, 2021). However, reasoners have been found to rate gender-based arguments as less persuasive when women make them than when men make them (Gervais & Hillard, 2014) and politically-charged arguments have a high variance in persuasiveness based on the arguer’s gender and the reasoner’s political ideology (Anderson-Nilsson & Clayton, 2021). These results indicate an unclear understanding of the actual impact of interpersonal bias on persuasion.

Therefore, we designed a study to determine whether an argument would be judged as less persuasive and whether the inter-rater reliability of the judgments would be lower if the argument were made by a woman or a Black person. We anticipate that minority arguers will be perceived as less persuasive overall but that the exact effects of interpersonal bias will be polarizing, as noted in the literature, leading to lower inter-rater reliability. Such a result would ordinarily be unexpected when presenting reasoners with the same argument. However, we hypothesize that both positive and negative implicit biases will skew the results.

Method

Participants A total of 158 participants were recruited from mTurk and asked to take one of four different surveys in Qualtrics. Each participant received either \$6 or \$12 for participation, depending on the length of the survey and the estimated time commitment. No demographic data on participants was collected.

Procedure Each participant was directed to complete a survey depicting ethical scenarios. Each scenario included a profession, an ethical rule, and an action performed by such a professional. The scenario then presented an initial argument, taken from the results of RQ1, either for or against the action following the ethical rule. The argument order was reversed from the player order in RQ1 where the argument structure allowed, i.e. where Player 2’s argument did not directly reference Player 1’s argument. Of 19 unique scenarios, 12 were presented in the original player order from RQ1 and 7 had the player order reversed. Participants were asked to judge the persuasiveness of the initial argument on a Likert-style scale (Likert, 1932) from 1 (very unconvincing) to 5 (very convincing). They were then presented with a counterargument and asked which argument they found more convincing overall. All four surveys were identical save for the inclusion of a fake name and profile icon to represent each arguer. This allowed us to compare participants’ judgments across racial and gender perceptions while keeping all else equal.

30 racially-specific names were selected by localizing the 100 most popular names for Black babies born between 2011 and 2016 in New York and Texas and eliminating names that were also in the 40 most popular names for White, Hispanic, or Asian babies. We also removed alternate spellings of the removed names (e.g., Chloe and Khloe), names that were in the top 100 for both boys and girls (e.g., Skylar and Taylor), and names that are diminutives of other names (e.g., Sam and Abby). The top eight boy names and eight girl names were then used. A converse method was used to select the top 14 White names.

Participants were not initially informed that the experiment considered racial and gender bias. Instead, the informed consent procedures stated that the experiment was considering how people judge the persuasiveness of arguments. This was intended to ensure participants’ honest responses despite the experiment’s sensitive nature. All participants were debriefed of the deception following the survey and given the option to withdraw their data without penalty. No participants withdrew from the experiment following the debrief.

Results

Across the 158 participants, we captured 4,112 persuasion ratings. We found two significant outliers in the persuasion ratings. The first was excluded for rating every argument as “very convincing.” The other participant’s ratings were significantly lower overall, but the variance and relative polarity of ratings was consistent with other participants. Therefore, we determined that this participant likely had a lower than average level of credulousness and their data was retained. There were no significant differences in persuasion across the remaining participants, indicating that individual variance in mood, confidence, or credulousness did not significantly impact the results.

Individuals’ average persuasion ratings ranged from 2.5 to 4.6 out of 5. As in RQ1, we found no significant indication towards the winning argument based on linguistic features.

However, we did note a significant preference for the first argument over the counterargument. We found this to be a direct result of the preference for Player 1’s argument reported in RQ1, although the result was weakened due to reversing some argument orders.

Figure 2 displays the average persuasiveness of each of the studied groups over all scenarios. The higher than average inter-rater reliability for White women could indicate that, in general, arguments made by White women were considered to be of higher quality than those made by other arguers. We will attempt to confirm this statistically.

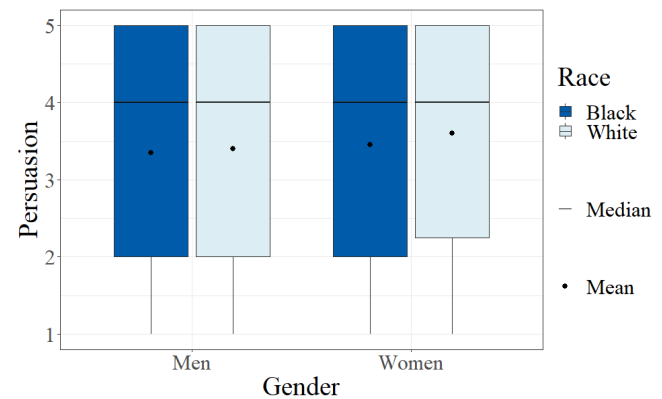


Figure 2: Persuasiveness of Studied Groups.

Statistical comparisons were performed in several ways. To determine whether the same argument would be more or less persuasive when made by a different arguer, we performed a paired Welch’s t-test for unequal variances (Welch, 1947). Next, we performed an unpaired Welch’s t-test to determine significant differences in the arguers’ general persuasiveness. For both paired and unpaired scenarios we calculated the coefficient of variation (CV) as a measure of inter-rater reliability and performed a z-test to determine whether the reliability was significantly different for different arguers. Finally, we performed a single-factor ANOVA (Fisher, 1925) on each scenario to determine whether certain scenarios had significant differences in persuasiveness when the arguers’ race or gender was changed. These results are given in Tables 2 and 3.

Table 2: Test Statistics for Paired Scenarios.

Test	Persuasion	CV
White Men : White Women	1.123	-1.3357
White Men : Black Men	-0.3576	0.0523
White Men : Black Women	-0.3481	-0.0493
White Women : Black Men	0.4542	-0.944
White Women : Black Women	1.5954	-1.2275
Black Men : Black Women	-3.5123**	2.9062**
White : Black	0.698	-0.9812
Men : Women	-1.5442	1.0794

* $P < .05$

** $P < .01$

Table 3: Test Statistics for Unpaired Scenarios.

Test	Persuasion	CV
White Men : White Women	-2.2306*	1.7499
White Men : Black Men	0.6415	-0.7606
White Men : Black Women	0.5385	-1.2858
White Women : Black Men	3.007**	-2.6144**
White Women : Black Women	1.6339	-3.1156**
Black Men : Black Women	-1.1921	-0.5767
White : Black	1.7509*	-2.752**
Men : Women	-2.4651**	0.7903

* $P < .05$

** $P < .01$

The average persuasion ratings for arguments made by Black men were significantly lower than for the same arguments made by Black women. Further, there was a significant difference in inter-rater reliability, implying an argument made by a Black man was considered systematically poorer than the same argument made by a Black woman. Additionally, an argument made by a White woman was considered slightly more persuasive than the same argument made by a Black woman, and an argument made by a woman was considered slightly more persuasive than the same argument made by a man. The latter may have been skewed by the preference for arguments made by Black women over Black men, but the preference for arguments made by women also increased significantly when considered in general, as opposed to across paired scenarios.

When considering unpaired scenarios, we also found participants considered arguments made by White people significantly more persuasive than those made by Black people, with a significantly lower CV. These findings hold when considering individual scenarios, as the most significant differences in persuasiveness and inter-rater reliability were in those scenarios comparing Black men with other arguers.







Can seeing an arguer’s face impact persuasion?

Although we can observe some impact of race and gender on persuasion, it is not immediately clear how reasoners classify arguers into demographic categories. Do they identify arguers based on visual cues, recognition of racially- or gender-specific names, or some combination therein? Previous research has indicated that reasoners seeing the arguer could increase the persuasive impact of poor arguments (Heim, Asting, & Schliemann, 2002). However, this was attributed to the social obligation of an arguer and a reasoner perceiving one another in a verbal exchange, as opposed to a reasoner independently judging a written argument. Further, no consideration was made for codependent biases.

Therefore, to establish the impact of simply seeing an arguer’s face on the persuasiveness of an argument, we utilized fake profile icons. Some icons depicted a face matching the race and gender implied by the arguer’s name, while others contained images of cats or artwork. Table 4 displays a sub-

set of the selected names and icons used. As the experiment in RQ2 utilized racially- and gender-specific names, we do not anticipate that seeing an arguer’s face will have a strong impact on persuasion but, rather, that reasoners will apply any presuppositions based on name-recognition.

Table 4: Subset of names and icons used for RQ3.

Men			Women		
					
Amir	Isaiah	Ethan	London	Nevaeh	Amelia

Procedure

Fake profile icons of faces, art, and cats were sourced from the StyleGAN2 open-source generator (Karras et al., 2020). No images of real people were used.

Results

We again performed a Welch’s t-test and calculated the CV for paired and unpaired scenarios. We found no significant differences in persuasion or inter-rater reliability for faces versus other profile icons. This result holds across race and gender variables.

Can emotionality impact inter-rater reliability?

Finally, emotionality in the reasoner may impact persuasion, both individually (Nabi, 2007; Petty & Briñol, 2015) and through the amplification of cognitive bias in argumentative environments. However, emotional discourse has also been found to increase engagement (Ksiazek, 2016; Villata, Benlamine, Cabrio, Frasson, & Gandon, 2018) and reduce the impact of peripheral biases on persuasive outcome (Petty, Cacioppo, & Goldman, 1981). As such, we also sought to determine how the interaction between bias and emotion affects persuasiveness. Although previous literature indicates that peripheral features are not relevant to emotional arguments (Petty et al., 1981), we actually anticipate that features such as race will exacerbate emotionality in reasoners.

Procedure

Of the 158 participants in RQ2, 58 were assigned to a pilot group and asked to review 10 randomly selected scenarios. The other 100 participants were asked to review 15 scenarios specifically selected as being either minimally or highly emotional. There was some overlap between scenarios and, as such, six scenarios were presented to all participants, and 19 unique scenarios were tested overall. The participants in the pilot group were split into a further two groups of 29; one group was asked to provide a 1-2 sentence justification for their judgments, and the other group was not. All participants in the main group were asked to provide justifications. There was no significant difference in persuasion ratings based on the presence of justifications.

Results

Across the 158 participants, we captured 3,532 written justifications. The participant whose data was removed for being an outlier was not in the group providing justifications and, therefore, was not relevant to this experiment. We first categorized scenarios as having high or low expected emotionality based on the level of controversy or conflict in the scenario. We then calculated the sample standard deviations and CV values for each scenario. Finally, we used a single-factor ANOVA to determine if there were significant differences in the CV values based on the emotionality categorizations. We found no significant differences in reliability based on emotionality, namely because our initial assumptions on which arguments participants would find emotional were mistaken. Of the eight scenarios put in place to evoke emotional responses, only five showed high levels of emotionality in the experiment. Additionally, two other scenarios meant to produce low levels of emotionality and two random scenarios from the pilot group evoked highly emotional responses.

As such, we re-categorized the scenarios based on the participants' justifications. Previous research indicates six main emotion categories: anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1993). The nature of the arguments used largely offset the presentation of happiness and surprise in the justifications. Therefore, we attempted to identify syntactic features which may indicate the remaining emotional features. The final categorization for "emotional" responses was: (1) Ad hominem attacks on the professional, the arguers, or us, as the experimenters; (2) Use of capital letters, exclamation points, or profanity to emphasize a personal opinion; (3) Direct appeals to fear, pity, empathy, or harmful consequences. Further, as participants were given an argument and a counterargument for each scenario, if either justification met these requirements, the entire scenario was categorized as having yielded an emotional response.

We averaged the number of emotional responses across each scenario and labeled the result as the *emotionality index* (E). Scenarios that had an overall emotionality index over 0.35, that is at least 35% of participants responded emotionally, were deemed as being highly emotional. This threshold was chosen as the approximate mean and median of the collected emotionality indices. Upon the recategorization, we found that scenarios that were considered highly emotional had significantly lower inter-rater agreement ($P < .001$) and, in fact, were significantly polarized in persuasiveness. Additionally, the converse holds and there is a high correlation ($P < .001$) between emotionality and inter-rater reliability. We found no significant differences in racial and gender biases based on a scenario's emotionality.

General Discussion and Future Work

This paper described an initial attempt to empirically study the evaluation of interpretive arguments and the impact of race and gender bias, as well as reasoner emotionality, on such arguments' persuasiveness. Although the Aporia

dataset is relatively small and cannot be used to reliably train transformer-based neural networks, we were able to analyze its statistical properties. We found no surface-level linguistic or textual features that may predict winning arguments from the text alone. However, our results suggest that a significant preference is given to arguments when perceived to have been made by White people or by women. The latter effect is interesting but not wholly unexpected given the ambiguous results of previous studies concerning gender and perceived competence.

Although we collected participant demographics in our initial experiment using Aporia, we elected not to for the subsequent studies to encourage participants to give their free and honest opinions. This removed our ability to consider in-group bias. However, we likely would not have had a statistically relevant sample size to control for the reasoner's race as the majority of players in Aporia self-identified as White or Asian. Further, we did not control for perceived trustworthiness or competence, the ratio of which is a significant determining factor in the presentation of in-group bias (E. Jones, Moore, Stanaland, & Wyatt, 1998; Khatib, 1989; Spence, Lachlan, Westerman, & Spates, 2013).

We found that initial arguments were highly more likely to be considered persuasive than their subsequent counterarguments. This is likely an artifact of Aporia's design, in that Player 1 was able to choose which side of the argument they wished to defend and was feasibly inclined to select the most easily defensible argument. This is clear in RQ2, where the preference for the initial argument is weakened by reversing the player order. However, in RQ2, there remained a strong preference for arguments made by Player 1 ($P < .001$), irrespective of argument order.

We also found that emotional reactions significantly impacted how arguments were received. Emotion-inducing scenarios had significantly lower and more polarized inter-rater agreements. However, we found no significant differences in racial and gender biases based on a scenario's emotionality. This is contrary to previous understanding that peripheral features only become relevant to arguments of low personal relevance (Petty et al., 1981), which would imply that less emotional scenarios would see higher presentations of bias. However, it is also contrary to our prediction that racial and gender features would exacerbate emotionality, leading to a positive interaction. This discrepancy may be worthy of further study.

Understanding the impact of racial and gender bias on interpretive arguments, particularly those that are written, has future implications for reducing misinformation in the media, minimizing unequal treatment of minority populations in the legal system, and mitigating bias in the acceptance of academic literature. Further, identifying the possibly unconscious cognitive biases that affect persuasiveness may improve our understanding of how the human mind processes external stimuli to influence decision-making.

Acknowledgments

This material is based upon work supported by the USF Grant Program for Understanding and Addressing Blackness and Anti-Black Racism in our Local, National, and International Communities and the Air Force Office of Scientific Research under award numbers FA9550-17-1-0191 and FA9550-18-1-0052. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of South Florida or the United States Air Force.

We thank the anonymous reviewers for their careful consideration and many insightful comments and suggestions which helped improve and clarify our paper.

References

- Anderson-Nilsson, G., & Clayton, A. (2021). Gender and policy persuasion. *Political Science Research and Methods*, 9(4), 818-831.
- Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology*, 95(4), 918-932.
- Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology*, 60(6), 817-830.
- Durmus, E., & Cardie, C. (2018). Exploring the role of prior beliefs for argument persuasion. In *Proceedings of NAACL-HLT 2018* (p. 1035-1045). New Orleans, LA: Association for Computational Linguistics.
- Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48(4), 384-392.
- El Baff, R., Wachsmuth, H., Al Khatib, K., & Stein, B. (2020). Analyzing the persuasive effect of style in news editorial argumentation. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (p. 3154-3160). Association for Computational Linguistics.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221-233.
- Gervais, S. J., & Hillard, A. L. (2014). Confronting sexism as persuasion: Effects of a confrontation's recipient, source, message, and context. *Journal of Social Issues*, 70(4), 653-667.
- Greve-Poulsen, K., Larsen, F. K., Pedersen, R. T., & Albæk, E. (2021). No gender bias in audience perceptions of male and female experts in the news: Equally competent and persuasive. *The International Journal of Press/Politics*.
- Heim, J., Astring, T., & Schliemann, T. (2002). Medium effects on persuasion. In *Proceedings of the second Nordic conference on human-computer interaction* (p. 259-261). Aarhus, Denmark: Association for Computing Machinery.
- Jones, E., Moore, J. N., Stanaland, A. J. S., & Wyatt, R. A. J. (1998). Salesperson race and gender and the access and legitimacy paradigm: Does difference make a difference? *The Journal of Personal Selling & Sales Management*, 18(4), 71-88.
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60(5), 493-502.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Girous.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition* (p. 8107-8116). Institute of Electrical and Electronics Engineers.
- Khatib, S. M. (1989). Race and credibility in persuasive communications. *Journal of Black Studies*, 19(3), 361-373.
- Ksiazek, T. B. (2016). Commenting on the news: Explaining the degree and quality of user comments on news websites. *Journalism Studies*, 19(5), 650-673.
- Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, 39(1), 19-37.
- Licato, J., Marji, Z., & Abraham, S. (2019). Scenarios and recommendations for ethical interpretive AI. In *Proceedings of the AAAI 2019 fall symposium on human-centered AI*. Arlington, VA: Association for the Advancement of Artificial Intelligence.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 5-55.
- Longpre, L., Durmus, E., & Cardie, C. (2019). Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th workshop on argument mining* (p. 167-176). Florence, Italy: Association for Computational Linguistics.
- Lucas, G., Stratou, G., Lieblch, S., & Gratch, J. (2016). Trust me: Multimodal signals of trustworthiness. In *Proceedings of the 18th ACM international conference on multimodal interaction* (p. 5-12). Tokyo, Japan: Association for Computing Machinery.
- Marji, Z., & Licato, J. (2021). Aporia: The argumentation game. In *Proceedings of the third workshop on argument strength*.
- McCroskey, J. C., & Young, T. J. (1981). Ethos and credibility: The construct and its measurement after three decades. *Central States Speech Journal*, 32(1), 24-34.
- McGinnies, E., & Ward, C. D. (1980). Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin*, 6(3), 467-472.
- Nabi, R. L. (2007). Emotion and persuasion: A social cognitive perspective. In D. R. Roskos-Ewoldsen & J. L. Monahan (Eds.), *Communication and social cognition: Theories and methods* (p. 377-398). New York: Routledge.

- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015 [Computer software manual]. Austin, TX.
- Petty, R. E., & Briñol, P. (2015). Emotion and persuasion: Cognitive and meta-cognitive processes impact attitudes. *Cognition and Emotion*, 29(1), 1-26.
- Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41(5), 847-855.
- Petty, R. E., Fleming, M. A., Priester, J. R., & Feinstein, A. H. (2001). Individual versus group interest violation: Surprise as a determinant of argument scrutiny and persuasion. *Social Cognition*, 19(4), 418-442.
- Petty, R. E., Fleming, M. A., & White, P. H. (1999). Stigmatized sources and persuasion: Prejudice as a determinant of argument scrutiny. *Journal of Personality and Social Psychology*, 76(1), 19-34.
- Sartor, G., Walton, D., Macagno, F., & Rotolo, A. (2014). Argumentation schemes for statutory interpretation: A logical analysis. In *Legal knowledge and information systems. (proceedings of jurix 14)* (p. 21-28).
- Schultz, J. R., & Maddox, K. B. (2013). Shooting the messenger to spite the message? exploring reactions to claims of racial bias. *Personality and Social Psychology Bulletin*, 39(3), 346-358.
- Spence, P. R., Lachlan, K. A., Westerman, D., & Spates, S. A. (2013). Where the gates matter less: Ethnicity and perceived source credibility in social media health messages. *Howard Journal of Communications*, 24(1), 1-16.
- Stanovich, K. E., & West, R. F. (2007). Natural Myside bias is independent of cognitive ability. *Thinking & Reasoning*, 13(3), 225-247.
- Villata, S., Benlamine, S., Cabrio, E., Frasson, C., & Gandon, F. (2018). Assessing persuasion in argumentation through emotions and mental states. In K. Brawner & V. Rus (Eds.), *Proceedings of the thirty-first international Florida artificial intelligence research society conference* (p. 134-139). Melbourne, FL: Association for the Advancement of Artificial Intelligence.
- Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1), 28-35.
- Xu, Z., Ellis, L., & Umphrey, L. R. (2019). The easier the better? comparing the readability and engagement of online pro- and anti-vaccination articles. *Health Education & Behavior*, 46(5), 790-797.