Online Assessment of Belief Biases and their Impact on the Acceptance of Fallacious Reasoning

Nicholas Diana[1], John Stamper[1], and Kenneth Koedinger[1]

Carnegie Mellon University
5000 Forbes Ave, Pittsburgh PA 15213, USA
ndiana@cmu.com, john@stamper.org, koedinger@cmu.edu

Abstract. Determining the impact of belief bias on everyday reasoning is critical for understanding how our beliefs can influence how we judge arguments. We examined the impact of belief bias on the user's ability to identify logical fallacies in political arguments. We found that participants had more difficulty identifying logical fallacies in arguments that aligned with their own political beliefs. Interestingly, this effect diminishes with practice. These results suggest that while belief bias is a potential barrier to correctly evaluating everyday arguments, interventions focused on activating rational engagement may mitigate its impact.

Keywords: Belief Bias · Informal Reasoning · Informal Logic · Logical Fallacies.

1 Introduction

For decades, research in formal logic has demonstrated that our prior-knowledge or beliefs can interfere with our ability to reason logically about an argument [2, 8, 4, 7, 5, 6]. This phenomenon, known as *Belief Bias*, is precisely defined as a tendency to "base [our] judgments on the believability of the conclusions" rather than the "logical form of the arguments" [8].

We used LIFTS (Logic and Informal Fallacy Tutoring System) [1] to test the impact of belief bias on one's ability to identify informal logical fallacies. Inside the tutor, we asked participants to identify informal fallacies in short arguments, given some context. Some of these arguments were political in nature, with conclusions designed to support either a typical conservative or typical liberal position on an issue. We hypothesized that participants would have more difficulty identifying fallacies in problems with conclusions that align with their own political beliefs (i.e., conclusions they may agree with), despite the fact that all of the arguments presented were fallacious.

2 Methods

Sixty-three participants were recruited for the experiment. In order to select a politically diverse sample, subjects were recruited using Amazon Mechanical

Turk with the restriction that they must reside in the United States. To mitigate concerns about data quality, we analyzed the log data to find participants who appeared to be "gaming" the tutor. Participants who provided an answer less than a second after seeing a problem were classified as "gamers" and excluded. Six participants were excluded for gaming, two for possessing clear outlier values (values above the 99.9th percentile) on the outcome variable (number of errors), and one for having a large time gap (more than an hour) between actions. Of the remaining 54 participants, 23 identified as female, 30 as male, and 1 as agender. The average age of participants was 31.31 years old (SD=6.67).

In this experiment, participants were asked to identify the fallacy present in an argument from a list of 3 different fallacies. The tutor consisted of 18 problems total, with each of the 3 fallacies being the correct answer 6 times. Problems were presented in a random order. Of the 18 problems, 9 were designed to contain a conservative conclusion and 9 were designed to contain a liberal conclusion. All of the arguments presented contained an informal logical fallacy, but we expected that participants would have more difficulty identifying the fallacy when their personal political orientation matched (or aligned with) the political orientation of the problem. Instruction was provided in expandable drop-down boxes that contained definitions and examples of the fallacies. After completing the tutor, participants were asked to complete a post-test questionnaire that included three questions that directly assessed beliefs about the specific political issues used in the study, and general demographics questions.

3 Results

We built a mixed linear regression model with *number of errors* as the outcome variable and *participant* as a random effect. Input into the model was a participant-by-problem table, such that each row represented one participant's performance on one problem. The mixed linear regression included the following as features: **Prior Opportunities at Fallacy (oppFallacy)** represents the number of times the participant has seen the fallacy in a question before. If learning occurs over the course of the experiment, we expect this feature's coefficient to be negative (i.e., inversely related to *number of errors*). **Prior Opportunities at Orientation (oppOrientation)** represents the number of times the participant has seen a problem with this political orientation (i.e., with a conservative or liberal conclusion) before. We do not expect this feature to be a significant predictor outside of an interaction. In other words, by itself, orientation should not add any difficulty to the problem. **Alignment** represents the degree to which the participant's political beliefs (as measured using the direct questions discussed above) align with the political orientation of the problem. We expect this feature to be a significant, positive predictor of *number of errors* outside of an interaction. **oppOrientation*Alignment** represents the interaction between the number of prior opportunities at an orientation and the degree to which the participant's political beliefs align with that orientation. It may be the case that belief bias has a strong effect on performance at the beginning

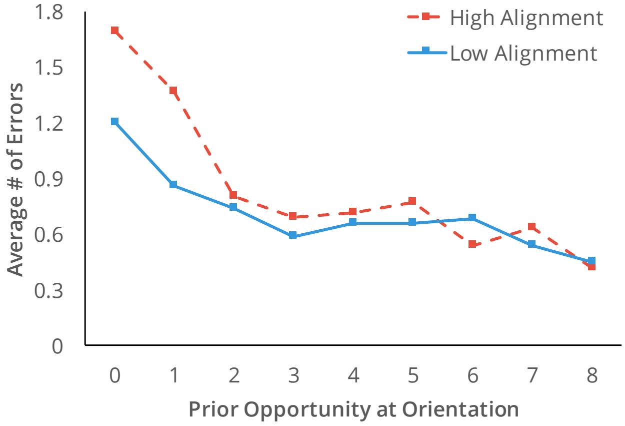

Fig. 1. In general, LIFTS succeeded at teaching fallacy identification (i.e., users made fewer errors with practice). We also see that users were biased by their beliefs, as evidenced by the differences between early performance on high alignment and low alignment problems. Alignment group was determined by score (1-5) on the self-reported alignment questions (High = 4 or 5, Low = 1 or 2). These results suggest that belief bias impacted performance early in the experiment, but that the effect diminished with practice. At least in this case, LIFTS was successful in reducing belief bias.

of the experiment, but as participants continue to practice identifying fallacies in arguments that align with their beliefs, the impact of belief bias diminishes. This term was designed to capture that interaction.

As predicted, the number of *Prior Opportunities at Fallacy* was a significant predictor ($\beta = -0.102, p < .05$) and inversely related to *number of errors*. As participants had more opportunities practicing a fallacy, their performance improved. In other words, learning occurred inside LIFTS.

We also found that the interaction between the number of *Prior Opportunities at Orientation* and the participant's *Alignment* to that orientation was a significant, negative predictor ($\beta = -0.021, p < .05$). If we plot this interaction (see Figure 1), it appears that belief bias impacts performance early in the experiment, but that the effect diminishes with practice. This interpretation is supported by the coefficient for *Alignment* by itself ($\beta = 0.124, p < .05$). We see that higher *Alignment* is associated with worse performance (i.e., higher *number of errors*) when *oppOrientation* is 0. Outside of the interaction, *oppOrientation* was not a significant predictor.

4 Discussion

There are two possible interpretations for the diminishing impact of belief bias with practice. First, it is possible that an improved understanding of the logical fallacies makes the fallacious features of an argument more salient. If this is the case, then reducing belief bias is a matter of better training in argument evaluation. However, it is also possible that it's not learning that is reducing belief bias, but rather that some typically dormant critical thinking faculties are coming online (as the task requires them) and overpowering the influence of belief bias. This interpretation seems to support the main assertion of Haidt's Social Intuitionist Model of moral reasoning [3], which argues that everyday moral reasoning happens quickly and is primarily based on intuitions (as opposed to a rational assessment of the argument). Rationalization enters into the model *after* a moral decision has been reached, to justify the decision (or conversely, to undermine an opposing position). With respect to the current experiment, it is possible that the belief bias effect seen early in the experiment is evidence of an intuitions-based moral reasoning, and performance improves as participants discover that the task requires rational reasoning. If this interpretation is correct, then performance on the earlier problems is representative of how we typically evaluate everyday arguments (i.e., in the absence of heightened critical thinking). Moreover, the difference observed between the *High Alignment* and *Low Alignment* groups on these early problems suggests that being susceptible to belief bias may be the typical case.

If this second hypothesis is true, then mitigating belief bias in everyday reasoning may not simply be a matter of better training in argument evaluation. Instead, systems designed to combat our susceptibility to weak arguments or misleading news stories should place a greater emphasis on understanding the user's beliefs and how those beliefs 1) relate to the beliefs present in the content they are consuming, and 2) impact their judgment of that content's validity.

5 Conclusion

We demonstrated that a participant's political beliefs impacted their ability to identify logical fallacies in arguments that aligned with those political beliefs. The larger impact of belief bias on earlier problems may be evidence of an intuitionist model of moral reasoning. As such, while our results suggest that the key to overcoming belief bias may be to simply think more critically about the argument in question, they also imply that we naturally forgo this critical evaluation when we agree with the argument. Combating the negative effects of belief bias in real-world contexts such as advertising and politics may benefit from some external agent that can relate the user's values to the values latent in the text, and prime us to think critically about invalid arguments that may intuitively seem true.

References

1. Diana, N., Stamper, J., Koedinger, K.: An instructional factors analysis of an online logical fallacy tutoring system. In: Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B. (eds.) Artificial Intelligence in Education. pp. 86–97. Springer International Publishing, Cham (2018)

2. Evans, J.S., Barston, J.L., Pollard, P.: On the conflict between logic and belief in syllogistic reasoning. Memory & cognition **11**(3), 295–306 (1983). https://doi.org/10.3758/BF03196976
3. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological review **108**(4), 814 (2001)
4. Henle, M., Michael, M.: The Influence of Attitudes on Syllogistic Reasoning. The Journal of Social Psychology **44**(1), 115–127 (1956). https://doi.org/10.1080/00224545.1956.9921907
5. Lefford, A.: The influence of emotonal subject matter on logical reasoning. The Journal of General Psychology **34**(July 2016), 127–151 (1946). https://doi.org/10.1080/00221309.1946.10544530, http://www.tandfonline.com/doi/pdf/10.1080/00221309.1946.10544530
6. Markovits, H., Nantel, G.: The belief-bias effect in the production and evaluation of logical conclusions. Memory & Cognition **17**(1), 11–17 (1989). https://doi.org/10.3758/BF03199552
7. Morgan, J.J.B., Morton, J.T.: The distortion of syllogistic reasoning produced by personal convictions. Journal of Social Psychology **20**(1), 39–59 (1944). https://doi.org/10.1080/00224545.1944.9918830
8. Revlin, R., Leirer, V., Yopp, H., Yopp, R.: The belief-bias effect in formal reasoning: the influence of knowledge on logic. Memory & cognition **8**(6), 584–592 (1980). https://doi.org/10.3758/BF03213778