Biased Misinformation Distorts Beliefs

Rani Moran (r.moran@qmul.ac.uk)
School of Biological and Behavioural Sciences (Queen Mary University of London),
Mile End Road London, E1 4NS, United Kingdom

Juan Vidal-Perez (juan.perez.21@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research,
10-12 Russell Square London, WC1B 5EH, United Kingdom

Raymond J. Dolan (r.dolan@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research,
10-12 Russell Square London, WC1B 5EH, United Kingdom

Abstract

Misinformation, particularly biased news, poses a growing threat to open societies by driving polarization and reinforcing false beliefs. This study explores how individuals process biased information through a reinforcement learning task. Participants (n=200) took part in a "bandit task," receiving feedback from biased (favorable, unfavorable) and unbiased sources. They first learned about the biases of these sources, then used this knowledge to adjust their belief updating. Although participants could identify and account for bias, their corrections were incomplete, leaving residual distortions in their beliefs. Exposure to biased sources also led participants to perceive unbiased sources as biased. These results underscore the difficulty of maintaining accurate beliefs in biased environments and offer strategies for combating misinformation.

Keywords: misinformation, reinforcement learning, bias

The widespread presence of biased information severely undermines informed societies, fueling polarization and the persistence of misbelief (Enders et al., 2023; Ruan et al., 2021). However, because biased information deviates from the truth in systematic and predictable ways it can, in principle, be corrected. For instance, the weight readings of a biased scale that consistently adds 10 kg to a person's weight can be corrected by simply subtracting 10 kg. Given the prevalence of biased information, a critical question emerges: Can individuals effectively learn and compensate for biases in information to form accurate beliefs?

Task design

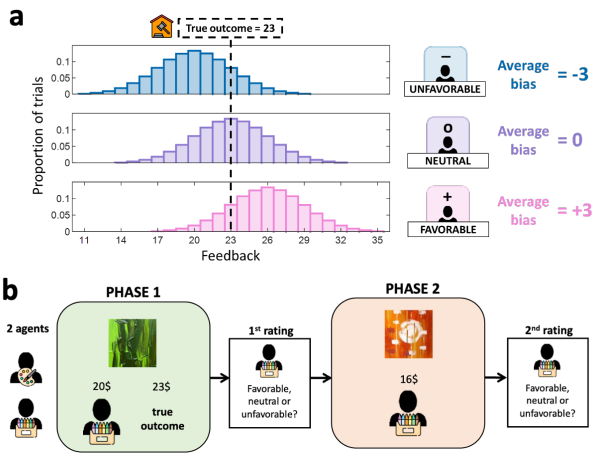

Figure 1: Task design

In this study, 200 participants played a “bandit task” under the cover story of managing a virtual art gallery. On each trial participants chose between two paintings of different average value. The chosen painting was sold for a variable price accumulated on behalf of participants earnings (true outcome). When a painting was selected, an agency provided an estimated selling price for its copy (agency feedback). These agencies were either favorably biased (overestimating on average by $3), unfavorably biased (underestimating on average by $3), or neutral

(unbiased) (Fig. 1a). Participants' bonuses were based on true outcomes, not agency feedback.

The experiment had six "superblocks," each featuring two agencies with different biases. Each superblock had two phases (Fig. 1b). In the first phase, participants saw both the agency's estimate and the true selling price, allowing them to learn each agency's bias. In the second phase (new paintings same agents), only the agency estimates were provided, requiring participants to use their learned bias knowledge to infer true values and make their painting choices.

Results

We used a reinforcement learning model that included a "debias parameter" for each information source. This parameter is subtracted, during phase 2, from the feedback provided by the source to produce "debiased feedback", which is then used to update beliefs about the painting values. Fitting this model to participants' choices revealed that corrections for bias were consistently incomplete (Fig. 2). Unfavorable feedback was adjusted upward by less than the optimal amount (-0.95 vs. the optimal -3), and favorable feedback was adjusted downward insufficiently as well (+1.90 vs. the optimal +3). Such insufficient corrections allowed residual biases to shape beliefs and decisions (e.g., more choice repetition following interaction with a favourable compared to a neutral agent).

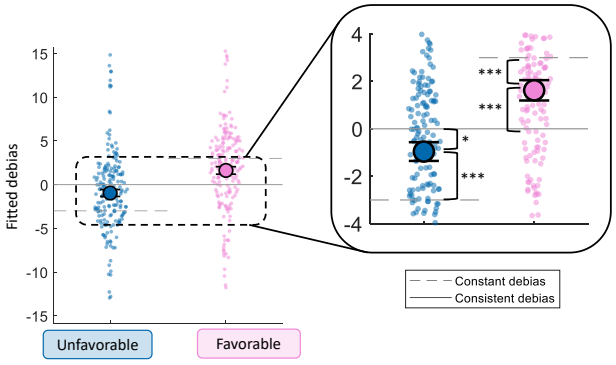

Figure 2: Corrections applied to biased sources

We asked participants to explicitly classify each agency as favourable, neutral, or unfavourable, following each Phase (Fig. 1b). Classification accuracy was above chance but decreased from Phase 1 to Phase 2 (Fig. 3a). Notably, neutral sources were more likely to be misclassified as having the opposite bias of their paired agent in the superblock (e.g., labelled as unfavourable when paired with a favourable agent). This tendency was present in the rating after Phase 2 (when only biased feedback was available), but not in the one after phase 1 (when ground truth was also available) (Fig. 3b).

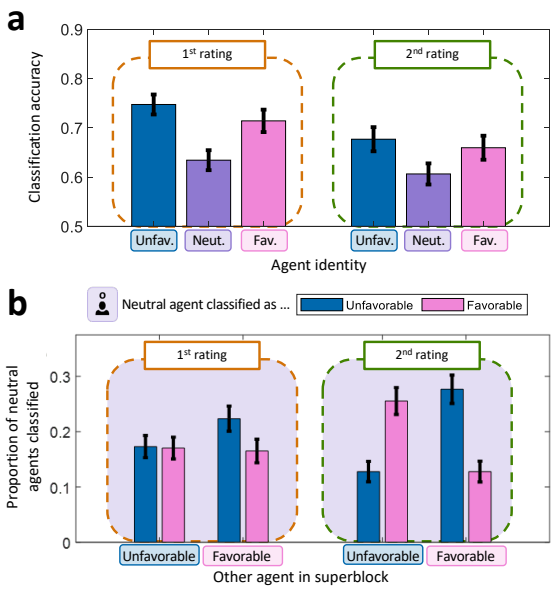

Figure 3: Classification of neutral sources.

Conclusion

Our findings show that individuals attempt to identify and correct for source biases. However, they under-correct for such biases, resulting in a shift in beliefs in the direction of the biased source, and perceive neutral sources as being oppositely biased. These processes persisted even with simple additive biases that were easily correctable, suggesting fundamental cognitive limitations are at play. In more complex real-world environments with subtler biases and less access to ground truth, these effects are likely to be even more pronounced, contributing significantly to the formation of biased belief.

References

Enders, A. M., Uscinski, J. E., Seelig, M. I., Klofstad, C. A., Wuchty, S., Funchion, J. R., Murthi, M. N., Premaratne, K., & Stoler, J. (2023). The Relationship Between Social Media Use and Beliefs in Conspiracy Theories and Misinformation. *Political Behavior*, *45*(2), 781–804. https://doi.org/10.1007/s11109-021-09734-6

Ruan, Q., Mac Namee, B., & Dong, R. (2021). Bias Bubbles: Using Semi-Supervised Learning to Measure How Many Biased News Articles Are Around Us. *AICS*, 153–164. https://ceur-ws.org/Vol-3105/paper40.pdf