Trust Calibration in Non-Dyadic Human-AI Teams

SANDER DE JONG, JOEL WESTER, and NIELS VAN BERKEL, Aalborg University, Denmark

This planned study aims to explore trust calibration in non-dyadic human-AI teams. We plan a mixed-method lab study to shed light on human-AI teaming beyond single human interaction. Both individuals and teams of three members solve sentiment analysis tasks, assisted by AI suggestions and explanations. In an attempt to reduce anchoring and group biases, half the participants are forced to engage with explanations for the alternatives before locking in their answers. The intended contributions are threefold. Analyzing group deliberation provides insights into the thought process behind AI-assisted decision-making and how tasks are distributed amongst a team performing this task. Comparing overreliance on AI suggestions between individuals and groups highlights whether teaming is beneficial for solving AI-supported decision tasks. Finally, measuring the effect of cognitive forcing on teams' reliance on AI systems can inspire directions to alleviate potential anchoring and group biases.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: human-AI teams, artificial intelligence, trust, collaboration

ACM Reference Format:
Sander de Jong, Joel Wester, and Niels van Berkel. 2023. Trust Calibration in Non-Dyadic Human-AI Teams. In . ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Depending on context and task, Artificial Intelligence (AI) systems have been shown to outperform humans in a range of decision-making tasks. On the other hand, numerous reports on the errors made by AI systems highlight the risk of excluding humans from the decision-making process. Integrating AI systems in a decision-support role in the real world requires that end-users place appropriate reliance on these systems [21]. To achieve complimentary human-AI team performance [1], the combination of humans and AI must outperform either party acting alone. To calibrate users' trust in AI systems, many contemporary solutions communicate confidence levels [30] and explanations [25]. While this has been shown to increase trust in AI systems, this can also result in overreliance [5, 13]. Prior work highlights that when humans are insufficiently able to calibrate their trust when presented with explanations, AI systems alone can outperform human-AI teams [1].

Recent attempts to improve trust calibration use cognitive forcing functions to make participants reflect on their decisions and reduce overreliance [18], albeit not enough to achieve complimentary human-AI team performance [4, 24, 28]. Cognitive forcing functions reduce overreliance more for those with high levels of need for cognition [4] (i.e., the tendency to engage in and enjoy cognitively effortful tasks [6]). Vasconcelos et al. argue that the effect of explanations on overreliance depends on the cost-benefit analysis made by humans, weighing costs (e.g., task and explanation difficulty) against benefits (e.g., monetary compensation) [28].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Association for Computing Machinery.
Manuscript submitted to ACM

Human-AI collaboration research has thus far focused on dyadic human-AI relationships. As groups have more problem-solving resources than individuals, they may be able to achieve better decision outcomes [11]. Moreover, we argue that group deliberation can be seen as cognitive forcing, as arguing for a particular answer out loud with group members makes people reflect on their thought processes, reducing overreliance. However, the literature on human teaming shows that collaborative decision-making can also introduce new biases (e.g., false consensus, groupthink) [16].

To investigate the role of group-based collaboration on Human-AI decision making, we outline the design for a mixed-method study in which human-AI teams solve a sentiment analysis task while being assisted by an AI system. Participants are placed into either a dyadic (one human, one AI system) or non-dyadic (three humans and an AI system) team to assess how group deliberation affects overreliance. Additionally, participants are challenged to consider the opposite [23] attempting to reduce anchoring bias and 'groupthink' [14] (i.e., making non-optimal decisions striving for consensus) in the case of non-dyadic interaction.

The intended contributions of this study are threefold: to obtain an understanding of team deliberations when confronted with AI explanations, to assess whether human teaming can reduce overreliance on AI suggestions, and to examine the effect of cognitive forcing on human-AI team reliance. We believe that the Workshop on Trust and Reliance in AI-Assisted Tasks can support us in further developing these ideas.

2 RELATED WORK

2.1 Overreliance on AI Systems

Algorithmic decision systems in various domains are able to provide suggestions that improve human-AI performance compared to humans acting alone [1, 5, 10, 13, 20]. However, the AI systems in these articles outperform both humans acting alone and human-AI dyads collaborating. In short, humans are unable to calibrate their trust based on AI suggestions to ensure appropriate trust in each AI suggestion [21].

Explainable AI (XAI) techniques typically provide confidence levels and explanations alongside AI suggestions. These techniques have been suggested as a means to make humans reason about AI suggestions before making final decisions. Presenting explanations alongside suggestions can improve human-AI performance. However, Bansal et al. outlined a range of related studies in which AI systems alone outperform human-AI dyads presented with AI explanations [1]. Their results show that explanations lead to more reliance on AI algorithms even when the suggestions provided were incorrect [1].

Bucinca et al. explored the use of cognitive forcing functions [9] to reduce overreliance on AI systems [4]. Cognitive forcing functions let people reflect on their answers by slowing down the process, letting the participants make an initial decision before showing the suggestions, or deciding when to see an explanation. The cognitive forcing functions reduced overreliance, but the AI system alone still outperformed human-AI dyads. The performance increase was higher for people with high levels of 'need for cognition' (NFC). NFC is a personality trait describing the 'tendency for an individual to engage in and enjoy critical thinking' [6].

Research beyond dyadic human-AI teaming is underexplored and could shed new light on overreliance on AI systems. We argue that group deliberation could elucidate people's considerations when making decisions using AI support.

2.2 Decision-making in Human Teams

The problem of achieving complementary team performance on collective tasks is not limited to human-AI interaction but has also been highly debated in human teaming research. Collins coined the 'assembly bonus effect' that describes

groups collectively outperforming each individual in the group working alone [8]. Kerr and Tindale claim that few studies have actually found the 'assembly bonus effect' and that the found effects were modest and debatable [19]. While one would think that a team's combined cognitive abilities can increase overall performance, prior work has highlighted that this is not necessarily the case.

Group decision-making processes introduce new biases alongside existing individual ones [16]. *False consensus bias*, for example, makes humans overestimate the amount of agreement with their opinion, stemming from the false belief that their views are typical. Groups suffering from *groupthink* are too concerned with reaching a consensus that they avoid sharing controversial opinions or ignore critical information. This hinders decision-making by not considering alternatives and making decisions without sufficient information [15]. *Group polarisation* steers the discussion to further explore an idea the group initially favours, reducing the group's efforts to consider other opinions [12]. *Escalation of commitment* describes a group's commitment to an initially agreed-upon strategy, even when it has failed [29]. This bias also influences individual decisions, but more frequently and severely impacts group decisions.

Group decision processes also tempt people to 'hide in the crowd'. Less dominant individuals might not want to speak up, limiting the diversity of group ideas. *Social loafing* describes the tendency of people to put less effort into a collective task [17] than an individual or co-operative task and is more prominent in larger groups.

The aforementioned problematic factors impacting group decision-making can be summarised as not sufficiently considering the available options. Assigning a team member to play 'devil's advocate' has been proven to reduce *groupthink* [26]. In this study we assign this role to the AI in the cognitive forcing condition. The 'consider-the-opposite' method lets individuals and teams engage with other possibilities before locking in their final answer [23]. Individuals with high NFC are less likely to resort to *social loafing* in collective tasks [27]. In team settings, we hypothesise that this brings the efforts to the fore of people with high NFC, who benefit most from cognitive forcing functions in AI-assisted decision-making [4].

3 METHOD

In our lab study, which follows a mixed-method approach, human-AI teams have to collectively solve multiple-choice sentiment analysis tasks. Our study follows a *2 (dyadic/non-dyadic) x 2 (cognitive forcing/base)* design. In the dyadic condition, participants individually solve the tasks with the help of AI suggestions and explanations. In the non-dyadic condition, three participants are paired up to form a team. The team has to collectively solve estimation tasks and is presented with an AI suggestion and explanation for each task. We will recruit participants from student groups, making sure that grouped participants know each other to limit the effect of group cohesion on the result.

Participants in the *cognitive forcing* condition will be presented with explanations for the alternative answers after giving their initial answer. They subsequently have to lock in their final answer to finish the task.

The study will be concluded with an interview asking participants about their considerations while working on the tasks, either alone or in groups. Participants will also be asked to fill out a form to gather demographic data and complete the NFC questionnaire [6]. NFC will be explored as a confounding factor as it has been shown to influence engagement with AI explanations [4].

End-users' reliance on AI recommendations depends on the type [7] and difficulty [28] of the task, therefore requiring careful selection of tasks for our study. We chose sentiment analysis as the task domain as it requires no specific domain knowledge.

3.1 Quantitative Assessment

The performance of the AI system will be compared across a dyadic setting, a non-dyadic setting, and with the AI acting on its own to assess the effect of teaming on task performance. We define overreliance as the number of incorrect AI suggestions the participant(s) follow. For the quantitative part of the study, we seek to answer the following questions:

- **Question 1:** Does team deliberation about AI explanations mitigate overreliance on the AI as compared to dyadic human-AI interaction?
- **Question 2:** Is overreliance on the AI reduced by forcing participants to engage with explanations for the alternatives to their initial choice?
- **Question 3:** Does participants' need for cognition affect overreliance?

3.2 Qualitative Assessment

We will observe teams of participants and subsequently interview them to gain insight into group dynamics and the reasoning behind their choices. Explaining a thought process might elucidate participants' deliberations while interacting with AI explanations. It also provides insight into how many of the team members are involved in the decision-making process and whether an individual 'takes the lead'. The observer's actions do not guide the groups (e.g., participants interact with a single computer and are not corrected when an individual carries out all the group tasks). We will analyse the data using thematic analysis [2] to identify relevant patterns in the deliberations while participants perform the tasks (only in the non-dyadic condition) and in the interviews afterwards.

- **Question 4:** How do human teams use AI explanations to reach a collective decision?
- **Question 5:** Are all team members involved in decision-making, or does 'social loafing' occur?

4 LIMITATIONS & OPEN QUESTIONS

Human team performance is improved by synchronising team members' mental models [22]. The lab study is relatively short and might give participants insufficient time to align mental models. As the study captures just one moment, long-term team developments are out of the scope of this study.

The proxy tasks performed during the study may not translate to actual decision-making tasks [3]. Moreover, Bach et al. found that clinicians are doubtful about the real-world application of interventions to reduce anchoring bias [18].

We present non-dyadic teaming as a critical research direction within human-AI decision-making. The Workshop on Trust and Reliance in AI-Assisted Tasks provides an inspiring place to discuss methodological approaches to best study these settings. The use of non-dyadic settings in human-AI interaction introduces a myriad of factors that HCI researchers can explore, including the effect of individual preferences, supporting information sharing, and encouraging deliberation between team members. We have outlined several factors from the human teaming literature that we expect to influence non-dyadic human-AI decision-making. These can be summarised as not sufficiently considering the available options, either as the result of biases that narrow the decision-making process or due to a lack of team member involvement in the decision process. We seek to explore possible interventions that can help teams avoid these oversights in order to improve non-dyadic human-AI decision-making.

ACKNOWLEDGMENTS

This work is supported by the Carlsberg Foundation project 'Algorithmic Explainability for Everyday Citizens'.

REFERENCES

[1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. https://doi.org/10.48550/ARXIV.2006.14779

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[3] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[5] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. 160–169. https://doi.org/10.1109/ICHI.2015.26

[6] John Cacioppo and Richard Petty. 1982. The Need for Cognition. *Journal of Personality and Social Psychology* 42 (01 1982), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

[7] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[8] Barry E Collins and Harold Steere Guetzkow. 1964. *A social psychology of group processes for decision-making*. New York: Wiley.

[9] Pat Croskerry. 2003. Cognitive forcing strategies in clinical decisionmaking. *Annals of emergency medicine* 41, 1 (2003), 110–120.

[10] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (nov 2019), 24 pages. https://doi.org/10.1145/3359152

[11] Simo Hosio, Jorge Goncalves, Niels van Berkel, Simon Klakegg, Shin'ichi Konomi, and Vassilis Kostakos. 2018. Facilitating Collocated Crowdsourcing on Situated Displays. *Human-Computer Interaction* (2018), 1–37. https://doi.org/10.1080/07370024.2017.1344126

[12] Daniel J Isenberg. 1986. Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology* 50, 6 (1986), 1141.

[13] Maia Jacobs, Melanie F.Pradier, Thomas McCoy, Roy Perlis, Finale Doshi velez, and Krzysztof Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational Psychiatry* 11 (02 2021). https://doi.org/10.1038/s41398-021-01224-x

[14] Irving Lester Janis. 1983. *Groupthink*. Houghton Mifflin Boston.

[15] Irving L Janis and Leon Mann. 1977. *Decision making: A psychological analysis of conflict, choice, and commitment*. Free press.

[16] Paul Jones and Peter Roelofsma. 2000. The potential for social contextual and group biases in team decision-making: Biases, conditions and psychological mechanisms. *Ergonomics* 43 (09 2000), 1129–52. https://doi.org/10.1080/00140130050084914

[17] Steven J Karau and Kipling D Williams. 1993. Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology* 65, 4 (1993), 681.

[18] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels van Berkel. 2023. "If I Had All the Time in the World": Ophthalmologists' Perceptions of Anchoring Bias Mitigation in Clinical AI Support. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems* (CHI'23). to appear. https://doi.org/10.1145/3544548.3581513

[19] Norbert L Kerr and R Scott Tindale. 2004. Group performance and decision making. *Annu. Rev. Psychol.* 55 (2004), 623–655.

[20] Vivian Lai and Chenhao Tan. 2018. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *CoRR* abs/1811.07901 (2018). arXiv:1811.07901 http://arxiv.org/abs/1811.07901

[21] John Lee and Katrina See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human factors* 46 (02 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

[22] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. 2000. The influence of shared mental models on team process and performance. *Journal of applied psychology* 85, 2 (2000), 273.

[23] Thomas Mussweiler, Fritz Strack, and Tim Pfeiffer. 2000. Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility. *Personality and Social Psychology Bulletin* 26 (11 2000), 1142–1150. https://doi.org/10.1177/01461672002611010

[24] Joon Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3 (11 2019), 1–15. https://doi.org/10.1145/3359204

[25] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 97–101. https://doi.org/10.18653/v1/N16-3020

[26] David M Schweiger, William R Sandberg, and James W Ragan. 1986. Group approaches for improving strategic decision making: A comparative analysis of dialectical inquiry, devil's advocacy, and consensus. *Academy of management Journal* 29, 1 (1986), 51–71.

[27] Brian N Smith, Natalie A Kerr, Michael J Markus, and Mark F Stasson. 2001. Individual differences in social loafing: Need for cognition as a motivator in collective performance. *Group Dynamics: Theory, Research, and Practice* 5, 2 (2001), 150.

[28] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2022. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. https://doi.org/10.48550/ARXIV.2212.06823

[29] Glen Whyte. 1993. Escalating Commitment in Individual and Group Decision Making: A Prospect Theory Approach. *Organizational Behavior and Human Decision Processes* 54, 3 (1993), 430–455. https://doi.org/10.1006/obhd.1993.1018

[30] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372852