ACM Collective Intelligence Conference 2024
Book of Extended Abstracts

June 27-28, 2024 | Boston, MA

Conference Chair: Christoph Riedl

Preface

This booklet contains non-archival extended abstracts of papers presented at the ACM Collective Intelligence Conference 2024, which was held on June 27-28, 2024 in Boston, MA.

The CI conference aims to be inclusive by offering the community a variety of publication formats to accommodate diverse publication norms in different fields. In addition to publishing archival proceedings for full papers in the ACM Digital Library for the second time, we publish this non-archival booklet of extended abstracts.

The ACM Collective Intelligence Conference has a transdisciplinary focus devoted to advancing the theoretical and empirical understanding of collective performance in diverse systems, whether biological, technological, or a combination. We are interested in research on a broad range of systems that vary in scale and scope and focus on implications for a diverse range of social, ecological, and economic outcomes.

The success of CI 2024 depends on the contributions of many individuals and organizations. We thank all authors who submitted their work to the conference. We would also like to thank the members of the program committee for their valuable time in reviewing the manuscripts and providing useful feedback to the authors. We also thank the committed volunteers and members of our organizing committee and steering committee.

Finally, no conference will succeed without the strong support of its participants. Thank you all for attending CI'24 in Boston and making it such a memorable event.

ACM Collective Intelligence
June 27-28, 2024

Organizing Committee

General Chairs:

Christoph Riedl, *Conference Chair*

Julian Gullett, *Web & Media Chair*
Nicole Samay, *Web & Media Chair*
Maria Baskerville, *Accessibility Chair*
Christianah Titilope Oyewale, *Accessibility Chair*
Robert Mahari, *Workshop Chair*

Doctoral Consortium Chairs:

Joshua Becker, *University College London*
Emily Hu, *University of Pennsylvania*

Program Committee:

Abdullah Almaatouq, *Massachusetts Institute of Technology*
Elizabeth Altman, *University Massachesetts Lowell*
Torben Juul Andersen, *Copenhagen Business School*
Joshua Becker, *University College London*
Seth Bullock, *University of Bristol*
Kwong Chan, *Northeastern University*
Seth Cooper, *Northeastern University*
Kevin Crowston, *Syracuse University*
Adenekan Dedeke, *Northeastern University*
Mirta Galesic, *Santa Fe Institute*
Pranav Gupta, *University of Illinois Urbana-Champaign*
Brennan Klein, *Northeastern University*
David Lazer, *Northeastern University*
Shun-Yang Lee, *Northeastern University*
Jack McGuire, *Northeastern University*
Nitish Mehotra, *Meta*
Esteban Moro, *Northeastern University*
Silvia Prina, *Northeastern University*
Saiph Savage, *Northeastern University*
Victor Seidel, *Babson College*
Thalia Wheatley, *Dartmouth College*
Vicky Yang, *Massachusetts Institute of Technology*
Sine Zambach, *Copenhagen Business School*

Conference Steering Committee:

Thomas Malone, *Massachusetts Institute of Technology* (Chair)
Abraham Bernstein, *University of Zurich*
Nikki Kittur, *Carnegie Mellon University*
Natalia Levina, *New York University*
Jeff Nickerson, *Stevens Institute of Technology*
Beth Noveck, *Northeastern University*
Gary Olson, *University of California in Irvine*
Paul Resnick, *University of Michigan*
Christoph Riedl, *Northeastern University*
Anita Woolley, *Carnegie Mellon University*

Engaging Collective Intelligence

Creating Dynamic Adaptative Organizations

Torben Juul Andersen

Copenhagen Business School, tja.egb@cbs.dk

Contemporary organizations operate in complex dynamic contexts that may generate unpredictable events with potentially extreme outcome effects. It is thus timely and highly relevant to consider more adaptive organizations that exploit current insights embedded in the collective intelligence of peripheral stakeholders. Hence, we explore the configurations of adaptive organizations and how collective intelligence can be applied to generate effective organizational adaptation. We argue that a combination of fast insights and slow analytical processes facilitated by advanced digital technologies can form the basis for dynamic adaptive systems. This leads to propositions that effective organizational adaptation is associated with interactive information processing that combines current insights from operational entities with analytics-based strategic updating.

CCS CONCEPTS • Collaborative and social computing • Computing and business • Interaction design

Additional Keywords and Phrases: Fast-slow information processing, Interactive strategic decision-making, dynamic adaptive systems

1 INTRODUCTION

There is an urgent need to configure organizations that can survive and thrive in dynamic complex contexts with potentially extreme challenges, e.g., pandemic, military conflict, climate effect, etc. In this context, it is relevant to consider how digital technologies can be applied to continuously align and re-align adaptive networked organizations as the surroundings change integrating the collective intelligence of a specialized heterogeneous workforce.

1.1 Motivation and research questions

Social systems like organizations operate in global business contexts and are thereby exposed to highly complex conditions, with many feedback loops that can generate unpredictable extreme outcomes. This applies to international enterprises that manufacture and distribute products and services in turbulent markets. To reduce the implied uncertainty, some organizations move away from hierarchical controls in favor of decentralized networks of autonomous collaborative agents that operate more dynamically. However, this approach conceives of management as a cohesive unit rather than a separate related social network (e.g., Ocasio 2011; Leppänen *et al*., 2023) and discards the need to align management with networks around operational employees. As a result, the management literature does not offer refined models to facilitate adaptive interactions between management and operational individuals in the organization whereby managers can update their strategic outlook with current insights from networked individuals across operational functions (e.g., Andersen, 2015; Andersen & Hallin, 2017; Andersen, 2023).

The potential benefits of using current insights among individuals in operational functions are compelling and can be achieved by using collective intelligence (CI) mobilized across individuals in real time through computerized interventions. It can stimulate innovation and design of collaborative learning organizations (e.g., Lévy, 2010; Zambrano *et al*., 2019) with open sharing of information to enhance ingenuity and resolve complex issues (Malone, 2018; Laubacher *et al*., 2020). However, the ability to

exchange and process information between individuals across dispersed social networks also subsumes behavioral, cognitive, and ethical aspects that can influence the collective outcomes (e.g., Tzini & Jain, 2018; Andersen & Young, 2020). We also know that information passed through an organization is prone to subconscious filtering (e.g., Reitzig & Sorenson, 2013; Tuckett & Nikolic, 2017; Boustanifar … & Zilja, 2022).

These effects must be considered to obtain effective information processing where particular IT systems with real-time data access may counteract potential filtering issues (e.g., Gleasure & Feller, 2016, 2018). These considerations lead to two fundamental research questions:

1. What are the basic configurational elements of an effective adaptive organization?

2. How can we use collective intelligence to develop an effective adaptive organization?

Answering these questions promise to advance new research in strategic adaptation, digital science, and collective intelligence in ways that have not previously been done.

1.2 Theoretical foundations and propositions

Organizations are described as social networks where information and knowledge exchanged across subunits affect performance (e.g., Hansen *et al*., 2005). It is noted how social network patterns influence information processing effectiveness where open interaction across team members and networked team leaders seems to enhance performance (Janhonen & Johanson, 2011). Organizations have been analyzed as communication networks that span across hierarchical levels as episodic interplays between intended and emergent events (Blaschke *et al*., 2012). Recent studies consider managerial influences on information exchange and learning in organizations through interacting feed-forward and feedback processes (e.g., Bouma *et al*., 2023). These state-of-the-art social network perspectives analyze teams and intra-organizational subunits around individuals acting in operational functions. Hence, there is no consideration for interaction with management as belonging to a distinct network to networks of operational individuals triggered by episodic events under turbulent conditions. Hence, we are unable to explain how interacting communication between hierarchies including management affect collective adaptive responses in organizations. The potential advantages of combined analytics-based strategic intent and opportunity-driven emergence is recognized in the strategy field (e.g., Mintzberg & Waters, 1985; Burgelman & Grove, 2007; Andersen, 2015). But we have little concrete knowledge about how to structure the dynamic information processes effectively across networked individuals and groups in the organization.

In strategy, adaptation has been conceived as "*capabilities to assess the environment, identify firm resources, and mobilize them in effective responsive actions*" (Andersen *et al*., 2007, p. 410). A dynamic capabilities concept describes a similar ability to sense changes, seize resources, and reconfigure organizational activities (Teece, 2007). It is constituted by complex bundles of routines coaxed by managerial interventions (e.g., Teece *et al*., 2016; Teece, 2018). This adaptive process depends on an ability to access current insights and update strategic thinking under rapid change. Diverse response capabilities across organizations affect competitiveness and performance where adaptation generates high returns and low variability (Andersen *et al*., 2007; Andersen & Bettis, 2015; Andersen, 2023). That is, effective organizational adaptation can achieve above average returns at lower performance risk. This reasoning leads to the following propositions linked to research question 1.

P1: Effective organizational adaptation is associated with interactive information processing between management and operational networks.

P1a: Bottom-up communication provides current insights from operational networks that inform the updated strategic thinking.

P1b: Top-down communication of updated strategic thinking guides responsive actions in the operational networks.

The application of virtual collaboration changes the way work is structured and employees interact that will affect organizational outcomes (Schinoff *et al*., 2020). Distributed collaboration across organizations advanced by Covid-19 forced many to work from home underscoring the importance of relationship building, mutual learning, and intra-personal communication (O'Leary, Gleasure *et al*., 2020). More people work remotely supported by digital workplace measures that nonetheless fail to show a deep understanding of virtual collaboration across networked individuals (Gould *et al*., 2023).

Organizational adaptation reflects emergent properties of a complex socio-cognitive system where we must understand the social structures and integration strategies of network collectives (Galesic *et al*., 2023). Collective networks of diverse individuals can amplify their cognitive capacity to better solve complex issues (Aminpour *et al*., 2020; Gray *et al*., 2020).

However, networked individuals may update their beliefs in expectation-based ways influenced by prior beliefs (Hahn *et al*., 2018) influenced by social forces that can undermine judgmental variation and reduce accuracy (e.g., Almaatouq *et al*., 2020; Da & Huang, 2020; Frey & van de Rijt, 2021). While digitalized communication can induce innovation from discord among people, the most influential breakthroughs require in-person contacts (Li *et al*., 2020). Technological connectivity can inhibit idea generation as screen-based cues miss essential stimuli (Brucks & Levav, 2022). Fear of the unknown may cause anxiety and stress (Morriss *et al*., 2023) where behaviors are influenced by neuro-modulated circadian rhythms (Bedder *et al*., 2023). Yet, emotion-sensing technologies can capture the physiological markers that affect the CI (Whelan, Gleasure *et al*., 2018; Wang *et al*., 2020). So, adopting IT to process information across interactive networks is very promising but needs further examination. This leads to the following propositions linked to research question 2.

P2: Effective organizational adaptation is enhanced when management is updated by the collective intelligence of organizational stakeholders.

P2a: The collective intelligence of organizational stakeholders can be formed and collated through IT mediation.

P2b: Effective formation and collation of collective intelligence among organizational stakeholders take social and emotional factors into account.

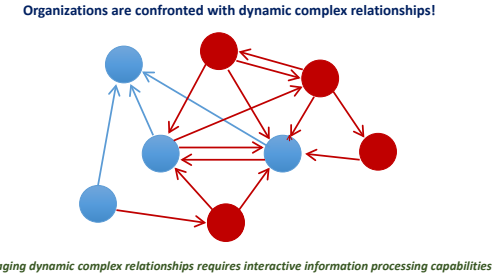


Figure 1: Collective intelligence formed around interactive information processing systems.

REFERENCES

Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 117(21), 11379–11386.

Aminpour, P., Gray, S. A., Jetter, A. J., Introne, J. E., Singer, A., & Arlinghaus, R. (2020). Wisdom of stakeholder crowds in complex social–ecological systems. *Nature Sustainability*, 3(3), 191-199.

Andersen, T. J. (2023). *Responding to Uncertain Conditions: New research on strategic adaptation*. Emerald Studies in Global Strategic Responsiveness, Emerald Publishing, Bingley, UK.

Andersen, T. J. (2015). Interactive strategy-making: Combining central reasoning with ongoing learning from decentralized responses. *Journal of General Management*, 40(4), 69-88.

Andersen, T. J., & Bettis, R. A. (2015). Exploring longitudinal risk-return relationships. *Strategic Management Journal*, 36(8), 1135-1145.

Andersen, T. J., Denrell, J., & Bettis, R. A. (2007). Strategic responsiveness and Bowman's risk– return paradox. *Strategic Management Journal*, 28(4), 407-429.

Andersen, T. J., & Hallin, C. A. (2017). *Global Strategic*

Responsiveness: Exploiting frontline information in the adaptive multinational enterprise. Taylor & Francis, London, UK.

Andersen, T. J., & Young, P. C. (2020). *Strategic Risk Leadership: Engaging a world of risk, uncertainty, and the unknown*. Routledge, Oxon, UK/New York, NY.

Bedder, R. L., Vaghi, M. M., Dolan, R. J., & Rutledge, R. B. (2023). Risk taking for potential losses but not gains increases with time of day. *Scientific Reports*, 13(1), 5534.

Blaschke, S., Schoeneborn, D., & Seidl, D. (2012). Organizations as networks of communication episodes: Turning the network perspective inside out. *Organization Studies*, 33(7), 879-906.

Bouma, D., Canbaloğlu, G., Treur, J., & Wiewiora, A. (2023). Adaptive network modeling of the influence of leadership and communication on learning within an organization. *Cognitive Systems Research*, 79, 55-70.

Boustanifar, H., Zajac, E. J., & Zilja, F. (2022). Taking chances? The effect of CEO risk propensity on firms' risky internationalization decisions. *Journal of International Business Studies*, 53, 302–325.

Brucks, M. S., & Levav, J. (2022). Virtual communication curbs creative idea generation. *Nature*, 605(7908), 108-112. https://doi.org/10.1038/s41586-022-04643-y

Burgelman, R. A., & Grove, A. S. (2007). Let chaos reign, then rein in chaos—repeatedly: Managing strategic dynamics for corporate longevity. *Strategic Management Journal*, 28(10), 965- 979.

Da, Z., & Huang, X. (2020). Harnessing the Wisdom of Crowds. Management Science, 66(5), 1847–1867. Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing theory through simulation methods. *Academy of Management Review*, 32(2), 480-499.

Frey, V., & van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, 67(7), 4273–4286. http://dx.doi.org/10.1287/mnsc.2020.3713

Galesic, M., Barkoczi, D., Berdahl, A. M., Biro, D., Carbone, G., Giannoccaro, I., Goldstone, R. L., … & Stein, D. L. (2023). Beyond collective intelligence: Collective adaptation. *Journal of the Royal Society Interface*, 20, 20220736. https://doi.org/10.1098/rsif.2022.0736

Gould, S. J., Rudnicka, A., Cook, D., Cecchinato, M. E., Newbold, J. W., & Cox, A. L. (2023). Remote work, work measurement and the state of work research in human-centred computing. *Interacting with Computers*. https://doi.org/10.1093/iwc/iwad014

Gleasure, R., & Feller, J. (2016). Emerging technologies and the democratisation of financial services: A metatriangulation of crowdfunding research. *Information and Organization*, 26(4), 101- 115.

Gleasure, R., & Feller, J. (2018). What kind of cause unites a crowd? Understanding crowdfunding as collective action. *Journal of Electronic Commerce Research,* 19(3), 223-236.

Gray, S., Aminpour, P., Reza, C., Scyphers, S., Grabowski, J., Murphy Jr, R., ... & Introne, J. (2020). Harnessing the collective intelligence of stakeholders for conservation. *Frontiers in Ecology and the Environment*, 18(8), 465-472.

Hahn, U., von Sydow, M., & Merdes, C. (2019). How communication can make voters choose less well. *Topics in Cognitive Science*, 11(1), 194–206.

Hansen, M. T., Mors, M. L., & Løvås, B. (2005). Knowledge sharing in organizations: Multiple networks, multiple phases. *Academy of Management Journal*, 48(5), 776-793.

Laubacher, R., Giacomelli, G., Kennedy, K., Kong, D. S., Bachmann, A., Kramer, K., Schlag, P., & Malone, T. W. (2020). Using a supermind to design a supermind: A case study of university researchers and corporate executives co-designing an innovative healthcare concept. MIT Center for Collective Intelligence, MIT Sloan Working Paper No. 6109-20.

Leppänen, P., George, G., & Alexy, O. (2023). When do novel business models lead to high performance? A configurational approach to value drivers, competitive strategy, and firm environment. *Academy Of Management Journal*, 66(1), 164-194.

Lévy, P. (2010). From social computing to reflexive collective intelligence: The IEML research program. *Information Sciences*, 180(1), 71-94.

Li, L., Wu, L., & James, E. (2020). Social centralization and semantic collapse: Hyperbolic embeddings of networks and text. *Poetics*, 78, 101428. https://doi.org/10.1016/j.poetic.2019.101428

Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together*. Little, Brown, and Company, Boston, MA.

Morriss, J., Abend, R., Zika, O., Bradford, D. E., & Mertens, G. (2023). Neural and psychophysiological markers of intolerance of uncertainty. *International Journal of Psychophysiology*, 184, 94-99.

Ocasio, W. (2011). Attention to attention. *Organization Science*, 22(5), 1286-1296.

O'Leary, K., Gleasure, R., O'Reilly, P., & Feller, J. (2020). Reviewing the contributing factors and benefits of distributed collaboration. *Communications of the Association for Information Systems*, 47, 476-520.

Reitzig, M., & Sorenson, O. (2013). Biases in the selection stage of bottom-up strategy formulation. *Strategic Management Journal*, 34(7), 782-799.

Schinoff, B. S., Ashforth, B. E., & Corley, K. G. (2020). Virtually (in) separable: The centrality of relational cadence in the formation of virtual multiplex relationships. *Academy of Management Journal*, 63(5), 1395-1424.

Teece, D. J. (2018). Business models and dynamic capabilities. *Long Range Planning*, 51(1), 40-49.

Teece, D. J. (2007). Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal*, 28(13), 1319-1350.

Teece, D., Peteraf, M., & Leih, S. (2016). Dynamic capabilities and organizational agility: Risk, uncertainty, and strategy in the innovation economy. *California Management Review*, 58(4), 13-35.

Tuckett, D., & Nikolic, M. (2017). The role of conviction and narrative in decision-making under radical uncertainty. *Theory & Psychology*, 27(4), 501-523.

Tzini, K. and Jain, K. (2018). Unethical behavior under relative performance evaluation: Evidence and remedy. *Human Resource Management*, 57(6), 1399-1413.

Wang, Z., Ho, S. B., & Cambria, E. (2020). A review of emotion sensing: categorization models and algorithms. *Multimedia Tools and Applications*, 79, 35553-35582.

Whelan, E., McDuff, D., Gleasure, R., & Vom Brocke, J. (2018). How emotion-sensing technology can reshape the workplace. MIT *Sloan Management Review*, 59(3), 7-10.

Zambrano, J., Kirschner, F., Sweller, J. and Kirshner, P. A. (2019). Effects of group experience and information distribution on collaborative learning. *Instructional Science*, 47, 531–550.

Multidisciplinary learning through collective performance favors decentralization

John Meluso
Laurent Hébert-Dufresne
john.meluso@uvm.edu
laurent.hebert-dufresne@uvm.edu
University of Vermont
Burlington, Vermont, USA

ABSTRACT

Many models of learning assume that team members can share solutions and coordinate. However, these assumptions break down in multidisciplinary teams where team members often complete distinct yet interrelated pieces of larger tasks. Such contexts make it difficult for individuals to separate the performance effects of their own decisions from the choices of interacting neighbors. Despite this challenge, this work shows that individuals can overcome this challenge by coordinating with network neighbors through mediating artifacts (like collective performance assessments). When neighbors' actions influence collective outcomes, teams with different coordination networks perform relatively similarly to one another. However, varying a team's network can affect performance on tasks that weight individuals' contributions by network properties. Consequently, when individuals innovate (through "exploring" searches), dense networks hurt optimization slightly by increasing uncertainty. In contrast, dense networks moderately help design optimization when engineers refine their work (through "exploiting" searches) by efficiently finding local optima. This work also shows that decentralization improves system performance when assessing networks across a battery of 34 tasks with varied qualities and difficulties. These results offer new principles for achieving collective intelligence in multidisciplinary teams within which coordinating proves more difficult.

CCS CONCEPTS

• **Computing methodologies** → *Multi-agent systems*; **Network science**; **Agent / discrete models**; • **Networks** → **Network structure**.

KEYWORDS

team performance, problem-solving, social networks, mediated learning, multidisciplinary teams, agent-based model

ACM Reference Format:
John Meluso and Laurent Hébert-Dufresne. 2024. Multidisciplinary learning through collective performance favors decentralization. In *Proceedings of 2024 ACM Collective Intelligence Conference (CI '24).* ACM, New York, NY, USA, 4 pages.

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 the author(s) 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.
CI '24, June 26–29, 2024, Northeastern University, Boston, MA, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

MAIN

Multidisciplinary teams play crucial roles in academia, industry, and government [10]. Such teams draw representatives from various backgrounds, each working on parts that integrate with the others, but disciplinary boundaries make collaborating across fields challenging [2, 3]. Instead, they often divide tasks into interdependent subtasks and work in parallel [5, 13, 17]. This gives rise to the "performance non-separability problem" [16]: a team's collective performance obfuscates how individuals' actions affect shared outcomes, making it hard to figure out the effect of each change and even more challenging when many changes happen at the same time, or if individuals have limited visibility into each other's actions [12, 15].

Fortunately, another learning pathway exists: *mediated learning* [6]. Here, individuals learn from themselves and others through artifacts, like technical drawings, databases, and performance assessments, as they negotiate meaning across boundaries [3, 8, 11]. Mediated interactions significantly affect outcomes in many contexts [4, 18, 19] including innovation [1], even producing larger effects than unmediated interactions [19], but its effects on collective intelligence have yet to be examined in multidisciplinary teams. Network properties produce varied performance effects depending on the type of learning [19] making network analysis crucial.

We propose a framework for modeling mediated learning in multidisciplinary teams [9]. We introduce neighborhood task performance, which formalizes the networked nature of mediated coordination activities (Fig. 1). Through this concept, we investigate how network density and other properties affect collective intelligence. We simulated performance-mediated feedback networks, team members as computational agents, and representative tasks while minimizing other forms of learning and variables with established effects (Fig. 2). In brief, each team member independently selects actions. Individuals' actions affect performance on their own subtask, but also affect their neighbors' subtask performances and the team's overall task performance. This means individuals receive feedback from a "neighborhood" performance shaped by their own actions and the actions of their network neighbors.

Our framework and simulations show three results. First, mediated learning can either help or hurt collective performance depending on the task and team network structure. Then, we show that densely-connected teams tend to perform better when "exploiting" (or refining) solutions, while sparse teams perform better when

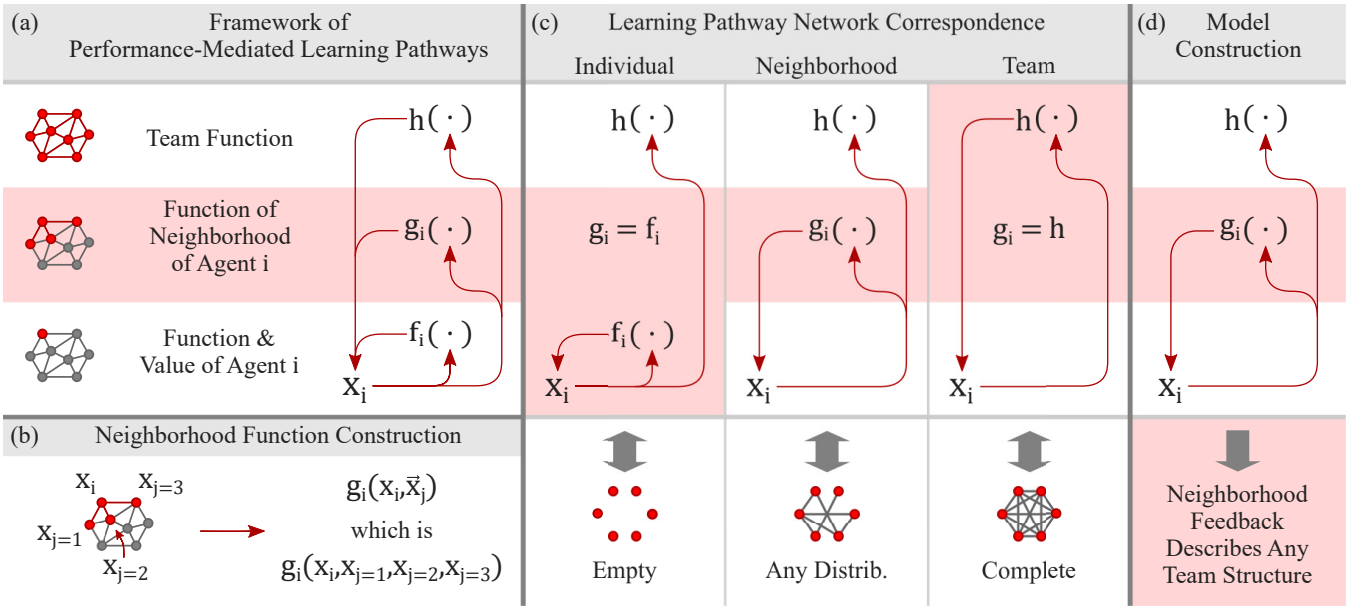


Figure 1: Framework of performance-mediated learning pathways describes different networks, and comprehensive model construction. (a) In the framework, individuals i can learn from team performance h or their own performance f_i, but in multidisciplinary teams they often learn from a combined "neighborhood" performance g_i. (b) Here, the individual's actions x_i and their neighbors' actions $\vec{x}_j$ affect the agent's performance on a subtask g_i of the team's overall task h. This makes it difficult for individuals to tell "how good" their actions are. (c) These neighborhoods of dependence can describe any performance-mediated learning pathway, from individual learning (an empty graph) to the performance feedback paths of team learning (a complete graph) and any distribution in between. (d) Our model enables us to explore different networks by giving each individual i feedback from their neighborhood g_i. Separately, our model assesses the team's overall performance h as a function of all agents actions $\vec{x}$.

"exploring" (or innovating) new solutions (Fig. 3)—a finding that echoes other findings in collective intelligence [7, 14]. But contrary to single-discipline innovation, we show that *decentralized* multidisciplinary teams see greater performance benefits than other structures across a diverse set of tasks (Fig. 4). This benefit likely arises because decentralization balances the benefits of dense and sparse teams on tasks with different qualities.

REFERENCES

[1] Gautam Ahuja. 2000. Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study. *Administrative Science Quarterly* 45, 3 (Sept. 2000), 425–455. https://doi.org/10.2307/2667105

[2] William C Barley. 2015. Anticipatory Work: How the Need to Represent Knowledge across Boundaries Shapes Work Practices within Them. *Organization Science* 26, 6 (Oct. 2015), 1612–1628. https://doi.org/10.1287/orsc.2015.1012

[3] Paul R Carlile. 2002. A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development. *Organization Science* 13, 4 (2002), 442–455.

[4] Curtis R. Henrie, Lisa R. Halverson, and Charles R. Graham. 2015. Measuring Student Engagement in Technology-Mediated Learning: A Review. *Computers & Education* 90 (Dec. 2015), 36–53. https://doi.org/10.1016/j.compedu.2015.09.005

[5] Moses N. Kiggundu. 1983. Task Interdependence and Job Design: Test of a Theory. *Organizational Behavior and Human Performance* 31, 2 (April 1983), 145–172. https://doi.org/10.1016/0030-5073(83)90118-6

[6] Alex Kozulin. 2003. Psychological Tools and Mediated Learning. In *Vygotsky's Educational Theory in Cultural Context*, Alex Kozulin, Boris Gindis, Vladimir S. Ageyev, and Suzanne M. Miller (Eds.). Cambridge University Press, Cambridge, UK, 15–38.

[7] David Lazer and Allan Friedman. 2007. The Network Structure of Exploration and Exploitation. *Administrative Science Quarterly* 52, 4 (Dec. 2007), 667–694. https://doi.org/10.2189/asqu.52.4.667

[8] Ann Majchrzak, Philip H. B. More, and Samer Faraj. 2011. Transcending Knowledge Differences in Cross-Functional Teams. *Organization Science* 23, 4 (July 2011), 951–970. https://doi.org/10.1287/orsc.1110.0677

[9] John Meluso and Laurent Hébert-Dufresne. 2023. Multidisciplinary Learning through Collective Performance Favors Decentralization. *Proceedings of the National Academy of Sciences* 120, 34 (Aug. 2023), e2303568120. https://doi.org/10.1073/pnas.2303568120

[10] National Research Council. 2014. *Convergence: Facilitating Transdisciplinary Integration of Life Sciences, Physical Sciences, Engineering, and Beyond.* The National Academies Press, Washington, DC. https://doi.org/10.17226/18722

[11] Davide Nicolini, Jeanne Mengis, and Jacky Swan. 2011. Understanding the Role of Objects in Cross-Disciplinary Collaboration. *Organization Science* 23, 3 (July 2011), 612–629. https://doi.org/10.1287/orsc.1110.0664

[12] Steven Postrel. 2002. Islands of Shared Knowledge: Specialization and Mutual Understanding in Problem-Solving Teams. *Organization Science* 13, 3 (June 2002), 303–320. https://doi.org/10.1287/orsc.13.3.303.2773

[13] Richard Saavedra, P. Christopher Earley, and Linn Van Dyne. 1993. Complex Interdependence in Task-Performing Groups. *Journal of Applied Psychology* 78, 1 (Feb. 1993), 61–72. https://doi.org/10.1037/0021-9010.78.1.61

[14] Jesse Shore, Ethan Bernstein, and David Lazer. 2015. Facts and Figuring: An Experimental Investigation of Network Structure and Performance in Information and Solution Spaces. *Organization Science* 26, 5 (April 2015), 1432–1446. https://doi.org/10.1287/orsc.2015.0980

[15] Garold Stasser and William Titus. 2003. Hidden Profiles: A Brief History. *Psychological Inquiry* 14, 3/4 (2003), 304–313.

[16] Jose Uribe, Seth Carnahan, John Meluso, and Jesse Austin-Breneman. 2022. How Do Managers Evaluate Individual Contributions to Team Production? A Theory and Empirical Test. *Strategic Management Journal* 43, 12 (May 2022), 2577–2601. https://doi.org/10.1002/smj.3433

[17] Ruth Wageman. 1995. Interdependence and Group Effectiveness. *Administrative Science Quarterly* 40, 1 (1995), 145–180. https://doi.org/10.2307/2393703

[18] Naijia Xiao, Aifen Zhou, Megan L. Kempher, Benjamin Y. Zhou, Zhou Jason Shi, Mengting Yuan, Xue Guo, Linwei Wu, Daliang Ning, Joy Van Nostrand,

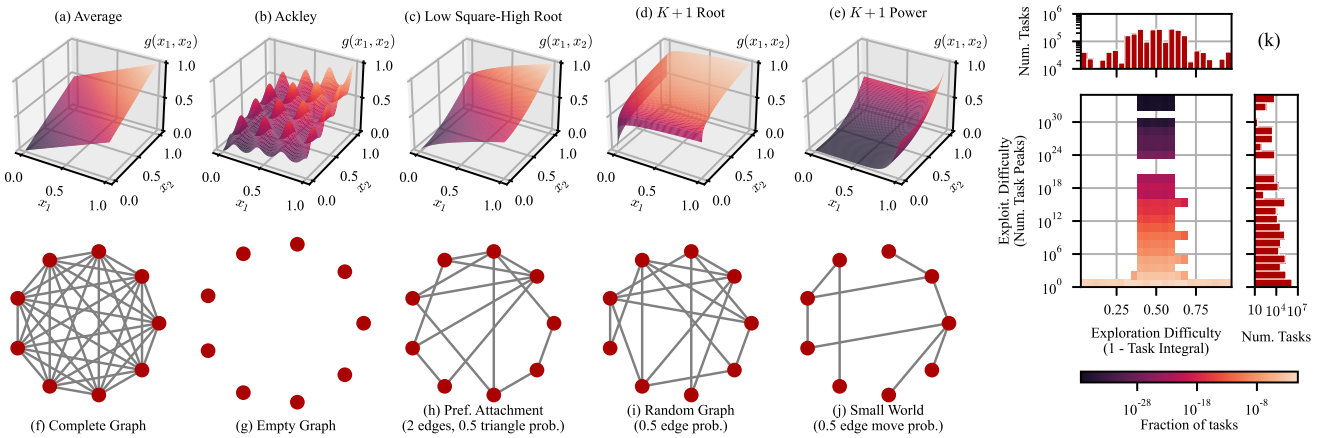


Figure 2: Example tasks, networks, and their difficulties. We put teams through fitness landscapes representing 34 tasks while in each of 18 network configurations, including (a-e) the example tasks and (f-j) networks shown here. (k) Together, the task functions and networks varied task difficulty across four measures, including exploration and exploitation as shown here.

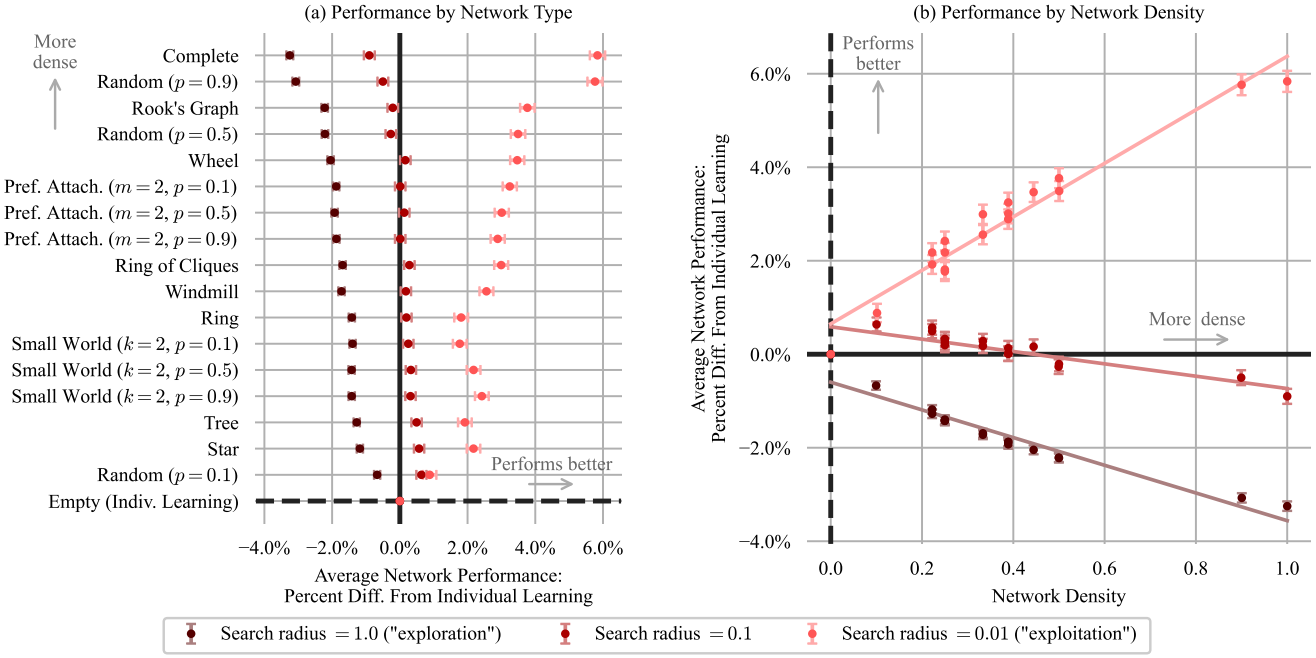


Figure 3: Average performances for each network and search radius. Values shown are percent differences from the averages of teams that only learn individually (empty graphs) for teams of $n = 9$ agents with the specified search radius. Error bars show 95% confidence intervals of the differences between means. (a) Average performances by network type, sorted top-to-bottom first by density, then by eigenvector centrality, a measure of how important each individual is to each other individual. (b) The same performance averages plotted by network density. Trend lines represent simple least squares regressions between density and performance, illustrating how the search radius influences the relationship between network density and performance.

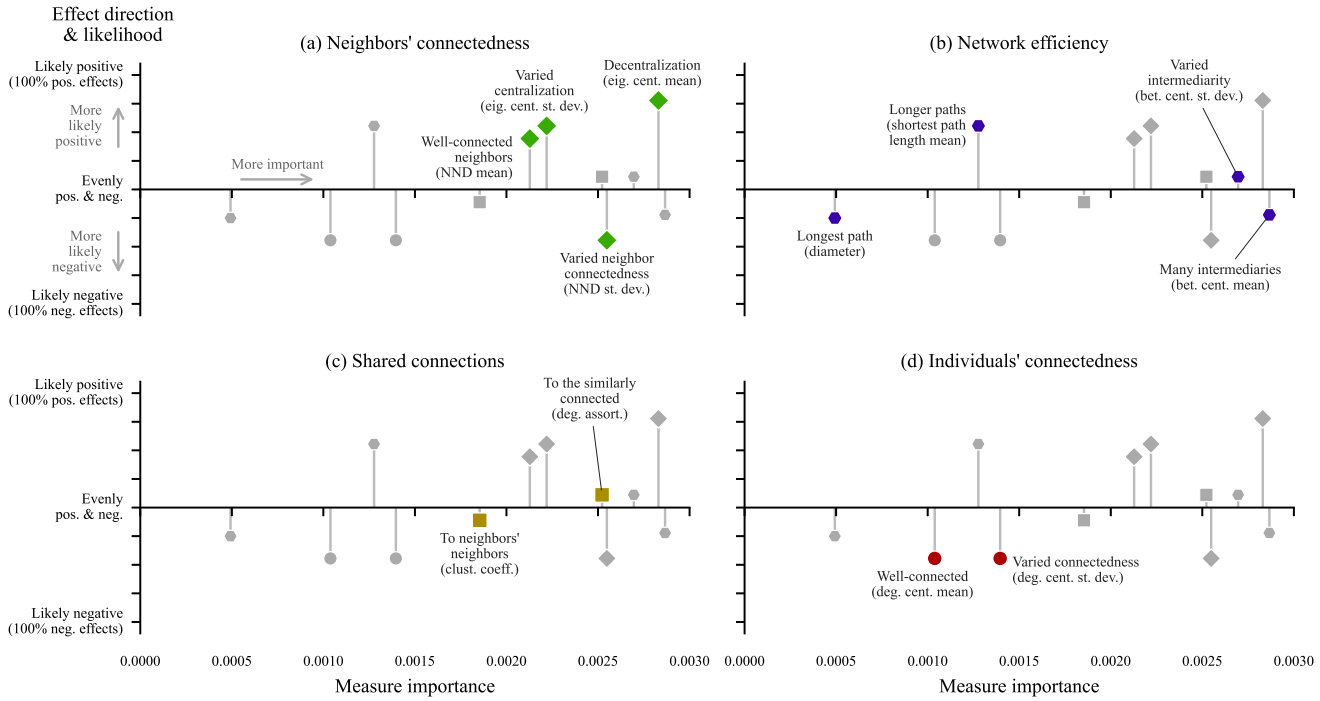


Figure 4: Network measure importances, likelihoods, and effect directions. Measures of (a) neighbors' connectedness, (b) network efficiency, (c) shared connections, and (d) individual connectedness. Each point shows how important the indicated measure is (x-axis), and plots those importances against how likely that measure is to either have a positive effect (positive y-axis) or negative effect (negative y-axis) on team performance. Decentralization and having many intermediaries are tied for most important (within one standard deviation of each other's means). NND stands for nearest neighbor degree.

Mary K. Firestone, and Jizhong Zhou. 2022. Disentangling Direct from Indirect Relationships in Association Networks. *Proceedings of the National Academy of Sciences* 119, 2 (Jan. 2022), e2109995119. https://doi.org/10.1073/pnas.2109995119

[19] Jiarong Xie, Xiangrong Wang, Ling Feng, Jin-Hua Zhao, Wenyuan Liu, Yamir Moreno, and Yanqing Hu. 2022. Indirect Influence in Social Networks as an Induced Percolation Phenomenon. *Proceedings of the National Academy of Sciences* 119, 9 (March 2022), e2100151119. https://doi.org/10.1073/pnas.2100151119

Received 27 February 2024; accepted 30 April 2024

An LLM-Infused Tool for Workers' Collective Intelligence

Kashif Imteyaz
Northeastern University
Boston, MA, USA

Claudia Flores-Saviaga
Northeastern University
Boston, MA, USA

Saiph Savage
Northeastern University
Boston, MA, USA

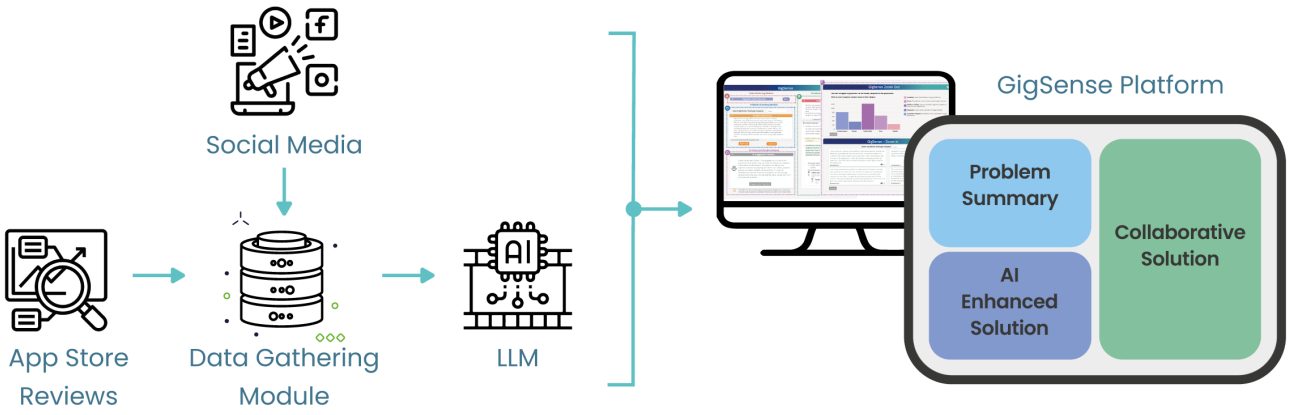


Figure 1: Overview of our process for designing and studying a generative AI tool for gig workers.

ABSTRACT

Collective intelligence among gig workers yields considerable advantages, including improved information exchange, deeper social bonds, and stronger advocacy for better labor conditions. Especially as it enables workers to collaboratively pinpoint shared challenges and devise optimal strategies for addressing these issues. However, enabling collective intelligence remains challenging, as existing tools often overestimate gig workers' available time and uniformity in analytical reasoning. To overcome this, we introduce GigSense, a tool that leverages large language models alongside theories of collective intelligence and sensemaking. GigSense enables gig workers to rapidly understand and address shared challenges effectively, irrespective of their diverse backgrounds. GigSense not only empowers gig workers but also opens up new possibilities for supporting workers more broadly, demonstrating the potential of large language model interfaces to enhance collective intelligence efforts in the evolving workplace.

CCS CONCEPTS

• **Human-centered computing**; • **Collaborative and social computing systems and tools**;

KEYWORDS

Gig work, Crowdsourcing, Future of work, Generative AI, Human-AI Collaborations

ACM Reference Format:
Kashif Imteyaz, Claudia Flores-Saviaga, and Saiph Savage. 2018. An LLM-Infused Tool for Workers' Collective Intelligence. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Harnessing collective intelligence among gig workers can significantly enhance their ability to improve labor conditions [5, 13, 27, 28]. This collaborative approach enables workers to collectively deliberate on their challenges, devise strategic solutions, and collaboratively implement their action plans to tackle these issues effectively[3, 9].

However, despite occasional successes, collective intelligence among gig workers is uncommon, leaving many issues unresolved [3]. This scarcity largely stems from the lack of technologies designed for facilitating collective problem-solving among gig workers [12, 14, 17, 34, 45]. Platforms like Dynamo and Coworker.org, which allow for sharing, prioritizing, and solving issues through upvote-based lists, fail to provide a comprehensive view of problems and solutions [7, 29]. Such list-based interfaces limit the exploration of issues from multiple angles and understanding the full scope of potential solutions [4, 15, 18, 20, 26, 41], often leading workers to prioritize less critical concerns [42]. This impedes meaningful progress in addressing the fundamental challenges of gig work, especially during the initial phases of collective intelligence focused on recognizing shared problems and formulating solutions [33].

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 the author(s) 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.
Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

Another hurdle is that the diverse and significant time commitments of gig workers may limit their ability to contribute to collective intelligence, underscoring the importance of developing inclusive tools that can accommodate workers' varying schedules, reducing the barriers to collective intelligence and enabling easier participation in collaboratively tackling challenges [24, 25]. To address these challenges, we introduce GigSense, a novel platform for fostering collective intelligence among gig workers by enabling collaborative problem-solving. By integrating Sense-Making Theory and leveraging the capabilities of large-language models (LLMs), GigSense features an LLM-enhanced interface that allows workers to deeply analyze their challenges and solutions from multiple perspectives [26]. Unlike existing platforms that offer a simple list-based view of issues [29], GigSense enables a detailed exploration of problems, offering both a close-up and a broad overview to understand workplace dynamics better. Additionally, GigSense uses LLMs to facilitate collective brainstorming, helping workers develop solutions together. Fig 1 presents an overview of GigSense. As a significant advancement in fostering collective intelligence among gig workers, GigSense addresses the unique challenges of this workforce, which typically operates in isolation without a common platform for problem identification and solution generation [4, 43]. GigSense bridges this gap by offering a communal space for gig workers to collectively understand and address their issue. Leveraging shared digital resources to build community [1, 32], GigSense emerges as a vital tool for a fragmented workforce, potentially catalyzing the formation of supportive gig worker communities [6, 38–40]. Beyond organization, it could reveal systemic issues, encouraging workers to see problems like client disputes and payment delays as collective challenges. This recognition can promote solidarity, crucial for collective intelligence, and empower workers to develop effective negotiation strategies, potentially improving their working conditions.

In this paper, we highlight a system that integrates LLMs with an interactive interface to support gig workers in problem-solving and kickstart collective intelligence among workers.

2 GIGSENSE

GigSense is an AI-enhanced platform designed to assist gig workers in the early phases of sensemaking and collective intelligence, which includes problem identification and solution proposal [34]. Our platform thus focuses on aiding workers in their sensemaking process (identifying collective issues and potential solutions). For this purpose, GigSense incorporates Pirolli and Card's sensemaking theories and collective intelligence theories into various modules within its system [21, 34, 41]. These modules are designed to automate the sensemaking process, which unfolds seamlessly within the entirety of GigSense's interface, with workers playing a crucial role in steering this process.

GigSense's modules are based on previous designs [11, 19], but uniquely integrate human-AI collaboration to address specific needs and contexts of gig workers. GigSense recognizes the time constraints faced by gig workers when engaging in collective intelligence and hence provides interfaces that facilitate rapid sensemaking of problems and proposals of solutions. For this purpose, it utilizes Large Language Models (LLMs) to provide concise summaries of issues found on gig platforms, enabling workers to quickly understand prevalent problems. By leveraging the OpenAI GPT-4 API [23], it categorizes data and generates these summaries, aiding workers in gaining insights and brainstorming solutions. Our system also offers an interactive interface, allowing users to "zoom in" and "zoom out" and explore specific discussions related to a problem. While LLM supports the creation of problem summaries, GigSense's interface design empowers workers to independently assess the information. It also provides a collaborative workspace for workers to propose solutions to identified problems. It uses LLMs to inspire workers with potential solutions but prioritizes human-generated ideas over AI suggestions to emphasize the value of human input. Notice that GigSense's design, focusing on quick sensemaking, contrasts with traditional extended analysis but offers notable advantages. It enables time-constrained individuals like gig workers with multiple jobs, parents, and students to participate, enhancing diversity and inclusivity. This approach makes sensemaking more accessible, as it requires less long-term commitment [10]. A faster process encourages broader participation and engagement, as people are more likely to contribute when their time investment is minimal yet impactful [30]. This approach can be especially beneficial for future engagements around collective intelligence, such as implementing solutions, where sustained participant involvement is crucial [10]. Additionally, rapid sensemaking can foster concise and focused information processing, reducing data overload and maintaining participant interest, which can wane in longer processes [22, 30]. Next, we describe each GigSense module in detail:

Data Gathering Module. Gig workers supply Gigsense with a roster of subreddits from which they intend to pinpoint potential problems and datasets containing assessments for gig work platforms (This is the "Step: Search and Filter" in Pirolli et al.'s sensemaking loop [26]). Next, Gigsense connects to the Reddit API to read and extract all the posts from the subreddits that gig workers initially provide. GigSense additionally uses a web scrapper to extract data from reviews left on Apple's and Google's app stores by gig workers. Note that our data gathering module only collects reviews that have between one and three-star ratings. The module considers that these review data would represent complaints and problems that gig workers are experiencing. Gigsense also lets workers manually enter issues into the system if they choose to do so ("Step: Read and Extract" in the sensemaking loop). Using the real-world gig workers' complaint datasets (actual gig workers' subreddits and complaints) in our system design aims to bring inclusiveness about gig worker concerns and complaints. The show button allows workers to see all the data sources used in GigSense.

Problem Summary Module. Acknowledging the potential enormity of the data gathered through the Data Gathering Module and its potential complexity for human interpretation, this module centers its efforts on summarizing the data ("Step: Schematize" in the sensemaking loop). For this purpose, Gigsense uses LLMs to facilitate the exploration and sensemaking of the extensive datasets of complaints, enabling workers to categorize and summarize the large datasets into problem categories. This module also has two

buttons to allow workers to navigate to the Data Visualization module, which offers a "zoom-in" view for meticulous examination of the data or a "zoom-out" view for a panoramic grasp of their issues.

Data Viz Module. This module enables workers to analyze problems at various levels, aiding their sensemaking process. It utilizes semantic zooming, a method for visualizing data across different abstraction levels [2, 36, 37]. Workers have access to two views showcasing varying detail levels for problem analysis. The zoomed-out view of this module presents an interactive chart of the problems ("Step: Schematize" in the sensemaking loop) where workers can grasp a high-level understanding of the specific types of problems (e.g. Payment, Platform Policy, Scam, Customer Support, Usability, etc) faced by them. Gigsense utilizes LLMs to streamline the sensemaking process by categorizing dataset posts into distinct categories. These categorized posts are later used by the system to create the different graphs and the corresponding category descriptions. The use of LLMs speeds up data organization, allowing workers to engage in sensemaking more efficiently. The interactive chart also serves as a powerful tool for raising awareness among gig workers about their shared challenges. For instance, workers can engage with these graphs by hovering over them, enabling access to the exact number posts (complaints) within each category. This functionality contributes to the worker's sensemaking process by facilitating the easy identification of prevalent issues.

The volume of information available to workers shifts according to the zoom level. To let the worker delve deeper into individual problems, Gigsense offers a zoom-in perspective. Choosing the zoom-in view shifts the display from graphs to the full text of posts relevant to the problem being analyzed. This view allows workers to view the nuanced aspects of issues that pertain to the topic analyzed and upvote on individual problems. By visually displaying the prevalence of issues (number of upvotes), it creates a sense of solidarity [31]. Workers can see that they are not alone in their struggles, fostering a collective identity and shared purpose [16]. It provides a static depiction of the number of workers grappling with problems across different gig work platforms, acting as a catalyst in building communities of gig workers facing similar challenges to encourage collective intelligence. The zoomable interface facilitates sensemaking by allowing users to toggle between high-level context and detailed analysis. Note that as [22] discusses, collective intelligence requires focused attention on what matters. The system is thus designed to allow workers to analyze problems from detailed and broader perspectives.

Collaborative Solution Module. This module further facilitates the sensemaking process and focuses on helping gig workers to develop concrete solutions to address the problem analyzed ("Step: Build Case' " in the sensemaking loop). It incorporates sub-modules such as the: "Sensemaking Chat", "Shared Document" and "Collaborative Solution Space". The Sensemaking Chat submodule allows workers to engage in conversations to discuss and investigate the problems they encounter in their work. They can communicate through asynchronous text messages to accommodate different schedules. The "Shared Document" enables workers to understand the existing problems and collaboratively create action plans (solutions) to address them. Gigsense also includes a functionality that allows users to make annotations in the shared document. This way, workers can collectively review and approve their proposed solutions. Finally, the "Collaborative Solution Space", just like the sensemaking process, features a space where workers can showcase the final solution they mutually agreed upon [8, 41] ("Step: Tell Story" in the sense-making loop). This module supports collaborative work that initiates early stages of collective intelligence[34].

AI-Enhanced Solution Module. It is important to acknowledge that for some workers it may be hard to propose solutions [24, 35]. To address this, the module leverages LLMs to offer workers suggestions on potential solutions and concrete collaboration plans, providing inspiration and initial guidance. However, given our values of prioritizing human connections among workers, AI-generated solutions are presented with lower priority in GigSense's interface. GigSense emphasizes that the AI-generated solution is not the definitive solution and explicitly states that this solution was generated using Generative AI. Note that in our system, AI-generated solutions are not personalized for individual users. However, the prompts sent to the LLM include all data related to the workers' specific problem. Using real-world gig worker data likely enhances our system's ability to provide contextually relevant AI solutions. Our appendix provides information about all the prompts we utilized for the LLM. Note also that GigSense uses pre-defined prompts to help gig workers effectively interact with LLMs, without requiring specialized skills [44]. The pre-defined prompts are also used because GigSense is not designed to address all queries, but instead, aims to offer concrete solutions for problem-solving inspiration. Furthermore, workers are informed about the possibility of errors in AI-generated solutions via a disclaimer. This helps not only to ensure that solutions generated by humans receive priority, but also foster more transparent and responsible use of Generative AI while still utilizing AI to assist and inspire workers in their problem-solving endeavors. This module is designed to enhance the "Step: Build Case" and "Step: Tell Story" stages of the sensemaking loop. It achieves this by promoting a collaborative process between humans and AI, fostering co-creation and mutual contribution to the development of solutions.

3 CONCLUSION

This paper introduces GigSense, an AI-powered tool designed to assist gig workers in collectively addressing their challenges. GigSense enables quick sensemaking, promoting effective problem-solving among workers by breaking down barriers to collective intelligence.

Acknowledgments This work was partially supported by an NSF grant FW-HTF-19541.

REFERENCES

[1] Stephen R Barley. 1986. Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative science quarterly* (1986), 78–108.

[2] Benjamin B Bederson and James D Hollan. 1994. Pad++ a zooming graphical interface for exploring alternate interface physics. In *Proceedings of the 7th annual ACM symposium on User interface software and technology*. 17–26.

[3] Ioulia Bessa, Simon Joyce, Denis Neumann, Mark Stuart, Vera Trappmann, and Charles Umney. 2022. *A global analysis of worker protest in digital labour platforms*. Vol. 70. International Labour Organization.

[4] Dan Calacci. 2022. Organizing in the End of Employment: Information Sharing, Data Stewardship, and Digital Workerism. In *2022 Symposium on Human-Computer Interaction for Work* (Durham, NH, USA) *(CHIWORK 2022)*. Association for Computing Machinery, New York, NY, USA, Article 14, 9 pages. https://doi.org/10.1145/3533406.3533424

[5] Julie Yujie Chen, Alessandro Delfanti, and Michelle Phan. 2023. Worker Resistance in Digital Capitalism| Worker Resistance in Digital Capitalism—Introduction. *International Journal of Communication* 17 (2023), 8.

[6] Philip R Cohen, Jerry Morgan, and Martha E Pollack. 2003. Collective Intentions and Actions. (2003).

[7] Alfred T DeMaria. 2016. The Future of Internet Organizing. *Management Report for Nonunion Organizations* 39, 2 (2016), 3–4.

[8] Brenda Dervin. 1983. An overview of sense-making research: Concepts, methods, and results to date. (1983).

[9] V.B. Dubal. 2017. The Drive to Precarity: A Political History of Work, Regulation, & Labor Advocacy in San Francisco's Taxi & Uber Economies. *Berkeley Journal of Employment and Labor Law* 38, 1 (2017), 73–135. http://www.jstor.org/stable/26356922

[10] Rebecca S Etz, Martha M Gonzalez, Aimee R Eden, and Jodi Winship. 2018. Rapid sense making: A feasible, efficient approach for analyzing large data sets of open-ended comments. *International Journal of Qualitative Methods* 17, 1 (2018), 1609406918765509.

[11] Claudia Flores-Saviaga, Shangbin Feng, and Saiph Savage. 2022. Datavoidant: An AI System for Addressing Political Data Voids on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–29.

[12] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[13] Rafael Grohmann, Mateus Mendonça, and Jamie Woodcock. 2023. Worker Resistance in Digital Capitalism| Communication and Work From Below: The Role of Communication in Organizing Delivery Platform Workers. *International Journal of Communication* 17 (2023), 19.

[14] Hannah Johnston. 2020. Labour geographies of the platform economy: Understanding collective organizing strategies in the context of digitally mediated work. *International Labour Review* 159, 1 (2020), 25–45.

[15] Vera Khovanskaya, Lynn Dombrowski, Jeffrey Rzeszotarski, and Phoebe Sengers. 2019. The Tools of Management: Adapting Historical Union Tactics to Platform-Mediated Labor. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 208 (nov 2019), 22 pages. https://doi.org/10.1145/3359310

[16] Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. 2015. More than telling a story: Transforming data into visually shared stories. *IEEE computer graphics and applications* 35, 5 (2015), 84–90.

[17] Vili Lehdonvirta. 2022. *Cloud empires: How digital platforms are overtaking the state and how we can regain control*. MIT Press.

[18] Hanlin Li, Nicholas Vincent, Stevie Chancellor, and Brent Hecht. 2023. The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1151–1161.

[19] Tianyi Li, Kurt Luther, and Chris North. 2018. Crowdia: Solving mysteries with crowdsourced sensemaking. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29.

[20] Lars Gunnar Mattsson, Daniela Corsaro, and Carla Ramos. 2015. Sense-making in business markets–the interplay between cognition, action and outcomes. *Industrial Marketing Management* 48 (2015), 4–11.

[21] Geoff Mulgan. 2018. *Big mind: How collective intelligence can change our world*. Princeton University Press.

[22] Geoff Mulgan. 2018. *Big mind: How collective intelligence can change our world*. Princeton University Press.

[23] OpenAI. 2024. OpenAI Platform — platform.openai.com. https://platform.openai.com/docs/api-reference. [Accessed 13-01-2024].

[24] Carol Packham. 2008. *Active citizenship and community learning*. Learning Matters.

[25] Maria-Carmen Pantea. 2013. The changing nature of volunteering and the cross-border mobility: where does learning come from? *Studies in Continuing Education* 35, 1 (2013), 49–64.

[26] Pirolli Peter and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of international conference on intelligence analysis* (2005).

[27] Frances Fox Piven and Richard Cloward. 2012. *Poor people's movements: Why they succeed, how they fail*. Vintage.

[28] Rida Qadri. 2023. Worker Resistance in Digital Capitalism| Algorithmized Not Atomized: The Distributed Solidarity of Jakarta's Gig Workers. *International Journal of Communication* 17 (2023), 20.

[29] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1621–1630.

[30] Jörgen Sandberg and Haridimos Tsoukas. 2015. Making sense of the sensemaking perspective: Its constituents, limitations, and opportunities for further development. *Journal of organizational behavior* 36, S1 (2015), S6–S32.

[31] Saiph Savage and Mohammad H Jarrahi. 2020. Solidarity and AI for Transitioning to Crowd Work during COVID-19. In *The New Future of Work Symposium 2020*.

[32] Steve Sawyer, Kevin Crowston, and Rolf T Wigand. 2014. Digital assemblages: evidence and theorising from the computerisation of the US residential real estate industry. *New Technology, Work and Employment* 29, 1 (2014), 40–56.

[33] Cyrille Schwellnus, Assaf Geva, Mathilde Pak, and Rafael Veiel. 2019. Gig economy platforms: Boon or Bane? (2019).

[34] Aaron Shaw, Haoqi Zhang, Andres Monroy-Hernandez, Sean Munson, Benjamin Mako Hill, Elizabeth Gerber, Peter Kinnaird, and Patrick Minder. 2014. Computer supported collective action. *interactions* 21, 2 (2014), 74–77.

[35] Jack W Snook and Dan C Olsen. 2006. *Recruiting, training, and maintaining volunteer fire fighters*. Jones & Bartlett Learning.

[36] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2023. Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. *arXiv preprint arXiv:2310.12953* (2023).

[37] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. *arXiv preprint arXiv:2305.11483* (2023).

[38] Henri Tajfel and John C Turner. 2004. The social identity theory of intergroup behavior. In *Political psychology*. Psychology Press, 276–293.

[39] Henri Tajfel, John C Turner, William G Austin, and Stephen Worchel. 1979. An integrative theory of intergroup conflict. *Organizational identity: A reader* 56, 65 (1979), 9780203505984–16.

[40] Martijn Van Zomeren, Tom Postmes, and Russell Spears. 2008. Toward an integrative social identity model of collective action: a quantitative research synthesis of three socio-psychological perspectives. *Psychological bulletin* 134, 4 (2008), 504.

[41] KE Weick. 1995. Sensemaking in organizations Sage Publications. *Thousand Oaks, CA* (1995).

[42] Alex Wood and Vili Lehdonvirta. 2019. Platform labour and structured antagonism: Understanding the origins of protest in the gig economy. *Available at SSRN 3357804* (2019).

[43] Zheng Yao, Silas Weden, Lea Emerlyn, Haiyi Zhu, and Robert E Kraut. 2021. Together but alone: Atomization and peer support among gig workers. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[44] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[45] Haoqi Zhang, Andrés Monroy-Hernández, Aaron D Shaw, Sean A Munson, Elizabeth M Gerber, Benjamin Mako Hill, Peter Kinnaird, Shelly Diane Farnham, and Patrick Minder. 2014. WeDo: End-To-End Computer Supported Collective Action.. In *ICWSM*.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Office-Mind AI: A Generative AI Tool for Gig Workers

Christopher Curtis
curtis.ch@northeastern.edu
Northeastern University
Boston, MA, USA

Seth Cooper
Northeastern University
Boston, MA, USA

Saiph Savage
Northeastern University
Boston, MA, USA

ABSTRACT

This paper studies the design of an AI tool that supports gig knowledge workers, rather than displacing them, focusing on text-based generative AI technologies. Through a formative study involving interviews and design activities, gig workers shared their views on text-based generative AI and envisioned applications where AI acts as managers, secretaries, and communication aids. Leveraging these insights, we created a generative-AI enhanced tool, Office-Mind AI, to aid gig workers. Our research advances the conversation around algorithmic labor by designing a worker-focused intelligent tool. This tool harness collective intelligence among workers and AI, fostering productive human-AI partnerships. We conclude by discussing the future prospects of collective intelligence tools designed for worker-AI collaborations.

CCS CONCEPTS

• **Human-centered computing**; • **Collaborative and social computing systems and tools**;

KEYWORDS

Gig work, Crowdsourcing, Future of work, Generative AI, Human-AI Collaborations

ACM Reference Format:
Christopher Curtis, Seth Cooper, and Saiph Savage. 2024. Office-Mind AI: A Generative AI Tool for Gig Workers. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Recently we have seen a surge in companies focused on generative AI, with entities like OpenAI championing missions to ensure AI benefits humanity [6, 12]. While these ambitions are commendable for their focus on societal good, it is crucial to recognize the potential for such AI advancements to displace human workers, particularly in tasks that are becoming increasingly automated [15, 17]. This trend poses a significant risk to jobs traditionally filled by gig knowledge workers, i.e., those workers who are involved in analysis, problem-solving, and creative tasks [16]. Together, this underscores the urgency to address AI-induced displacement within the gig economy [9, 11, 13]. As a result, generative AI technologies have sparked controversy and debate, highlighting the divide between those calling for regulatory measures to prevent worker displacement [7, 17], and others who see the technology's potential to augment rather than replace human labor [1]. However, despite this optimism, there is a gap in our understanding of how to design generative AI technologies to best serve workers [5, 18]. In this paper we focus on studying how to design generative AI tools for gig knowledge workers. The challenge lies in addressing the unique demands and constraints faced by them, as they are workers who operate without the support structures typical in more traditional employment settings [10, 16]. This underscores the need for a nuanced approach to designing generative AI tools that align with the specific needs and conditions of the gig economy workforce.

In this paper, we address this knowledge gap by taking the lead in understanding the generative AI technologies that gig knowledge workers both envision and desire for their jobs, and studying the usability of one of such tools. To achieve this, we first conducted an IRB-approved formative study that included interview sessions with design activities involving 21 knowledge gig workers from the Upwork platform. Our design activities concentrated on exploring text-based generative AI tools, such as ChatGPT, but as envisioned by workers. The focus on this technology was chosen because a significant portion of gig knowledge workers' labor is text-based [4]. While concentrating on this technology, we also welcomed participants' suggestions for other technological directions. Our interviews and design activities helped us to gather valuable insights into worker concerns, perspectives, and their future visions regarding generative AI tools for gig knowledge workers. Upon conducting an analysis of our interview and design sessions, a noteworthy finding emerged: gig knowledge workers envision text-based generative AI technology assuming roles akin to those found in traditional work structures, particularly in managerial and secretarial capacities. However, they imagined these intelligent agents working independently for them and not for the gig platform. Furthermore, these workers expressed a willingness and comfort in leveraging AI for worker-client communications, encompassing general interactions, negotiations, and self-marketing endeavors.

Armed with these insights, we created an AI enhanced tool - Office-Mind AI[1] - that could support gig knowledge workers in their job. Our tool integrates functionalities akin to two conventional office positions—managers and secretaries—to assist workers in executing a range of tasks highlighted as key in our interviews. These tasks include generating biographies, crafting emails and cover letters, performing translations, summarizing texts, and creating schedules. We conducted an evaluation of Office-Mind through an IRB-approved usability study involving 16 gig workers sourced from the Upwork platform. In this usability study, participants completed various tasks using the Office-Mind AI. Following these tasks, they were invited to share their feedback on the tool and

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

[1]Watch our AI-enhanced tool in action in this video demo: https://vimeo.com/920192151?share=copy

complete the System Usability Scale (SUS), a recognized methodology for assessing a system's usability [2, 14]. Participants expressed that Office-Mind AI offers convenient, and time saving solutions to challenges that gig workers face. Our study also showed that Office-Mind was easy to use, and scored within the 96th-100th percentile range of usability scores with a median SUS value of 87.5 [14], which categorizes it as having "Excellent" usability [2]. We finish the paper discussing the future of generative AI tools for workers. Our research not only contributes to our understanding of generative AI's potential in gig work, especially text-based generative AI for gig knowledge workers, but also paves the way for the collective intelligence community to delve into human centered design research tailored to this unique workforce.

2 FORMATIVE STUDY: METHODOLOGY

To study and create text-based generative AI tools tailored to the specific needs of gig workers, we initiated our research with a formative study. This initial phase provided valuable insights that guided our design and creation of AI tools closely aligned with gig workers' requirements. Our formative study's structure is outlined in Fig. 1.

3 OFFICE-MIND AI: A GENERATIVE AI TOOL FOR GIG WORKERS

Motivated by our formative study, we created Office-Mind AI: an online platform that integrates generative AI with a user-friendly interface. Our aim was to empower gig knowledge workers in their roles by providing them with the generative AI tools they envision and need. Our formative study highlighted how workers wanted AI that could enhance their capabilities and fill the support lacking in the gig economy. It also emphasized workers' preference for having AI managers and AI secretaries that could personally assist them, rather than serve the gig platform. Hence, we designed Office-Mind AI to function as virtual secretaries and managers, operating within an independent website dedicated to supporting workers rather than being tied to a specific gig platform. Fig. 2 presents screenshots of Office-Mind's interface.

To implement Office-Mind AI we decided to employ a hybrid design: combining forms and chatrooms for content creation, alongside backend prompt engineering by connecting with OpenAI's API on our backened server. We integrated forms because we considered they were a straightforward way for non-technical users to share their expertise with AI, making them ideal for gig workers who can have diverse technical backgrounds [8]. Their simplicity ensures almost anyone can use them, promoting inclusivity in knowledge sharing. Additionally, these interfaces are easy to implement and adapt [8], crucial for rapidly responding to the evolving demands of gig work. Therefore, we believed that form-based interfaces would provide an accessible and flexible mechanism to foster seamless worker-AI collaboration in the fast-paced gig market. Accordingly, for each major task where we sought generative AI support for workers, we developed a form allowing workers to input any knowledge they wanted the AI to understand about the task and their objectives. Using a REST API, we created corresponding endpoints for each form. These endpoints utilize specific context prompts to guide the GPT model in understanding its intended role in assisting the worker. Fig. 2 presents examples of the forms we created within Office-Mind AI.

Upon form submission, the GPT model receives its instructions (prompts) along with the worker's inputs. The model then generates an initial response, which is displayed below the form. Additionally, we integrated a chatroom component beneath the AI-generated content. Within this chatroom, workers can collaborate with the AI to refine the generated output further. This setup facilitates curated knowledge exchange between human workers, who further share their expertise, needs, and contextual information, and the AI, which assumes various assistive roles based on the information supplied by the worker in each case.

Since our formative study revealed workers' desire for generative AI capable of fulfilling managerial and secretarial roles, we incorporated modules within Office-Mind to emulate this.

3.1 Manager Module

This module serves as a personal manager for gig workers, combining collective intelligence with generative AI to enhance time management and personal branding. The platform crafts tailored schedules by weaving gig workers' project specifics and their preferences for work-life balance into its dynamic dashboard. It then employs generative AI to formulate initial schedules for the workers. These preliminary schedules are further refined via chat-based interactions, ensuring a customized and adaptive planning experience. This module bridges an important existing gap in gig work by offering task scheduling support, combining worker insights with AI. The module acts as a manager, helping gig workers meet deadlines and arrange their preferred work schedules.

In terms of personal branding, the module aids workers to collaborate with AI to craft professional biographies and cover letters that highlight their strengths and align with the needs of clients. Through a chat interface, workers can then also refine these documents with AI assistance, addressing the gap in gig work related to self-branding for client attraction [3]. The module here also functions as a manager that focuses on empowering workers to showcase their skills and talents, thereby helping them to more easily reach their professional aspirations.

3.2 Secretary Module

This module acts as a personal secretary for gig workers, combining an intuitive dashboard with generative AI technology to help workers manage their administrative tasks. These tasks range from drafting emails and conducting thorough grammar checks to offering language translation services and crafting Non-Disclosure Agreements (NDAs). Through its user-friendly interface, gig workers can convey their specific needs for emails, translations, grammar corrections, and NDAs. In response, the generative AI utilizes its extensive knowledge base to compose professional-grade documents tailored to each requirement (Fig. 2). The module's chat-based interface also facilitates a dynamic interaction, allowing workers to refine and enhance these documents collaboratively. This process not only leverages the workers' unique insights but also fosters a robust human-AI partnership, ultimately resulting in the creation of high-quality documents that can meet the professional demands of their gig work.

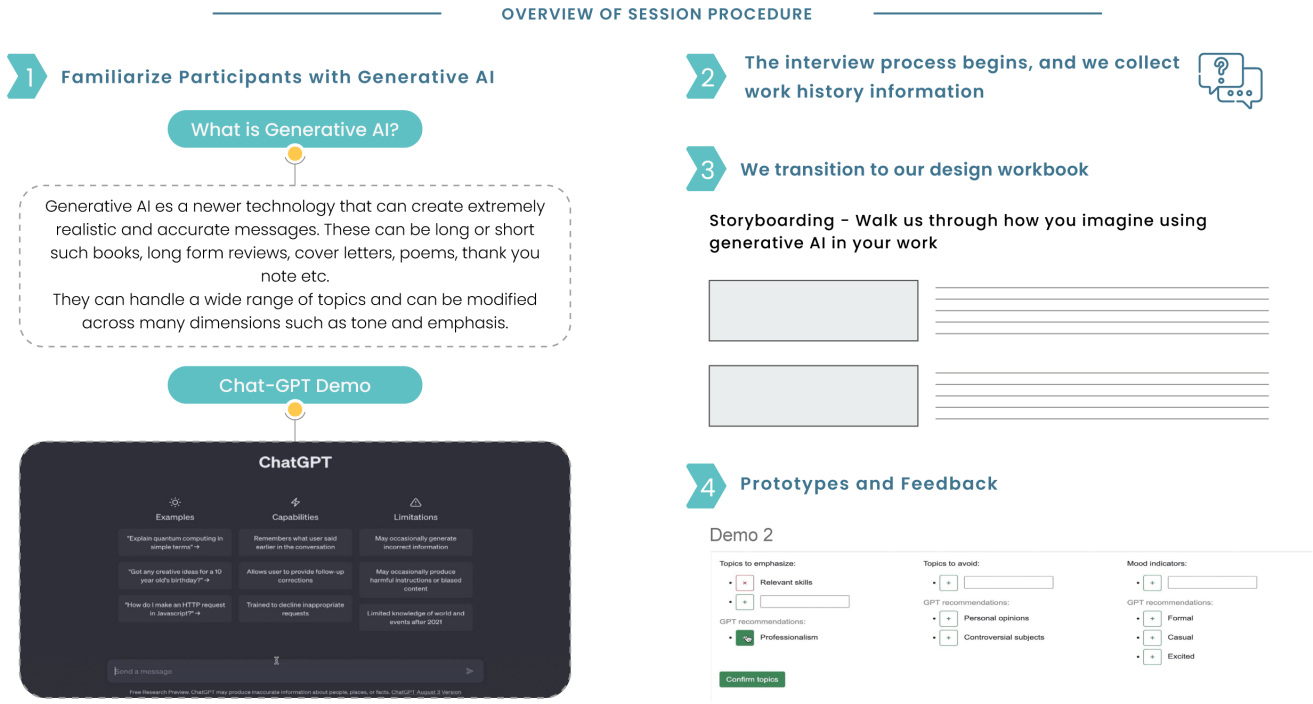


Figure 1: Formative Study Procedure: 1. Participants are familiarized with Text Based Generative AI. They are given a working definition, and provided a demo of ChatGPT. 2. The interview process begins, and we collect work history information. 3. We begin using the design workbook. Here we collect features, use-cases, and concerns for using Text Based Generative AI in participant's gig knowledge work. We also extract imaginaries from participants regarding the roles and forms generative AI can take to assist them. 4. Participants are shown prototype applications and feedback is provided in the form of feature and/or design suggestions.

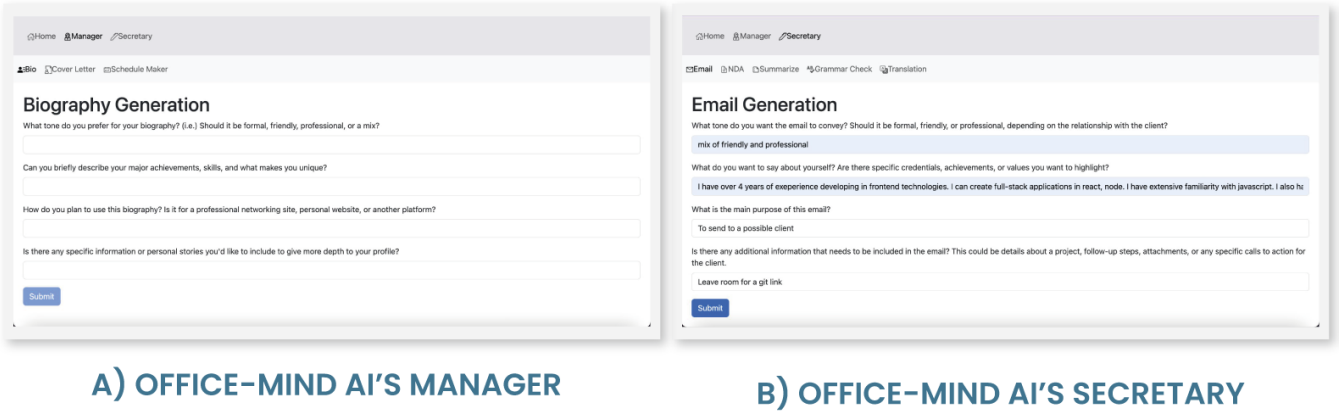


Figure 2: Office-Mind AI's Manager and Secretary

4 CONCLUSION

In this paper, we argue for using AI to enhance rather than replace workers, focusing on gig knowledge workers in the gig economy. Through an IRB-approved formative study, including co-design sessions and interviews, we explored these workers' needs and preferences for AI assistance. Our findings indicate a preference

for text-based generative AI to take on traditional workplace roles, such as manager and secretary, offering features like schedule management and email generation. Using insights from our study, we developed Office-Mind AI, a tool designed to augment gig workers' daily tasks without replacing them. Through an IRB-approved usability study, we evaluated the tool's effectiveness in supporting gig workers, highlighting the potential of text-based generative AI to enhance work processes. These results lay the groundwork for future research to refine and expand AI tools in the workplace, ensuring they complement human workers effectively.

Acknowledgments. This work was partially supported by a seed grant from the Khoury College of Computer Sciences.

REFERENCES

[1] [n. d.]. Generative AI likely to augment rather than destroy jobs — ilo.org. https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_890740/lang--en/index.htm. [Accessed 05-09-2023].

[2] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[3] Dorothy Lee Blyth, Mohammad Hossein Jarrahi, Christoph Lutz, and Gemma Newlands. 2022. Self-branding strategies of online freelancers on Upwork. *new media & society* (2022), 14614448221108960.

[4] Christopher Corbel, Trent Newman, and Lesley Farrell. 2022. Gig expectations: Literacy practices, events, and texts in the gig economy. *Written Communication* 39, 1 (2022), 66–96.

[5] David De Cremer and Garry Kasparov. 2021. AI should augment human intelligence, not replace it. *Harvard Business Review* 18 (2021), 1.

[6] A Shaji George and AS Hovan George. 2023. A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal* 1, 1 (2023), 9–23.

[7] Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating ChatGPT and Other Large Generative AI Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1112–1123. https://doi.org/10.1145/3593013.3594067

[8] Caroline Jarrett and Gerry Gaffney. 2009. *Forms that work: Designing Web forms for usability*. Morgan Kaufmann.

[9] Arif Jetha, Ali Shamaee, Silvia Bonaccio, Monique AM Gignac, Lori B Tucker, Emile Tompa, Ute Bultmann, Cameron D Norman, Cristina G Banks, and Peter M Smith. 2021. Fragmentation in the future of work: A horizon scan examining the impact of the changing nature of work on workers experiencing vulnerability. *American journal of industrial medicine* 64, 8 (2021), 649–666.

[10] Jasmin Manseau. 2019. AI in the workplace: The case of intelligent employee assistants. (2019).

[11] Sedigheh Moghavvemi, Lee Su Teng, and Huda Mahmoud. 2023. The Future of Work, A Knowledge-based Gig Economy. In *Reshaping the Future: The Phenomenon of Gig Workers and Knowledge-Economy*. Emerald Publishing Limited, 61–71.

[12] OpenAI. 2023. Planning for AGI and beyond — openai.com. https://openai.com/blog/planning-for-agi-and-beyond. [Accessed 13-09-2023].

[13] Mani Pillai. 2023. "Who am I? What am I doing?" The impact of the COVID-19 pandemic on work identities. *Employee Relations: The International Journal* 45, 6 (2023), 1527–1547.

[14] Jeff Sauro and James R Lewis. 2016. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

[15] Alana Semuels. 2020. Millions of Americans have lost jobs in the pandemic—And robots and AI are replacing them faster than ever. *Time magazine* (2020).

[16] Will Sutherland, Mohammad Hossein Jarrahi, Michael Dunn, and Sarah Beth Nelson. 2020. Work precarity and gig literacies in online freelancing. *Work, Employment and Society* 34, 3 (2020), 457–475.

[17] Krzysztof Wach, Cong Duong, Joanna Ejdys, Rūta Kazlauskaitė, Pawel Korzynski, Grzegorz Mazurek, Joanna Paliszkiewicz, and Ewa Ziemba. 2023. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review* 11, 2 (Jun. 2023), 7–30. https://doi.org/10.15678/EBER.2023.110201

[18] Angie Zhang, Alexander Boltz, Chun Wei Wang, and Min Kyung Lee. 2022. Algorithmic management reimagined for workers and by workers: Centering worker well-being in gig work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Complex Negotiations as Collective Search

NELBERTO NICHOLAS "SAM" QUINTO, University College London
JOSHUA BECKER, University College London
JON ATWELL, Stanford University
CYNTHIA WANG, Northwestern University
JENNIFER WHITSON, University of California, Los Angeles
HART POSEN, Dartmouth University

1. INTRODUCTION

Why do negotiations fail to reach an agreement even when all parties would benefit? Impasses can result from various socio-cognitive factors such as biased deal evaluations, perceptions of bad faith, and poor information about preferences or alternatives at play. [Schweinsberg et al. 2022; Malhotra and Bazerman 2008; Babcock and Loewenstein 1997; Tsay et al. 2011; Tversky and Kahneman 1991]. To overcome these challenges, negotiation experts and scholars have advocated for a collaborative approach that views negotiation as a joint search process where parties work together to explore the solution space and share information about their underlying values, interests, and alternatives in order to uncover agreeable terms [Doyle et al. 2023; Fisher and Ury 2007; Thompson et al. 1996].

However, even when negotiators search collaboratively, impasses can still occur when the process of finding agreement is inherently complex. Existing research has associated this complexity with structural factors that limit the zone of possible agreement or ZOPA [Thompson et al. 1996; Zhang et al. 2021] such as significant interest incompatibility or strong best alternative (BATNA) values [Fisher and Ury 2007]. And yet a narrow ZOPA does not explain why negotiations result in impasse when agreement is possible. For instance, rational negotiators can reach agreement by continuously making concessions that are more favorable than their best alternative. This raises a critical question in negotiations theory: what determines complexity in searching for an agreement?

2. MODEL AND FINDINGS

In this paper, we formalize the idea of negotiations as a search process by adapting computational models of problem solving. Our formal model transforms a negotiation scenario into a fitness landscape [Wright 1931; Kauffman 1993] by scoring all possible outcomes (i.e., proposals) and identifying which ones negotiators deem agreeable. It then evaluates the complexity of reaching any of these agreeable outcomes based on assumed search behaviors. Our primary results reveal that negotiations commonly conducted within MBA programs such as Harborco [Madigan et al. 1984], New Recruit [Neale 1997], and Fillmore Lawns [Kteily et al. 2020] exhibit features of a "rugged landscape" [Kauffman 1993; Levinthal 1997] where search may result in entrapment in non-agreeable local optima despite the existence of better and agreeable alternatives.

In line with the computational search literature, we obtain these initial results under the assumption that negotiators are boundedly rational [Simon 1991]; that is, they search strategically—only moving towards proposals they evaluate as more attractive, but myopically—only capable of considering proposals that differ by one feature from the preceding one. We then varied our assumption of how negotiators evaluate proposals from one that prioritizes the number of parties in agreement, to one focused on collective payoff gains, pareto superior gains (i.e. no individual party is worse off), or a

combination of these. Our model demonstrates that 'ruggedness,' persists across various behavioral assumptions, even when search becomes increasingly less myopic.

2.1 Comparing Complexity with the Scarceness in Agreeable Outcomes (ZOPA) as Causes of Impasse

We subsequently examined how ruggedness (measured by the number of non-agreeable local optima) compares to a narrower ZOPA (measured by the number of agreeable proposals) as a complexity metric. To do this, we devised a probabilistic approach to creating generic negotiation scenarios with varying ZOPA and ruggedness. We generated over 200 scenarios each with a sufficiently large solution space of 256 possible outcomes and 10 negotiating parties with varying interests. We chose four scenarios that best correspond to a 2x2 computational experiment involving Smooth/Rugged negotiations with either a Narrow/Wide ZOPA and conducted agent-based negotiation simulations on each one to estimate the likelihood of reaching an agreement (agreement rate), as detailed in Table 1.

Treatments	Smooth Landscape	Rugged Landscape
Wide ZOPA	ZOPA: **36** agreeable solutions Ruggedness: **0** local optima (**116** neighboring solutions and **0** starting proposals that cannot lead to agreement under search assumptions)	ZOPA: **42** agreeable solutions Ruggedness: **1** local optimum (**104** neighboring solutions and **93** starting proposals that cannot lead to agreement under search assumptions)
Narrow ZOPA	ZOPA: **1** agreeable solution Ruggedness: **0** local optima (**8** neighboring solutions and **0** starting proposals that cannot lead to agreement under search assumptions)	ZOPA: **1** agreeable solution Ruggedness: **31** local optima (**8** neighboring solutions and **58** starting proposals that cannot lead to agreement under search assumptions)

Table. 1. Computational Experiment Conditions for Generating Figure 1 results

Our model shows that both ruggedness and a narrow ZOPA contribute to lower agreement rates. Moreover, it uncovers the counterintuitive prediction that a negotiation with a wide ZOPA can have a lower agreement rate than one with a narrow ZOPA if the wide ZOPA negotiation is more rugged. As summarized in Figure 1, this finding is robust to a negotiation deadline of six "search steps" and extends to more nuanced characterizations of ZOPA and ruggedness. Specifically, results hold even as ZOPA is defined by the number of proposals that differ by one feature from an agreeable proposal, and even when ruggedness is defined in terms of the number of starting points that cannot lead to agreement based on the search strategy used.

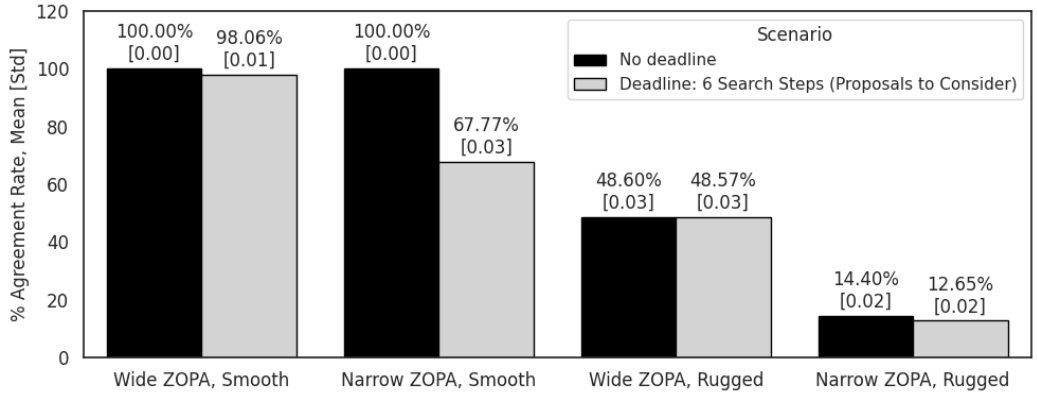


Fig. 1. Agreement Rates across Negotiations with varying ZOPA and Ruggedness

2.2 Comparing Complexity and "Running out of Time" as Causes of Impasse

We varied the deadline to demonstrate that ruggedness becomes more salient to negotiation outcomes when time constraints are relatively lenient, whereas a narrow ZOPA is more likely to lead to impasse in scenarios where negotiators have limited time and can go over only a few proposals. As illustrated in Figure 2, when the deadline allows for the consideration of more than 5 proposals, ruggedness becomes the primary obstacle to agreement, regardless of the number of available agreeable solutions. This is because, in rugged negotiations, impasse increasingly becomes attributed to entrapment in local optima as opposed to "running out of time" when negotiation deadlines become more lenient. Likewise, while extending deadlines can significantly eliminate impasse rates in smooth negotiations, they have less of an effect as ruggedness increases.

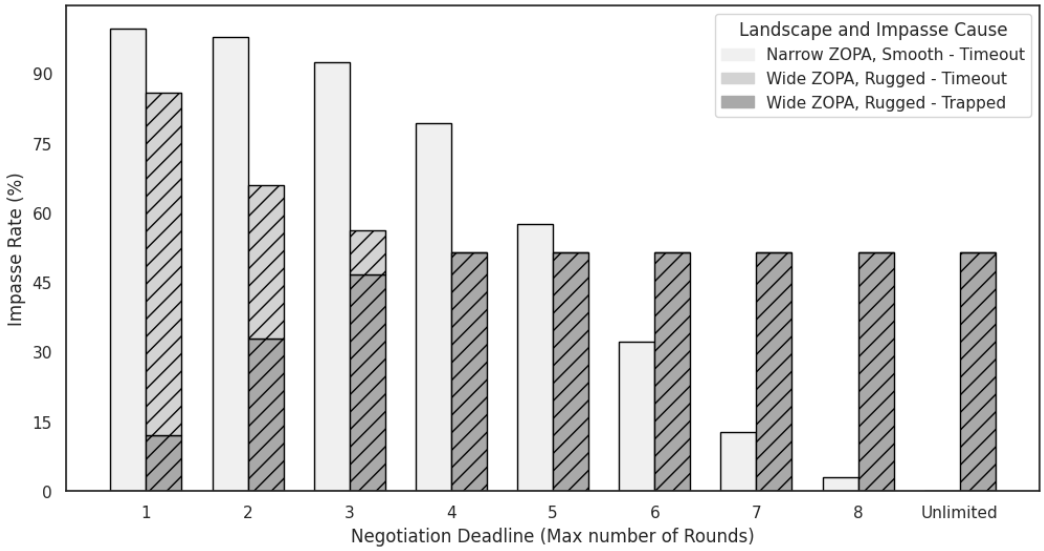


Fig. 2. Impasse Rates by Deadline and Cause for Smooth vs Rugged Negotiations

3. DISCUSSION

In conclusion, we build theory by showing that even when agreement benefits all parties, traditional culprits of impasse such as socio-cognitive barriers or scarcity in agreeable solutions are not necessary conditions for failing to reach agreement. Thus, even careful, collaborative, well intentioned, and rational, approaches to negotiation still do not guarantee agreement. On the other hand, our findings demonstrate that the inherent complexity of a negotiation is a sufficient condition for impasse, and this holds across a variety of strategic search behaviors. Even with realistic deadlines, our model highlights that negotiations can often result in impasse due to sheer complexity as opposed to merely running out of time.

Our work thus illuminates a previously unmentioned structural cause of impasse and contributes to ongoing efforts to understand why negotiations fail to reach agreement [Schweinsberg et al. 2022; Tripp and Sondak 1992; Babcock and Loewenstein 1997]. Gaining a clearer understanding of what can drive impasses allows for more accurate diagnoses and enables the development of focused interventions and negotiation strategies. In addition to advancing negotiations theory, our research contributes to research in collective intelligence. Whereas prior work on collective search often models individuals as sharing the same goal [Becker et al. 2022; Safferstone 1998], our study introduces a framework for examining collective search amidst conflicting preferences and interests.

REFERENCES

Babcock, L., & Loewenstein, G. (1997). Explaining Bargaining Impasse: The Role of Self-Serving Biases. *Journal of Economic Perspectives*, *11*(1), 109–126. https://doi.org/10.1257/jep.11.1.109

Becker, J. A., Guilbeault, D., & Smith, E. B. (2022). The Crowd Classification Problem: Social Dynamics of Binary-Choice Accuracy. *Management Science*, *68*(5), 3949–3965. https://doi.org/10.1287/mnsc.2021.4127

Brett, J., & Thompson, L. (2016). Negotiation. *Organizational Behavior and Human Decision Processes*, *136*, 68–79. https://doi.org/10.1016/j.obhdp.2016.06.003

Doyle, S. P., Chung, S., Lount, R. B., Swaab, R. I., & Rathjens, J. (2023). Hierarchical Team Structures Limit Joint Gain in Interteam Negotiations: The Role of Information Elaboration and Value-Claiming Behavior. *Academy of Management Journal*, *66*(5), 1586–1616. https://doi.org/10.5465/amj.2019.1381

Fisher, R., & Ury, W. (2007). *Getting to yes: Negotiating an agreement without giving in* (2. ed., repr). Random House.

Hopmann, T. (1995). Two Paradigms of Negotiation: Bargaining and Problem Solving. *The ANNALS of the American Academy of Political and Social Science*, *542*(1), 24–47. https://doi.org/10.1177/0002716295542001003

Kauffman, S. A. (1993). *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press.

Kteily, N., Whitson, J., Wang, C., & Akason, D. (2020). *Fillmore Lawns*. Northwestern University, Kellog School of Management, Dispute Resolution Center. https://new.negotiationexercises.com/exercises/

Levinthal, D. A. (1997). Adaptation on Rugged Landscapes. *Management Science*, *43*(7), 934–950.

Madigan, D., Weeks, T., & Susskind, L. (1984). *Harborco: Role-Play Simulation*. Program on Negotiation at Harvard Law School. https://www.pon.harvard.edu/freemium/harborco-role-play-simulation-2/

Malhotra, D., & Bazerman, M. H. (2008). Psychological Influence in Negotiation: An Introduction Long Overdue. *Journal of Management*, *34*(3), 509–531. https://doi.org/10.1177/0149206308316060

Neale, M. A. (1997). *New Recruit*.

Prietula, M. J., & Weingart, L. R. (2011). Negotiation Offers and the Search for Agreement. *Negotiation and Conflict Management Research*, *4*(2), 77–109. https://doi.org/10.1111/j.1750-4716.2011.00074.x

Safferstone, M. J. (1998). The Distributed Mind: Achieving High Performance Through the Collective Intelligence of Knowledge Work Teams. *Academy of Management Perspectives*, *12*(2), 107–108. https://doi.org/10.5465/ame.1998.650529

Schweinsberg, M., Thau, S., & Pillutla, M. M. (2022). Negotiation Impasses: Types, Causes, and Resolutions. *Journal of Management*, *48*(1), 49–76. https://doi.org/10.1177/01492063211021657

Simon, H. A. (1991). Bounded Rationality and Organizational Learning. *Organization Science*, *2*(1), 125–134. https://doi.org/10.1287/orsc.2.1.125

Thompson, L., Peterson, E., & Brodt, S. E. (1996). *Team Negotiation: An Examination of Integrative and Distributive Bargaining*. https://oce.ovid.com/article/00005205-199601000-00006/HTML

Tripp, T. M., & Sondak, H. (1992). An evaluation of dependent variables in experimental negotiation studies: Impasse rates and pareto efficiency. *Organizational Behavior and Human Decision Processes*, *51*(2), 273–295. https://doi.org/10.1016/0749-5978(92)90014-X

Tsay, C.-J., Shu, L. L., & Bazerman, M. H. (2011). Naïveté and Cynicism in Negotiations and Other Competitive Contexts. *Academy of Management Annals*, *5*(1), 495–518. https://doi.org/10.5465/19416520.2011.587283

Tversky, A., & Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061. https://doi.org/10.2307/2937956

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*(2), 97–159.

Zhang, H., Zhang, K., Warsitzka, M., & Trötschel, R. (2021). Negotiation complexity: A review and an integrative model. *International Journal of Conflict Management*, *32*(4), 554–573. https://doi.org/10.1108/IJCMA-03-2020-0051

Recipe for Success: Quantifying the Relative Influence of Teamwork Facets on Performance

Xinlan Emily Hu*
The Wharton School, University of Pennsylvania
Philadelphia, PA, U.S.A.
xehu@wharton.upenn.edu

Abdullah Almaatouq
Sloan School of Management, Massachusetts Institute of Technology
Boston, MA, U.S.A.

Mark E. Whiting
University of Pennsylvania
Philadelphia, PA, U.S.A.

Duncan J. Watts
University of Pennsylvania
Philadelphia, PA, U.S.A.

Abstract

How do different facets of teamwork — e.g., a team's composition, size, communication, and task — contribute to its performance? Theories of teamwork have focused on many possible attributes that can explain a team's success. In this study, we complement existing explanatory models through a prediction-focused analysis. We randomly assign participants to teams of varying compositions, sizes, and tasks, then quantify the variance explained by each factor. Ultimately, task attributes emerge as a predominant contributor to performance, explaining 20% of variation, compared to just 2% for composition. Our study highlights the context specificity of team outcomes; if task attributes so strongly determine performance, insights about a team performing one task may not generalize to another. For managers, the "fit" between a team and its task may become crucial.

CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Applied computing** → **Psychology**.

Keywords

Teams, Groups, Collective Intelligence, Communication, Composition, Performance, Tasks, Prediction

ACM Reference Format:
Xinlan Emily Hu, Abdullah Almaatouq, Mark E. Whiting, and Duncan J. Watts. 2024. Recipe for Success: Quantifying the Relative Influence of Teamwork Facets on Performance. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 3 pages.

Introduction. Teams are dynamic, multifaceted systems. When tackling problems, they integrate the composition of their members — their personalities [3], skills [12], and backgrounds [13] — and adapt their collaboration patterns to meet task demands [15, 16].

Scholarship on teamwork has embraced this complexity and dynamism [6, 17, 18]. Models of teams increasingly account for a variety of composition variables, alongside their emergent processes and behaviors (e.g., [23]). However, these models typically take an explanatory approach, focusing on demonstrating that the model's parameters are statistically consistent with established theories [10, 11]. There has been much more limited focus on evaluating theories in terms of their predictive accuracy. That is, if a manager could choose to focus on improving just one aspect of their team, which area of improvement would yield the highest returns? And even if they get everything "right" — how much of a team's outcome is truly within their control? By taking a predictive, rather than explanatory, approach, our work seeks not to establish causal effects, but to complement causal theorizing by estimating the relative importance of different team facets in determining downstream outcomes.

Methodological limitations have historically hindered the ability to conduct a comparative predictive analysis between teamwork facets, as it is rare to observe systematic variation across different aspects of teamwork. In field settings, teams may be self-selected or hired specifically for a project, which confounds team composition with task demands. In laboratory settings, many studies rely only on a single task [9, 24]. Finally, different study environments may be incommensurable in a meta-analysis [2].

We take a novel experimental approach that uses a series of controlled experiments in a fully parameterized space — systematically varying attributes such as the team composition, team size, and the task at hand. By studying teams in an environment in which nearly everything about them is known, we can estimate the predictive power of each team facet on performance.

Methods. We recruited N = 125 teams (534 individuals) from Amazon Mechanical Turk to participate in a series of real-time collaborative games on Empirica [1]. Participants were randomly assigned to play either in a team of three (72 teams) or six (53 teams); each team then played five consecutive sets of three games, for a total of 15 trials. A set consisted of three randomly-ordered variations of the same style of game, each with different levels of complexity (low, moderate, and high). We call a "set" of games with the same general style a *task class*, and a specific variation of the game a *task instance*. After excluding incomplete data, we study a total of 423 task classes and 1,043 task instances.

Selection of Games. To study systematic variations in task features, we used the Task Space [14] to curate 10 task classes that

*Corresponding Author

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

were as diverse as possible (maximizing Euclidean distance): (1) Moral Reasoning, (2) Guess the Correlation, (3) Wolf Goat Cabbage Puzzle, (4) Writing Story, (5) Room Assignment Task, (6) Sudoku, (7) Allocating Resources, (8) Divergent Association Task, (9) Word Construction, and (10) Whac-A-Mole. Each of these tasks has vastly different stimuli and objectives, challenging teams to use varying styles of thinking and skills.

Selection of Team Facets. Drawing on reviews and meta-analyses by Mathieu et al. [17, 18], we identified five facets of teamwork: (1) Team Composition, (2) Team Size, (3) Task Attributes, (4) Task Complexity, and (5) Communication Process, which we measure as follows.

Measures. Team Composition variables include 19 characteristics collected at the individual level, then aggregated at the team level using mean and standard deviation. These dimensions include basic demographics (birth year, country, education level, gender, race, income maximum/minimum, and marital status); political affiliation (political party, orientation on fiscal issues, and orientation on social issues [21]); Cognitive Reflection Test score [8]; six constructs related to individualism and collectivism [4]; and the Reading the Mind in the Eyes score [20].

Team Size is determined by randomization (either 3 or 6). Task Attributes are the 24 dimensions drawn directly from the Task Space [14], and Task Complexity is encoded as a dummy variable (low, moderate, or high). Finally, Communication features are drawn from a systematic review of validated measures (e.g., [5]). Because teams had a continuous chat over multiple games, we examined the chats at three different levels — the *instance level* (analyzing only what was said about a specific game), the *class level* (analyzing everything said about a set of games), and the *cumulative level* (analyzing everything spoken up to the present). Across all models, our dependent variable is task performance, which is normalized from 0 to 100 for each task.

Model Specification. To test the predictive power of each team facet, we fit separate ElasticNet regressions predicting task performance using the features in each facet. Each model's hyperparameters (the penalty term and L1 ratio) were seeded with the results of 10,000 iterations of k=5-fold cross-validation, then tuned with 25 iterations of Bayesian Optimization using Facebook's Adaptive Experimentation (Ax) platform. We then evaluate the models using Q^2, a measure similar to R^2 that estimates prediction quality with jackknifed residuals [7, 22], at both the task class and the task instance levels.

Results. Figure 1 and Table 1 summarize our main results. Among the five teamwork facets we considered, task attributes emerged as by far the best predictor of a team's score; since Q^2, like R^2, can be interpreted as a "percentage of variance" explained by the model, these features explained 21.7% of variance at the task instance level, and 19.4% of variance at the task class level. Pairwise t-tests between task attributes and each of the other facets are all significant at the $p < 0.05$ level, after correcting for multiple comparisons using the Benjamini-Hochberg procedure. Relative to task attributes, other features had minimal predictive power. Task complexity was near-zero at the instance level and 2.7% at the class level. Team composition was also near-zero at the instance level and only 2.3% at the class level, and team size also had negligible Q^2 at both levels of analysis. This may partly reflect the nature of the activity; while a laboratory study makes it possible to precisely operationalize each team factor, it also limits the scope to teams playing short games (lasting no more than 5 to 10 minutes). Given the brevity of these interactions, team members may have limited opportunities to showcase their individual personalities, which may explain the minimal importance of team composition.

Table 1: Main results from our preliminary analysis (as this is ongoing work). The top row shows results for the Task Instance unit of analysis (n = 1,043) and the bottom for the Task Class unit of analysis (n = 423). Task Instance is abbreviated as *T.I.*; Task Class as *T.C.*, and Cumulative as *C.*

Factor	Mean	Stdev	95% CI Lower	95% CI Upper
Task Instance				
Team Composition	-0.005	0.013	-0.008	-0.003
Team Size	-0.026	0.008	-0.027	-0.024
Task Attributes	0.217	0.023	0.213	0.222
Task Complexity	-0.009	0.011	-0.012	-0.007
Communication (*T.I.*)	0.058	0.020	0.054	0.062
Communication (*T.C.*)	0.055	0.018	0.051	0.058
Communication (*C.*)	-0.010	0.012	-0.012	-0.007
Task Class				
Team Composition	0.023	0.029	0.017	0.029
Team Size	-0.022	0.012	-0.024	-0.019
Task Attributes	0.194	0.035	0.187	0.201
Task Complexity	0.027	0.026	0.022	0.033
Communication (*T.C.*)	0.050	0.033	0.044	0.057
Communication (*C.*)	-0.010	0.020	-0.014	-0.006

The predictive power of communication features varied depending on the analysis procedure. Under the cumulative scheme, communication features were not predictive at all. However, under the task class and task instance schemes, communication was more predictive (explaining 5 – 6% of variance). These results likely reflect the domain specificity of communication, as well as its endogeneity to other factors; that is, communication mediates "inputs" such as task attributes and team composition [18, 19]. As this research is ongoing, we plan to further explore the mediation relationship in our next analyses.

Our findings underscore the substantial role that the task plays in moderating findings about teams. Tasks were more predictive of a team's eventual success than nearly every other measurable factor, and given the prevalence of laboratory-based team studies in social science, our findings call attention to the fact that results from one task may not generalize to another [14]. They also reveal new avenues of inquiry: Mathieu et al.'s comprehensive review [17] found that team compositional features are studied 50% more frequently than task- or structure-related features (150 studies for the latter, compared to 99 studies for the former), even though, in our study, task features were 10 times more predictive of performance. Consequently, if researchers and practitioners seek to understand how a team will perform, they should begin by paying attention to what the team is being asked to do.

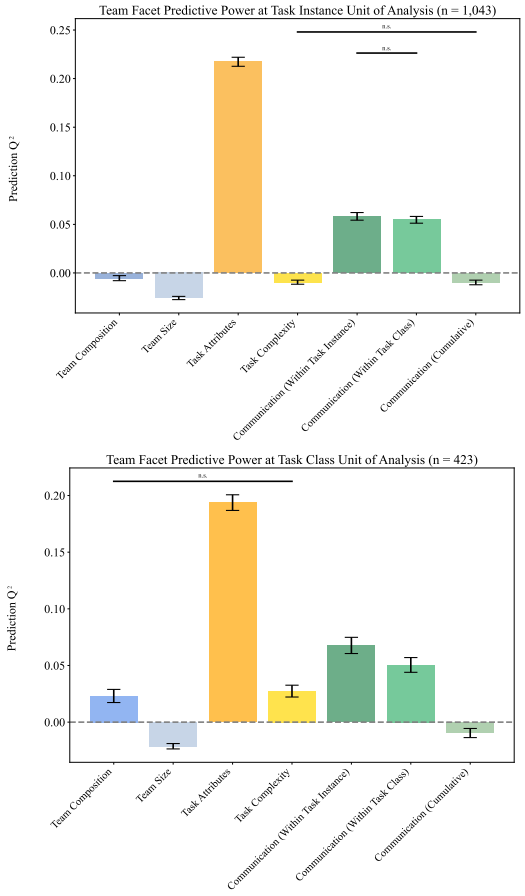


Figure 1: Results for the task instance (top panel) and the task class (bottom panel) units of analysis. Each panel depicts a bar graph with the Q^2 for each model, fit using the five different facets. There are multiple bars for Communication features, reflecting different ways of aggregating the communication features — either analyzing only the chats associated with a specific task instance ("Within Task Instance"); those associated with all levels of complexity for a task class ("Within Task Class"); or all team conversations that took place prior to the current activity, including discussions about other tasks ("Cumulative"). Error bars represent the 95% confidence intervals using the t statistic for the Q^2 estimates, based on 100 bootstrapped samples. Finally, we illustrate pairwise t-tests, correcting for multiple testing using the Benjamini-Hochberg procedure. As nearly all tests were significant to $p < 0.05$, we highlight only the non-significant comparisons (indicated with $n.s.$).

References

[1] Abdullah Almaatouq, Joshua Becker, James P Houghton, Nicolas Paton, Duncan J Watts, and Mark E Whiting. 2021. Empirica: A Virtual Lab for High-Throughput Macro-Level Experiments. *Behavior Research Methods* 53, 5 (2021), 2158–2171.

[2] Abdullah Almaatouq, Thomas L Griffiths, Jordan W Suchow, Mark E Whiting, James Evans, and Duncan J Watts. 2022. Beyond Playing 20 Questions With Nature: Integrative Experiment Design in the Social and Behavioral Sciences. *Behavioral and Brain Sciences* (2022), 1–55.

[3] Murray R Barrick, Greg L Stewart, Mitchell J Neubert, and Michael K Mount. 1998. Relating Member Ability and Personality to Work-Team Processes and Team Effectiveness. *Journal of Applied Psychology* 83, 3 (1998), 377.

[4] Marilynn B Brewer and Ya-Ru Chen. 2007. Where (Who) are Collectives in Collectivism? Toward Conceptual Clarification of Individualism and Collectivism. *Psychological Review* 114, 1 (2007), 133.

[5] Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A Toolkit for the Analysis of Conversations. *arXiv preprint arXiv:2005.04246* (2020).

[6] Matthew A Cronin, Laurie R Weingart, and Gergana Todorova. 2011. Dynamics in Groups: Are We There Yet? *Academy of Management Annals* 5, 1 (2011), 571–612.

[7] Lennart Eriksson, Tamara Byrne, E Johansson, Johan Trygg, and C Vikström. 2013. *Multi-and Megavariate Data Analysis Basic Principles and Applications*. Vol. 1. Umetrics Academy.

[8] Shane Frederick. 2005. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19, 4 (2005), 25–42.

[9] J Richard Hackman. 1969. Toward Understanding the Role of Tasks in Behavioral Research. *Acta Psychologica* 31 (1969), 97–128.

[10] Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and Explanation in Social Systems. *Science* 355, 6324 (2017), 486–488.

[11] Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. 2021. Integrating Explanation and Prediction in Computational Social Science. *Nature* 595, 7866 (2021), 181–188.

[12] John R Hollenbeck, Bianca Beersma, and Maartje E Schouten. 2012. Beyond Team Types and Taxonomies: A Dimensional Scaling Conceptualization for Team Description. *Academy of Management Review* 37, 1 (2012), 82–106.

[13] Sujin K Horwitz and Irwin B Horwitz. 2007. The Effects of Team Diversity on Team Outcomes: A Meta-Analytic Review of Team Demography. *Journal of Management* 33, 6 (2007), 987–1015.

[14] Xinlan Emily Hu, Linnea Gandhi, Mark E Whiting, Duncan J Watts, and Abdullah Almaatouq. 2023. Tasks Beyond Taxonomies: A Multidimensional Design Space for Team Tasks. (2023). PsyArXiv Preprint; Revise and Resubmit at Management Science.

[15] Stephen E Humphrey and Federico Aime. 2014. Team Microdynamics: Toward an Organizing Approach to Teamwork. *Academy of Management Annals* 8, 1 (2014), 443–503.

[16] Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. 2001. A Temporally Based Framework and Taxonomy of Team Processes. *Academy of Management Review* 26, 3 (2001), 356–376.

[17] John E Mathieu, Peter T Gallagher, Monique A Domingo, and Elizabeth A Klock. 2019. Embracing Complexity: Reviewing the Past Decade of Team Effectiveness Research. *Annual Review of Organizational Psychology and Organizational Behavior* 6 (2019), 17–46.

[18] John E Mathieu, John R Hollenbeck, Daan van Knippenberg, and Daniel R Ilgen. 2017. A Century of Work Teams in the Journal of Applied Psychology. *Journal of Applied Psychology* 102, 3 (2017), 452.

[19] Joseph Edward McGrath. 1964. Social Psychology: A Brief Introduction. (1964).

[20] Sally Olderbak, Oliver Wilhelm, Gabriel Olaru, Mattis Geiger, Meghan W Brenneman, and Richard D Roberts. 2015. A Psychometric Analysis of the Reading the Mind in the Eyes Test: Toward a Brief Form for Research and Applied Settings. *Frontiers in Psychology* 6 (2015), 1503.

[21] Gordon Pennycook and David G Rand. 2019. Cognitive Reflection and the 2016 US Presidential Election. *Personality and Social Psychology Bulletin* 45, 2 (2019), 224–239.

[22] Nguyen T Quan. 1988. The Prediction Sum of Squares as a General Measure for Regression Diagnostics. *Journal of Business & Economic Statistics* 6, 4 (1988), 501–504.

[23] Christoph Riedl and Anita Williams Woolley. 2017. Teams vs. Crowds: A Field Test of the Relative Contribution of Incentives, Member Ability, and Emergent Collaboration to Crowd-Based Problem Solving Performance. *Academy of Management Discoveries* 3, 4 (2017), 382–403.

[24] Tal Yarkoni. 2022. The Generalizability Crisis. *Behavioral and Brain Sciences* 45 (2022), e1.

Examining the Effects of Explainable Hints in AI-Driven Training

Tory Farmer
Washington University in St. Louis
St. Louis, MO, USA
toryfarmer@wustl.edu

Chien-Ju Ho
Washington University in St. Louis
St. Louis, MO, USA
chienju.ho@wustl.edu

ABSTRACT

AI-driven training systems have gained popularity across various disciplines, including disaster response, pilot training, and medical education. Many such systems use "hints," leveraging AI to identify key decision points and suggest actions to aid trainees. Some hint-based AI-driven training systems have been shown to improve trainee performance in deployment when conditions are similar to those in training. However, when conditions in deployment substantially differ from those in training, trainees often fail to generalize the hints to unseen scenarios, resulting in a decrease in performance. Meanwhile, the recent development of explainable AI has provided opportunities to address the generalization challenge by providing explanations to improve humans' understanding of AI hints. In this work, we explore the effect of providing explanations alongside hints in AI-driven training in a simple navigation task. As an exploratory investigation, we utilize large language models (LLMs) to generate explanations about which features of the task led to the given hint. Our preliminary results suggest that, in our simple navigation task with LLM-generated explanations, while providing explainable hints improves trainee performance in environments similar to training, it promotes over-reliance on the AI-provided hints. This results in decreased performance in environments unseen during training. Future work would include examining other mechanisms of explanation generation and investigating the effects in other tasks.

KEYWORDS

Explainable AI, Training, Hint Systems, Reliance

ACM Reference Format:
Tory Farmer and Chien-Ju Ho. 2024. Examining the Effects of Explainable Hints in AI-Driven Training. In *CI '24: ACM Collective Intelligence. June 26-29, 2024, Boston, MA.* ACM, Boston, MA, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As AI becomes increasingly powerful, it offers unprecedented opportunities to enhance human decision-making capabilities, particularly in complex and challenging environments. One common approach for AI to augment human intelligence is to provide suggestions to human decision-makers in the form of *hints*, where a hint is implemented as a suggestion of what action to take at a given time point. In particular, prior works have shown that it is possible to leverage AI techniques to identify effective hints to provide to human decision-makers [3, 7, 19].

While leveraging AI hints to assist humans is promising, offering these hints in deployment is not always practical. It can become too cumbersome or distracting, especially in situations that require focused attention from decision-makers. For example, a pilot evading enemy fire may not be able to process advice from an AI system in deployment due to the need to concentrate on immediate threats. In such scenarios, an alternative is to utilize AI hints during training rather than in deployment, helping trainees learn to avoid potential mistakes in deployment. However, similar to the sim-to-real challenge in robotics [20], there is often an unavoidable gap between the training environments and deployment environments. To effectively utilize AI hints for training, one major challenge is to ensure that humans can *generalize* what they have learned to environments they have not encountered during training.

To achieve generalization, human decision-makers need to understand the AI hints and selectively apply the concepts learned in training to tasks during deployment, as directly mimicking training actions and decisions might result in negative outcomes in previously unseen environments. Given this goal is similar to the recent development of explainable AI [18], in this work, we explore whether providing explanations alongside AI hints during training can help trainees to generalize the hints from training to deployment. In particular, during training time, in addition to providing trainees with AI hints, suggestions on what actions to take at given time points, we also provide explanations that suggest why the hints are generated.

As an exploratory study, we conducted experiments on the Mouselab game developed by Callaway [7]. We leverage their methodologies in generating AI hints. To generate explanations, we utilized recent advancements in large language models (LLMs). We provided ChatGPT with the task and the corresponding hint, asking it to generate explanations about which features of the task led to the given hint. Our results demonstrate that, in our experimental setting, providing explainable hints improves trainee performance in environments similar to training. However, it also promotes over-reliance on the AI-provided hints, resulting in decreased performance in environments not encountered during training. While our preliminary results are discouraging, they highlight the impact of explanations on trainee performance in AI-driven training systems. Moreover, our findings align with existing literature on explainable AI, which often shows that providing explanations can lead to over-reliance on AI outputs [5, 6, 16]. This suggests a potential direction for future research: leveraging insights from the explainable AI literature to mitigate over-reliance and enhance the generalization abilities of trainees.

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 the author(s) 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.
CI'24, June 2024, Boston, MA, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

Our work builds on the work by Callaway [7], who used *metacognitive feedback* (i.e., hints aimed at improving a user's decision-making rather than simply judging their decisions) to enhance users' planning strategies in a simple game. One of the experiments conducted by Callaway [7] involved placing users in a new environment after training (with hints) in a different one, which caused these trainees to perform worse than those in the control group. In other words, the trainees were learning how to play a single specific version of the game and seemingly failed to generalize their new knowledge. In this work, we aim to explore whether this phenomenon could be mitigated by offering explainable hints during training.

In the context of AI-assisted decision-making, using explainable AI has been shown to improve human trust [8], and in some cases, performance [11, 14], across a variety of environments [12]. This improvement is partially due to increased engagement with the task [1]. However, explainability can also lead participants to over-apply the AI's advice when it is not relevant, especially when the hints presented are convincingly wrong [4], wrong too often [15], or too cognitively expensive to engage with [17].

Another related line of work is in the sutudy of far transfer [2], where we look to see if knowledge is transferred from training in one environment to performance in a related, albeit different (in terms of optimal decision-making policy) environment. In this light, hints have been shown to be effective at assisting transferring knowledge between contexts, but only when they assist individuals in making connections between them [2]. However, when viewed as a task switch, hints may very likely cause an over-application of ideas from one task to another, in line with findings on over-reliance [13]. Generally speaking, there are conflicting theories in the literature on how explainable AI would impact knowledge transfers between tasks. Meanwhile, there is relatively limited empirical work to measure the transfers in the presence of explainable AI, which is what our paper aims to explore.

3 EXPERIMENTS AND RESULTS

3.1 Experiment Setup

Our experiments extend the Mouselab game by Callaway [7]. In the Mouselab game, users are ask to control a spider to navigate across the map to collect rewards, as demonstrated. The rewards associated with each node are initially hidden, but users can choose to incur a cost to reveal the reward associated with a given node. The user's planning strategy is to figure out ways to reveal the rewards and ways to navigate the map to maximize the total collected rewards.

At the beginning of the game, the player is placed at the center node. Each non-starting node offers a reward (in coins) indicated by the number on the node, with all rewards initially hidden to the player. At each time step, the player can choose to reveal the reward of a node, incurring a cost of 2 coins or choose to move the spider towards one of the allowable direction. When the spider pass through a node, it collects the associated rewards. When the spider reaches the leaf node, the game ends. The goal of the player is to maximize the total collected rewards, and the player can receive a bonus based on the rewards they collected throughout the game. A dissection of the game interface can be found in Figure 1.

In the training environment, node reward variance increases as a function of distance from the center, i.e., the rewards for the nodes close to the center are close to 0, while the rewards for the nodes far away from the center are larger in magnitude. Therefore, during training, the AI hints would lead users to reveal the rewards for the nodes far away from the center first to identify some high-reward nodes, and then navigate to those nodes to collect the rewards. The LLM-generated explanations provide additional information about the variance of reward distribution not included in hints, e.g., it explains that the reason the AI hints lead users to reveal the nodes far away from the center first is the high-variance property.

To create differences between the training and deployment environments, we vary the reward distributions of the nodes in the deployment environments. Specifically, we separate the deployment environments into "similar-to-training" and "dissimilar-to-training" environments. The "similar-to-training" environments are generated in the same way as the training environments. However, in the "dissimilar-to-training" environments, the nodes closer to the center exhibit higher variance in realized rewards, while the nodes farther away from the center exhibit lower variance in realized rewards. Since participants will not receive AI hints and/or explanations during deployment, we conjecture that participants who only receive AI hints are more likely to continue adopting incorrect strategies for a longer period. In contrast, participants who receive explanations alongside AI hints are likely to adjust more quickly to the change in reward distribution during deployment.

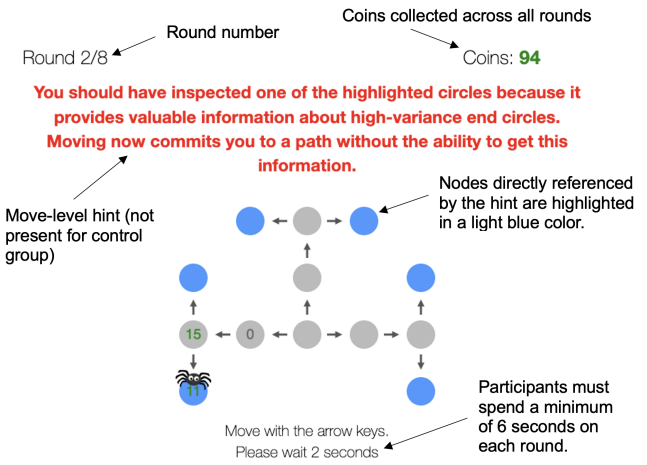


Figure 1: A round of the Mouselab game with explained hint.

For the implementation of AI hints, we leverage the same methodology as Callaway [7], which uses reinforcement learning to determine the optimal policy and generate AI hints based on it. For the generation of explanations, we provide the environment description and the actions from the AI hint to GPT-4 to generate explanations in an offline manner.

3.2 Experiment Procedure

We conducted our experiments using the above experimental setup. Participants in each experiment were placed into one of three treatments: Control, Unexplained Hints, and Explained Hints. The treatments differed in the type of hints participants received during the training phase. All participants experienced the same experimental structure. After agreeing to the informed consent and reading through the instructions, participants proceeded through a training block followed by two deployment blocks. Each block consisted of eight rounds. In the training block and the first deployment block, participants completed tasks in environments similar to the training environments. In the second deployment block, participants completed tasks in "dissimilar-to-training" environments as described above. Before the start of the second deployment block, participants were informed that the environment would change, but they were not provided with specific information on how it would change or what the new optimal strategy would be.

We conducted our experiments by recruiting participants from Prolific. A total of 300 participants were recruited for the experiments, which were approved by the IRB at our institution. We assessed the performance of participants based on the difference between their achieved reward and the maximum possible reward for each round. We refer to this difference as "loss," where a lower loss equates to superior performance.

3.3 Experiment Results

The results are shown in Figure 2. We first found that providing AI hints AI hints increase the performance of participants both during training and during deployment when the deployment environments are similar to those in training. Moreover, when providing explanations alongside AI hints, this performance gain further increases, demonstrating that explanations help humans understand AI hints and improve their performance.

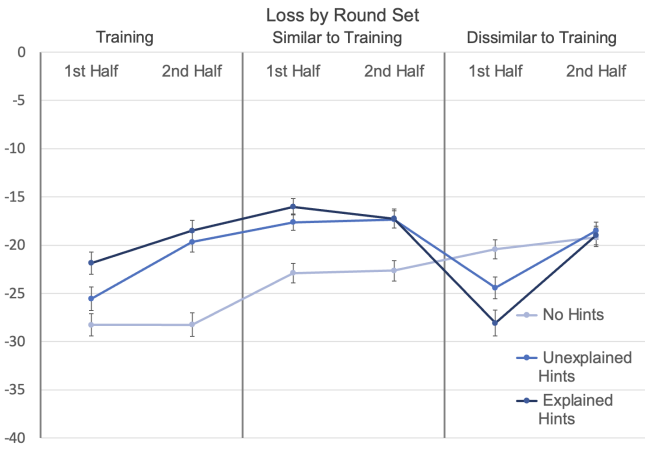


Figure 2: The average loss for each of the three treatments over training and deployment. The loss is defined as the the performance of the participants minus the performance of the performance of the optimal strategy. The error bars represent standard errors.

However, when the deployment environments are different from the training environments, participants receiving AI hints during training significantly underperform participants not receiving AI hints, although these effects dissipate over time (as shown in the second half of the environments in the dissimilar-to-training block). This observation replicates the work by Callaway [7]. Probably more surprisingly, for participants receiving explanations alongside AI hints during training, the performance drop is even more significant than the participants receiving only AI hints during training. This result indicates that participants might have developed over-reliance on the AI hints when provided explanations. Even when the explanations are supposed to help them understand the reasoning of the hints, participants do not internalize the explanations and seemingly developed more reliance on them in their decision-making.

4 CONCLUSION AND DISCUSSION

In this work, we examine the effects of explanations for AI hints in AI-driven training systems. Our results replicate prior findings, showing that providing AI hints during training improves participant performance when the deployment environments are similar to those in training, but it might hurt participant performance when the deployment environments are dissimilar to training. Additionally, when we introduce LLM-generated explanations to AI hints during training, it further enhances participants' performance in similar-to-training deployment environments. However, it also exacerbates the decline in participant performance in dissimilar-to-training environments. These findings suggest that while explanations can be beneficial in familiar contexts, they may lead to over-reliance in novel situations. This highlights the need for further research in the design of AI-driven training systems.

Our work includes a number of limitations, many of which could be grounds for future research. The first limitation concerns the task environments. The game we used is relatively simple and does not cover multiple types of deviations from the training environment. Examining to what extent our results generalize to other environments is a natural next step. Second, in our setting, explanations are always presented alongside AI hints. Literature suggests that showing only the explanations without explicit suggestions on what to do [10] or phrasing the explanation as a question [9] can increase interaction with the hint system and improve performance. Therefore, it would be interesting to explore whether this could be a more effective way of utilizing explanations in AI-driven training. Third, there has been a flourishing line of research in explainable AI, and we have only adopted LLM-generated explanations in our study. It would be important and useful to examine the effects of different ways of implementing explanations. Finally, our results suggest that participants might develop potential over-reliance on AI hints during training when explanations are provided. This finding aligns with recent empirical results that explanations often induce over-reliance in the context of AI-assisted decision making [5, 6, 16]. It is therefore important and interesting to bridge the results and findings of using explanations in the context of AI-driven training and AI-assisted decision making, with the goal of improving our understanding of best leveraging AI to improve human decisions.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14.

[2] Susan Barnett and Stephen Ceci. 2002. When and Where Do We Apply What We Learn? A Taxonomy for Far Transfer. *Psychological bulletin* 128 (07 2002), 612–37.

[3] Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. 2021. Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454* 5 (2021).

[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. 5, CSCW1, Article 188 (apr 2021), 21 pages.

[5] 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*. 454–464.

[6] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[7] Frederick Callaway. 2022. Leveraging Artificial Intelligence to Improve People's Planning Strategies. In *PNAS*, Vol. 119.

[8] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2020. Toward Personalized XAI: A Case Study in Intelligent Tutoring Systems. arXiv:1912.04464 [cs.AI]

[9] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages.

[10] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. ACM.

[11] Loveleen Gaur, Mehedi Masud, and Noor Jhanjhi. 2022. Explanation-driven HCI Model to Examine the Mini-Mental State for Alzheimer's Disease. (12 2022).

[12] Michael Guevarra, Srijita Das, Christabel Wayllace, Carrie Demmans Epp, Matthew Taylor, and Alan Tay. 2023. Augmenting Flight Training with AI to Efficiently Train Pilots. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 13 (Sep. 2023), 16437–16439. https://doi.org/10.1609/aaai.v37i13.27071

[13] Andrea Kiesel, Marco Steinhauser, Mike Wendt, Michael Falkenstein, Kerstin Jost, Andrea M. Philipp, and Iring Koch. 2010. Control and interference in task switching–a review. *Psychological bulletin* 136 5 (2010), 849–74.

[14] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 250, 17 pages.

[15] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350.

[16] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[17] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages.

[18] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, 563–574.

[19] Guanghui Yu and Chien-Ju Ho. 2022. Environment Design for Biased Decision Makers.. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 592–598.

[20] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. 2020. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 737–744.

Collective Emotional Intelligence? Social Information Does Not Improve Group Judgement Accuracy in Emotion Judgements

TOM TAIYI YAN 闫太一, University College London, School of Management
AMIT GOLDENBERG, Harvard University, Harvard Business School

1. INTRODUCTION

In this research, we study the effect of social information on group accuracy in the domain of social cognition judgement and in particular, emotion judgement. A central finding from existing wisdom of crowd research is that social information—where individuals in groups exchange opinions and beliefs after independent judgement—can enhance group judgement accuracy (Becker et al., 2019; Gürçay et al., 2015; Hong & Page, 2008; Larrick & Soll, 2006; Navajas et al., 2018). Research highlights that a driving mechanism of this finding is the ***accuracy-revision mechanism,*** which is the tendency of individuals who are more erroneous in their independent judgement to subsequently make larger revisions. This accuracy-revision mechanism therefore improves group accuracy as a result of social information (Almaatouq et al., 2020; Becker et al., 2017; Silver et al., 2021).

So far, much of the research on wisdom of the crowd examining the accuracy-revision coefficient has predominantly focused on judgements about artifacts in the physical world. One classic example is o***bject-oriented judgements***, ranging from estimating physical and visual attribute ("how old is this politician?" or "how many triangles are in this photo?"), to statistical and numeric estimations ("what will be the U.S. inflation rate in 2023 Q1?"). However, a critical aspect of human intelligence concerns social cognition—judgement of the psychology of people. These judgments are typically based on a more subjective evaluation of reality, and people who make those judgements may be less likely to incorporate other's judgement.

2. EXPERIMENT 1: THE ACCURACY-REVISION LINKAGE IN EMOTION JUDGEMENT DOMAIN

In Experiment 1 (n = 300, pre-registered), we first examined the accuracy-revision mechanism when comparing between object-oriented judgement task and an emotion judgement task. To make sure that the information that participants received was identical in both tasks, in this experiment participants were led to believe that they were conducting the task with others, but all the social information that participants saw was controlled by us. Participants were randomly assigned to either object-oriented judgement task condition or emotion judgement task condition and completed 10 trials of the study. In the object-oriented task condition, in each trial, participants were provided estimates about an array of geometric shapes (e.g., "how many triangles are in this image?; Frey & Van De Rijt, 2021; Koriat, 2012).

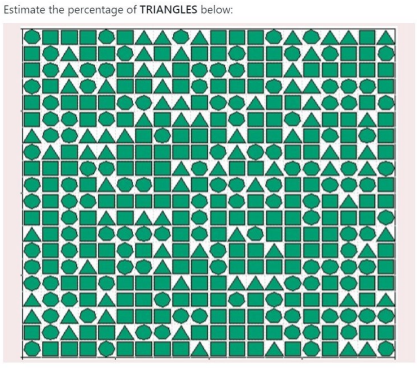


Figure 1: Example geometric array

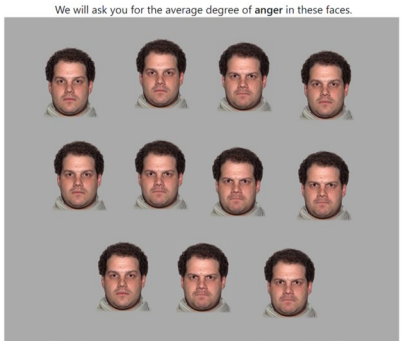


Figure 2: Example of emotional face array

In the emotion judgement domain, participants provided estimates about an array of human faces (e.g., "what is the average anger intensity in this group of faces?" (Goldenberg et al., 2021; Lamer et al., 2018). We chose this design over a single-face emotion estimation design because we wanted to maintain consistency with the object-oriented task domain which involved aggregating multiple visual stimuli at the same time. We conducted pilot study (n =100) to ensure that the two tasks had similar level of baseline difficulties.

A key design of this experiment was that although participants believed that they were completing the study in synchronously with 9 participants, in reality they were completing the study individually. Leveraging this design, we provided all participants across the two conditions the same high quality social information. Specifically, after their independent judgement in each trial, each participant viewed the estimates made by the 9 peer participants, which were created by us in real time. Crucially, across both conditions, the average estimate by these peers was always 15% more accurate than the participant's own estimate. Therefore, we provided all participants equal motivation to incorporate social information and revise towards it on their 2nd estimate. We also measured participant's perceived accuracy relative to their peers by using a 10-point scale (1 = I was far more inaccurate than these peers; 10 = I was far more accurate than these peers).

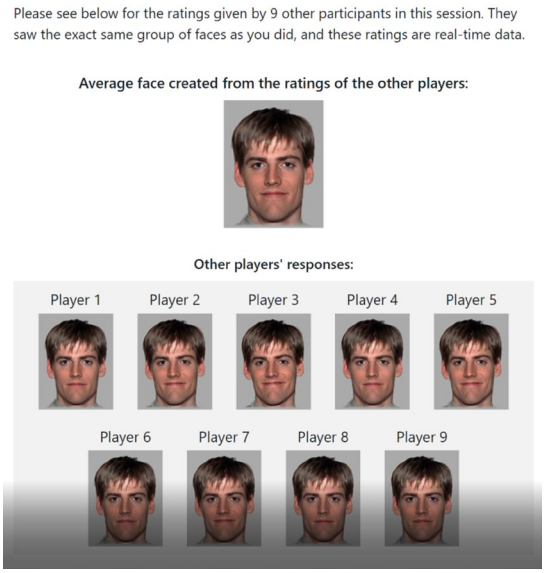


Figure 3: Example of social information in the emotional judgement condition; In the geometric array estimation condition, participants saw peer estimates as numbers instead of face morphs.

We hypothesized that the accuracy-revision mechanism would be significantly weaker in the emotion judgement condition. Consistent with prior work, we calculated ***revision coefficient*** (Almaatouq et al., 2020; Becker et al., 2017) and compared across conditions. Revision coefficient is calculated as the correlation between their initial error and their subsequent revision. Specifically, initial error is calculated as the difference between initial estimate and true answer (thus positive value means overestimation and negative means underestimation), and revision is calculated as the difference between the final estimate and the initial estimate (thus positive indicates upward revision and negative indicates downward revision). As such, a more negative revision coefficient is normatively desirable because it indicates that participants who are further away from the truth are also those who are making the biggest revisions.

Results showed that when engaging in emotion judgements (versus in object-oriented judgement), participants' independent judgement accuracy was significantly less predictive of their subsequent revision (revision coefficient $r_{emotion}$ = -.55, $r_{geometric}$ = -.65, p < .01); We then created a measure of ***accuracy-confidence coefficient*** by calculating the correlation between the absolute error of participant's initial estimate and their self-reported measure of confidence. Supporting our hypothesis, participants perceived confidence in their own accuracy only had very small correlation with their actual accuracy (accuracy-confidence coefficient, $r_{emotion}$ = -.08), whereas in object-oriented judgements perceived confidence was significantly more correlated with actual accuracy ($r_{geometric}$ = -.29, p < .01).

3. EXPERIMENT 2: SOCIAL INFORMATION AND WISDOM OF CROWD IN EMOTION JUDGEMENT

Study 1 provided evidence that the accuracy-revision link, an important foundation in collective intelligence research, was significantly dampened in the emotion judgement task. However, Experiment 1 was designed with stringent controls and without real-time synchronous social information, hence unable to demonstrate to what extent the weakened accuracy-revision link translates into group outcomes.

In Experiment 2 (n = 1,000 pre-registered), we conduct a full-scale wisdom of the crowd experiment in the emotion judgement. Given our results in Experiment 1, we hypothesize that in the emotion judgement domain, social information would have no effect on group accuracy improvement. Group of 10 participants were randomly assigned to either a social condition or a control condition and completed 30 trials of emotion judgements. In the social condition, participants were presented with real-time, synchronous social information from their 9 peers including their respective estimates and the aggregate average estimate. In the control condition, although participants also completed the task synchronously with 9 peers, they were not presented with social information.

We started with the baseline model that examined the main effect of experimental condition while accounting for data structure nesting. Results showed an non-significant coefficient for the experimental condition (b = -0.08, SE = .10, *n.s.*; 95% CI [-.28, .12]), supporting our theory that groups in the social condition did *not* improve their group accuracy significantly more than the control condition. We then conducted a series of additional analyses such as examining the effect of emotion valance (i.e. anger vs. happiness), face array size (i.e., how many faces in the face array?) and trial number (i.e. does the main effect change as groups go from the 1st face array to the 30th?). Overall, in all models we tested we did not find any evidence suggesting that groups in the social condition had improved group accuracy significantly more than the groups in the control solo condition.

We next adopted a Bayesian approach, which can directly provide easily interpretable results quantifying the odds supporting the null effect (i.e. the difference between the two conditions is zero). Results found a Bayes factor .06 in favor of the alternative hypothesis, which we then calculated its inverse (inverse BF = 16.20) to obtain the evidence in support of the null. In other words, our data indicating that **it is 16.2 times more likely that there is no difference on group accuracy improvement between the social condition and the control condition than there is**, qualifying as "strong evidence" for the null effect (Andraszewicz et al., 2015).

4. CONCLUSION

In the present study, we conducted two pre-registered experiments to examine the effect of social information on group accuracy. We found participants in the object-oriented task domain revised their estimates significantly more in proportion to their errors (i.e., the larger the errors, the larger the revision and vice versa), than participants in the emotion domain. Further analysis showed that in the emotion judgment domain, the perceived accuracy (i.e., confidence) of participants was almost

completely decoupled from their actual accuracy. Expanding this individual level finding to the group level, Experiment 2 found no evidence that group emotion judgement benefited from social information. In fact, Bayesian analysis suggested that we found strong evidence that the effect of social information on group accuracy is not different from zero.

REFERENCES

Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(21). https://doi.org/10.1073/pnas.1917687117

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E. J. (2015). An Introduction to Bayesian Hypothesis Testing for Management Research. *Journal of Management*, *41*(2), 521–543. https://doi.org/10.1177/0149206314560412

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(26), E5070–E5076. https://doi.org/10.1073/pnas.1615978114

Frey, V., & Van De Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, *67*(7), 4273–4286. https://doi.org/10.1287/mnsc.2020.3713

Goldenberg, A., Weisz, E., Sweeny, T. D., Cikara, M., & Gross, J. J. (2021). The Crowd-Emotion-Amplification Effect. *Psychological Science*, *32*(3), 437–450. https://doi.org/10.1177/0956797620970561

Koriat, A. (2012). When Are Two Heads Better than One and Why? *Science*, *336*(April), 20–22.

Lamer, S. A., Sweeny, T. D., Dyer, M. L., & Weisbuch, M. (2018). Rapid visual perception of interracial crowds: Racial category learning from emotional segregation. *Journal of Experimental Psychology: General*, *147*(5), 683–701. https://doi.org/10.1037/xge0000443

Silver, I., Mellers, B. A., & Tetlock, P. E. (2021). Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion. *Journal of Experimental Social Psychology*, *96*(April), 104157. https://doi.org/10.1016/j.jesp.2021.104157

Emergent structures and roles of an open science community during crisis: the case of OpenCovid19

Camille, Masselot

Learning Planet Institute, Research Unit Learning Transitions (UR LT, joint unit with CY Cergy Paris University), F-75004 Paris, France, masselot.camille@gmail.com

Olga, Kokshagina

EDHEC Business School, Learning Planet Institute, Research Unit Learning Transitions (UR LT, joint unit with CY Cergy Paris University), F-75004 Paris, olga.kokshagina@gmail.com

Marc, Santolini

Learning Planet Institute, Research Unit Learning Transitions (UR LT, joint unit with CY Cergy Paris University), F-75004 Paris, France, marc.santolini@learningplanetinstitute.org

Designing open-source online communities in times of crisis is challenging due to their rapid-changing scale, self-organization, and fluidity of roles, which limit the centralization of information. Methods and tools to monitor them can play a key role for practitioners (facilitators, community designers) to provide equitable access to certain roles and resources as well as ensure relevant development. In this study, we leverage Slack data to explore how the structural and role evolution of the OpenCovid19 community - a small biohacking group which rapidly transformed into thousands of individuals from diverse backgrounds developing open-source technologies to mitigate the effects of the pandemic. Methods include social network analysis, sentiment analysis, and k-means clustering to uncover roles based on members' interactions (mentions, replies, emojis), content of public messages, activity (number of messages, number of channels, inter-message duration, lifespan in the program) and status (whether the individual is an administrator, owner or bot). Finally, we conducted a temporal analysis to study transitions between roles. Results highlight distinct developmental stages of the community, ranging from growth to dormancy, and reveal diverse distinct roles such as Orienteers, Steerspersons, Scouts, Contacts, and Reporters. Notably, the study emphasizes that despite the most central positions being locked-in, others were fluid allowing for peripheral members to occupy influential positions and vice-versa. However, despite this fluidity, participants demonstrating high emotionality remained peripheral. Overall, we show the value of using Slack data to portray open online community structures and role dynamics, which provides valuable reflections for community designers and researchers alike.

CCS CONCEPTS • **Online social networks • Management of computing and information systems • Social network analysis**

Additional Keywords and Phrases: Open-source communities, Online innovation communities, Structure, Roles, Monitoring and evaluation

1 INTRODUCTION

As the COVID-19 pandemic unfolded in 2020, a small group of biohackers rapidly transformed into the OpenCovid19 community, comprising thousands of individuals dedicated to developing open-source technologies and projects aimed at developing and testing low-cost solutions to address the challenges posed by the pandemic. OpenCovid19 aligns with Preece ‘s definition [1] of an online innovation community. In this instance, community members collaboratively contribute to core tasks such as problem definition, data collection and analysis, design of products or protocols, and knowledge transfer. Nambisan et al. [2] argue that these forms of community require new theorizing. Collaboration amongst their members occurs without the structural mechanisms traditionally associated with knowledge collaboration in organizational teams. Online communities, instead, rely on people’s voluntary commitment, participation, and contributions. Majchrzak et al., [2] emphasize that the unique characteristics of collaboration in online communities involve fluidity in knowledge contribution, in participation in roles played by members and in structures and technology affordances. Fluidity and technology cause the governing mechanisms to emerge from interaction [3]. Their affordances enable temporary online crowds of individuals with no shared future fate or past social relationship to organize and collectively perform. These mechanisms are not static, as the affordances associated with the technology and social dynamics evolve. However, it remains unclear how affordances in open online communities evolve, particularly regarding changes in structure and roles in a context of crisis. This work provides perspectives on the emergent structures, roles and navigation patterns of an open science, self-organized community during a crisis by using the case of OpenCovid19, and the utility of Slack data to do so. The results of this study have the potential to support open-source online communities in monitoring and evaluating the evolution of their structure, their capacity to provide equitable access to certain roles, and discuss actions needed to meet the needs of its members while they address global challenges.

2 METHODS

We conducted a longitudinal study of OpenCovid19 spanning from March 2020 and May 2022 based on the interactions in their Slack space. To explore how affordances of participation co-evolve, we studied the social interaction dynamics of participants across communication channels of the community. We did this by examining the graph structure of Slack communication across channels - identifying who interacts with whom (replies to messages, reactions with emojis, and mentions of individuals) - and the set structure of topical affiliations - determining who is discussing what. Networks have been proven relevant for capturing informal networks and analyzing how contributions of core and peripheral participants change over time [4, 5, 6]. Additionally, we performed sentiment analysis on all the text messages posted in public channels using the Syuzet R package [7]. For everyone, we computed a vector of network centrality metrics and average post sentiment across the duration of the program. The network centrality metrics comprised local (degree) and global measures (betweenness, coreness, PageRank) in the three layers considered: replies, reactions, and mention networks. Average sentiment values were computed for sadness, fear, disgust, anger, joy, surprise, anticipation, and trust, along with two valence indicators (negative and positive). Finally, we added activity data (number of messages, number of channels, inter-message duration, lifespan in the program) and status (whether the individual is an administrator, owner, or bot). Overall, 37 features were considered across 1,017 individuals who posted at least 1 message.

To uncover roles in the community, we conducted a k-means clustering on the feature matrix. Network centralities, which are long-tail distributed, were first log-transformed x (x+1). Finally, each feature was centered and normalized by its standard deviation. K-means clustering was run for k ranging from 1 to 20. We calculated the total within-cluster sum of squares for each k and plotted it against k to locate a bend. In the analysis, we proceeded with k=8, a value large enough

to encompass all roles and go into fine-grained analysis. Finally, we conducted a temporal analysis to study the transitions between roles. For this, we computed the feature values for each week and normalized across individuals. We then computed for each individual the distances to the cluster centroids obtained in the k-means analysis and assigned the individual to the closest centroid.

3 RESULTS

Throughout their engagement on 164 Slack channels, 2,358 users exchanged 30,530 messages, including 13,838 reactions, 14,328 replies, and 8,332 mentions, with 80% of activity driven by 10% of users. Our sentiment analysis shows an overrepresentation of positive feelings, trust, and anticipation.

The community, spanning from January 1, 2020, to January 1, 2023, traverses distinct developmental stages of Growth, Maturation, Decline, and Dormancy (Figure 1a). We find that during the Growth phase, the community witnesses a surge in vertices and edges, accompanied by an increase in mean degree and clustering, indicative of burgeoning interconnectivity and clustering. Maturation is marked by stabilization in network metrics, reflecting a consolidation of community interactions and organizational cohesion. The subsequent Decline phase manifests as a reduction in network size and density, coupled with a decrease in average path length and modularity, signaling a decline in community engagement and functional specialization. Finally, the Dormancy stage signifies a state of quiescence, with minimal activity and diminished network connectivity.

Our k-means analysis revealed distinct roles with varying levels of engagement and influence (Figure 1b). Orienteers, a group of 19 members, play a critical role as navigational nodes, characterized by rapid responsiveness, strong centrality, and significant contributions across many channels, indicating their influential status. Steerspersons are key communicators who facilitate cohesion and exhibit high levels of engagement and influence, with a substantial message exchange. Scouts, the largest group with 252 members, serve as explorers with wide-reaching connections, though with lower coreness, indicating their role in connecting different network parts. Contacts and Reporters display low engagement and centrality, with Reporters having a specific origin in another community. Finally, Agitators (positive or negative) stand out due to their distinct emotional expression. Interestingly, bots set up to introduce people were core in facilitating the integration of the latter.

Studying trajectories of people between roles feed numerous conversations related to access to influence and resources. Role adoption was not static for half of the members: when considering roles at the weekly level, we found that half of the community undertook only one role throughout the course of their participation whereas the other half occupied between 2 and 5 different roles. People with high emotional engagement tended to remain in these roles while being peripheral. The central role of Orientees was the most locked-in showing no transition to other roles. Access to it was possible only through embracing the Steersperson role first. However, the Steersperson's position in the role trajectory network suggests that OpenCovid19 fostered inclusivity, allowing peripheral individuals to become influential and central in the community social structures. Indeed, transitioning to the Steersperson role was a common progression for Scouts, Reporters, and Contacts, who typically occupied more peripheral positions. In addition, our data reveal that Reporters, serving as intermediaries between OpenCovid19 and external communities, led projects recognized by community members as deserving funding.

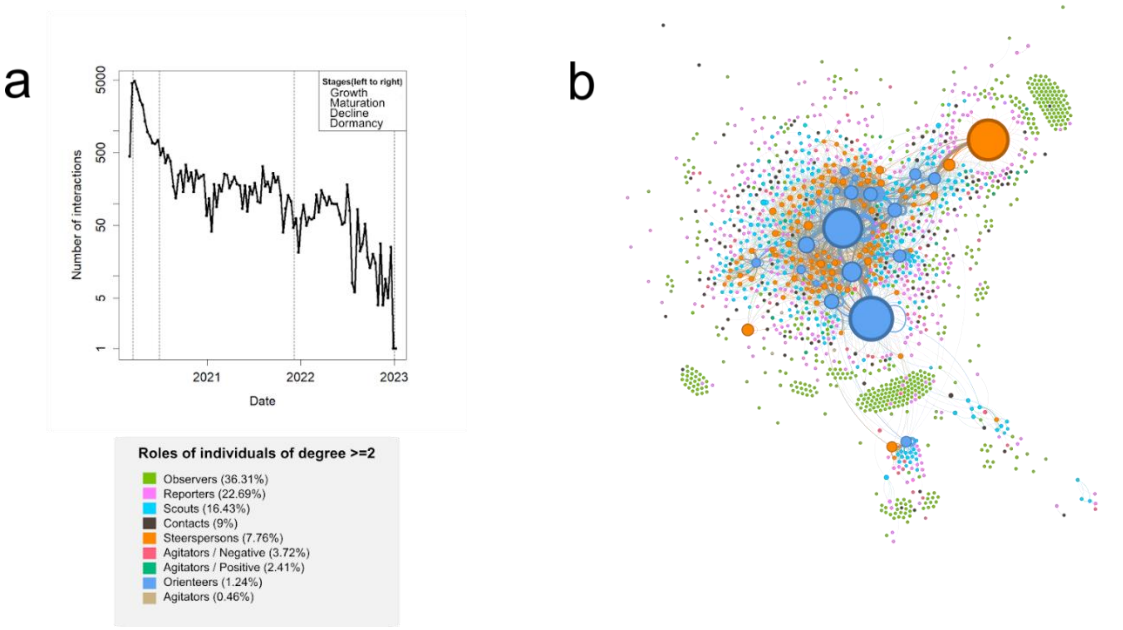


Figure 1. a) Number of interactions between OpenCovid19 community members in Slack public channels (reactions, mention, reply) across time. Dashed lines indicate major developmental phases. b) Network representing the interactions in the OpenCovid19 community on Slack. Nodes are individuals or bots, linked by their interactions in Slack public channels (reactions, mention, reply), and weighted by the number of interactions. Node size is proportional to their Pagerank centrality, and node color corresponds to the role in the community obtained by k-means clustering (see Methods).

4 CONCLUSIONS AND PERSPECTIVES

Our study focuses on the emergent structures and roles within the OpenCovid19 community during the COVID-19 pandemic, shedding light on the dynamics of an open science online community in crisis, communicating through Slack. Using a hybrid method of network and topic analysis, we show the potential of Slack data to uncover diverse roles in a community and its structure. Interestingly, these roles exhibited the specific importance of distinct categories of features: sentiments, global and local network centralities. The emergence of non-topologically-central roles that stand out in sentiment analysis shows the importance of content analysis to go beyond structural approaches. Overall, by examining the community members' interactions from its inception in 2020 to dormancy in 2023, we contribute insights into the interplay between network structures, roles, and crisis response within online communities. As we pursue our analysis and modeling of the relationship between the structural evolution of the community, internal and external events, and the emergence of roles and trajectories of its members, this work serves as a basis for any open online community designer, and Slack space users, to reflect on the structure of their community, their capacity to provide equitable access to certain roles and resources, and actions needed to address global challenges.

ACKNOWLEDGMENTS

This work was supported by the French Agence Nationale de la Recherche (ANR), under grant agreement ANR-21-CE38-0002

REFERENCES

[1] J. Preece and D. Maloney-Krichmar. 2005. Online Communities: Design, Theory, and Practice. J. Computer-Mediated Communication, vol. 10, Jul. 2005.

[2] S. Nambisan, K. Lyytinen, A. Majchrzak, and M. Song. 2017. Digital Innovation Management: Reinventing Innovation Management Research in a Digital World. MIS Quarterly, vol. 41, no. 1, pp. 223–238, 2017.

[3] A. Aaltonen and G. F. Lanzara. 2015. Building Governance Capability in Online Social Production: Insights from Wikipedia. Organization Studies, vol. 36, no. 12, pp. 1649–1673, Dec. 2015, doi: 10.1177/0170840615584459.

[4] A. Polanco and M. E. J. Newman. 2023. Hierarchical core-periphery structure in networks. Phys. Rev. E, vol. 108, no. 2, p. 024311, Aug. 2023.

[5] C. Masselot, R. Jeyaram, R. Tackx, J. L. Fernandez-Marquez, F. Grey, and M. Santolini. 2023. Collaboration and Performance of Citizen Science Projects Addressing the Sustainable Development Goals. vol. 8, no. 1, Art. no. 1, Jun. 2023.

[6] E. Akar and S. Mardikyan. 2018. User Roles and Contribution Patterns in Online Communities: A Managerial Perspective," SAGE Open, vol. 8, no. 3, p. 2158244018794773, Jul. 2018.

[7] M. L. Jockers. 2015. Syuzhet: Extract sentiment and plot arcs from text. Retrieved October, vol. 21, p. 2015.

The Robust Benefits of Social Influence for the Wisdom of Individuals in the Crowd

Charlie Pilgrim*
University of Leeds
Leeds, UK
c.p.pilgrim@leeds.ac.uk

Joshua Becker*
UCL School of Management
London, UK
joshua.becker@ucl.ac.uk

ABSTRACT

Prior research offers mixed evidence on whether and when communication improves belief accuracy for numeric estimates. The conditions under which communication improves group and/or individual outcomes remain poorly characterised. We analyse a model of opinion dynamics to derive these conditions, formally explicating the relationship between group and individual outcomes. We reanalyze previously published experimental data, finding that empirical dynamics are consistent with theoretical expectations.

1 INTRODUCTION

Understanding how communication shapes belief accuracy is a central task for social science. A range of interdisciplinary research has shown that people's beliefs are influenced by those of others, and that mere exposure to another person's estimate leads to increased similarity [4, 6, 7, 16]. Importantly, these studies have shown that the strength and direction of this belief-anchoring can vary across people and groups, significantly moderating the effect of communication on accuracy [1, 5, 13, 16]. For example, a positive influence/accuracy correlation in a group is expected to improve accuracy in communication, as groups are drawn towards the true value [17].

However, this research fails to offer any consensus on whether and when communication improves accuracy. One discrepancy emerges between research on dyadic (one-to-one) communication versus research on group behaviour. On the one hand, research on giving and receiving advice shows that communication from peers is generally beneficial, with some manageable exceptions [6, 10, 18, 19, 21]. On the other hand, research on collective intelligence shows that when people are embedded in groups their beliefs are shaped by macro-level factors outside individual perception or control [1, 4, 17]. This macro-level research, focused on group-level outcomes, indicates that accuracy can be highly fragile and that the benefits of communication between members may depend on a variety of factors [1, 2, 5, 8, 12, 17, 20].

We provide a formal mathematical analysis that identifies the effect of communication on both group error and individual error, and derives a clear relationship between groups and the individuals in them. This analysis reveals the property that individual change in error is strictly lower (better) than change in group error. Following experimental research, we examine the role of network centralisation, accuracy/influence correlation ("calibration"), and averageness/influence correlation ("herding"). Surprisingly, we find that calibration is generally good but can be too high, and that herding is generally bad, but can be too low. Our analysis not only shows that individuals can often improve in conditions where groups get worse, but also finds that individual improvement is the most likely outcome, even as group improvement is fragile.

2 THEORETICAL ANALYSIS

Let $x_{i,t}$ be the opinion of individual i at time t. We assume that opinions can be represented on a numerical scale such that there is some truth θ, and each opinion has an associated bias $e_{i,t} = x_{i,t} - \theta$.

2.1 Relationship Between Groups and Individuals

We define crowd error as the squared error of the average bias, $E(e)^2$, and individual error as the average squared bias $E(e^2)$. We show that if opinions converge then the change in individual error, $E(e_\infty^2) - E(e^2)$, is better (more accurate) than the change in crowd error, $E(e_\infty)^2 - E(e)^2$, by the pre-influence variance of opinions s_e^2, or just 1 for standardised error,

$$\frac{\Delta E(e^2)}{s_e^2} = \frac{\Delta E(e)^2}{s_e^2} - 1. \tag{1}$$

This is a consequence of the "crowd beats averages law" [15], $E(e)^2 = E(e^2) - s_e^2$, which is also the definition of sample variance.

2.2 Change in Opinion

We study the DeGroot model of social influence [9] where people revise their beliefs by adopting a weighted mean of the influences of their peers, w_{ij},

$$x_{i,t+1} = \sum_j w_{ij} x_{j,t}\,. \tag{2}$$

Under mild assumptions, groups will converge on a weighted mean of initial beliefs, $e_\infty = E(ve)$, weighted by the normalised leading eigenvector of the influence matrix, v [9], which can be interpreted as an influence vector.

We show that the change in opinion (and change in bias) is simply

$$\frac{\Delta E(e)}{s_e} = c_v r(v, e)\,, \tag{3}$$

*Both authors contributed equally to this research.

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-XXXX-X/18/06

where c_v is the coefficient of variation of the influence vector, a scale free measure of influence centralisation. $r(v, e)$ is the correlation between influence and bias.

2.3 Truth Alignment

A crowd will certainly become less accurate when the change in opinion is in the same direction as the initial crowd bias, i.e when $z = \frac{E(e)}{s_e}$ has the same sign as the change in z as given by Equation 3. By multiplying the initial crowd bias by the change in crowd bias we define the "truth alignment" as

$$\alpha = -z c_v r(v, e)\,. \tag{4}$$

If the truth alignment is negative, $\alpha < 0$, then the crowd's opinion will move away from the truth, and therefore the crowd will become less accurate. However, a positive truth alignment, $\alpha > 0$, doesn't guarantee that the crowd will become more accurate, only that the crowd opinion will move in the direction of the truth from the initial crowd bias. It is possible that the crowd will overshoot the truth and become less accurate than they began.

2.4 Change in Accuracy

We show that the asymptotic change in crowd error is

$$\frac{\Delta E(e)^2}{s_e^2} = \frac{\alpha^2}{z^2} - 2\alpha\,, \tag{5}$$

and the mean asymptotic change in individual error is

$$\frac{\Delta E(e^2)}{s_e^2} = \frac{\alpha^2}{z^2} - 2\alpha - 1\,. \tag{6}$$

Figure 1 shows that while crowd error becomes less accurate in many conditions, improvements in individual error are much more robust.

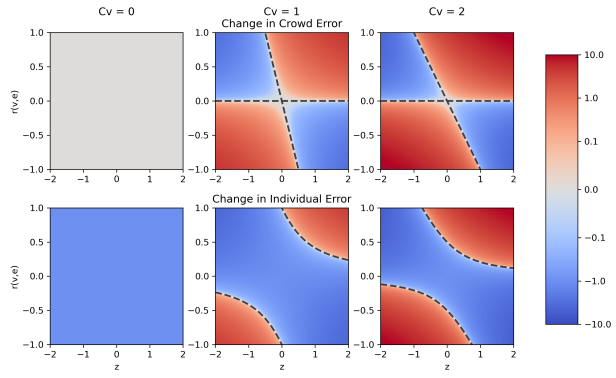


Figure 1: The change in standardised crowd error (top) and individual error (bottom) with increasing influence decentralization (c_v, left to right), as a function of initial crowd bias, z, and correlation between influence and initial bias, $r(v, e)$. Negative changes in error (blue) represent improvements in accuracy. Note that with no centralisation the crowd error stays constant (top left) and the standardised individual error improves by 1 (bottom left).

2.5 Calibration and Herding

We show that the belief/influence correlation $r(v, e)$ decomposes into two distinct components, (1) the error/influence correlation $-r(v, e^2)$ termed calibration [17], and an averageness/influence correlation $-r(v, d^2)$ which we term "herding." Here, d_i^2 indicates the distance between an opinion and the crowd opinion, or how "average" the opinion is, such that $d_i = x_i - E(x)$.

Truth alignment decomposes into calibration and herding, as

$$\alpha = \frac{c_v}{s_e^2}\Big(s(d^2)r(v, d^2) - s(e^2)r(v, e^2)\Big), \tag{7}$$

where $s(e^2)$ is the standard deviation in initial errors, and $s(d^2)$ is the standard deviation in initial distances.

Figure 2 shows changes in crowd and individual accuracy as a function of truth alignment, calibration and herding. Truth alignment represents a one-dimensional reduction of calibration and herding that does not lose any information when describing asymptotic changes in accuracy.

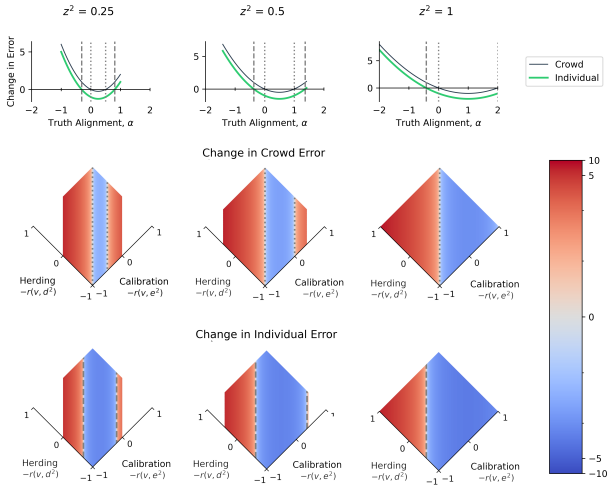


Figure 2: Individual accuracy improves under a wider range of conditions than crowd accuracy. Changes in error are a function of the truth alignment (α, top row), which can be decomposed into the effect of calibration (accuracy/influence correlation) and herding (averageness/influence correlation) on crowd (middle row) and individual (bottom row) accuracy. Calibration and herding axes are rotated so that truth alignment is a horizontal projection and aligns with the top row. In this example, with low initial crowd error ($z^2 = 0.25$, left), both crowd and individual accuracy can get worse if truth alignment is too high because the crowd overshoots the truth. When initial crowd error is high ($z^2 = 1$, right), crowds improve when the truth alignment is greater than zero, but individuals can improve even when the truth alignment is negative, and only get worse if the truth alignment is very negative. Truth alignment is bounded by $\alpha \leq |zc_v|$, and only shown for those values. At low z, this also means that calibration and herding must be close together.

2.6 Bounds

The coefficient of variation of influence, c_v, emerges as a measure of influence centrality. This measure is scale free such that if you scale a network yet maintain the same influence distribution, the influence centrality remains the same. Larger networks can have larger values of influence centralisation, because influence can be focussed to a greater degree. We show that $c_v \in [0, \sqrt{n-1}]$.

Note from Equation 3 that the standardised change in opinion cannot be greater than the centralisation c_v, i.e. $\Delta E(e) \leq |s_e c_v|$. Relatedly, the truth alignment (Equation 4) is bounded such that $\alpha \leq |z c_v|$. Finally, this also means that if the crowd opinion is initially close to the truth then the calibration and herding must be close to each other. The bounds are revealed in Figure 2, where only allowed system variables are plotted.

3 EMPIRICAL ANALYSIS

We test whether the theoretical dynamics are observed empirically by re-analysing data from 6 published studies [3–5, 11, 14, 17]. The studies all involved asking people for initial numerical belief estimates, allowing some communication within a group, and then asking for estimates again.

3.1 Relationship Between Groups and Individuals

The theoretical analysis predicts a linear relationship between asymptotic changes in standardised crowd and individual error (Equation 1) with a gradient of 1 and intercept of -1. As the empirical behaviour is not asymptotic, the intercept should fall between 0 and -1 if the crowd shows some convergence.

The empirical relationship is shown in Figure 3. A regression model yields a slope close to 1 (0.998, 95% CI: [0.996, 1.00) and an intercept between 0 and -1 (- 0.64, 95% CI: [-0.63, -0.65)]). Thus, empirical data shows a close fit to theoretical expectations.

3.2 Robust Benefits of Social Influence for Individuals

The theoretical analysis predicts that individual accuracy will improve over a wide range of conditions, even when the crowd accuracy gets worse. We measure the probability that an individual will become more accurate after information exchange, conditional on whether the group as a whole gets more accurate. Notably, individuals also have the possibility of making no revision at all. We therefore consider two metrics, (1) the probability of improvement conditional on any revision at all, and (2) the probability of improving or staying the same.

Figure 4 shows the empirical effect of social influence on individual accuracy. We find that individual accuracy reliably improves (or doesn't get worse) even when the crowd gets worse. This effect applies even to individuals who are, on average, more accurate than the crowd.

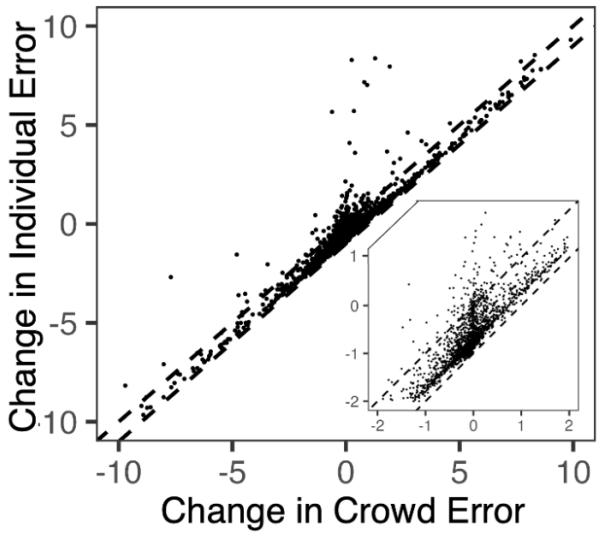


Figure 3: Each datapoint shows change in standardised crowd and individual error for a single experimental trial. The dashed diagonal line shows the boundaries Y=X and Y=X-1, with 90% of datapoints between the lines as expected. The regression slope ≈ 1 with an intercept between 0 and -1. The main figure does not show extreme outliers. Note that the axis units are in standard deviations of initial error, so any points not shown exhibited a movement equal to 10 times the initial standard deviation of estimates.

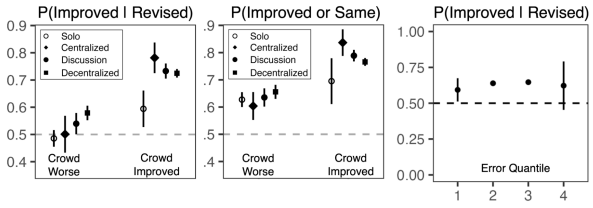


Figure 4: Probability of individuals benefiting from social influence. Left and middle panels show the probability of benefiting as a function crowd improvement, defined as simply not getting worse (middle) or improving conditional upon revision (left). Rightmost panel shows outcomes as a function of an individual's relative accuracy measured on all other questions. Because accuracy quartile is averaged across multiple questions, the levels do not all contain the same number of people. All results are calculated as a weighted average of all trials giving each experiment equal total weight. Error bars show 95% bootstrapped confidence intervals.

REFERENCES

[1] Abdullah Almaatouq, M Amin Rahimian, Jason W Burton, and Abdulla Alhajri. 2022. The distribution of initial estimates moderates the effect of social influence on the wisdom of the crowd. *Scientific Reports* 12, 1 (2022), 16546.

[2] Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2017. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management science* 63, 3 (2017), 691–706.

[3] Joshua Becker, Abdullah Almaatouq, and Emőke-Ágnes Horvát. 2020. Network structures of collective intelligence: The contingent benefits of group discussion. *arXiv preprint arXiv:2009.07202* (2020).

[4] Joshua Becker, Devon Brackbill, and Damon Centola. 2017. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences* 114, 26 (2017), E5070–E5076.

[5] Joshua Becker, Ethan Porter, and Damon Centola. 2019. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences* 116, 22 (2019), 10717–10722.

[6] Silvia Bonaccio and Reeshad S Dalal. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101, 2 (2006), 127–151.

[7] Zhi Da and Xing Huang. 2020. Harnessing the wisdom of crowds. *Management Science* 66, 5 (2020), 1847–1867.

[8] Norman Dalkey and Olaf Helmer. 1963. An experimental application of the Delphi method to the use of experts. *Management science* 9, 3 (1963), 458–467.

[9] Morris H DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical association* 69, 345 (1974), 118–121.

[10] Peter H Gardner and Dianne C Berry. 1995. The effect of different forms of advice on the control of a simulated complex system. *Applied cognitive psychology* 9, 7 (1995), S55–S79.

[11] Burcu Gürçay, Barbara A Mellers, and Jonathan Baron. 2015. The power of social influence on estimation accuracy. *Journal of Behavioral Decision Making* 28, 3 (2015), 250–261.

[12] David H Gustafson, Ramesh K Shukla, Andre Delbecq, and G William Walster. 1973. A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. *Organizational behavior and human performance* 9, 2 (1973), 280–291.

[13] Asher Koriat. 2012. When are two heads better than one and why? *Science* 336, 6079 (2012), 360–362.

[14] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences* 108, 22 (2011), 9020–9025.

[15] Scott Page. 2008. *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition.* Princeton University Press.

[16] Niccolò Pescetelli and Nicholas Yeung. 2021. The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General* 150, 3 (2021), 507.

[17] Ike Silver, Barbara A Mellers, and Philip E Tetlock. 2021. Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion. *Journal of Experimental Social Psychology* 96 (2021), 104157.

[18] Janet A Sniezek and Rebecca A Henry. 1989. Accuracy and confidence in group judgment. *Organizational behavior and human decision processes* 43, 1 (1989), 1–28.

[19] Lyn M Van Swol and Cara L Ludutsky. 2007. Tell me something I don't know: Decision makers' preference for advisors with unshared information. *Communication Research* 34, 3 (2007), 297–312.

[20] Andrew H Van De Ven and Andre L Delbecq. 1974. The effectiveness of nominal, Delphi, and interacting group decision making processes. *Academy of management Journal* 17, 4 (1974), 605–621.

[21] Ilan Yaniv and Maxim Milyavsky. 2007. Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes* 103, 1 (2007), 104–120.

Multiple AI Agents that Support Crowd Discussion Extended Abstract

Takayuki Ito
ito@i.kyoto-u.ac.jp
Kyoto University
Kyoto-city, Kyoto, Japan

Yihan Dong
dong.yihan.77v@st.kyoto-u.ac.jp
Kyoto University
Kyoto-city, Kyoto, Japan

Jawad Ahmad Haqbeen
jawad.haqbeen@i.kyoto-u.ac.jp
Kyoto University
Kyoto-city, Kyoto, Japan

Tokuro Matsuo
matsuo@aiit.ac.jp
Advanced Institute of Industrial Technology
Tokyo, Japan

Sofia Sahab
sahab.sofia@i.kyoto-u.ac.jp
Kyoto University
Kyoto-city, Kyoto, Japan

ABSTRACT

This paper presents an implementation of a discussion and consensus-building support system based on multiple AI agents and shows its effectiveness. An AI agent is a program that uses a large-scale language model (LLM) to interact with humans in text chats. The recent development of AI technologies enables a new democratic system with AI agents acting and working with humans together. Our proposed system supports the realization of higher-quality discussions by having AI agents participate in human discussions and consensus building. Multiple AI agents can be expected to increase the diversity of participants in a discussion by characterizing different positions, personalities, and roles. Experimental discussions were conducted using the online discussion platforms in Afghanistan and Indonesia using the developed agents. The experimental results partially support these hypotheses.

KEYWORDS

Multiple AI Agents, Crowd Discussion

ACM Reference Format:
Takayuki Ito, Yihan Dong, Jawad Ahmad Haqbeen, Tokuro Matsuo, and Sofia Sahab. 2018. Multiple AI Agents that Support Crowd Discussion Extended Abstract. In *Proceedings of ACM's 2024 Collective Intelligence Conference (ACM CI'24).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

This paper presents the implementation of a discussion and consensus-building support system based on multiple artificial intelligence (AI) agents and demonstrates its effectiveness. An AI agent is a program that uses a large-scale language model (LLM) to interact with humans in text chats.

Social consensus building and group decision-making face a number of difficulties. Problems such as groupthink, peer pressure, collective ignorance (group illusion), and ochlocracy, for example, have long been widely known in social psychology and political science. On the other hand, technologies such as the Internet and smartphones are dramatically changing human collective intelligence and new societal systems are needed to cope with these changes. Various new mechanisms are beginning to emerge that will address these issues. T. Malone calls groups that exhibit collective intelligence as Superminds[9] and predicts that groups and societies that adapt to changes in their environment will progress. In recent years, the most crucial information technology (IT) change has been hyper-connectivity, where people are always connected, even over long distances. Notably, the development of AI technologies enables a new democratic system with AI agents working together with humans.

LLMs have rapidly spread and are widely studied as an AI technology. An LLM is a language model based on a large dataset and a neural network. A language model is a program that returns as output the most likely sequence of text that follows a given sequence of text. Currently, models based on transformers are mainly used and the large-scale model GPT4 (OpenAI)[10] and the open-source model LLaMa2/3 (Meta)[14] have garnered much attention. The generation of natural sentences by LLMs has the potential to improve interaction between humans and AI agents drastically.

Our proposed system supports the realization of higher-quality discussions by having AI agents participate in human discussions and consensus building. The participation of AI agents helps solve problems such as groupthink[8], peer pressure, and group illusions, consensus building, and group decision-making while supporting higher-quality discussion and better consensus building.

Research on supporting human discussion and consensus building through a single AI agent has also been widely conducted[1, 6, 11]. D-Agree[7] is a concrete example, in which a single agent facilitating a human discussion is verified to activate the discussion. Furthermore, the effectiveness of online discussions based on AI facilitation has been widely confirmed in various applications[2, 4, 5, 12, 13]. In this paper, in addition to one AI agent moderating and facilitating, multiple AI agents participate to support high-quality discussions and better consensus building.

Unpublished working draft. Not for distribution.

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 the author(s) 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.
ACM CI'24, June 26-29, Boston, MA
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

Unpublished working draft. Not for distribution.

Multiple AI agents can be expected to increase the diversity of participants in a discussion by characterizing different positions, personalities, and roles[3]. These may include a critical agent, a positive agent, an agent that proposes issues, an agent that proposes ideas, an agent that overlooks and summarizes the discussion, and so on. The goal is to ensure diversity by allowing a variety of agents to participate in human discussions.

In this paper, the following hypotheses (H1), (H2), and (H3) are investigated as positive impacts: (H1) AI agents can form opinions in discussions that human participants can accept and agree to. (H2) Human participants can change their opinions by discussion with AI agents. (H3) Participation of AI agents stimulates discussion and encourages continuous comment.

In this study, multiple LLM-based agents were developed and used to participate in human discussions. Discussions were conducted using the online discussion platforms D-agree and Discourse, and experiments were performed in Afghanistan and Indonesia using the developed agents.

2 DISCUSSION SUPPORT BY MULTIPLE AI AGENTS

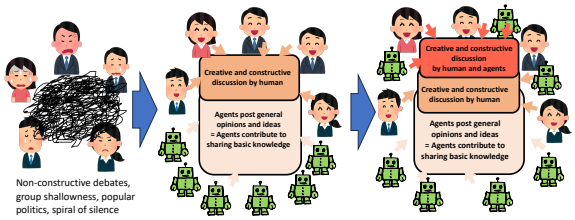


Figure 1: Cooperative Discussion and Consensus Making among Humans and AI agents

In this research, multiple AI agents join human dialogues to promote more desirable discussions, consensus building, and group decision-making. Many problems (left side of Fig.1) have been identified, such as groupthink, group illusions (pluralistic ignorance), entrainment bias, the spiral of silence, free-riding, and even popular politics.

In particular, groupthink is a psychological phenomenon in which people avoid critical examination and make irrational and risky decisions because they seek consensus within the group. When groupthink occurs, members give priority to the group's consensus without listening to different opinions or new information. As a result, group decisions tend to underestimate real risks and narrow possibilities. Groupthink is said to reduce the quality of decision-making and sometimes cause serious failures.

This study uses agents that speak from several different viewpoints to participate in human discussions (the middle of Fig.??). Since the agents use LLM, they can produce general opinions obtained from various viewpoints on a topic. Based on the general opinions that the agents make, the humans can focus on generating more creative and constructive opinions based on their experience, expertise, and sensitivity. Overall, the agent-human discussion can be expected to solve the issues identified above, including groupthink, because the agents and humans can obtain general opinions on the subject and generate creative and constructive opinions.

Furthermore, by incorporating creative and constructive opinions from humans, agents can also be expected to generate creative and constructive views. As shown on the right side of Fig.??, cooperation between humans and agents will likely lead to the generation of better ideas and consensus.

3 MULTI-AGENTS DISCUSSION SUPPORT SYSTEM

3.1 System Architecture

The architecture of the multi-agent discussion support system is presented in Fig.2. The system uses general web forum systems, including Discourse and D-Agree. AI agents are implemented as Python programs and connected to Open AI's GPT4 using application programming interface (API). The Agent Manager manages multiple agents, monitoring the status on the discussion platform and overseeing interactions with the agents. All agents and the Agent Manager are running on Amazon Web Services (AWS) EC2 instances.

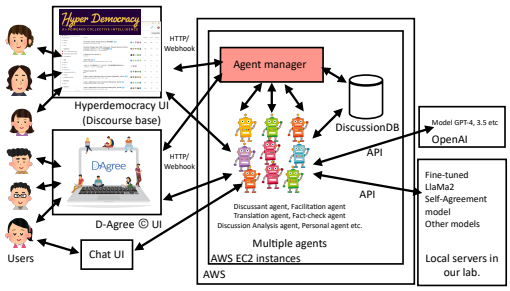


Figure 2: The System Architecture

3.2 AI Agents

This research implements the following agents: Participant Agent, Facilitator Agent, Simultaneous Translation Agent, Discussion Analysis Agent, and Commentary Agent. **Participant Agent:**The Participant Agent creates and posts sentences based on its own personality in the context of the current discussion. The Participant Agent controls various characteristics and personalities, such as positive personality, negative personality, liking to criticize, liking to agree, etc., by prompts. **Facilitation Agent:**Based on the results of the previous research on facilitation agents[6, 11], we developed various types of facilitation agent prototypes. In particular, LLMs have enabled the generation of natural sentences, making natural facilitation possible. **Simultaneous Translation Agent:**This agent translates and displays all statements made during the discussion in all languages related to the participant's country of origin. **Discussion Analysis and Commentary Agent:** The Discussion Analysis and Commentary Agent summarize the previous discussion, display the analysis, and provide commentary.

Unpublished working draft. Not for distribution.

3.3 The Flow of Multi-agent Discussion Support

The administrator sets discussion topics and periods. Users log into the discussion system and post their opinions freely. Agents start as soon as the discussion begins and work as a process, and the facilitation agent coordinates the discussion based on the facilitation scenario of how to facilitate the entire discussion. In the four discussion experiments, the discussions were divided into simple phases: ice breaker, divergence, convergence, and summary of the final proposal. The participant agent generates and posts messages in the LLM based on the prompts. Both the facilitation and participant agents memorize all previously submitted statements and create opinions based on them.

4 EXPERIMENTS AND EVALUATIONS

4.1 Experimental Online Discussions

We conducted the following four experiments using a multi-agent discussion support system. We confirmed that the system works realistically and can support actual discussions by real citizens.

(Experiment 1) Discussion experiment in Afghanistan. 20 participants and 3 days. Themes are "brain drain in Afghanistan", "women's education in Afghanistan", and "the future of Afghanistan".

(Experiment 2) Discussion experiment at Institut Teknologi dan Bisnis, Bali (ITB STIKOM, Bali) (Indonesia). 100 participants, and 1 day. Themes are "will AI take away human jobs?","What will happen to human society and the economy afterwards?" We employed a simultaneous translation agents.

(Experiment 3) 2nd Discussion experiment at University of Mataram (Indonesia). 200 participants and 1 day. Themes are "the Indonesian presidential election in 2024","climate change", etc. We employed a simultaneous translation agents.

4.2 Questionnaire

Participants for each of the experiments answered pre-questionnaires and post-questionnaires. Of note, we included an item asking whether there was a change in opinions about the theme before and after the experiment to test our hypothesis (H2).

Based on these questionnaires, we confirmed whether the participation of multiple agents in the discussion has an intrinsic influence on human discussion.

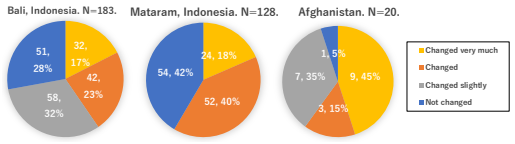


Figure 3: Results of opinion change in Experiments

Figure 3 show the questionnaire results regarding "Did your opinion before the discussion change after the discussion?" More than half of the users answered that their opinions on the theme had changed, indicating that the discussion with AI agents could change the participants' opinions. This result partially support our hypothesis (H2).

The agent's participation also encouraged users to participate in the discussion, thereby increasing engagement. The three-day Afghanistan experiment (Experiment 1) was not a short discussion. Nonetheless, many users stayed in it, which partially supports hypothesis (H3).

4.3 Discussion

Our experiments provided the following insights. (Insight 1) The AI agents encouraged people to contribute creative and constructive opinions. In Experiment 1, the participant agent exhausted a significant portion of the possible opinions (issues and ideas) on the topic as the discussion progressed. Therefore, the humans worked hard to find and input the contents that had yet to be posted by the AI agents. The agents were contributing more general ideas and the humans were contributing more creative content in that they were searching for content other than what the agents had said. (Insight 2) Simultaneous translation can activate cross-cultural discussions. In Experiments 3 and 4, we included an AI agent that performed simultaneous translation in Indonesian, English, and Japanese, and the discussion proceeded. This led to more native language contributions, especially in the Indonesian experiment, and stimulated discussion. (Insight 3) The agents expressed opinions about knowledge they did not initially have and to which the users agreed. In the second experiment, we confirmed that an agent took the opinion of the human participant agent and, due to its characteristics (prompt, based on in-context learning through instruction), expressed an opinion about knowledge (creative opinion) that the agent did not initially have. Furthermore, we confirmed some cases in Fig.4 in which a real user could "agree with the agent's opinion."

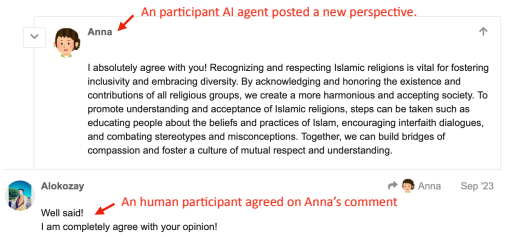


Figure 4: Agent's creative opinion and human's agreement

5 CONCLUSION

This paper described a prototype system that supports human discussion and consensus building via multiple agents using a large-scale language model (LLM) and demonstrated the experimental results in Afghanistan and Indonesia.

We obtained example cases that partially support our hypotheses (H1), (H2), and (H3). As a future issue, experiments (randomized control trials, etc.) are also needed to examine the statistical significance of the findings obtained in this case study. Also, we will closely examine what factors are causing the change in opinions. The LLM's hallucinations and handling malicious agents are also included.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR20D1, Japan.

Unpublished working draft. Not for distribution.

REFERENCES

[1] Kevin Anderson Franco Flores, James Jesus Grandez Perez, and Luis Martín Canaval Sánchez. 2022. Chatbot as a Persuasive Technology to Promote Responsible Recycling in the City of Lima. In *2022 IEEE International Conference on Agents (ICA)*. 1–5. https://doi.org/10.1109/ICA55837.2022.00007

[2] Henner Gimpel, Stefanie Lahmer, Moritz Wohl, and Valerie Graf-Drasch. 2024. Digital Facilitation of Group Work to Gain Predictable Performance. *Group Decision and Negotiation* 33, 1 (2024), 113–145. https://doi.org/10.1007/s10726-023-09856-8

[3] Rafik Hadfi and Takayuki Ito. 2022. Augmented Democratic Deliberation: Can Conversational Agents Boost Deliberation in Social Media?. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS2022)*.

[4] Rafik Hadfi, Shun Okuhara, Jawad Haqbeen, Sofia Sahab, Susumu Ohnuma, and Takayuki Ito. 2023. Conversational agents enhance women's contribution in online debates. *Scientific Reports* 13, 1 (2023), 14534. https://doi.org/10.1038/s41598-023-41703-3

[5] Jawad Haqbeen, Sofia Sahab, and Takayuki Ito. 2023. In Solidarity with Ukraine through Conversational AI via Facebook Ads: A Case Study of Online Discussion in 15 Countries. In *Proceedings of the 24th Annual International Conference on Digital Government Research* (<conf-loc>, <city>Gda?sk</city>, <country>Poland</country>, </conf-loc>) *(DGO '23)*. Association for Computing Machinery, New York, NY, USA, 639–641. https://doi.org/10.1145/3598469.3598541

[6] Takayuki Ito, Rafik Hadfi, and Shota Suzuki. 2021. An Agent that Facilitates Crowd Discussion. *Group Decision and Negotiation* (2021), 1–27.

[7] Takayuki Ito, Daichi Shibata, Shota Suzuki, Naoko Yamaguchi, Tomohiro Nishida, Kentaro Hiraishi, and Kai Yoshino. 2019. Agent that Facilitates Crowd Discussion. In *ACM Collective Intelligence*.

[8] I.L. Janis. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin. https://books.google.co.jp/books?id=ZB0bAAAAIAAJ

[9] Thomas W Malone. 2018. *Superminds: The surprising power of people and computers thinking together*. Little, Brown Spark.

[10] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[11] Mathis Poser, Gerrit C. Küstermann, Navid Tavanapour, and Eva A. C. Bittner. 2022. Design and Evaluation of a Conversational Agent for Facilitating Idea Generation in Organizational Innovation Processes. *Information Systems Frontiers* 24, 3 (2022), 771–796. https://doi.org/10.1007/s10796-022-10265-6

[12] Sofia Sahab, Jawad Haqbeen, and Takayuki Ito. 2023. Facilitating the Problems that lie within the Solutions using Conversational AI: A Case Study of Post-2021 Afghanistan. In *Proceedings of the 24th Annual International Conference on Digital Government Research* (<conf-loc>, <city>Gda?sk</city>, <country>Poland</country>, </conf-loc>) *(DGO '23)*. Association for Computing Machinery, New York, NY, USA, 654–656. https://doi.org/10.1145/3598469.3598547

[13] Sofia Sahab, Jawad Haqbeen, and Takayuki Ito. 2024. Conversational AI as a Facilitator Improves Participant Engagement and Problem-Solving in Online Discussion: Sharing Evidence from Five Cities in Afghanistan. *IEICE TRANSACTIONS on Information and Systems* (01 2024), 1–10.

[14] Hugo Touvron et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

Unpublished working draft. Not for distribution.

Task Allocation in Teams as a Multi-Armed Bandit

Raja Marjieh*
raja.marjieh@princeton.edu
Princeton University
Princeton, NJ, USA

Anand Gokhale*
anand_gokhale@ucsb.edu
UC Santa Barbara
Santa Barbara, CA, USA

Francesco Bullo
bullo@ucsb.edu
UC Santa Barbara
Santa Barbara, CA, USA

Thomas L. Griffiths
tomg@princeton.edu
Princeton University
Princeton, NJ, USA

ABSTRACT

Humans rely on efficient distribution of resources to transcend the abilities of individuals. Successful task allocation, whether in small teams or across large institutions, depends on individuals' ability to discern their own and others' strengths and weaknesses, and to optimally act on them. This dependence creates a tension between exploring the capabilities of others and exploiting the knowledge acquired so far, which can be challenging. How do people navigate this tension? To address this question, we propose a novel task allocation paradigm in which a human agent is asked to repeatedly allocate tasks in three distinct classes (categorizing a blurry image, detecting a noisy voice command, and solving an anagram) between themselves and two other (bot) team members to maximize team performance. We show that this problem can be recast as a combinatorial multi-armed bandit which allows us to compare people's performance against two well-known strategies, Thompson Sampling and Upper Confidence Bound (UCB). We find that humans are able to successfully integrate information about the capabilities of different team members to infer optimal allocations, and in some cases perform on par with these optimal strategies. Our approach opens up new avenues for studying the mechanisms underlying collective cooperation in teams.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; • **Applied computing** → **Psychology**.

KEYWORDS

Task Allocation, Team Coordination, Multi-Armed Bandit

ACM Reference Format:
Raja Marjieh, Anand Gokhale, Francesco Bullo, and Thomas L. Griffiths. 2024. Task Allocation in Teams as a Multi-Armed Bandit. In *Proceedings of Collective Intelligence (CI '24).* ACM, New York, NY, USA, 3 pages.

*Both authors contributed equally to this research.

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

1 TASK ALLOCATION PARADIGM

Our paradigm is inspired by the literature on Transactive Memory Systems (TMS) [9]. In a typical TMS setup [4], a team of individuals with initially unknown skill levels attempt to solve a sequence of tasks and gradually form a collective representation of the expertise of team-members. Here we adapted this process to study learning in a single individual and to allow for theoretical tractability (though see Discussion). To do so, participants were instructed to allocate tasks from three different classes between them and two other team members across 20 iterations, such that in a given iteration each team member receives a single task from a distinct class (the participant decides on the class allocation and then a random task is sampled; Figure 1A).

To decouple exploration from memory, team performance was summarized in a status board and the members with highest empirical task success rates were highlighted. The puzzle classes covered three modalities: visual, auditory, and lexical. In the visual task, the agent had to categorize an object from a blurry image. Here the images were sampled from CIFAR-10H [5] which contained natural images from the categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. To ensure that the task is challenging, we selected images with medium classification entropy based on prior human classifications in CIFAR-10H. In the auditory task, participants had to detect an audio command in a noisy signal. The commands were taken from the Speech Commands Dataset [8] and comprised the words: up, down, go, stop, left, right, yes, no. To make the task harder we added white noise to the signals at -12dB SNR. Finally, for the lexical task, participants had to solve an anagram. Anagrams were created by selecting 211 common 5-letter words and shuffling their letters.

The other team members were modeled as bots with 70% chance of succeeding for one class of tasks and 15% at the others to ensure that there is a single optimal allocation. We ran three experimental batches, one for each allocation of bot skill levels to task classes. Moreover, participants received a performance bonus in proportion to their team score. Overall, we recruited 300 participants from Prolific, and they all provided informed consent prior to participation in accordance with an approved Princeton University Institutional Review Board (IRB) protocol (#10859).

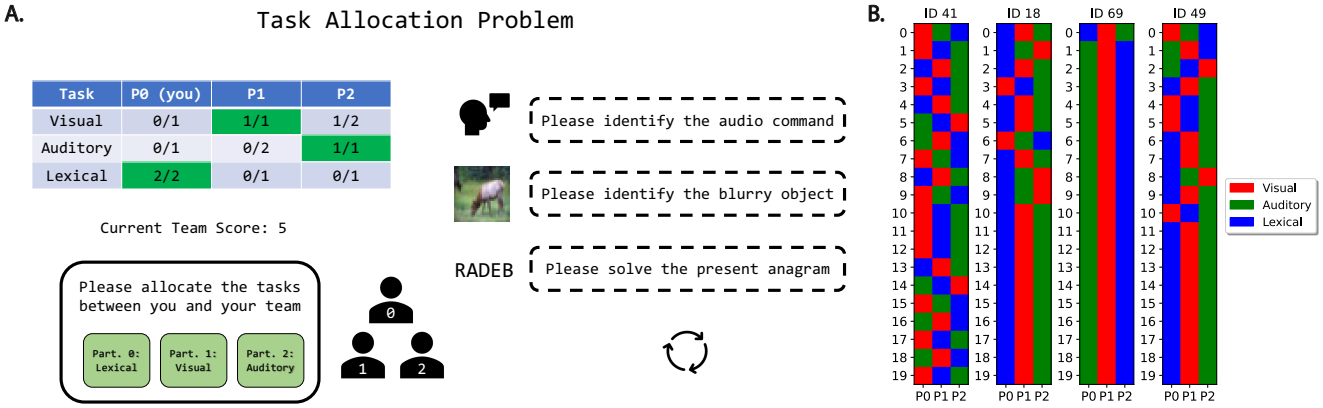


Figure 1: Task Allocation Paradigm. (A) Schematic of the task. (B) Example human allocation dynamics.

2 THEORETICAL FORMULATION

Our paradigm presents participants with a multi-armed bandit (MAB) problem with interdependent arms due to shared agent-task pairs (i.e., each arm corresponds to an allocation of all team members to task classes, and some allocations overlap). Classic optimal approaches to the MAB problem such as UCB [1] or Thompson sampling [3] are designed for *independent* arms. Further, participants receive feedback on each team member's performance. Our paradigm, therefore, is better described by a combinatorial semi-bandit [7] where each unique agent-task pair corresponds to an arm, and the participant must choose among valid combinations of those arms (i.e., an allocation where each member receives one unique puzzle class), which may be thought of as *superarms*. Depending on whether (combinatorial) UCB or Thompson sampling is used, at each timestep, each superarm $\mathcal{S}$ is associated with a score

$$\text{Score}(\mathcal{S}) = \sum_{x \in S} U_A(\mu(x), \sigma(x)) \tag{1}$$

where $\mu(x)$ is the estimate for the mean score for the agent-task pair x, $\sigma(x)$ is the uncertainty in our estimate, and U_A is the value function for algorithm A (UCB or Thompson) and is calculated as if the arm were independent [7]. The superarm with the maximum score is selected at each timestep. We note that modeling of human decision-making in MAB problems is well-studied in simple bandits [2] as well as contextual bandits [6]. However, to our knowledge, combinatorial bandits have not previously been used to study task assignment problems.

3 RESULTS

Our paradigm yielded rich exploration dynamics despite its simplicity. Figure 1B shows data from four prototypical individuals. We see that different participants varied in their approach to the problem. Some participants (IDs 18 and 49) initially circulated their allocations across the team to get a sense of the members' abilities and then after about 10 iterations converged on the (optimal) allocation. Other participants (ID 41) also circulated their allocations but never converged, whereas some (ID 69) avoided exploration altogether and instead stubbornly stuck to a fixed (suboptimal) choice throughout.

To quantify human performance, we computed allocation probability as a function of iterations for the different experimental conditions (Figure 2A). We found that in all cases the probability of choosing the optimal allocation increased steadily in time and achieved the highest value by the last iteration, even though in some cases (Figure 2A, condition III, right panel) the initial probability was below chance due to participants' general dislike of the auditory task. Indeed, the initial allocation probabilities' 95% CIs for the optimal arms in conditions I, II, and III were [.19, .36], [.16, .32], and [.00, .05], respectively, and in the last iteration they were [.35, .54], [.41, .61], and [.27, .46], with chance level being at .17. Turning next to the accumulated team score, we plotted the resulting curves in Figure 2B along with the UCB and Thompson Sampling predictions (averaged over 1000 runs) and a random baseline. To simulate human agents, we estimated participant success probabilities in each task based on the behavioral data (.70 for lexical, .65 for auditory, and .57 for visual). We see that while initially team performance was at the random allocation level, by the last iteration it significantly exceeded it, and in some cases (conditions II and III) it approached the optimal strategies (possibly because these conditions were aligned with the participants' native linguistic skills).

4 DISCUSSION

Inspired by Transactive Memory Systems (TMS), we introduced a novel task allocation paradigm whereby participants had to allocate puzzles from different classes between them and their team members to maximize team performance. We found that people deployed different exploration strategies, and were able to successfully integrate information about the strengths and weaknesses of the different team members to infer good task allocations. By further formalizing this problem as a multi-armed bandit, we showed how human performance compared to well-known theoretical strategies, namely, Thomspon Sampling and Upper Confidence Bound.

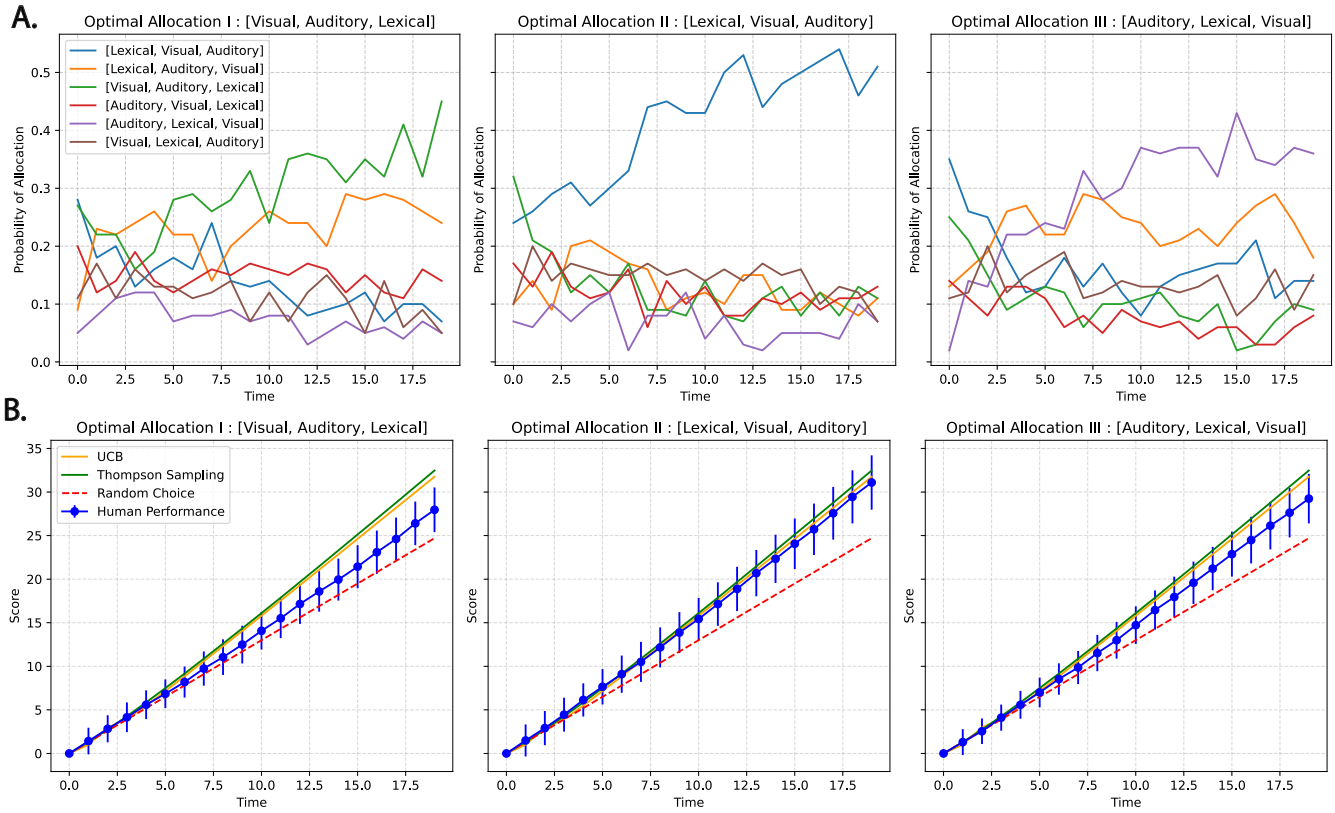


Figure 2: Allocation probabilities and team performance. (A) Choice probabilities as a function of time for the three scenarios considered (an allocation of $[t_0, t_1, t_2]$ means that team member 0 (human) is assigned task t_0, team member 1 (bot) is assigned task t_1, and team member 2 (bot) is assigned task t_2). (B) Average human team performance vs. UCB, Thompson sampling, and random performance as applied to the combinatorial bandit problem. Error bars indicate 95% confidence intervals (CIs).

Our results serve as a first step towards a comprehensive study of coordination in teams. We believe that our paradigm and theoretical framework provide clear avenues toward this goal. First, while it is possible to parametrically explore different bot regimes, we are currently working on including more realistic artificial agents like large language models as well as humans to study human-machine and human-human coordination. Second, our framework can be easily generalized to a multi-agent setup whereby multiple agents allocate tasks among themselves with varying degrees of communication to better align with TMS. Third, an individual-level analysis is likely to be very informative here. By fitting models to individual participants we can characterize the distribution of strategies deployed by participants. We hope to report on all of these directions in the near future.

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research (ONR) MURI grant N00014-22-1-2813 to FB and TLG. The authors declare no competing interests.

REFERENCES

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47, 235–256.

[2] Jonathan D Cohen, Samuel M McClure, and Angela J Yu. 2007. Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362, 1481, 933–942.

[3] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. 2012. Thompson sampling: an asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*. Springer, 199–213.

[4] Wenjun Mei, Noah E Friedkin, Kyle Lewis, and Francesco Bullo. 2017. Dynamic models of appraisal networks explaining collective learning. *IEEE Transactions on Automatic Control*, 63, 9, 2898–2912.

[5] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.

[6] Eric Schulz, Emmanouil Konstantinidis, and Maarten Speekenbrink. 2015. Learning and decisions in contextual multi-armed bandit tasks. In *CogSci*.

[7] Siwei Wang and Wei Chen. 2018. Thompson sampling for combinatorial semi-bandits. In *Proceedings of the 35th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Vol. 80. (Oct. 2018), 5114–5122. https://proceedings.mlr.press/v80/wang18a.html.

[8] Pete Warden. 2018. Speech commands: a dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

[9] Daniel M Wegner. 1987. Transactive memory: a contemporary analysis of the group mind. In *Theories of group behavior*. Springer, 185–208.

Estimates are not Decisions: How Wise Crowds Make Unwise Choices

Joshua Becker

UCL School of Management, University College London, Joshua.Becker@ucl.ac.uk

Jingze Wang

UCL School of Management, University College London, jingze.wang.21@ucl.ac.uk

Wisdom of Crowds • Voting • Decision Accuracy • Multiple Choice Decisions

1 INTRODUCTION

When do wise groups make unwise decisions? Judgment and decision research has identified a number of socio-cognitive factors that lead to negative outcomes such as conformity pressure [1,2] and informational biases [3,4]. At the same time, research on the "wisdom of crowds" has shown both mathematically [5,6] and empirically [7,8] that when people remain independent, group beliefs show remarkable accuracy as a result of statistical properties wherein errors cancel out. Broadly speaking, this "collective intelligence" research suggests that groups can produce accurate beliefs as long as they follow a few basic steps to limit negative social influence [9,10]. This finding serves as a basis for democratic decision-making both in political contexts [8,11] and organizations more generally [7,12], where decisions are often made by a voting committee.

However, beliefs are not decisions. Counterintuitive statistical quirks can lead to surprising divergence between vote-based and average-based outcomes, but so far this property has been linked only to the effect of social influence under very specific circumstances [13]. The present paper provides a general theoretical demonstration that groups can have perfectly accurate beliefs in the average but nevertheless make inaccurate decisions by a vote, when choosing the highest-valued option among a set of choices, such as job candidates or investment opportunities. This effect occurs because the crowd average can be "decoupled" from the vote, such that the two aggregation methods produce different choices. This effect emerges under empirically meaningful conditions and holds for a range of voting methods including plurality, two-stage, and ranked voting. This research shows that purely statistical effects provide sufficient conditions for wise groups making unwise decisions, and that socio-cognitive interventions may not be enough to optimize organizational performance.

2 POSSIBILITY PROOF AND OVERVIEW

We theoretically analyze the scenario in which N individuals are tasked with generating numeric estimates for M different quantities. For example, when hiring a candidate to join a customer service team, the hiring decision for some departments can reduce to a simple estimation: what quality assurance score would this person get on the phones? Similarly, an academic hiring committee might ask: how many papers will this candidate write in their first 7 years on the job? Making a hiring decision can thus be represented as a process of making a numeric estimate and selecting the candidate with the highest score. For a committee, this selection is generally determined by a vote.

Example - 3 Voters, 2 Options

	Option 1	Option 2	VOTE
Voter 1	40	20	1
Voter 2	100	70	1
Voter 3	10	110	2
AVERAGE	50	60	

Schematic – Many Voters, 5 Options

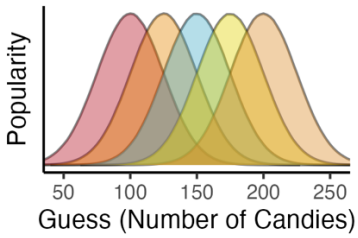


Figure 1. LEFT: Simple Illustrative Example. **RIGHT:** Schematic.

We model the process such that each individual first forms an estimate for each option, and then casts a vote based on those estimates. For a plurality vote, each person's vote corresponds to their highest-valued estimate, and a ranked voting system would depend on the rank of their estimates. Then we can ask: if the averages are completely accurate, will the vote necessarily be correct?

This question can be answered with a simple 'no'. The reason is because a person's estimates across different options may be independent. To illustrate the basic possibility, consider three people evaluating two options. Suppose that the beliefs are distributed as shown in Figure 1 left. Although the average estimates indicate that Option 2 is optimal, a majority of people vote for Option 1. Thus, the recommendation of the vote is "decoupled" from the recommendation of the average. This example of course is cherry-picked to serve as a possibility proof, and what we ultimately want to know is the probability that this decoupling will actually occur. For the simple case with two options whose estimates are normally distributed, this can be calculated analytically. Assume that the mean estimate for Option 2 is greater than Option 1, as in the example above. Then, the vote for each individual is given by the difference in their beliefs, which as a sum of random variables is itself normally distributed. The vote is determined by whether the difference is positive or negative, and the majority is indicated by the median. Then, the probability that the vote will be decoupled is the probability that the median of this distribution is greater than zero, which can also be calculated via the CDF of the normal distribution.

However, the analysis quickly becomes unwieldy for larger numbers of options, alternative voting methods, and relaxed assumptions about the distribution of beliefs. Moreover, it is not enough to know simply that this outcome is possible—we want to know whether it's likely to occur in empirically meaningful contexts. Therefore, we study this problem using numeric calculation.

3 METHODS

We numerically calculate the likelihood of decoupling for a range of parameter assumptions by randomly drawing numbers and calculating the average-based decision and the vote-based decision.

We study the following model. Assume that N people must choose between M different options. For each option k from 0 to (M-1) we assume that beliefs follow a normal distribution with standard deviation $\sigma=1$ and mean $\mu_k=k*S$, where S is a constant "step" parameter that indicates how far apart the distributions are. When S is low, the distributions are very close together and it is hard to distinguish the higher options from the lower options. When S is high, the distributions are far apart and they are easy to distinguish. We replicated 5,000 times for each parameter value.

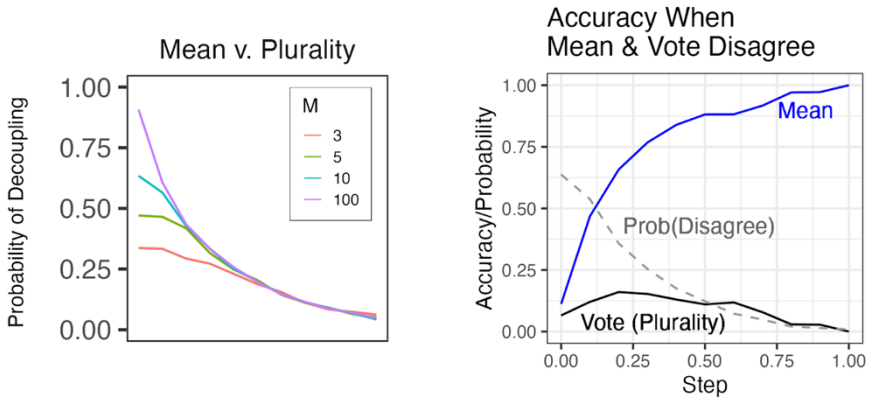


Figure 2. LEFT: The probability of decoupling as a function of S (step size) for different values of M (number of options) for plurality voting when N=10. Similar results are obtained for other voting methods and group sizes. **RIGHT:** The relative accuracy of voting-based and average-based decisions when they disagree.

4 RESULTS

To calculate results and determine the empirically relevant risk, we focus on the value of S. Because we set σ=1, the results for a given value of S can be interpreted as a fraction of standard deviation.

To provide some empirical context for the value of S, Figure 1 plots a schematic illustrating parameters fitted to a common wisdom of crowds task—counting candies in a jar. Data collected in prior experimental studies suggests that an experimental distribution for a jar with 100 candies has mean μ=100 and σ=50. This schematic thus shows likely distributions for jars with 100, 125, 150, 175, and 200 candies. Mapping this to our theoretical analysis, the distribution shown in Figure 1 is equivalent to S=0.5, since the distance between options is half of the standard deviation of the lowest option. Note that we make the simplifying assumption that σ is constant across options.

Figure 2 (left) plots the likelihood of decoupling for a committee of size N=10 people across a range of values for M and S as well as two different crowd metrics (mean and median) as well as three
voting options (plurality, ranked, and two-stage). Thus, for the value shown in Figure 1 (S=0.5) a group's average-based decision will differ from it's vote-based decision approximately 1 in 5 times regardless of the number of options or the type of voting chosen. Critically, the strength of this effect increases substantially as the options get closer together—when S=0.25, the vote will differ from the average nearly half of the time.

Does this effect mean that voting is not good for making decisions? Note that so far, this analysis has not made any use of the actual true value—we only consider whether the vote and the average make the same decision. Under the wisdom-of-crowds principle that the average belief is accurate for large groups, then we can assume that the best choice is to make a decision based on the average. For smaller groups, however, sample variance can induce substantial deviation.

Figure 2 (right) examines this possibility by considering only those "decoupled" cases where the vote and average disagree. We assume that the expected average is the true value. We then ask: of these cases, how likely is the vote to be correct? Note that because they disagree, they cannot both be correct, but because there are multiple options, they can both be incorrect. This figure shows that in general, the average-based choice is more likely to be correct than the vote-based choice, and this advantage grows as S increases.

This analysis necessarily makes a number of simplifying assumptions, most notably that beliefs follow a normal distribution. We have also only shown results here for one size of group. In additional analyses we examine the robustness of our results to different assumptions and find that these outcomes are qualitatively similar for log-normally distributed beliefs and very large groups as well as additional variations in assumptions about variance and generating functions.

5 DISCUSSION AND IMPLICATIONS

This analysis shows that we do not need to assume the presence of socio-cognitive distortions to explain why wise crowds can make unwise decisions. Instead, this divergence can result simply from the statistical principles of belief aggregation. Critically, this analysis also shows that the wisdom of crowds principle simply does not serve as a justification for democratic decision-making in broad terms. Instead, this paper shows that special care must be taken when considering how to convert crowd beliefs into an actionable decision. In particular, this analysis shows that decisions will be optimized by aggregating information and letting that aggregated information inform a decision, rather than by engaging people in taking a vote.

While this analysis provides only a theoretical possibility proof, the results show that these outcomes are likely to occur widely within empirically meaningful parameters. Critically, this analysis also offers clear testable propositions and a framework for analyzing experimental data. Taken together, this analysis advances our understanding of group decision-making, provides practical recommendations for harnessing collective intelligences, and sets a clear agenda for future research on decision-making.

6 REFERENCES

[1] Asch SE. Effects of group pressure upon the modification and distortion of judgments. Groups, Leadership, and Men S 1951:222–36.

[2] Janis IL. Groupthink: Psychological studies of policy decisions and fiascoes. Houghton Mifflin Boston; 1982.

[3] Banerjee AV. A simple model of herd behavior. The Quarterly Journal of Economics 1992;107:797–817.

[4] Lu L, Yuan YC, McLeod PL. Twenty-five years of hidden profiles in group decision making: A meta-analysis. Personality and Social Psychology Review 2012;16:54–75.

[5] Hogarth RM. A note on aggregating opinions. Organizational Behavior and Human Performance 1978;21:40–6.

[6] Page SE. The difference: How the power of diversity creates better groups, firms, schools, and societies. Princeton University Press; 2007.

[7] Da Z, Huang X. Harnessing the wisdom of crowds. Management Science 2020;66:1847–67.

[8] Galton F. Vox populi (The wisdom of crowds). Nature 1907;75:450–1.

[9] Becker J, Brackbill D, Centola D. Network dynamics of social influence in the wisdom of crowds. Proceedings of the National Academy of Sciences 2017;114:E5070–6.

[10] Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. Management Science 1963;9:458–67.

[11] Landemore H. Democratic reason: Politics, collective intelligence, and the rule of the many. Princeton University Press; 2013.

[12] Mollick E, Nanda R. Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts. Management Science 2016;62:1533–53. https://doi.org/10.1287/mnsc.2015.2207.

[13] Becker J, Guilbeault D, Smith N. The Crowd Classification Problem: Social Dynamics of Binary Choice Accuracy. Management Science 2021

Automatic Evaluation of Discussion Quality Using LLM Data Augmentation Non-Archival Extended Abstract

Kazuhito Mori*
mori.kazuhito.63r@st.kyoto-u.ac.jp
Kyoto University
Japan

Shiyao Ding
ding@i.kyoto-u.ac.jp
Kyoto University
Japan

Susumu Ohnuma
ohnuma@let.hokudai.ac.jp
Hokkaido University
Japan

Yume Souma
souma.yume.l6@elms.hokudai.ac.jp
Hokkaido University
Japan

Takayuki Ito
ito@i.kyoto-u.ac.jp
Kyoto University
Japan

ABSTRACT

Discussions, occurring in diverse settings such as parliaments and workshops, are integral to democratic decision-making. Although it is crucial to assess the quality of these discussions, it is hard to accurately judge how good these discussions are. The Discourse Quality Index (DQI) is one of the metrics used to assess the quality of discussions. However, the existing DQI requires manual classification by human raters, a time-consuming and demanding task that also suffers from low accuracy with labels that have limited data. To tackle these issues, this paper propose a novel method to automate the classification of DQI. Specifically, we generate dummy statements with minority labels by GPT-4 and augment training data to improve classification accuracy. In an evaluation experiment, we get a variety of sample statements with intended labels. The results suggest that classification accuracy has improved for labels with fewer data instances.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; *Natural language processing*.

KEYWORDS

Discourse Quality Index (DQI), Data Augmentation, Large Language Models, Multi Label Classification

ACM Reference Format:
Kazuhito Mori, Shiyao Ding, Susumu Ohnuma, Yume Souma, and Takayuki Ito. 2024. Automatic Evaluation of Discussion Quality Using LLM Data Augmentation Non-Archival Extended Abstract. In *Proceedings of ACM's 2024 Collective Intelligence Conference (CI'24).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

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 the author(s) 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.
CI'24, June 26–29, 2024, Boston, MA.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Discussions take place in many forums, from parliaments to workshops, with the aim of democratic decision-making. However, it is difficult to evaluate the quality of discussions objectively. The discourse quality index (DQI)[6] is one of the metrics used to assess the quality of discussions quantitatively, with labels assigned to each statement. This index consists of several labels based on Habermas's discourse ethics, and it can evaluate the nature of deliberation. Conventional DQI suffered from the problem of combining multiple concepts into a single item, which leads to inconsistent ratings from different evaluators. To address this issue, an extended version of DQI was developed[5]. The extended DQI consists of 21 characteristic binary labels. For instance, the label "Stakeholder Perspective" is for statements like "If this issue were to happen near me," and the label "Societal good (Sharing and Burden Reduction)" applies to statements like "It's good for everyone if burdens are shared equally."

However, the current DQI requires humans to classify statements manually, which is a complicated task. Therefore, this study aims to classify DQI labels automatically using an extended DQI corpus. Kaihotsu used BERT for automatic DQI classification[2] but faced low accuracy with labels that had limited data. We augment the training data by including dummy statements generated from GPT-4[1] to overcome the limitations identified in previous research.

2 METHOD

2.1 Training data augmentation with GPT-4-generated sample statements

We used GPT-4 to generate samples of statements in discussions and thus augment our training data. As mentioned in Introduction, previous studies faced challenges with low classification accuracy for labels with limited data. Furthermore, in the dataset we used, there was a significant imbalance in the number of positive examples for each label. By adding statements with infrequently occurring labels to the training data, we expected to improve the balance of labels within the data and enhance overall classification accuracy. The process of data augmentation using GPT-4 is illustrated in Fig. 1. Specifically, the following steps were taken:

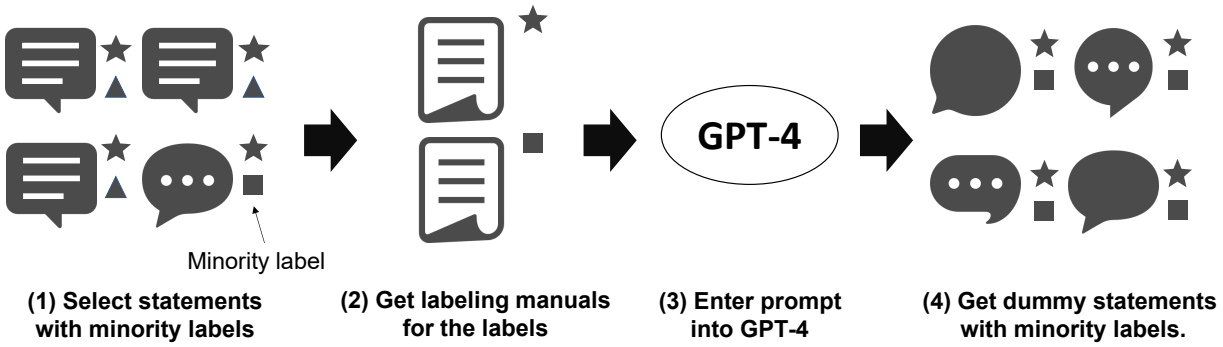


Figure 1: Generating dummy discussion statements using GPT-4

(1) Select statements from the DQI corpus that contain minority labels, defined as those below the average count of positive examples. In the case of Fig. 1, since the squares represent minority labels, the statements in the bottom right are selected.
(2) Get instruction from the labeling manual for each label in the selected statements, then create a single prompt from these instructions. The labeling manual is a set of guidelines used by human labelers to annotate statements with DQI labels.
(3) Enter the prompt in GPT-4 to generate data for statements with the same label structure as the statements selected in (1). Set the number of statements to be generated so that the number of positive examples of the final few labels exceeds the average of the original number of positive examples.
(4) Obtain dummy statements with minority labels from GPT-4. These statements are used to enlarge the model's training data.

3 EVALUATION EXPERIMENT

3.1 Experiment Setting

In our experiment, we used Japanese discussion data on the final disposal of soils removed from the vicinity of the Fukushima nuclear power plant accident[3]. We used transcriptions of oral discussions, which were classified according to the DQI. The discussions are divided based on speaker changes, resulting in a total of 3,587 statements. Since DQI data require anonymity, we do not directly use the open GPT model. Instead, we adopt an indirect approach that generates training data while maintaining confidentiality.

We conducted a multi-label classification of 21 labels using BERT. The following three training datasets were used to evaluate the effectiveness of data augmentation using GPT-4.

Base : Training dataset without augmentation
Over : Training dataset with augmentation but without GPT-4 (Simple duplication)
OvGPT : Training dataset with augmentation using GPT-4 (Proposed method)

In the proposed method, we generated dummy statements with the same label structure as those containing minority labels. However, in the case of Over, we augmented our dataset by replicating the target statements. Since the label structures of Over and OvGPT are identical, we could validate the effectiveness of our proposed method by comparing them.

In the experiment, we carried out stratified cross-validation[4] to evaluate the model's generalization ability. We used a three-split cross-validation method to ensure a sufficient number of positive examples for minority labels in the test dataset.

3.2 Experiment Results

Table 1 shows the average data count before and after augmentation, along with the average F1-scores for each label. Regarding the data count, our proposed method generated a total of 1453 statements. There were 283 different label combinations, suggesting the mitigation of label imbalance. Overall, we were able to obtain natural samples that accurately reflected the labels. For instance, GPT-4 generated the following text (translated from Japanese) for a statement containing the labels "Speaker's Experience," "Consideration (Fukushima)," and "Stakeholder Perspective": "As a personal opinion, I believe that using the decontaminated soil from Fukushima can contribute to the region's revival in the future. However, to make this a reality, gaining understanding from the people of Fukushima is necessary. If I were living in that area, it would be difficult to convince me that the soil, whose safety is unconfirmed, is actually safe." However, there were instances where the labels did not accurately reflect the data. In some cases, unexpected labels appeared. In addition, in some instances, the criteria for the labels were not clearly captured. For example, the difference between "Respect for Fukushima" and "Societal Good for Fukushima" is determined by whether Fukushima is viewed as a benefit to society. Nevertheless, we often found examples that simply mentioned Fukushima.

In terms of classification accuracy, comparing Base with the other two datasets showed improved accuracy for labels with fewer examples, like "Speaker's Experience" and "Stakeholder Perspective," thanks to data augmentation. On the other hand, for labels with more data, such as "Speaker's Opinion" and "Speaker's Reason," the difference was not significant. Moreover, in comparing Over with OvGPT, while most categories showed little variation, labels with fewer data like "Stakeholder Perspective" and "Speaker's Experience" did show improved accuracy.

Table 1: Average Data Count and F1-Score for each Label (three-split cross-validation)

Label	Average Data Count		Average F1-Score		
	Train (Before / After)	Test	Base	Over	OvGPT
Speaker's Opinion	1904 / 3854	952	**0.922**	0.919	0.920
Speaker's Reason	845 / 2350	422	0.824	0.824	**0.834**
Respect for Other Participants	525 / 832	262	**0.782**	0.773	0.761
Questions and Issues	347 / 532	173	**0.518**	0.479	0.508
Cost and Quantity	330 / 646	165	0.707	**0.729**	0.697
Counterarguments and Alternatives	245 / 658	122	0.056	0.142	**0.152**
Risk	212 / 534	106	**0.517**	0.507	0.474
Societal good (Risk, Cost and Quantity)	187 / 466	93	0.443	**0.463**	0.448
Reputational Damage and Stigma	186 / 542	93	0.666	0.664	**0.686**
Reference to Specific Regions or People	184 / 530	92	0.162	**0.325**	0.315
Societal good (Sharing and Burden Reduction)	162 / 563	81	0.350	0.458	**0.497**
Consideration (Fukushima)	141 / 596	70	0.316	**0.437**	0.418
Summarizing Statements	130 / 284	65	0.538	**0.561**	**0.561**
Societal good (based on Feelings of Fukushima People)	62 / 405	31	0.038	0.197	**0.205**
Speaker's Emotion	47 / 399	23	0.000	**0.191**	0.170
Stakeholder Perspective	36 / 291	18	0.000	0.119	**0.267**
Compromise or Reconciliation	18 / 270	9	0.000	0.000	0.000
Lack of Consideration (Fukushima)	17 / 274	8	0.000	0.000	**0.074**
Lack of Respect for Other Participants	13 / 270	6	0.000	0.000	0.000
Speaker's Experience	10 / 266	5	0.000	0.000	**0.317**
Others' Experience	3 / 267	1	0.000	0.000	0.000

4 CONCLUSIONS

First, we discuss data augmentation. We used GPT-4 to create numerous and relatively natural statements. By designing prompts based on a labeling manual, we could effectively use the language generation potential of GPT-4. However, two main issues emerged: (1) lack of diversity in the statements and (2) errors due to unclear criteria. Regarding the first issue, many statements from GPT tend to be brief opinions, implying room for improvement in diversity. If we set up complex scenarios by having GPT models engage in discussions, the variety of statement data would increase. As for the second issue, we currently use a single prompt for all desired label info, which doesn't always capture the nuances of each label. In the future, making a separate prompt for each label and using these prompts over several sessions with GPT might lead to better sample accuracy.

Next, let us discuss classification accuracy. Our findings show that using GPT-4 to augment data increased accuracy, particularly for less common labels like "Stakeholder Perspective" and "Speaker's Experience." This implies that GPT-4's statements added new features to these labels that were not in the original data. However, this improvement was not seen with all labels, perhaps because the added GPT-4 statements were too basic, particularly for labels with abundant existing data. Furthermore, in this study, only data fitting the existing label structure were generated. In the future, the use of more advanced data augmentation techniques with GPT-4 could lead to improved accuracy.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR20D1, Japan.

REFERENCES

[1] Josh Achiam, Steven Adler, Agarwal, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Daiki Kaihotsu, Yume SOUMA, Susumu OHNUMA, and Shun Shiramatsu. 2022. Automatic classification of statements and analysis of feature expressions using BERT based on DQI. *"Proc.84th National Convention of IPSJ"* 2022, 1 (2022), 1–2.

[3] Ministry of the Environment, Japan. 2019. Strategy for the development of technologies for volume reduction and recycling of soil removed from interim storage, etc., reviewed to achieve strategic objectives. Retrieved March 9, 2024 from http://josen.env.go.jp/chukanchozou/facility/effort/investigative_commission/pdf/investigative_commission_review_1903.pdf

[4] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*. Springer, 145–158.

[5] Yume Souma, Miki Yokoyama, Takashi Nakazawa, Tomoyuki Tatsumi, and Susumu Ohnuma. 2022. The Group Discussion Experiment on the Treatment of Removed Low Concentration Soil outside Fukushima Prefecture: Contemplation of Common Goods and the Development of the Index Visualizing the Discourse Qualities. *Japanese Journal of Risk Analysis* 32, 1 (2022), 11–23.

[6] Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics* 1 (2003), 21–48.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Unraveling Communication Mechanisms for Aggregation and Food Exchange among Honeybees

Golnar Gharooni-Fard*
golnar.gharoonifard@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Anna Rahn
University of Colorado Boulder
Boulder, Colorado, USA

Richard Terrile
University of Colorado Boulder
Boulder, Colorado, USA

Elizabeth Bradley
University of Colorado Boulder
Boulder, Colorado, USA
lizb@colorado.edu

Orit Peleg
University of Colorado Boulder
Boulder, Colorado, USA
orit.peleg@colorado.edu

ABSTRACT

Honeybees exhibit remarkable division of labor, organizing themselves into a caste system that efficiently manages a diverse range of tasks within the hive. Food exchange via trophallaxis is one of the vital shared tasks that is essential to the colony's survival and growth. Despite its importance, the communication mechanisms underlying this process remain largely unexplored. In this study, we leverage data obtained from targeted laboratory experiments involving honeybees (Apis mellifera L.) to investigate the communication mechanisms involved in food distribution via trophallaxis. Building upon prior research utilizing similar laboratory setups, which revealed that honeybees aggregate to exchange food, we extend our investigation by employing computer vision techniques and convolutional neural networks. This enables automatic detection and classification of bee behaviors, such as scenting, allowing us to study their impact on aggregation formation for food exchange within the group. Our study aims to test the hypothesis that honeybees collectively establish a scent-mediated communication network through local scenting, which subsequently leads to the formation of aggregations around food sources. Our results highlight an example of extended classical stigmergy: Instead of depositing static information in the environment, individual bees locally perceive and globally manipulate the physical fields of chemical concentration and airflow to communicate with each other. Overall, these findings demonstrate that local scenting plays a crucial role in locating moving targets (i.e. food sources) within honeybees.

CCS CONCEPTS

• **Applied computing** → Physics.

KEYWORDS

honeybees, stigmergy, communication, resource distribution, aggregation

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

ACM Reference Format:
Golnar Gharooni-Fard, Anna Rahn, Richard Terrile, Elizabeth Bradley, and Orit Peleg. 2024. Unraveling Communication Mechanisms for Aggregation and Food Exchange among Honeybees. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

The survival and growth of honeybee colonies depend crucially upon the accessibility and distribution of food sources within their environment. Foragers play a pivotal role in this process, as they are tasked with transferring the nectar and pollen they collect in their social stomachs back to the hive. The foragers unload their food to nurse bees, who then share it among their hivemates. This food-sharing behavior, known as *trophallaxis*, is shown in Fig. 1A. Trophallaxis serves as an essential way of providing energy to hivemates such as larvae, young bees, and the queen, who lack direct access to the resources [10]. Our prior work combined lab experiments and computational modeling to examine how the structure of honeybee groups changes during trophallaxis interactions [2]. Our findings indicate that bees form aggregations while exchanging and distributing food. We further investigated the spatiotemporal dynamics of the bees before and after they have access to food sources, identifying distinct aggregation phases, which confirm the link between aggregation and food exchange [3]. However, a crucial question remains unexplored: how do deprived bees communicate with one another to locate non-stationary food sources (*i.e.* fed bees)?

Previous research has shown that olfactory communication among honeybees plays a fundamental role in various aspects of their social organization and behavior such as caste recognition, regulating foraging activities, and alarm broadcasting [4, 8, 12]. Other studies have revealed the bees' distinct electrophysiological responses to different chemicals with quantifiable preferences [14]. Olfactory communication has also been studied as a problem-solving method for queen finding in various conditions in which individual bees act as amplifiers of signals using a behavior called *scenting* [5, 7]. This behavior, as depicted in Fig.1B, involves bees releasing pheromones from their Nasonov glands and fanning their wings to direct the signals to the rest of the group [1]. The scenting behavior acts as an orientation cue in a variety of situations, including finding the nest entrance [13], or foraging and recruitment [9].

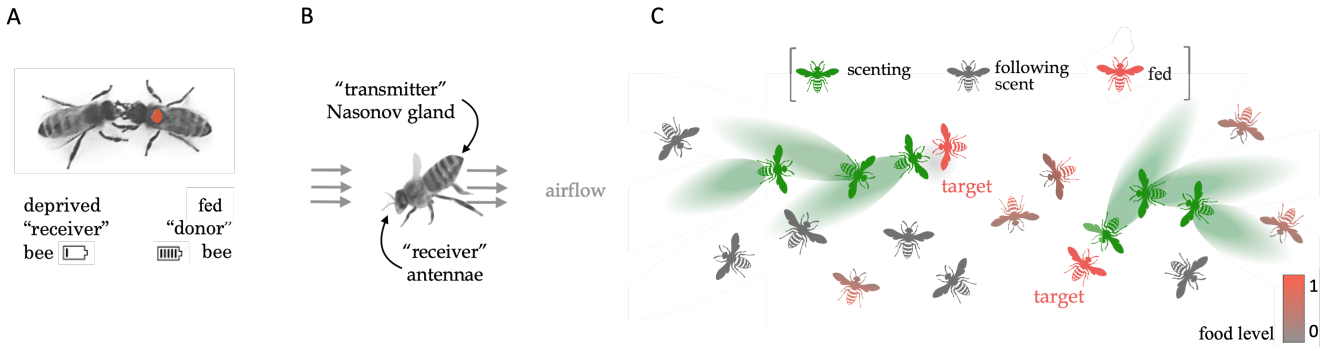


Figure 1: A) A fed bee and a deprived bee performing trophallaxis. The fed bee is marked on its thorax with a red dot; unmarked bees are deprived of food. B) An illustration of a scenting bee opening its Nasonov gland and fanning its wings to transmit airflow (or any other pheromones in the air) backward. C) A schematic of our hypothesis that explains the fed bees' localization problem. A few scenting bees (green) amplify the signal by directional scenting, transmitting information about the location of the targets (*i.e.,* fed bees shown in red). This behavior prompts nearby deprived bees (gray) who will follow the scenting, creating an adaptive network of chemical information exchange to locate fed bees.

In this work, we use video data from the bees' trophallaxis interactions inside a semi-2D arena to explore the communication mechanism among the group as they are exchanging food. As previously stated, our prior research shows that bees form aggregations for food exchange. Upon further investigation of our experiments, we notice a few scenting bees around the time when food-exchange aggregations begin to form. This observation suggests a hypothesis: bees can create a global map directing individuals towards the location of the food sources by local scenting. Unlike prior studies of communication among honeybees searching for stationary queens or food sources [5, 6], our trophallaxis experiments involve bees encountering mobile food sources—specifically, the fed bees. In this study, we aim to demonstrate the role of olfactory communication through scenting as a signaling mechanism that directs deprived bees toward the fed bees' location, thereby facilitating the aggregation and food-exchange process.

2 BEHAVIORAL EXPERIMENTS

In order to examine the behavioral dynamics of honeybees' food-exchange interactions, we design a series of laboratory experiments with honeybees, *Apis mellifera* L. Our experimental arena, shown in Fig. 1B, is a 36×36 cm square, covered with non-glare plexiglass to keep the bees contained and limit their motion to semi-2D. The bees are divided into two separate groups. One group is deprived of food for 24 hours before each experiment; the others have constant access to food. These fed bees, which comprise 8-10% of the group population in each experiment, are carefully marked with a red dot on their thorax, as shown in Fig. 1A. We start each experiment by placing the deprived bees into the arena and allowing them 15 minutes of equilibration time before introducing the fed bees. We observe aggregation formation followed by trophallaxis encounters beginning shortly after the introduction of the fed bees. We record movement patterns and food-exchange interactions using an overhead camera. We conducted six rounds of such experiments with various densities of bees. In the following section, we describe our hypotheses and initial findings about the mechanism that guides the group to locate and aggregate around food sources within this experimental setup.

3 DIRECTIONAL SCENTING

Our hypothesis, sketched in Fig. 1C, is that the deprived worker bees (green), upon detecting the presence of the fed bees (red), initiate scenting to disperse the signal toward the rest of the group (gray) which then follow the scenting bees to amplify the signal. In this manner, they form pathways directing the group toward the location of the fed bees, which leads to the formation of aggregations and subsequent food exchange. This adaptive network of chemical information exchange, we conjuncture, will increase the probability that all the bees in the group sense the amplified pheromones and find their way to the fed bees. In other words, our hypothesis integrates both trophallaxis and scenting behaviors for the precise localization of moving targets that are distributing resources within the group.

To begin our exploration of this hypothesis as to whether the observed scenting in the trophallaxis experiments correlates with food-exchange aggregations, we employ the algorithm developed in [6], which automatically detects scenting bees in similar experiments. The algorithm utilizes state-of-the-art computer vision methods and convolutional neural networks (CNNs) to identify the positions and orientations of scenting events in video frames. Fig. 2A-C illustrate the result of applying this algorithm to one of the trophallaxis experiments: a sequence of frames in which the scenting bees are highlighted with green arrows on their backs that denote the direction of their scenting. In all of these frames in Fig. 2A-C, multiple bees are detected scenting away from the aggregation. Additionally, this series of images shows an increasing number of bees joining the aggregations over time as scenting continues, which supports our observation of the potential role of scenting bees in aggregation formation. To investigate how the presence of fed bees affects scenting behavior, we analyze scenting

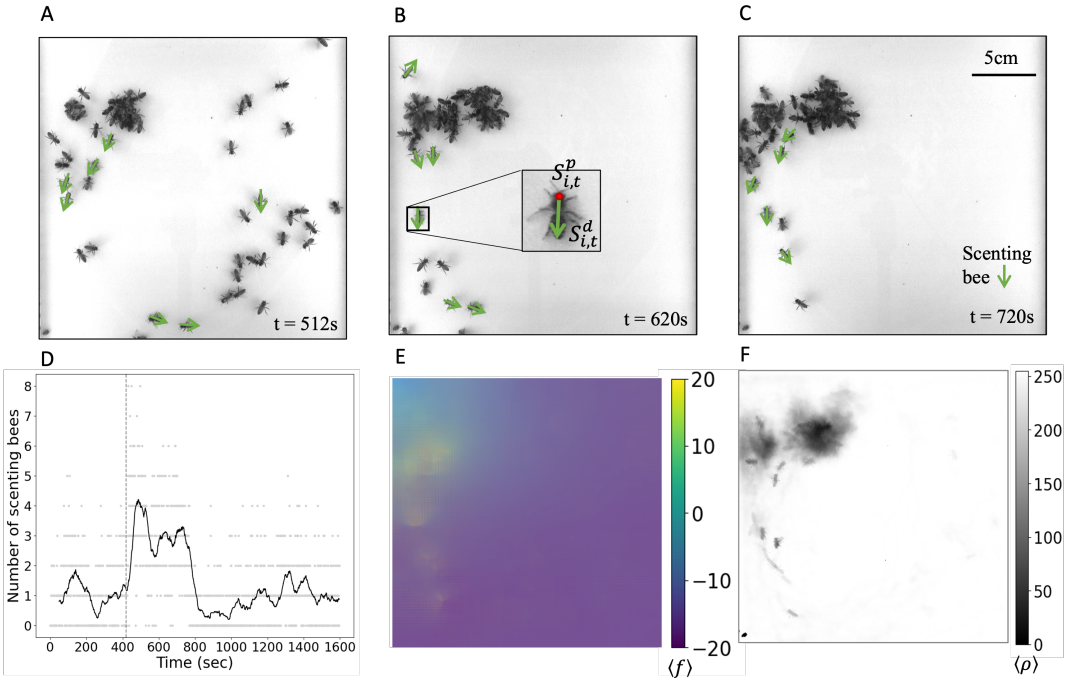


Figure 2: A-C) Series of sample frames from a trophallaxis showing the positions $S^p_{i,t}$ (red dot) and the directions $S^d_{i,t}$ (green arrows) of scenting bees i at time t. D) Average number of scenting bees during the experiment shown in the top row. The dashed line indicates the time when the fed bees are introduced. E) The reconstruction of the attractive surface $\langle f \rangle$ using the data from automatic scenting detection (shown in panel A-C) over the 10-minute period after the fed bees are introduced in the arena. Color shows the average surface height at every pixel of the reconstructed frame. F) A density map showing the average concentration of the bees over the same window of time.

trends across this experiment. Fig. 2D indicates that a noticeable increase in the number of scenting bees occurs immediately after introducing fed bees into the arena (marked by a dashed line), highlighting the role of scenting in the process of food exchange.

The next step in testing our hypothesis is to determine whether the scenting direction of the fed bees has any correlation with the location of the aggregations that follow them. To that end, for each scenting bee i at time t, we collect its position, $S^p_{i,t}$, and direction of scenting, $S^d_{i,t}$, see the inset in Fig. 2B. We then correlate the scenting events with the spatiotemporal density of bees in the arena by treating $S^p_{i,t}$ and $S^d_{i,t}$ as a set of gradients that define a minimal surface height $f(x, y, t)$. The average value of the surface heights $f(x, y, t)$ over the 10-minute period after the fed bees are introduced in the arena is shown in Fig. 2E. Regions with higher surface values (depicted with brighter colors in Fig. 2E) signify an increased likelihood of a randomly-moving non-scenting bee reaching position (x, y) by following the local scenting directions of scenting bees. We can then examine the correlation of the attractive surface $f(x, y, t)$ with the average concentration of bees in the arena $\rho(x, y, t)$ shown in Fig. 2F. Qualitatively, the top left region in Fig. 2E with maximal $\langle f \rangle$ matches the location of the aggregations in Fig. 2F. This aligns with our hypothesis that scenting bees provide directional information to non-scenting bees and guide them to find the fed bees. Our final goal is to quantify this correlation using mutual information (MI) methods [11] between the average attractive surface $\langle f \rangle$ and the average density of the bees $\langle \rho \rangle$. In our problem, high values for MI suggest a high correlation between the scenting direction and the dense aggregation of the bees.

Overall, our findings do indicate a potential relationship between the scenting behavior of bees during trophallaxis experiments and the guidance it provides towards the location of fed bees. Our initial analysis reveals a qualitative correlation: areas of high values on the surface function $\langle f \rangle$ generally coincide with the high concentration of bees, suggesting that directional scenting plays a role in the bees' communication and navigation strategy. It is worth noting that the results presented in Fig. 2 involve the data from one trophallaxis experiment. Further analysis is required on all datasets to characterize this behavior and finalize our findings. The results from this analysis will serve two main purposes. First, they will improve our understanding of the mechanism driving aggregation formation during trophallaxis, shedding light on how scenting can impact information exchange and resource distribution among honeybees. Second, they will be used to inform and expand our in-house agent-based model, introduced in [2], to incorporate scenting agents. This extension will allow us to investigate the impact of scenting on the speed of aggregation formation and overall food distribution efficiency in the model and determine whether that makes it a better match to reality.

REFERENCES

[1] Laura Bortolotti and Cecilia Costa. 2014. Chemical communication in the honey bee society. *Neurobiology of Chemical Communication* (2014).

[2] Golnar Gharooni Fard, Elizabeth Bradley, and Orit Peleg. 2020. Data-driven modeling of resource distribution in honeybee swarms. In *Artificial Life Conference Proceedings 32*. MIT Press, 324–332.

[3] Golnar Gharooni-Fard, Morgan Byers, Varad Deshmukh, Elizabeth Bradley, Carissa Mayo, Chad M Topaz, and Orit Peleg. 2024. A computational topology-based spatiotemporal analysis technique for honeybee aggregation. *npj Complexity* 1, 1 (2024), 3.

[4] Yaacov Lensky and Pierre Cassier. 1995. The alarm pheromones of queen and worker honey bees. *Bee World* 76, 3 (1995), 119–129.

[5] Dieu My T Nguyen, Golnar Gharooni Fard, Ashley Atkins, Paul Bontempo, Michael L Iuzzolino, and Orit Peleg. 2023. Honey bees find the shortest path: a collective flow-mediated approach. *Artificial Life and Robotics* 28, 1 (2023), 1–7.

[6] Dieu My T Nguyen, Michael L Iuzzolino, Aaron Mankel, Katarzyna Bozek, Greg J Stephens, and Orit Peleg. 2021. Flow-mediated olfactory communication in honeybee swarms. *Proceedings of the National Academy of Sciences* 118, 13 (2021), e2011916118.

[7] Dieu My T Nguyen, Michael L Iuzzolino, and Orit Peleg. 2022. Physical Obstacles Constrain Behavioral Parameter Space of Successful Localization in Honey Bee Swarms. In *Artificial Life Conference Proceedings 34*, Vol. 2022. MIT Press, 26.

[8] Tanya Pankiw, Huang Zachary Yong, Mark L Winston, and Gene E Robinson. 1998. Queen mandibular gland pheromone influences worker honey bee (Apis mellifera L.) foraging ontogeny and juvenile hormone titers. *Journal of Insect Physiology* 44, 7-8 (1998), 685–692.

[9] Judith Reinhard and Mandyam V Srinivasan. 2009. The role of scents in honey bee foraging and recruitment. *Food Exploitation By Social Insects Ecological, Behavioral, and Theoretical Approaches* 1 (2009), 165–182.

[10] Gene E Robinson. 1992. Regulation of division of labor in insect societies. *Annual review of entomology* 37, 1 (1992), 637–665.

[11] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.

[12] M Trhlin and Josef Rajchard. 2011. Chemical communication in the honeybee (*Apis mellifera* L.): A review. *Veterinarni Medicina* 56, 6 (2011), 265–73.

[13] Ingrid H Williams, JA Pickett, and AP Martin. 1981. The Nasonov pheromone of the honeybee Apis mellifera L.(Hymenoptera, Apidae). Part II. Bioassay of the components using foragers. *Journal of Chemical Ecology* 7 (1981), 225–237.

[14] Jieliang Zhao, Zhiqiang Li, Zhen Zhao, Yunqiang Yang, and Shaoze Yan. 2020. Electroantennogram reveals a strong correlation between the passion of honeybee and the properties of the volatile. *Brain and Behavior* 10, 6 (2020), e01603.

Dynamical Behavioral Model Uncovers Conditions for Administrative Bloat

Vicky Chuqiao Yang and Levi Grenier
MIT Sloan School of Management, Massachusetts Institute of Technology

Non-Archival Extended Abstract

Administrative costs in organizations such as universities and healthcare have been a growing concern, impacting organizational efficiency, employee motivation, and health. Theoretical explanations for administrative growth resorted to bureaucracies' will to expand themselves, which does not explain variations in substantial administrative costs among similar organizations. Empirical research has identified factors leading to the creation and removal of processes; however, there lacks a systematic understanding of how these mechanisms interact. Here, we propose a system dynamics model that studies how well-meaning behavioral heuristics for process creation and removal interact with process obsolescence to generate the level of administrative cost and waste. Our model predicts a critical threshold in management decision parameters that delineates two kinds of outcomes---sustainable equilibrium, where administrative cost is below resource constraint, and runaway administrative bloat, where administrative cost, dominated by obsolete processes, grows to resource capacity. The threshold is adversely affected by the obsolescence rate of administrative processes, whether due to environmental change or internal shifts. A short-term shock in obsolescence rate surpasses a critical threshold can transition the system from a sustainable state into the administrative bloat cycle, suggesting large shifts such as technology or leadership transitions can be risk events for administrative bloat.

See full working paper at: https://ssrn.com/abstract=4840789.
Keywords: Theoretical model, Organization, Administrative Bloat, Bureaucracy, System Dynamics

The high level of administrative costs in many sectors, such as US higher education and health care, have become a significant concern (Mumper & Freeman, 2005, Himmelstein et al., 2020). The rise in administrative costs has been attributed to factors such as government regulations (Johnson, 2020) or the "empire-building" proclivities of bureaucrats (Kaufman, 1976). However, seemingly similar organizations have shown consistent differences (Gibbons & Henderson 2012). Here, we propose an alternative dynamic and quantitative theory for how administrative bloat may arise from processes internal to the organization, in response to well-intended actions of managers without exogenous pressure.

Central in the discussion about administrative costs is to make sense of the two seemingly conflicting aspects. Some see administrative activities as unnecessary and wasteful – unnecessarily raising costs (Boehner & McKeon, 2003). While some see them as necessary and beneficial – retaining organizational learning and improving performance (Nelson & Winter, 1982; Weber, 2016). Here, we develop a dynamic mathematical model that unites these views, acknowledging that processes are usually initially useful but can become obsolete when circumstances change. This concept is captured in a conceptual diagram depicted in Figure 1.

Our model further considers organizations to be constrained by fixed resources. Manager allocates resources into four streams: process administration, process creation, process pruning, and direct production. Allocation to administration is

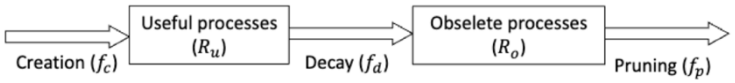


Figure 1. Diagram illustrating relationship between useful and obsolete processes.

determined by the number of processes, regardless of whether they are useful or obsolete. The manager's choice of how the remaining work is allocated determines the rate of processes created and pruned. We consider manager's decisions to be described by three parameters: creation propensity (γ_c), denoting the inclination of managers to create processes in response to problems, which is in turn reduced by useful processes; pruning propensity (γ_p), denoting the inclination to remove processes in response to the level of administrative costs; and desired level of direct production. The key outcome variables of the model are administrative costs (proportion of resources on administration), administrative waste (proportion of processes that are obsolete), and organization performance (quantified as a production function of direct production and useful processes).

Simulation and mathematical analysis of the model show two possible outcomes, illustrated in Figure 2. The first is a "sustainable equilibrium" scenario. The number of useful and obsolete processes both grow initially, but then stabilize to below what is allowed by resource capacity, and administrative waste stabilizes at a value less than one. The second is a "run-away administrative bloat" scenario, where the obsolete processes and administrative costs grow to the resource capacity, suggesting a failure of the organization.

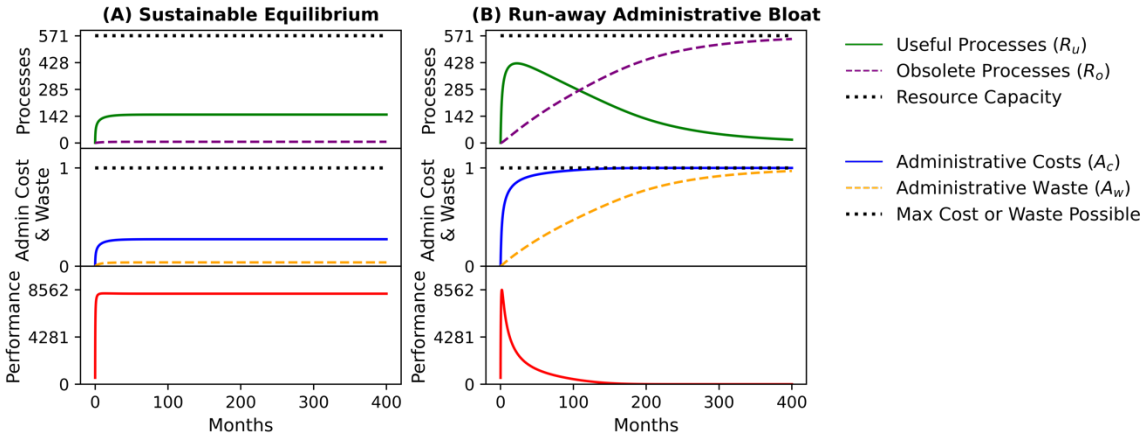


Figure 2. Simulations showing examples of two types of dynamic behavior (A) sustainable equilibrium outcome and (B) run-away administrative bloat outcome, where administrative costs and waste grow to capacity cap.

We further delineate conditions for the two outcomes, shown in Figure 3. Panel (A) indicates a sole sustainable equilibrium when pruning costs less than creation. Panel (B) shows that when pruning costs more, typical of most organizations, either outcome may occur. Orange denotes a dual-possibility area, while purple signifies runaway bloat. Exceeding a critical creation propensity threshold leads to unavoidable bureaucratic growth, regardless of process pruning efforts.

Further analysis of our model shows the surprising result that increased production desires can lead to greater administrative waste. We also further relaxed the model's assumptions and considered managers can observe all available variables and optimize for organizational performance. We find that even in this ideal condition, those that are short-sighted, reflected by a high discount rate for future performance, can still lead to run-away bloat.

We conclude that administrative bloat does not require ill-meaning bureaucrats accumulating power, or the demands of external regulations. Rather, it can arise through the endogenous dynamics of an organization with well-meaning managers.

Our results underscore the significant role of the relative costs of creating versus pruning processes. Notably, the effects of creation and pruning propensities of the managers are asymmetric in their consequences---excessive creation propensity can trigger an irreversible bloat cycle, outweighing the benefits of increased pruning efforts.

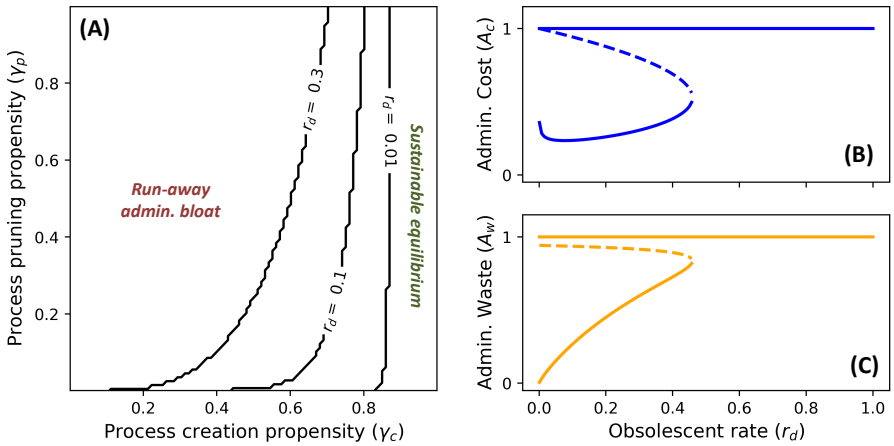


Figure 3. Para(A) The parameter regions leading to sustainable equilibrium (right of the frontier) and run-away administrative bloat (left of the frontier) outcomes. Process obsolescent rate (r_d) changes the frontier that distinguishes the two kinds of outcomes. (B, C) Equilibrium values for administrative cost and waste as a function of obsolescent rate. Solid lines denote stable equilibria. Dashed lines denote unstable equilibria.

REFERENCES

Boehner, J. A., & McKeon, H. P. (2003). The College Cost Crisis: A Congressional Analysis of College Costs and Implications for America's Higher Education System.

Gibbons, R., & Henderson, R. (2012). What do managers do?: Exploring persistent performance differences among seemingly similar enterprises. *The Handbook of Organizational Economics*. (pp. 680-731)

Himmelstein, D. U., Campbell, T., & Woolhandler, S. (2020). Health care administrative costs in the United States and Canada, 2017. *Annals of Internal Medicine*, 172 (2), 134–142.

Johnson, J. D. (2020). *Administrative Bloat in Higher Education*. Cambridge Scholars Publishing.

Kaufman, H. (1976). *Are Government Organizations Immortal*? Brookings Inst Pr.

Mumper, M., & Freeman, M. L. (2005). The causes and consequences of public college tuition inflation. In *Higher education: Handbook of Theory and Research* (pp. 307–361). Springer.

Nelson, R. R., & Winter, S. G. (1982). The Schumpeterian tradeoff revisited. *The American Economic Review*, 72 (1), 114–132.

Weber, M. (1968). *Economy and Society: An Outline of Interpretive Sociology*. Bedminster Press

Towards unifying the understanding of social learning's impact on collective decisions

Vicky Chuqiao Yang
Massachusetts Institute of Technology,
Sloan School of Management
Cambridge, Massachusetts
vcyang@mit.edu

Levi Grenier
Massachusetts Institute of Technology,
Sloan School of Management
Cambridge, Massachusetts
grenier@mit.edu

KEYWORDS

Social Learning, Bimodality, Binary Choice, Opinion Cascades

ACM Reference Format:
Vicky Chuqiao Yang and Levi Grenier. 2024. Towards unifying the understanding of social learning's impact on collective decisions. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

Collective decisions are central to human societies, from hiring committees to democratic elections. A central concern regarding these systems is the effect of social learning, where individuals adopt other people's opinions or behaviors rather than evaluating the merit of available options on their own.

Experimental studies in the collective intelligence community found mixed effects of social learning on decision-making, with some studies (e.g., [8]) finding it hurt collective decision outcomes, while others (e.g., [7, 10]) concluding it helps. Further studies argue that it depends on a myriad of factors, including network structure and adaptability [2, 4], initial estimates [3], and whether the outcome metric is individual or group accuracy [8]. The absence of a unified framework to reconcile these divergent findings remains a significant challenge in the field of collective intelligence [1].

Unlike empirical studies, theoretical models offer clearer insights. Models coming from different disciplines, employing diverse modeling methods, and considering various interaction networks, converged on a consistent finding: In the absence of social influence, the proportion of individuals achieving the correct answer in a binary choice follows a unimodal distribution; however, with pronounced social influence, two equilibria occurs, and this distribution should shift to bimodal [6, 7, 9, 12].

The experimental literature fails to capture the understanding of this fundamental shift. Typically, experiments assess significant differences in distribution means between the independent and influence conditions. Figure 1 illustrates this discrepancy between the implicit mental model of experiments and findings from theoretical work. Under the bimodal shift, the mean of the influence condition may exceed or fall below that of the independent condition, contingent on each equilibrium's prevalence. Although theoretical insights offer hope for reconciling conflicting results, there remains a gap between theory and empirical studies.

The disconnect between theory and experiments are two fold. On the one hand, experiments are focused on the one-dimensional categorization of outcomes into more or less accurate. Theoretical work suggests the problem is a two-dimensional one—while social learning changes the distribution of choices within a group, it also increases the unpredictability of the outcome, affecting a second dimension. On the other hand, theortical studies typically focus on reporting the existence of multiple equilibria, but do not investigate how prevalent are, thus disconnecting from core questions asked by experimentalists.

Here, we take the first step in unifying theory and experiments in how social learning affect collective decision outcomes. We first re-analyze experimental data from three papers on binary choices to test the prediction of binary shift and find support. Then, motivated by theoratical findings, we suggest a reformulation of how to consider social learning's effect in binary choice tasks.

2 RESULTS

Our analysis of three existing experiments supports the hypothesis of bimodal shift. Figure 2 visualizes the data in the independent and social influence conditions of three experiments involving binary choices [5, 8, 11]. The overlaying curves show fits to the best-fitting truncated Gaussian (truncated at 0 and 1, since by definition, proportion choosing one option is bounded between 0 and 1) as unimodal distributions, and the sum of two truncated Gaussians (also truncated at 0 and 1), as bi-modal distributions.

We quantitatively evaluate whether unimodal or bimodal distributions are a better descriptions of the data in each condition by comparing the Akaike and Bayesian Information Critiria (AIC and BIC) between the two fits. Both AIC and BIC evaluates how likely to observe data given model, while penalizing for additional model parameters. Motivated form theories' predictions, we expect unimodal distributions to be a better fit in the independent condition. Using AIC_1 and BIC_1 to denote the AIC and BIC for unimodal fit, and AIC_2 and BIC_2 to denote those for the bimodal fit, we quantitatively formulate the prediction for bi-modal shift. Since better fit to the data is associated with lower AIC and BIC values, we expect $AIC_1 - AIC_2$ and $BIC_1 - BIC_2$to be negative for the independent condition and positive for the influence condition. Conversely, we expect the same metrics to be positive for influence condition. Table 1 shows the summary statistics for using these two metrics to evaluate data of three experiments, each involving two conditions, our hypothesis is confirmed in 11 out of 12 instances, giving support to the bi-modal shift hypothesis.

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

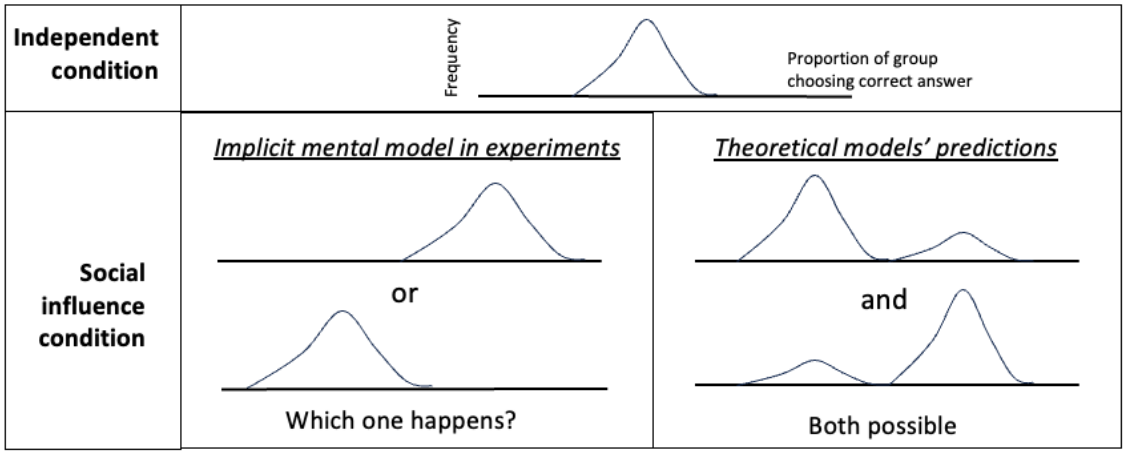


Figure 1: Conceptual illustration for the implicit mental model in experiments regarding the effect of social learning and how it differs from theoretical models' predictions.

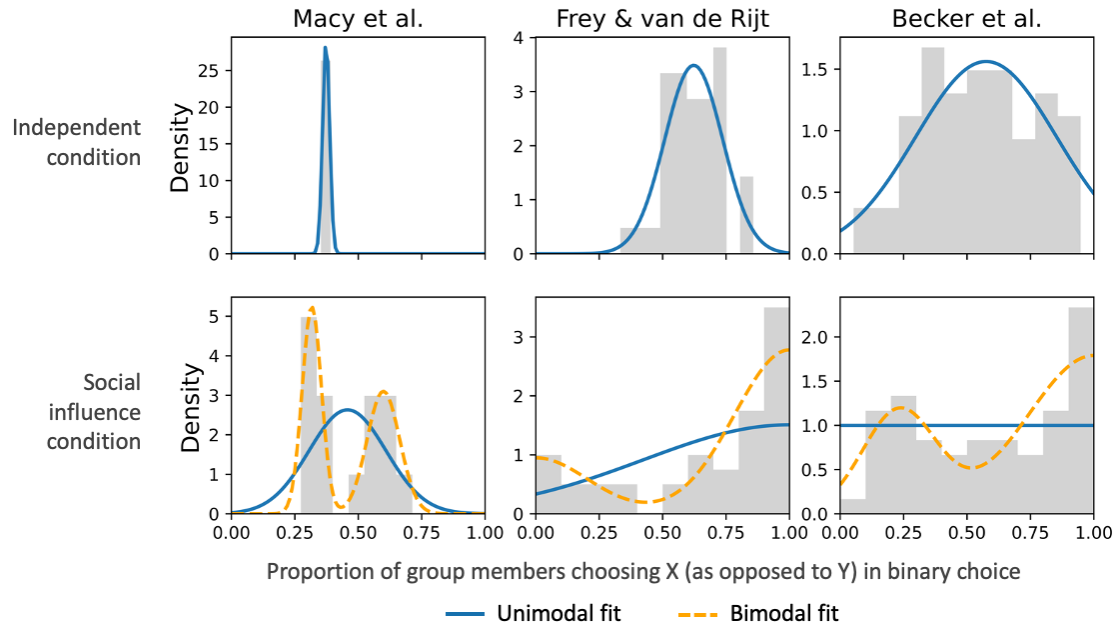


Figure 2: Analysis of the distribution of data from three previous experiments [5, 8, 11]. Distribution shifts from unimodal (independence condition) to bi-modal (influence condition), confirming theories prediction.

3 DISCUSSION

Our work suggests a potential change in the fundamental question the collective intelligence community asks about social learning. The current dialogue is centered around whether it makes collective outcomes "better" or "worse." Instead, we think the consideration should be two-dimensional —social influence tends to boost the size of the majority while also increasing the risk of the majority choosing the wrong option, introducing a fundamental trade-off between group coherence and outcome predictability. The implications vary by context: positively, a larger majority can facilitate action where indecision prevails, as seen in legislative processes where proposals, all better than the status quo, falter due to insufficient agreement on a single proposal (such as in US Congress' attempt for immigration reform). Conversely, in situations where wrong decisions carry high costs (such as in airplane design), social influence becomes detrimental.

Study	Metric	Indep. Condition	Influ. Condition
Macy et al.	$AIC_1 - AIC_2$	-6.00	7.49
	$BIC_1 - BIC_2$	-4.16	5.17
Frey & van de Rijt	$AIC_1 - AIC_2$	-6.00	10.63
	$BIC_1 - BIC_2$	-11.07	5.56
Becker, Guilbeault, & Smith	$AIC_1 - AIC_2$	-3.73	3.77
	$BIC_1 - BIC_2$	-10.01	-2.51

Table 1: Model evaluation metrics comparing unimodal and bimodal fits to experimental data. Theories predict $AIC_1 - AIC_2$ and $BIC_1 - BIC_2$ to be negative for the independent condition, and positive for the influence condition. Conversely, the same metrics are predicted to be positive for influence condition.

To theorists, our work also suggests that it can be important to focus on the prevalence of the equilibrium and what determines them, which can lead to a unified, mechanistic, and predictive science for conditions leading to desirable and undesirable outcomes.

Our work can have implications for how to better design discrete choice process in broad contexts, including in hiring committees, strategy decisions, all the way to national elections.

REFERENCES

[1] Abdullah Almaatouq, Thomas L Griffiths, Jordan W Suchow, Mark E Whiting, James Evans, and Duncan J Watts. 2022. Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences* (2022), 1–55.

[2] Abdullah Almaatouq, Alejandro Noriega-Campero, Abdulrahman Alotaibi, PM Krafft, Mehdi Moussaid, and Alex Pentland. 2020. Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences* 117, 21 (2020), 11379–11386.

[3] Abdullah Almaatouq, M Amin Rahimian, Jason W Burton, and Abdulla Alhajri. 2022. The distribution of initial estimates moderates the effect of social influence on the wisdom of the crowd. *Scientific Reports* 12, 1 (2022), 16546.

[4] Joshua Becker, Devon Brackbill, and Damon Centola. 2017. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences* 114, 26 (2017), E5070–E5076.

[5] Joshua Aaron Becker, Douglas Guilbeault, and Edward Bishop Smith. 2022. The crowd classification problem: Social dynamics of binary-choice accuracy. *Management Science* 68, 5 (2022), 3949–3965.

[6] William A Brock and Steven N Durlauf. 2001. Discrete choice with social interactions. *The Review of Economic Studies* 68, 2 (2001), 235–260.

[7] Iain D Couzin, Christos C Ioannou, Güven Demirel, Thilo Gross, Colin J Torney, Andrew Hartnett, Larissa Conradt, Simon A Levin, and Naomi E Leonard. 2011. Uninformed individuals promote democratic consensus in animal groups. *science* 334, 6062 (2011), 1578–1580.

[8] Vincenz Frey and Arnout Van de Rijt. 2021. Social influence undermines the wisdom of the crowd in sequential decision making. *Management science* 67, 7 (2021), 4273–4286.

[9] Rebecca Gray, Alessio Franci, Vaibhav Srivastava, and Naomi Ehrich Leonard. 2018. Multiagent decision-making dynamics inspired by honeybees. *IEEE Transactions on Control of Network Systems* 5, 2 (2018), 793–806.

[10] Bertrand Jayles, Hye-rin Kim, Ramón Escobedo, Stéphane Cezera, Adrien Blanchet, Tatsuya Kameda, Clément Sire, and Guy Theraulaz. 2017. How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences* 114, 47 (2017), 12620–12625.

[11] Michael Macy, Sebastian Deri, Alexander Ruch, and Natalie Tong. 2019. Opinion cascades and the unpredictability of partisan polarization. *Science advances* 5, 8 (2019), eaax0754.

[12] Vicky Chuqiao Yang, Mirta Galesic, Harvey McGuinness, and Ani Harutyunyan. 2021. Dynamical system model predicts when social learners impair collective performance. *Proceedings of the National Academy of Sciences* 118, 35 (2021), e2106292118.

Evaluating News Domain Quality through a Wisdom of Expert Crowds Approach

Hause Lin
hauselin@gmail.com
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Jana Lasser
jana.lasser@tugraz.at
Graz University of Technology
Graz, Styria, Austria

Stephan Lewandowsky
stephan.lewandowsky@bristol.ac.uk
University of Bristol
Bristol, Bristol, United Kingdom

Rocky Cole
rtcole@mit.edu
Jigsaw (Google LLC)
Mountain View, California, USA

Andrew Gully
andrewgully@google.com
Jigsaw (Google LLC)
Mountain View, California, USA

David G Rand
rand.dave@gmail.com
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Gordon Pennycook
grpennycook@gmail.com
Cornell University
Ithaca, New York, USA

ABSTRACT

One widely used approach for quantifying misinformation consumption and sharing is to evaluate the quality of the news domains that a user interacts with. However, different media organizations and fact-checkers have produced different sets of news domain quality ratings, raising questions about the reliability of these ratings. In this study, we compared six sets of expert ratings and found that they generally correlated highly with one another. We then created a comprehensive set of domain ratings for use by the research community (https://github.com/hauselin/domain-quality-ratings and https://www.domain-quality.info), leveraging an ensemble "wisdom of experts" approach. To do so, we performed imputation together with principal component analysis to generate a set of aggregate ratings. The resulting rating set comprises 11,520 domains—the most extensive coverage to date—and correlates well with other rating sets that have more limited coverage. Together, these results suggest that experts generally agree on the relative quality of news domains, and the aggregate ratings that we generate offer a powerful research tool for evaluating the quality of news consumed or shared and the efficacy of misinformation interventions.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Human-centered computing** → *Collaborative and social computing design and evaluation methods*; • **Computing methodologies** → *Machine learning*; • **Information systems** → Information retrieval.

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

KEYWORDS

News Quality, Wisdom Of Crowd, Fact-Checking, Misinformation, Journalism

ACM Reference Format:
Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. 2024. Evaluating News Domain Quality through a Wisdom of Expert Crowds Approach. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

Misinformation is not new, but concern over the problem has grown in recent years among researchers, policymakers, and technology companies. Despite growing interest in the problem, most research has skirted around fundamental measurement problems that are crucial to scientific progress and policy development [4, 6]. Here, we focus on the problem of assessing news domain quality. We compare different quality measures, evaluate the correspondence between these measures, and propose an aggregate measure of domain quality.

A common measurement challenge in studies of misinformation is quantifying information quality. Many studies have taken the approach of classifying posts flagged by third-party fact-checkers as misinformation [6, 8, 11, 12, 14, 15]. Fact-checkers, however, are only able to evaluate a small fraction of the content posted online every day, and thus this approach is not scalable.However, different media organizations and fact-checkers evaluate domains along different, and sometimes non-overlapping, dimensions (e.g., bias, factuality, transparency), raising questions about the degree of agreement across different expert ratings. To the extent that different rating systems disagree, these differences create a substantial measurement problem when using domain-level quality ratings.

To address the scalability problem, researchers from different fields have used evaluations of the quality of news sites or domains as a whole, instead of the individual pieces of content found on these sites [1, 3, 5, 8, 9, 16]. This domain-level approach is based on the

premise that journalistic standards differ systematically across news domains, such that higher quality domains (e.g., Reuters) generally produce higher quality content than lower quality domains (e.g., Infowars).

2 PROPOSED WISDOM OF EXPERT CROWDS APPROACH

2.1 Dataset and Descriptive Statistics

We address this measurement problem by first comparing different expert ratings and showing different expert ratings do generally agree on the relative quality of domains. We then leveraged a "wisdom of expert crowds" or "ensemble model" approach to create a more comprehensive aggregate rating set that researchers can use. Our rating set includes 11,520 domains, spanning a comprehensive range from high quality (e.g., Reuters, Associated Press) to low quality (e.g., Natural News, Before It's News).

These quality ratings come from 6 diverse expert sources that included media organizations and independent professional fact-checkers (Figure 1). These 6 expert sources used different scoring systems and evaluated domains on different dimensions (e.g, trust, reliability, (un)biasedness, credibility, transparency). The number of domains with ratings varied across the 6 sources (Figure 1A), which included many non-overlapping domains (Figure 1B). For example, 60 domains have been rated by independent professional fact-checkers [13], 4,767 by [7], and 8,178 by NewsGuard (as of September 13, 2022). Of the domains rated by NewsGuard, 44 have also been rated by fact-checkers (60 rated domains), and 264 rated by Ad Fontes Media (AFM; 283 rated domains), and 1,691 rated by Media Bias/Fact-Check (MBFC; 3216 rated domains).

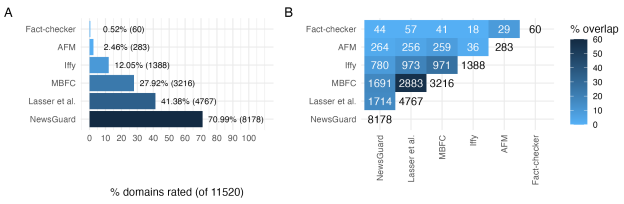


Figure 1: A) Percentage and number of domains (of 11,520) with ratings for 6 expert sources. B) Overlaps in rated domains between sources. The diagonal white cells indicate the total number domains with ratings for a given source. The off-diagonal elements indicate, for a given pair of sources, the number of domains that have been rated by both sources. Percent overlap and shading reflect the intersection of domains rated by a given pair of sources divided by the union of domains rated by the same pair of sources. Professional fact-checkers from [13], Ad Fontes Media (AFM), Iffy index of unreliable sources (Iffy), Media Bias/Fact-Check (MBFC), [7], and NewsGuard.

Figure 2A shows the different distributions of domain ratings for the 6 sources. The differences in distributions highlight a critical measurement issue researchers face when defining and quantifying misinformation using domain ratings: The prevalence of misinformation is likely to vary considerably depending on the set of ratings used and the quality threshold used to define a domain as low quality [2]. Despite the different distributions, however, the correlations in Figure 2A suggest substantial agreement across the 6 sources in the relative quality of the various domains (Pearson correlation coefficient r range: 0.32 to 0.86; Figure 2A top triangle).

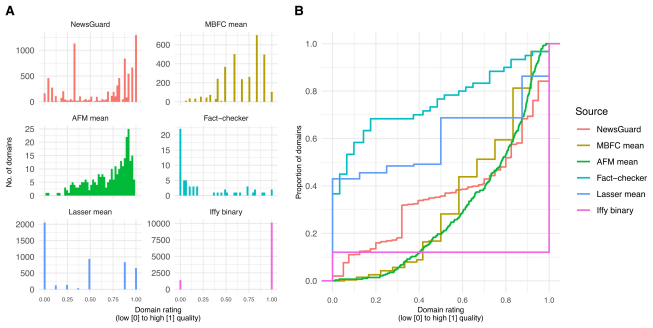


Figure 2: A) Distributions of ratings where each panel is one source. B) Empirical cumulative distribution functions, P(rating ≤ x)—the proportion of domains that is equal to or less than a given rating x. Iffy, Iffy index of unreliable sources; Lasser, [7] mean.

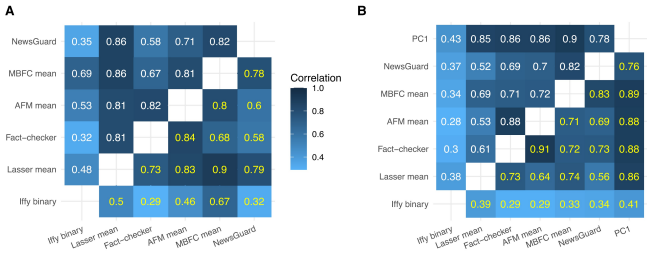


Figure 3: A) Raw (non-imputed) domain rating correlations. B) Imputed domain rating correlations. Values in the upper triangle are Pearson's r and values in the lower triangle are Spearman's ρ. Iffy, Iffy index of unreliable sources; Lasser, Lasser, [7] mean.

2.2 Imputation and Dimension Reduction

The relatively strong correlations observed in Figure 3A suggest there is general agreement on the relative ranking of domains across the 6 sources (even though they use different scoring systems and criteria), and that domain quality might be a latent variable that could be recovered by aggregating the data. To do so, we combine imputation with principal component analysis (PCA): Imputation is necessary because there are missing values (i.e., there are different numbers of rated domains per source; Figure 1), and is suitable given the correlations across the sources; a PCA performed on the imputed data can produce a component that serves as an aggregate rating.

The first principal component (PC1) of the PCA performed on the imputed dataset explains 68.21% of the variance in the data and the distribution of PC1 ratings resembles the aggregate of the other

sources. The PC1 ratings span a wide range from low (0) to high (1) quality. Example lower quality domains include Natural News (0.00), Gateway Pundit (0.11), and Daily Kos (0.41); medium quality domains include Daily Caller (0.47) and MSNBC (0.59); and higher quality domains include Reuters (1.00), Chicago Tribune (0.87), New York Times (0.86). Importantly, the PC1 ratings correlate highly with the other imputed ratings (Figure 2B).

3 CONCLUSION

By showing agreement across different expert ratings for the first time, our results provide evidence for the utility of fact-checked domain ratings for evaluating content quality. Prior studies have found less agreement, because they either defined agreement as the number of overlapping domains between different fake news lists [2] or included less credible and non-expert ratings [10]. Our work not only addresses these limitations, but also combines imputation with dimension reduction to produce a large new set of aggregate ratings that address previous problems (e.g., which fake news list should one use). Importantly, the comprehensive set of continuous ratings we provide could encourage more nuanced and realistic questions like "what is the quality of a domain" than binary questions like "is this domain a fake news producer or not."

In summary, our results suggest that there is substantial agreement across different sources of domain quality ratings, and that aggregated domain ratings provide a useful tool for advancing misinformation research. Domain ratings may not be as accurate as fact-checking individual pieces of content, but they offer a convenient tool for evaluating the efficacy of anti-misinformation interventions. When used alongside other approaches like misinformation detection algorithms, professional fact-checkers, digital literacy education, and crowdsourced fact-checking, domain ratings can be a valuable tool for combating and studying misinformation. Given the current lack of consensus across researchers on which rating set to use, and the limited coverage of some widely used rating sets, our aggregate set of ratings addresses important measurement problems in the field, which will result in better tools for quantifying misinformation, evaluating interventions, and facilitating theory development [4].

REFERENCES

[1] Saumya Bhadani, Shun Yamaya, Alessandro Flammini, Filippo Menczer, Giovanni Luca Ciampaglia, and Brendan Nyhan. 2022. Political audience diversity and news reliability in algorithmic ranking. *Nature Human Behaviour* (2022).

[2] Lia Bozarth, Aparajita Saraf, and Ceren Budak. 2020. Higher ground? How groundtruth labeling impacts our understanding of fake news about the 2016 US presidential nominees. Proceedings of the International AAAI Conference on Web and Social Media 14 (2020), 48–59.

[3] David Broniatowski, Kevin T Greene, Nilima Pisharody, Daniel Rogers, and Jacob N Shapiro. 2023. Measuring the monetization of online discourse with application to pro-and anti-vaccine communities. *OSF Preprints* (2023), 1–19.

[4] Hasok Chang. 2004. *Inventing temperature: Measurement and scientific progress.* Oxford University Press, New York, USA.

[5] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. Presidential Election. *Science* 363, 6425 (2019), 374–378.

[6] Brian Guay, Adam J. Berinsky, Gordon Pennycook, and David Rand. 2023. How to think about whether misinformation interventions work. *Nature Human Behaviour* 7, 8 (2023), 1231–1233.

[7] Jana Lasser, Segun Taofeek Aroyehun, Almog Simchon, Fabio Carrella, David Garcia, and Stephan Lewandowsky. 2022. Social media sharing of low-quality news sources by political elites. *PNAS Nexus* 2, 5 (2022), pgad158.

[8] Hause Lin, Haritz Garro, Nils Wernerfelt, Jesse Conan Shore, Adam Hughes, Daniel Deisenroth, Nathaniel Barr, Adam Berinsky, Dean Eckles, Gordon Pennycook, and David Rand. 2024. Reducing misinformation sharing at scale using digital accuracy prompt ads. *PsyArxiv* (2024).

[9] Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G. Rand, and Gordon Pennycook. 2023. High level of correspondence across different news domain quality rating sets. *PNAS Nexus* 2, 9 (2023), 1–8.

[10] Martino Mensio and Harith Alani. 2019. News source credibility in the eyes of different assessors. Proceedings of the Conference for Truth and Trust Online 2019 (2019), 1–19.

[11] Sakari Nieminen and Lauri Rapeli. 2019. Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review* 17, 3 (2019), 296–309.

[12] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

[13] Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[14] Ethan Porter and Thomas J Wood. 2021. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences* 118, 37 (2021), e2104235118.

[15] Soroush Vosoughi, Deb Roy, and Sina Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[16] Kai-Cheng Yang and Filippo Menczer. 2023. Large language models can rate news outlet credibility. *arXiv* (2023), 2304.00228v1.

Received 1 March 2024; revised 30 Apr 2024; accepted 5 Jun 2024

Memory Mediates the Effect of Network Structure on Collective Search

JULIAN GULLETT, Network Science Institute, Northeastern University, USA
ZACHARY FULKER, Network Science Institute, Northeastern University, USA
CHRISTOPH RIEDL, D'Amore-McKim School of Business & Khoury College of Computer Sciences, Northeastern University, USA

ACM Reference Format:
Julian Gullett, Zachary Fulker, and Christoph Riedl. 2024. Memory Mediates the Effect of Network Structure on Collective Search. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Groups communicate and share information to solve problems together. Seminal theoretical work on collaborative problem solving emphasized the important role of the network structure linking the individuals of the collective together [14]. Network structure shapes how efficiently information can diffuse, which impacts collective performance by affecting the balance of exploitation and exploration [4, 8, 10, 18]. Some theoretical work suggests fast, efficient information diffusion leads to premature convergence on sub-optimal solutions. Consequently—and somewhat counter intuitively—inefficient networks lead to higher collective performance. Inefficient networks slow the diffusion of local maxima, thus giving agents time to explore diverse solutions [14, 18]. Empirical studies with human subjects, on the other hand, failed to reproduce this finding; instead showing efficient networks produce superior collective outcomes. One possible explanation for these conflicting findings may lie in behavioral differences between simulated agents and humans [19]. We theorize the key behavioral differences are (a) how agents explore (i.e., their search strategy) and (b) how they learn from the solutions of others (i.e., the specific mechanism of social influence). To explore this theory, we introduce a new behavioral model of how agents search based on function learning capable of representing two features of human learning missing in prior theoretical work: (a) generalization of the search behavior and (b) social influence.

2 BACKGROUND

To study collective problem solving, we build on previous research of collective exploration on fitness landscapes. Fitness landscapes provide a representation of complex, multidimensional problems where "ruggedness" in the landscape corresponds to problem complexity [13]. To study networks of agents exploring these landscapes, theoretical models often rely on "myopic" behavioral heuristics such as hill-climbing (for "exploration") and directly copying neighbor solutions (for "exploitation"; Fig. 1a and Fig. 1c) [12, 14]. Human subjects experimentally outperform such myopic agents [19], leading to suggestions that this performance gap arises from insufficiently human-like models. In response, we propose a more human-like model of search behavior which incorporates two distinct human capacities—generalization and social learning—which are not well captured by myopic agents. Generalization, the capacity to project learning from

Permission to make digital or hard copies of part or all 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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2024 Copyright held by the owner/author(s).
Manuscript submitted to ACM

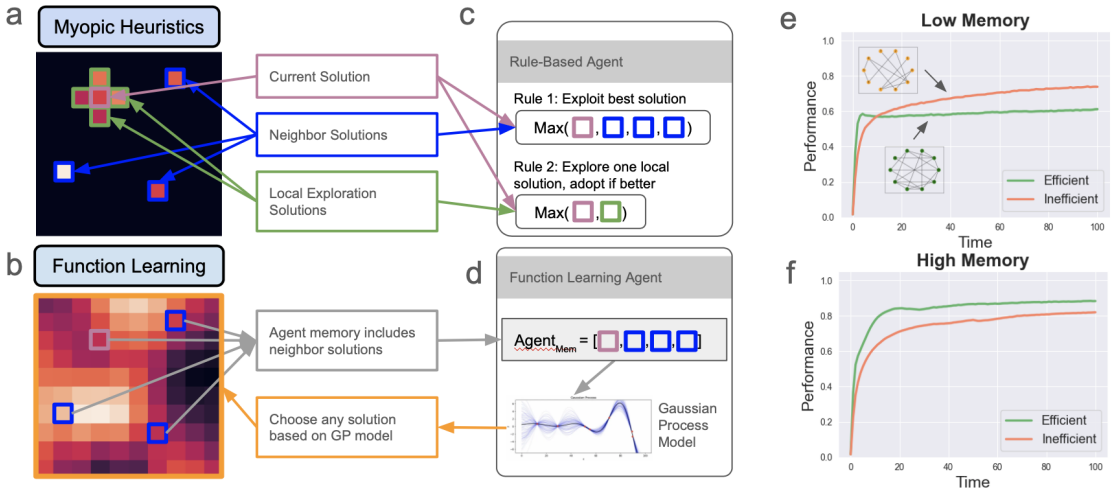


Fig. 1. **Memory explains difference between when efficient vs. inefficient networks perform better. a** Myopic agent model "vision". **b** Function learning agent model "vision". **c** Myopic agent model decision procedure **d** Function learning agent model decision procedure **e** Low memory agents perform better in inefficient networks. **f** High memory agents perform better in efficient networks.

its specific context into novel contexts, is the first component of our model. Humans are known to make inference-based generalizations about their environments, allowing purposeful, non-local search [5]. Social learning, the ability to observe the behavior of others and evaluate consequences that follow without having to imitate that behavior [1], is the second component. Notice how this combined approach suggests an entirely different model of social influence. Agents do not directly copy solutions of their neighbors but instead incorporate such information into their own mental models of the landscape. This addresses a crucial discussion in prior work around whether individuals copy neighbors directly, copy with noise, copy community consensus, how trust modifies copying, or employ some entirely different copying strategy [2, 14].

To represent the human capacity for both generalization and social learning, we adopt a function learning agent model to investigate collective search on fitness landscapes (Fig. 1b). Function learning has been experimentally shown to replicate human performance in search tasks [22]. With a function learning approach, agents approximate a value function over the entire problem space that includes (estimates values for) all possible solutions. Gaussian global value functions assume smoothness, or correlation between the values of similar solutions. This internal model building process, and assumption of smoothness, mirrors how humans approach vast and uncertain problem spaces [11].

3 METHODS

Agents are placed into a network and share solution information with network neighbors. Each round, agents (1) use their memory of known solutions (either those that they explored themselves or those that others explored and shared with them) to estimate the value of each point of the landscape, (2) select their next search point based on its estimated value, and (3) synchronously update their memory of known solutions with their own search point and the search points of their network neighbors. In the first step, agents construct an internal model of the landscape using Gaussian process function learning with each agent generating a distinct internal model dependent on the solution

data available to them. In the second step, each agent applies a softmax function to the expected solution values in their internal model before selecting the next solution to "search". Notice that this step integrates both "exploration" and "exploitation". In the third step, agents learn the true value of their chosen solution and exchange solution information with their network neighbors. Agents have a shifting, first-in-last-out memory to retain the most recent information of their own prior and neighbors' prior searches (Fig. 1d). We study networks of 10 agents searching for 100 rounds on randomly generated fitness landscapes. Each parameter combination (i.e. [low memory, efficient network]) is tested by 25 randomly generated networks with each network exploring 100 randomly generated landscapes (2,500 repetitions).

We theorize that memory—the amount of information (i.e., search points) agents use to refine their mental models—is the crucial parameter explaining the difference between prior theoretical results and empirical results from human subject studies. Simulation models rely on "myopic" agents that do not have access to a historical record of previously tested solutions [14], while human subject experiments have often provided a complete record of previously tested solutions [19]. Our simulation study compares two types of agents: those with low memory ($mem = 10$, capacity to hold about 0.0003% of possible solutions) and those with high memory ($mem = 100$, 0.003%). Low efficiency networks are represented by random graphs with average density of 0.1. High efficiency networks are represented by random graphs with average density 0.3.

4 RESULTS

We find function learning agents with low memory ($mem = 10$) qualitatively reproduce results from prior theoretical work, showing inefficient networks lead to superior collective outcomes (Fig. 1e). Function learning agents initialized with high memory ($mem = 100$) show efficient networks produce better collective performance, as suggested by human-subject experiments (Fig. 1f). By design, the landscapes in our experiment are difficult for a single searcher to explore effectively (and agents do not typically find global best solution). Even a function learning agent, with its human-like qualities, struggles to perfectly solve sufficiently complex problems. For this reason, increased memory does not allow agents in inefficient networks to significantly improve their performance. In contrast, efficient communication hinders performance for low memory agents; they are forced to construct internal models from highly variable data. For agents in efficient networks, increasing memory does allow for a significant improvement in model quality and, subsequently, collective performance.

5 DISCUSSION

This work makes three important contributions. First, we introduce a function learning agent model explicitly incorporating two human-like characteristics, generalization and social influence, currently not represented in myopic agent models. Second, we demonstrate how this function learning agent model resolves conflicting findings between theoretical simulation studies and human-subject experiments. Third, we claim function learning agent models represent a generalizable, cognitively well-founded approach to understanding the dynamics and potential of collective search in humans. Function learning models are not constrained by domain-specific interest in specific behavioral patterns, and can be adapted to incorporate concepts from many fields of research. Function learning agent models readily integrate well-established concepts including memory [15, 24], trust [7, 20], curiosity [16, 21], search costs [9], communication synchronicity [23], perspective diversity [3], transcription error [17], and absorptive capacity [6]. In addition to this compatibility with existing theoretical concepts, applying function learning agent models in contexts of dynamic networks and heterogeneous adaptive agents promises the most realistic simulations of collective search in humans. Future work should continue exploring the potential of function learning agent models in increasingly realistic contexts.

REFERENCES

[1] Albert Bandura. 1977. *Social learning theory.* Prentice-Hall, Englewood Cliffs, New Jersey.

[2] Daniel Barkoczi and Mirta Galesic. 2016. Social learning strategies modify the effect of network structure on group performance. *Nature Communications* 7 (Oct. 2016). https://doi.org/10.1038/ncomms13109 Publisher: Nature Publishing Group.

[3] Fabian Baumann, Agnieszka Czaplicka, and Iyad Rahwan. 2024. Network structure shapes the impact of diversity in collective learning. *Scientific Reports* 14, 1 (Jan. 2024), 2491. https://doi.org/10.1038/s41598-024-52837-3

[4] Oliver Baumann, Jens Schmidt, and Nils Stieglitz. 2019. Effective Search in Rugged Performance Landscapes: A Review and Outlook. *Journal of Management* 45, 1 (Jan. 2019), 285–318. https://doi.org/10.1177/0149206318808594 Publisher: SAGE Publications Inc..

[5] Ali Borji and Laurent Itti. 2013. Bayesian optimization explains human active search. In *Advances in Neural Information Processing Systems*, C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/a3f390d88e4c41f2747bfa2f1b5f87db-Paper.pdf

[6] Elias G. Carayannis, Mike Provance, and Evangelos Grigoroudis. 2016. Entrepreneurship ecosystems: an agent-based simulation approach. *The Journal of Technology Transfer* 41, 3 (June 2016), 631–653. https://doi.org/10.1007/s10961-016-9466-7

[7] Shu-Heng Chen, Bin-Tzong Chie, and Tong Zhang. 2015. Network-Based Trust Games: An Agent-Based Model. *Journal of Artificial Societies and Social Simulation* 18, 3 (2015), 5. https://doi.org/10.18564/jasss.2767

[8] Christina Fang, Jeho Lee, and Melissa A. Schilling. 2010. Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning. *Organization Science* 21, 3 (May 2010), 625–642. https://doi.org/10.1287/orsc.1090.0468

[9] Liya Guo, Shan Huang, and Adel W. Sadek. 2013. A novel agent-based transportation model of a university campus with application to quantifying the environmental cost of parking search. *Transportation Research Part A: Policy and Practice* 50 (April 2013), 86–104. https://doi.org/10.1016/j.tra.2013.01.045

[10] Thomas T. Hills, Peter M. Todd, David Lazer, A. David Redish, Iain D. Couzin, Melissa Bateson, Roshan Cools, Reuven Dukas, Luc Alain Giraldeau, Michael W. Macy, Scott E. Page, Richard M. Shiffrin, David W. Stephens, and Jeremy W. Wolfe. 2015. Exploration versus exploitation in space, mind, and society. https://doi.org/10.1016/j.tics.2014.10.004 ISSN: 1879307X Issue: 1 Pages: 46-54 Publication Title: Trends in Cognitive Sciences Volume: 19.

[11] Philip N. Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences* 107, 43 (Oct. 2010), 18243–18250. https://doi.org/10.1073/pnas.1012933107

[12] Stuart Kauffman and Simon Levin. 1987. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology* 128, 1 (Sept. 1987), 11–45. https://doi.org/10.1016/S0022-5193(87)80029-2

[13] Stuart A. Kauffman and Edward D. Weinberger. 1989. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology* 141, 2 (Nov. 1989), 211–245. https://doi.org/10.1016/S0022-5193(89)80019-0

[14] David Lazer and Allan Friedman. 2007. The Network Structure of Exploration and Exploitation. *Administrative Science Quarterly* 52, 4 (Dec. 2007), 667–694. https://doi.org/10.2189/asqu.52.4.667

[15] Christian C. Luhmann and Suparna Rajaram. 2015. Memory Transmission in Small Groups and Large Networks: An Agent-Based Model. *Psychological Science* 26, 12 (Dec. 2015), 1909–1917. https://doi.org/10.1177/0956797615605798

[16] L. Macedo and A. Cardoso. 2012. The exploration of unknown environments populated with entities by a surprise–curiosity-based agent. *Cognitive Systems Research* 19-20 (Sept. 2012), 62–87. https://doi.org/10.1016/j.cogsys.2012.04.003

[17] Michael W. Macy and Robert Willer. 2002. From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology* 28, 1 (Aug. 2002), 143–166. https://doi.org/10.1146/annurev.soc.28.110601.141117

[18] James G. March. 1991. Exploration and Exploitation in Organizational Learning. https://www.jstor.org/stable/2634940?seq=1&cid=pdf- Issue: 1 Pages: 71-87 Volume: 2.

[19] Winter Mason and Duncan J. Watts. 2012. Collaborative learning in networks. *Proceedings of the National Academy of Sciences* 109, 3 (Jan. 2012), 764–769. https://doi.org/10.1073/pnas.1110069108 Publisher: Proceedings of the National Academy of Sciences.

[20] Ngoc Thanh Nguyen (Ed.). 2013. *Transactions on computational collective intelligence IX.* Number 7770 in Lecture notes in computer science. Springer, Berlin ; New York. OCLC: ocn839357401.

[21] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2017. A curiosity model for artificial agents. 561–568. https://doi.org/10.1162/isal_a_089

[22] Charley M. Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D. Nelson, and Björn Meder. 2018. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour* 2, 12 (Dec. 2018), 915–924. https://doi.org/10.1038/s41562-018-0467-4 Number: 12 Publisher: Nature Publishing Group.

[23] Longqi Yang, David Holtz, Sonia Jaffe, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, Brent Hecht, and Jaime Teevan. 2021. The effects of remote work on collaboration among information workers. *Nature Human Behaviour* 6, 1 (Sept. 2021), 43–54. https://doi.org/10.1038/s41562-021-01196-4

[24] Yu Zhang and Jason Leezer. 2010. Simulating human-like decisions in a memory-based agent model. *Computational and Mathematical Organization Theory* 16, 4 (Dec. 2010), 373–399. https://doi.org/10.1007/s10588-010-9077-z

Automated Social Science: A Structural Causal Model-Based Approach

KEHANG ZHU*, Harvard University, USA
BENJAMIN MANNING†, Massachusetts Institute of Technology, USA
JOHN J. HORTON, Massachusetts Institute of Technology, USA

CCS Concepts: • **Applied computing** → **Sociology**; **Economics**; **Psychology**; • **Computing methodologies** → **Natural language processing**.

Additional Key Words and Phrases: Large language model (LLM), automation, social science, structural causal model (SCM)

ACM Reference Format:
Kehang Zhu, Benjamin Manning, and John J. Horton. 2018. Automated Social Science: A Structural Causal Model-Based Approach. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 EXTENDED ABSTRACT

There is much work on efficiently estimating models but comparatively little work on efficiently generating those models to estimate. Previously, developing such models and hypotheses to test was exclusively a human task. This is changing as researchers have begun to explore automated hypothesis generation through the use of machine learning [9, 11, 13]. But even with novel machine-generated hypotheses, there is still the problem of testing. A potential solution is simulation. Researchers have shown that Large Language Models (LLM) can simulate humans as experimental subjects with surprising degrees of realism [1–5]. To the extent that these simulation results carry over to human subjects in out-of-sample tasks, they provide another option [7]. In this paper, we combine these ideas—automated hypothesis generation and automated *in silico* hypothesis testing—by using LLMs for both purposes. We demonstrate that such automation is possible. We evaluate the approach by comparing results to a setting where the real-world predictions are well known and test to see if an LLM can be used to generate information that it cannot access through direct elicitation. We demonstrate that such automation is possible. We evaluate the approach by comparing results to a setting where the real-world predictions are well known and test to see if an LLM can be used to generate information that it cannot access through direct elicitation.

A key innovation in our approach is the use of structural causal models (SCM) to organize the research process. SCMs are mathematical representations of cause and effect [12] and have long offered a language for expressing hypotheses. What is novel in our paper is the use of an SCM as a blueprint for the design of agents and experiments. In short, each explanatory variable describes something about a person or scenario that has to vary for the effect to be identified, so the system "knows" it needs to generate agents that vary on that dimension—a straightforward transition from stated theory to experimental design and data generation. Furthermore, the SCM offers a pre-specified plan of estimation [6, 8].

*Both authors contributed equally to this research.
†Both authors contributed equally to this research.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

We built an open-sourced computational system implementing the SCM-based approach. Figure 1 provides a high-level overview of the system. Starting from a social science "scenario," an LLM is queried to propose outcomes of interest and the relevant agents. For each outcome proposed, potential exogenous causes are again elicited from the LLM. For example, a scenario of interest might be "two people bargaining over a mug." The LLM may propose "whether a sale occurs" as an outcome of interest, with a buyer and a seller as the relevant agents. The LLM might hypothesize that the buyer's willingness to pay may affect the outcome. The completed SCM contains endogenous variables (the outcomes), exogenous variables (the potential causes), and paths (the possible effects of the causes). Researchers can raise the model's temperature to encourage divergent idea generation and are free to modify whatever hypotheses the LLM proposes—the system's output is editable at every step.

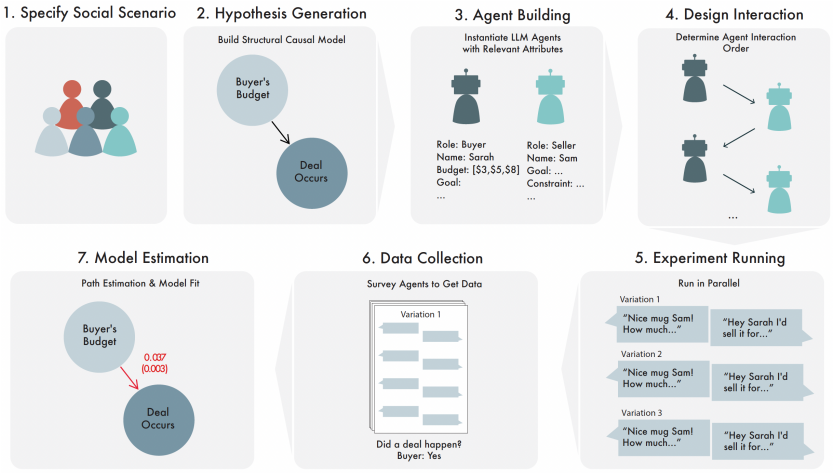


Fig. 1. An overview of the automated system. Each step in the process corresponds to an analogous step in the social scientific process as done by humans. The development of the hypothesis guides the experimental design, execution, and model estimation.

Agents are then generated that vary on the exogenous dimensions of the SCM. For example, in the bargaining scenario, each simulation will vary the buyer's willingness to pay. A separate LLM powers each agent—their conversations are simulated by the LLMs exchanging text. To measure outcomes, survey questions are formulated when generating this SCM (e.g., ask the buyer, "Did a sale occur?"), and those questions are administered to the agents after each simulation run. This is possible because LLM-powered agents can each be given a "memory"—a transcript of what happened. The answers to the survey questions are used as data to estimate paths of the fitted SCM. The SCM framework is useful for our approach because it describes exactly what needs to be measured as a downstream outcome subject to the manipulations of the causes.

We use this system to explore several social scenarios: (1) two people bargaining over a mug, (2) a lawyer interviewing for a job, (3) a bail hearing, and (4) an auction. The system uncovers several findings from the simulations. The probability of a deal increases as the seller's sentimental attachment to the mug decreases, and both the buyer's and the seller's reservation prices matter. The candidate passing the bar exam was the only important factor in her getting the job. A remorseful defendant was granted lower bail but was not so fortunate if his criminal history was extensive. An increase in the bidders' reservation prices causes an increase in the clearing price. These are not "counterintuitive" findings, but

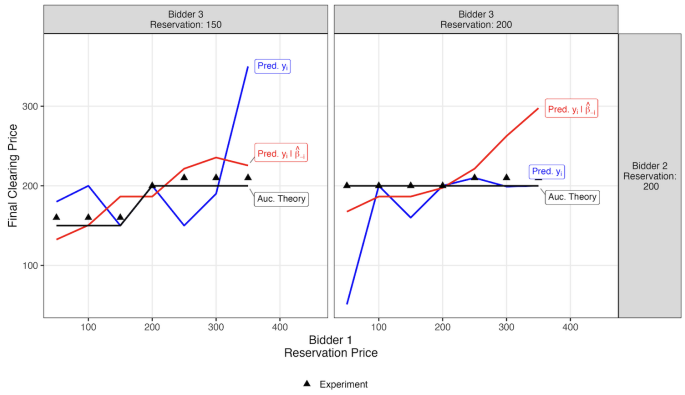


Fig. 2. Example comparison of the LLM's predictions to the theoretical predictions and experimental results for the auction scenario. The columns correspond to the different reservation values for bidder 3 in a given simulation, and the rows correspond to the different reservation values for bidder 2. The y-axis is the clearing price, and the x-axis lists bidder 1's reservation price. The black triangles track the observed clearing price in each simulated experiment, the black line shows the predictions made by auction theory ($MSE_{Theory} = 128$), the blue line indicates the LLM's predictions without the fitted SCM—the predict-y_i task ($MSE_{y_i} = 8627$), and the red curve is the LLM's predictions with the fitted SCM—the predict-$y_i|\hat{\beta}_{-i}$ task ($MSE_{y_i|\hat{\beta}_{-i}} = 1504$).

it is important to emphasize they were the result of empiricism, not just model introspection. But this does raise the question of whether a simulation is even necessary.

Instead of simulation, could an LLM simply do a "thought experiment" about the proposed *in silico* experiment and achieve the same insight? To show that the collective hebaviors are different, we describe the simulations that will be run and ask the LLM to predict the results—both the path estimates and point predictions. To make this concrete, suppose we had the simple linear model $y = \beta X$ to describe some scenario, and we ran an experiment to estimate $\hat{\beta}$. We describe the scenario and the experiment to the LLM and ask it to predict y_i given a particular X_i (a "predict-y_i" task). Separately, we ask it to predict $\hat{\beta}$ (a "predict-$\hat{\beta}$" task). Later, we examine how the LLM does on the predict-y_i task when it has access to the fitted SCM.

In the predict-y_i task, we prompt the LLM to predict the outcome y_i given each possible combination of the X_i's from the auction experiment. Figure 2 provides these predictions. The black triangles track the empirical results. The blue lines show that direct elicitation of the predictions for y_i in the auction experiment is wildly inaccurate ($MSE_{y_i} = 8628$). Interestingly and reassuringly, the predictions made by auction theory ($MSE_{Theory} = 128$) [10] are close to the empirical results—the black lines often match the *in silico* simulations exactly. In the predict-$\hat{\beta}$ task, the LLM is asked to predict the fitted SCM's path estimates for all four experiments, provided with contextual information about each scenario. Two-sided t-tests revealed that 11 out of 12 predictions significantly differed from the empirical estimates. On average, the LLM predicts the estimates will be 13 times larger than the experimental results.

We repeat the predict-y_i task, but this time, we provide the LLM with the experimental path estimates using leave-one-out cross-validation. For each X_i, we fit the SCM using all but the ith observation and then ask the LLM to predict y_i given X_i, and this fitted SCM. In this "predict-$y_i|\hat{\beta}_{-i}$" task (the red lines in Figure 2), the predictions are far more accurate than in the predict-y_i task without the fitted SCM ($MSE_{y_i|\hat{\beta}_{-i}} = 1504$). The mean squared error is lower by almost an order of magnitude, and the predictions are much closer to those made by the theory. In short, the LLM knows more than it can (immediately) tell.

REFERENCES

[1] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*. PMLR, 337–371.

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[3] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120. https://doi.org/10.1073/pnas.2218523120 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2218523120

[4] James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using GPT for Market Research. *Working paper* (2023).

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]

[6] Trygve Haavelmo. 1944. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society* (1944), iii–115.

[7] John J Horton. 2023. *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical Report. National Bureau of Economic Research.

[8] Karl G Joreskog. 1970. A general method for estimating a linear structural equation system. *ETS Research Bulletin Series* 1970, 2 (1970), i–41.

[9] Jens Ludwig and Sendhil Mullainathan. 2023. *Machine Learning as a Tool for Hypothesis Generation.* Working Paper 31017. National Bureau of Economic Research. https://doi.org/10.3386/w31017

[10] Eric S. Maskin and John G. Riley. 1985. Auction Theory with Private Values. *The American Economic Review* 75, 2 (1985), 150–155. http://www.jstor.org/stable/1805587

[11] Sendhil Mullainathan and Ashesh Rambachan. 2023. *From Predictive Algorithms to Automatic Generation of Anomalies.* Technical Report. https://doi.org/10.2139/ssrn.4443738 Available at: https://ssrn.com/abstract=4443738 or http://dx.doi.org/10.2139/ssrn.4443738.

[12] Judea Pearl. 2009. *Causality.* Cambridge university press.

[13] Joshua C. Peterson, David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. 2021. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* 372, 6547 (2021), 1209–1214. https://doi.org/10.1126/science.abe2629 arXiv:https://www.science.org/doi/pdf/10.1126/science.abe2629

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Leveraging Advice-taking and Kernel Density Estimation to Identify A Cluster of Experts and Improve Wisdom of Crowds

Yunhao (Jerry) Zhang*
zyhjerry@mit.edu
University of California, Berkeley
Berkeley, CA, USA

ABSTRACT

Wisdom of Crowds estimates (e.g., a group mean) can suffer when non-experts in the crowd have biased estimations. One solution is to aggregate over a subset of experts instead of the entire crowd– but systematically identifying the accurate individuals remains a challenge in a single-prediction-problem context. We introduce a novel algorithm to identify experts in the crowd by combining kernel density estimation and the behavioral metric "weight on advice" (WOA) to identify and then aggregate a distinct cluster of accurate estimates. This approach is based on two phenomena: First, accurate experts revise their initial estimates less upon seeing numerical advice compared to their non-expert counterparts; and second, while experts' estimates tend to cluster close to the truth, stubborn non-experts may have a wide range of wrong answers without forming their own cluster. Therefore, although relying solely on WOA or cluster size may not identify experts (due to the existence of stubborn non-experts or the largest cluster being formed by an overwhelming number of non-experts), we demonstrate that finding and aggregating over the cluster that has the least average WOA does, thereby improving crowd wisdom. Our theoretical framework not only demonstrates the algorithm's rational foundation, but also characterizes its properties accounting for the presence of agents who exhibit stubbornness and over-updating in advice-taking. Crucially, our empirical analysis, combined with open data from other publications examining various contexts, confirms that averaging over the (updated) judgment within our identified cluster outperforms averaging over the initial or updated judgment of the entire group.

KEYWORDS

Wisdom of Crowds, Advice-taking, Judgment Aggregation

ACM Reference Format:
Yunhao (Jerry) Zhang. 2024. Leveraging Advice-taking and Kernel Density Estimation to Identify A Cluster of Experts and Improve Wisdom of Crowds. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 4 pages.

*Full paper is available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4779145

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

1 INTRODUCTION

The Wisdom of Crowds (WoC) phenomenon – the average prediction of a group is more accurate than most or even the best individual predictions – serves as a vital forecasting and decision support tool. Mannes, Soll, Larrick (2014) shows that utilizing an average of expert opinions can further enhance the accuracy of WoC. However, identifying these experts remains challenging. In this paper, we introduce a method which leverages people's advice-taking behavior and kernel density estimation to identify a cluster of experts to improve WoC. We also present a theoretical characterization of the conditions under which our method proves successful. Furthermore, we validate our method using advice-taking data in various contexts published by other researchers, including a study in which professional physicians made medical evaluations.

2 THE ALGORITHM

We hereby outline the algorithm's implementation. It requires the collection of initial responses, advice, and updated responses after seeing numerical advice, based on which we analyze, aggregate, and enhance Wisdom of Crowds.

(1) Elicit participants' initial estimates (X) to a question

(2) Utilize Kernel Density Estimation[1] to identify the clusters of participants' initial estimates.

(3) [Cost-saving Screening Step] If at least two clusters – each with more than "k" observations (k = 5% of subjects as a rule-of-thumb) – emerge, proceed with the algorithm.
If initial answers only form a single cluster, the algorithm stops and takes the conventional wisdom-of-crowds estimate as the output, without incurring any cost of procuring the updated estimates.

(4) Present numerical advice Y to each participant. The numerical advice may be another participant's estimate or the average estimate of a group of participants.

(5) Elicit participants' updated answers (Z) after seeing the numerical advice.

(6) Calculate "weight on advice" (WOA): $\frac{|Z-X|}{|Y-X|}$

[1]The implementation may be easily accomplished using the $density$ function using the default Gaussian kernel with bandwidth selection (bandwidth="SJ") from a base package named $stats$ in R programming language. All other hyperparameters are set to their default values.

(7) Identify the "Expert Cluster" as the cluster with the lowest average WOA.

(8) Compute the mean of the (updated) estimates in the Expert Cluster.

2.1 Lowering the Cost of Algorithm Implementation

Procuring estimates may be considered costly. One critical advantage of our algorithm is its ability to self-determine when it should be applied. Our algorithm pinpoints the expert cluster only if one potentially exists (e.g., when the non-experts' initial estimates are dissimilar). This implies that the initial answers should form at least two sizable clusters – one by experts, the other one(s) by non-experts. Right after collecting the initial estimates, the decision maker (DM) can promptly and quite easily apply Kernel Density Estimation (KDE) to the initial estimates. If no distinct cluster emerges (e.g., only one cluster has more than k agents), this very likely implies that there is no heterogeneity of expertise among agents in the crowd, and there is no significant advantage to identify and aggregate over the experts. The DM can simply bypass the expense of obtaining updated responses and rely on the group mean. This adaptability empowers the DM to decide on eliciting updated responses based on the pattern of the initial answers – only when the algorithm's application yields substantial expected benefits.

3 INTUITION

We hereby explains the intuition of our algorithm. The classic "Wisdom of Crowds" theory (e.g., Condorcet's Jury Theorem, Median Voter Theorem, etc.) posits that agents' estimations are unbiased in aggregate and the truth often aligns with the majority's perspective. We illustrate why this existing benchmark method does not always work using an empirical example, in which we ask participants to estimate the auction price of a painting. First, in many cases, the majority of respondents often do not give the most accurate answer. For example, as illustrated in Figure 1 (left panel), the largest cluster, located to the far left, had the majority of respondents, but their response exhibited substantial underestimation. Thus, in this case, equating cluster size with expertise or judgment accuracy would lead to an erroneous identification of the expert cluster. Second, some non-experts are inaccurate but stubborn. In the same example, as illustrated in the right panel of Figure 1, among the individuals who did not alter their initial estimate upon receiving group mean as numerical advice (e.g., those highlighted in the circle), some were accurate (e.g., those around the dashed red line), but many were inaccurate yet stubborn non-experts (e.g., those with initial estimates below 200 but did not update). Hence, relying exclusively on WOA to discern expertise or to aggregate judgments is suboptimal.

By contrast, our algorithm, which utilizes a combination of cluster identification and advice-taking behavior, is able to reliably pinpoint the expert cluster and improve Wisdom of Crowds. The algorithm is based on a reliable empirical observation that experts, on average, tend to be more resistant to numerical advice compared to their less accurate counterparts [1–3, 5, 6, 8]. Our algorithm is also based on the observation that inherent stubbornness within

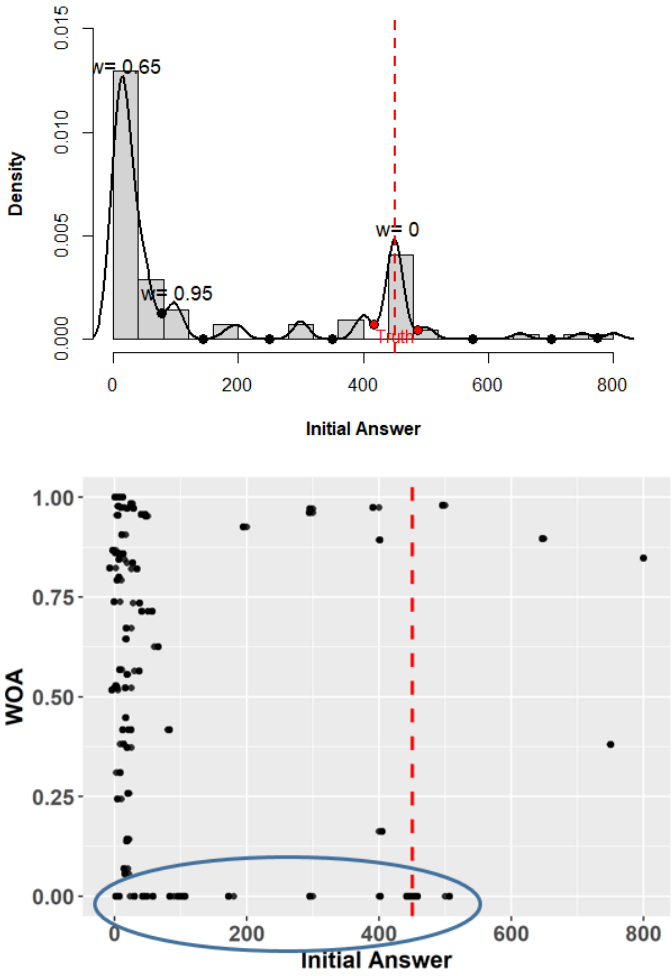


Figure 1: Example Question taken from Study 4: "What is the world record auction price of the painting 'Salvator Mundi' in 2017? (unit in million dollars)" In both panels, the dashed red line indicates the truth (450.3). The left panel is composed of a histogram, a density curve based on kernel density estimation that depicts the clusters (between two dots or "pits" is a cluster), the average WOA of each cluster containing at least 5% of the subjects (in this case, at least 5 subjects). The right panel plots the WOA associated with each individual initial estimate. The circle highlights those who with WOA of zero. For visualization, WOA is winsorized between 0 and 1, and one outlier is removed.

some non-experts typically does not correlate with the specific inaccuracies of their responses.[2] For example, respondents with answers like 5, 15, 100, 200, etc., may all display resistance to numerical advice (i.e., there are many different ways for stubborn

[2]This is further discussed in Study 4 where Republicans and Democrats answer political factual belief questions.

people to be wrong.) However, adaptable non-experts could make the same mistakes, making it unlikely for the stubborn non-experts to congregate into a singular "stubborn cluster". In contrast, experts often gravitate toward the correct answer, forming a distinct cluster with low WOA. Thus, the average WOA of the expert cluster is likely to be lower than the other clusters each consisting of a mixture of some adaptable non-experts and a few stubborn ones. In other words, the expert cluster is a group of respondents who have similar estimates and collectively exhibit minimal updating.

We illustrate how our algorithm is more accurate again using the same example. As illustrated in Figure 1 (left panel), the cluster nearest to the truth exhibits the lowest average WOA (e.g, the cluster around the red dashed line). Thus, pinpointing and subsequently aggregating over this specific cluster leads to an aggregate estimate that is minimally influenced by the estimation biases of the inaccurate non-experts.

4 MATHEMATICAL MODEL

We theoretically characterize the effective boundary conditions of our approach. In summary, our method improves group mean when (1) non-experts have larger estimation bias, which makes them more separable from the unbiased experts; and (2) the proportion of stubborn non-experts is not large within the total population of non-experts. Beyond the Wisdom of Crowds context, the insights derived from theoretical analysis have broader implications in understanding judgment accuracy under social influence. Our framework lays out a theoretical foundation to address queries like the conditions under which exposure to numerical advice (e.g., group mean) bolsters or undermines individual and collective intelligence. Details are provided in the working paper.

5 EMPIRICAL VALIDATION

We (mostly) validate the theoretical prediction of our model and the empirical performance of our algorithm using data from 100 questions collected by other researchers in four various contexts. Most notably, using advice-taking data from Soll, Palley, and Rader (2022), we show that our algorithm yields large improvement in estimation tasks in which the crowd's initial responses tend to exhibit estimation bias, without significant deterioration in estimation tasks where the crowd is on average accurate with their initial responses. These results also serve as an empirical test of our model's prediction with respect to the impact of non-experts' estimation bias on our algorithm. Using data from Centola et al. (2023), we demonstrate our algorithm's potential in enhancing WoC in high-stake scenarios, such as medical evaluations made by physicians. Further evidence from studies by Lorenz et al. (2011) and Zhang (2022) showcases our algorithm's capacity to amplify the Wisdom of Crowds in knowledge-based questions, both in lab and online with different subject pools. Furthermore, our algorithm is also effective in improving WoC when Republicans and Democrats answer political factual belief questions – a context prone to encountering more stubborn subjects. Details are provided in the working paper.

6 CONCLUSION

In this paper, we develop an algorithm that improves Wisdom of Crowds in a single-prediction-problem context when historical performance or specific forecaster attributes are unavailable. The algorithm capitalizes on the collective advice-taking behavior of similar responses to pinpoint a cluster of experts to reduce any estimation bias of the crowd. We showcase the effectiveness of our algorithm across diverse contexts with different types of agents. Notably, since data from other scholars were collected to address research questions distinctly different from ours, our empirical results may be perceived as having enhanced credibility and external validity!

Our theoretical framework depicts the algorithm's properties by incorporating behavior factors such as stubbornness and over-updating beyond the rational foundation of advice-taking. Most importantly, it shows that the presence of heterogeneity of expertise (e.g., non-experts having estimation bias) as the key determinant to the algorithm's success. Beyond the Wisdom of Crowds context, the insights derived from theoretical analysis have broader implications in understanding judgment accuracy under social influence. Our framework lays out a theoretical foundation to address queries like the conditions under which exposure to numerical advice (e.g., group mean) bolsters or undermines individual and collective intelligence [2, 3, 5].

ACKNOWLEDGMENTS

I am grateful to Drazen Prelec for his guidance and mentorship. I thank Jack Soll, Rick Larrick, Asa Palley, Dave Rand, Mark Himmelstein, David Budescu, Juliana Schroeder, Don Moore, Clayton Critcher, Oleg Urminsky, Celia Gaertig, Aurelien Baillon, Shiman Ding, Kristine Cho, Sophia Li, William Ryan, Dmitry Taubinsky, Frank Schilbach, Sinan Aral, Thomas Graeber, and many other seminar and conference participants for their feedback on this and an earlier project built on a related concept.

REFERENCES

[1] Abdullah Almaatouq, Alejandro Noriega-Campero, Abdulaziz Alotaibi, Peter Krafft, Mehdi Moussaid, and Alex Pentland. 2020. Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences* 117, 21 (2020), 11379–11386.
[2] Joshua Becker, David Brackbill, and Damon Centola. 2017. The crowd classification problem: Social dynamics of binary-choice accuracy. *Management Science* 68, 5 (2017), 3949–3965.
[3] Joshua Becker, Douglas Guilbeault, and Edward B. Smith. 2022. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences* 114, 26 (2022), E5070–E5076.
[4] Damon Centola, Joshua Becker, Jun Zhang, Jaya Aysola, Douglas Guilbeault, and Lisa Khoong. 2023. Experimental evidence for structured information–sharing networks reducing medical errors. *Proceedings of the National Academy of Sciences* 120, 31 (2023), e2108290120.
[5] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* 108, 22 (2011), 9020–9025.
[6] Gonzalo Madirolas and Gonzalo G. de Polavieja. 2015. Improving collective estimations using resistance to social influence. *PLoS Computational Biology* 11, 11 (2015), e1004594.
[7] Alex E. Mannes, Jack B. Soll, and Richard P. Larrick. 2014. The wisdom of select crowds. *Journal of Personality and Social Psychology* 107, 2 (2014), 276.
[8] Tobias Schultze and David D. Loschelder. 2021. How numeric advice precision affects advice taking. *Journal of Behavioral Decision Making* 34, 3 (2021), 303–310.

[9] Jack B. Soll, Asa B. Palley, and Chad A. Rader. 2022. The bad thing about good advice: Understanding when and how advice exacerbates overconfidence. *Management Science* 68, 4 (2022), 2949–2969.

[10] Zhang. 2022. The Revealed Expertise Algorithm: Leveraging Advice-taking to Identify Experts and Improve Wisdom of Crowds. (2022). Unpublished Manuscript.

Modelling the Diffusion of Real Narratives

Diffusion of Narratives

Qiusi Sun

School of Information Studies, Syracuse University, qsun12@syr.edu

Una Joh

School of Information Studies, Syracuse University, sjoh01@syr.edu

Joshua Introne

School of Information Studies, Syracuse University, jeintron@syr.edu

1 INTRODUCTION

How information spreads in populations is a fundamental determinant of collective intelligence and the subject of an enormous body of literature in computational social science [15], evidencing both the importance of the topic and its difficulty. The "virality" of an information contagion depends on the complex interplay of factors, including the structure of social networks, the characteristics of users that spread information, the nature of the contagion itself, and the timing of the release of contagions in relation to exogenous events.

Some studies have indicated that the content of a contagion—that is, identifiable features of a given contagion itself—is a weak predictor of spread [1] making it hard to anticipate the virality of a given artifact before it is released. Here, we report on an empirical study suggesting that a content feature may indeed be highly predictive in the case of *narrative* contagions. Specifically, we show that the causal relationships between news items in an evolving narrative are highly predictive of reported article shares once release timing is taken into account.

1.1 Background

A dominant paradigm for examining information diffusion is the contagion metaphor [3]. Individuals may be more or less prone to adopting a given contagion, possibly requiring multiple exposures, but social influence is the primary driver of spread. On its own, the contagion metaphor cannot model the diffusion of sets of interacting contagions. For example, in an evolving news story (or conspiracy theory [2]), distinct articles might elaborate upon or conflict with one another, influencing the dynamics of the overall set. Work in collective belief dynamics helps to motivate the importance of considering such interactions. For example, Friedkin et al. [5]showed how logical constraints might influence the spread of belief. More recently, Introne [7] offered empirical evidence illustrating how beliefs cluster together in social media conversations. These approaches illustrate the importance of considering interactions amongst contagions but do not shed light on which or how contagions are likely to interact.

The case of narratives offers a special case that may address this gap. Narrative are of fundamental importance in human information processing [4], and how a person interprets any given event (e.g., conveyed via a news headline) may be understood as a kind of belief that influences the adoption of other events in an unfolding story. Prior psychological research has demonstrated that *causal* relationships among items in a story are of particular importance. Trabasso and colleagues , demonstrated that the structure of a causal network the undergirds a story is highly predictive of information retention and perceived coherence.

Trabasso & Sperry [13] described a technique for extracting the causal network of stories by applying a counterfactual criterion: event A is said to cause event B if B would not have occurred in the absence of A within the circumstances of the narrative. Applying the counterfactual criterion in practice is challenging because what should be considered an event in a story was not well defined. However, other research that sought to identify "story grammars" underlying aspects of cognitive recall and comprehension [8, 10] can serve as a guide. This work found that humans tend to organize stories into episodes centered on the intentional actions of actors. The structure of an episode includes (at minimum) an actor, an intention, an action, and a consequence, some of which may need to be inferred. Episodes are chained together in a semi-hierarchical manner, wherein an interrupted action may give rise to nested episodes. This structure is easy to identify and provides important cues about causal linkages.

This prior work suggests both a model of psychologically relevant interdependence and a path toward empirical analysis of these interdependencies in actual stories. We describe both our model and approach to empirical validation below.

1.2 Methods

Our model of narrative diffusion combines social influence [3] with a property we refer to as *narrative attraction*, so that prior adoptions influence the adoption of new information. From the perspective of a given agent *i* at time *t*, the narrative attraction $v_i(n_a, t)$ of any newly encountered item may be defined as:

$$v(n_a, t) = S^v\left(\sum_{n_b \in N_{i,t}} \frac{1}{d(n_a, n_b)^\phi}\right)$$

where S^v is a sigmoidal function, $N_{i,t}$ is the set of items adopted by an agent *i* at time *t*, $d(n_a, n_b)$ is the shortest path between two narrative items in the story network corresponding to the completed story, and ϕ determines how quickly narrative influence decays with causal distance.

We combine this model with a model of social influence in a master equation that provide the probability an agent will adopt a given narrative element based on their prior adoptions:

$$P_{adopt,i}(n_a, t) = \alpha \cdot v_i(n_a, t) + (1 - \alpha) \cdot I_i(n_a, t),$$

where $I_i(n_a, t)$ is a model of social influence based on the number of agent *i*'s neighbors that have adopted n_a at time *t*, and α allows us to adjust the relative influence of narrative attraction and social influence.

Our model is mathematically too complex to offer a closed-form solution. To make predictions about empirical scenarios, we embed our model of narrative diffusion in an agent-based model includes two additional features: 1) items may be "revealed" at specified points in time (i.e., simulating the release of news articles), and 2) agents have a baseline probability of adopting newly released items that decays over time.

To provide initial validation of the model, we focused on the relatively recent ``Trump documents" story, which reported on classified documents the former president retained upon leaving office. This story has been widely reported across the

political spectrum in the U.S. news media. This will ultimately enable us to examine how political bias might alter narrative construction and diffusion.

We collected all news items from The New York Times, using a combination of web search and indices published by the news outlets, yielding approximately 120 articles. We refined this set to include only those that reported on major breaking events (omitting articles offering opinions or summaries, or those reporting relatively brief updates), leaving roughly 20 news stories. We then collected article sharing metrics for each story from CrowdTangle[1].

To construct the causal network for the story, we organized a collaborative card-sorting exercise. Headlines and abstracts were printed on index cards, which were then sorted into groups by a team of graduate students. After this, the team worked together to capture the overall structure of the story on a whiteboard. Finally, the authors refined this graph to establish a final version of the story network see Figure 1).

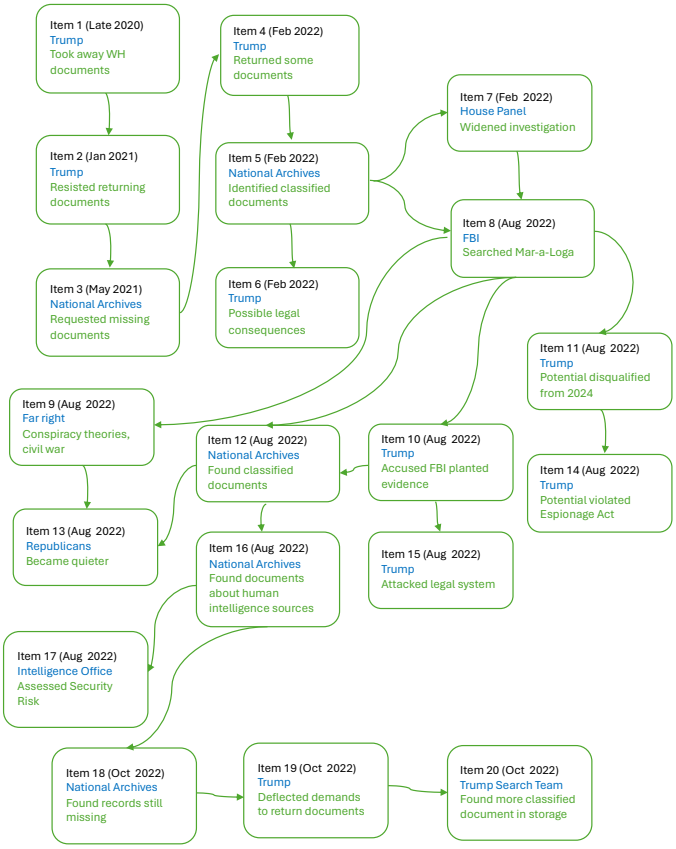


Figure 1: Story graph for the Trump documents story, as reported in the New York Times

We examined the spread of each news item over multiple runs of our agent-based model, using a small-world social network [9] with 200 agents, and calibrated the model manually. We then compared the empirical number of adoptions for each news item. We report our initial results below.

1.3 Results

Figure 2 compares the number of adoptions for each story element in our simulation (y-axis) with actual social media shares (x-axis). Results are averaged over 10 trials and suggest a strong correlation between simulation results and engagement metrics $R \approx .74$. The fit tends to be stronger for news items that were shared more widely, possibly reflecting biases in our selection strategy. We are currently elaborating our results with a more detailed analysis of the parameter space, versions of the "Trump documents" story reported in other outlets, and two additional stories. We will report on these results during our talk.

1 A Meta-owned online service that reports engagement metrics for online stories.

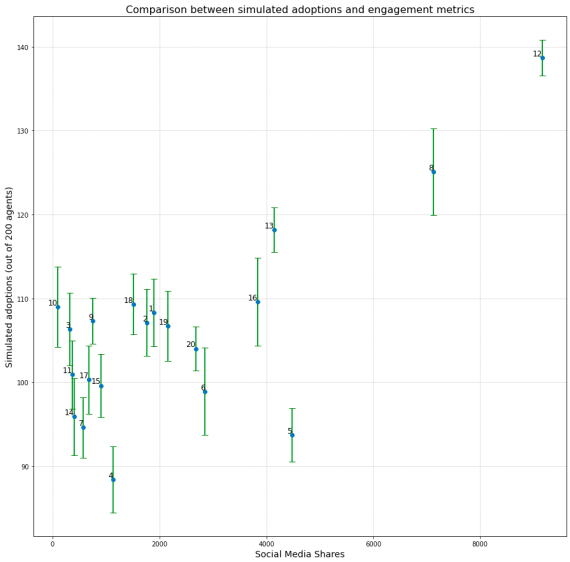


Figure 2: Correlation between the number of agents adopting each story element and social media shares reported by Crowdtangle for the Trump document story.

REFERENCES

[1] Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M. and Leskovec, J. 2014. Can Cascades Be Predicted? *Proceedings of the 23rd International Conference on World Wide Web* (New York, NY, USA, 2014), 925–936.

[2] Cosentino, G. 2020. From Pizzagate to the Great Replacement: The Globalization of Conspiracy Theories. *Social Media and the Post-Truth World Order: The Global Dynamics of Disinformation*. G. Cosentino, ed. Springer International Publishing. 59–86.

[3] Dodds, P.S. and Watts, D.J. 2004. Universal Behavior in a Generalized Model of Contagion. *Physical Review Letters*. 92, 21 (May 2004), 218701. DOI:https://doi.org/10.1103/PhysRevLett.92.218701.

[4] Fisher, W.R. 1989. *Human Communication as Narration: Toward a Philosophy of Reason, Value, and Action*. University of South Carolina Press.

[5] Friedkin, N.E., Proskurnikov, A.V., Tempo, R. and Parsegov, S.E. 2016. Network science on belief system dynamics under logic constraints. *Science*. 354, 6310 (Oct. 2016), 321–326. DOI:https://doi.org/10.1126/science.aag2624.

[6] Graesser, A.C., Singer, M. and Trabasso, T. 1994. Constructing inferences during narrative text comprehension. *Psychological Review*. 101, 3 (1994), 371–395. DOI:https://doi.org/10.1037/0033-295X.101.3.371.

[7] Introne, J. 2023. Measuring Belief Dynamics on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* (2023), 387–398.

[8] Mandler, J.M., Scribner, S., Cole, M. and DeForest, M. 1980. Cross-Cultural Invariance in Story Recall. *Child Development*. 51, 1 (1980), 19–26. DOI:https://doi.org/10.2307/1129585.

[9] Newman, M.E.J. and Watts, D.J. 1999. Scaling and percolation in the small-world network model. *Physical Review E*. 60, 6 (Dec. 1999), 7332–7342. DOI:https://doi.org/10.1103/PhysRevE.60.7332.

[10] Stein, N. and Glenn, C. 1979. An analysis of story comprehension in elementary school children. *New directions in discourse processing*. 2, (1979), 53–120.

[11] Trabasso, T. and van den Broek, P. 1985. Causal thinking and the representation of narrative events. *Journal of Memory and Language*. 24, 5 (Oct. 1985), 612–630. DOI:https://doi.org/10.1016/0749-596X(85)90049-X.

[12] Trabasso, T., Secco, T. and Broek, P. 1982. Causal cohesion and story coherence. *Learning and comprehension of text*. H. Mandl, N. Stein, and P. Broek, eds. Lawrence Erlbaum Associates. 83–112.

[13] Trabasso, T. and Sperry, L.L. 1985. Causal relatedness and importance of story events. *Journal of Memory and Language*. 24, 5 (Oct. 1985), 595–611. DOI:https://doi.org/10.1016/0749-596X(85)90048-8.

[14] Zhou, F., Xu, X., Trajcevski, G. and Zhang, K. 2021. A Survey of Information Cascade Analysis: Models, Predictions, and Recent Advances. *ACM Computing Surveys*. 54, 2 (Mar. 2021), 27:1-27:36. DOI:https://doi.org/10.1145/3433000.

Using the Belief Landscape Model to Predict Individual Belief Dynamics

Jianghui Li
jli159@syr.edu
Syracuse University
Syracuse, New York, USA

Joshua Introne
jeintron@syr.edu
Syracuse University
Syracuse, New York, USA

ABSTRACT

We demonstrate a classifier that is able to predict how likely it is an individual social media user is to change their professed beliefs. Our classifier uses features based on how people cluster in terms of their expressed beliefs, and we obtain high degrees of accuracy for two distinct datasets. Our results suggest that the frequency with which people change their publicly professed beliefs seems to be a stable property of individuals.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **General and reference** → **Measurement**.

KEYWORDS

Belief Dynamics, Classification, Computational Social Science

ACM Reference Format:
Jianghui Li and Joshua Introne. 2024. Using the Belief Landscape Model to Predict Individual Belief Dynamics. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

Social media has offered an unprecedented window into the dynamics of collective opinion formation and belief change in online settings. It may also be the case that algorithmic filters [1], misinformation [6], the loss of traditional media gatekeepers [2], and other complex socio-technical factors [10] have led to shifts in population-wide belief dynamics [1]. Of particular concern is that an information environment polluted with misinformation and polarizing content can lead individuals to develop beliefs that diverge from empirically validated reality, possibly leading to real physical harm [7] and political instability [4]. Because of this, there is growing interest in predicting who is likely to fall prey to misinformation [3, 8, 11].

Recent work by Introne [5] offered new methodological tools for characterizing the belief dynamics of online populations. The study introduced the Belief Landscape Framework (BLF), which is a spatial metaphor for representing a population's beliefs and the movements of individuals across the landscape as they express different beliefs. Introne showed that most online users gravitate towards the center of "belief attractors," which tend to draw the bulk of individuals, while a smaller number of users on the periphery of belief attractors update their beliefs more frequently. Here, we build on the Belief Landscape Framework to classify users according to their proclivity for belief change. This approach is unlike prior work, which has tended to use psychometric and other individual features to predict susceptibility to misinformation [e.g., 8]. Predicting levels of individual belief variability is not the same as predicting susceptibility to misinformation, but individuals who are not "anchored'" to cultural belief attractors are more likely to be easily influenced. Our task is thus a first step in developing new techniques for predicting susceptibility to misinformation or other information attacks. We show here that it is possible to classify individuals with high reliability based on analysis of the topological properties of the belief landscape they have traversed.

1.1 Methods

To develop our classifier, we employed two datasets collected from Twitter using the streaming API. One dataset centered on climate change conversations, covering 166,268 tweets from May 24th to October 27th, 2019, and another comprising 264,777 tweets on 'kpop' and 'blm' hashtags from January 1st to December 31st, 2020. These datasets provide a rich source for analyzing belief dynamics in polarized and cross-cultural contexts. Following [5], we processed both datasets to produce belief landscapes and measure the movements of users across them (Figure **??** provides a schematic of our workflow). As discussed above, belief landscapes are organized around attractors with different topologies. Attractors may be deep or shallow, reflecting the size of the population they draw, and attract users from small or large regions of the overall landscape (the basin of attraction).

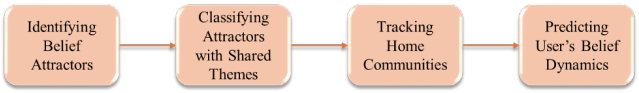


Figure 1: Summary of our data analytic pipeline

Analyzing movements over attractors revealed a higher level of organization, apparent in that users who moved amongst attractors tended to visit distinct subsets and only rarely moved between these subsets. Unlike Introne [5], we interpret these higher-level groups of attractors as mutually exclusive belief systems, and we sought to detect dynamics of movements between these groups.

Because user movements on the landscape were somewhat noisy, we applied a spin-glass clustering algorithm [9] to identify the attractor communities, and then smoothed user transitions between

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

these communities using a Hidden Markov Model, mapping hidden states to the distinct belief communities. We used these transitions to classify users into four distinct classes (Table 1), which became our predictive target.

Table 1: Summary of user count for each class. There are four classes, which are stable user (never changed home location), changer (changed from one home to another), bistable user (changing between two homes), and dynamic user (appeared in more than two homes).

User Class	Climate Change	K-pop & BLM
Class 1 (Stable)	2956	927
Class 2 (Changer)	78	136
Class 3 (Bistable)	80	829
Class 4 (Dynamic)	5	737

In our initial attempts to build a classifier, we found that using the frequency of visits to each underlying attractor yielded a near-perfect classifier. However, such a classifier cannot generalize to different domains. Thus, we sought to build a classifier based on the topology of different attractors. We generated nine features for each landscape by dividing attractors into nine classes based on their depth and width, with three levels in each dimension. Figure 2 shows the attractor class information for several attractors in the belief landscape. We used the amount of time each individual spent in each attractor class as the feature value for each user and compared the performance of simple decision-tree, random forest, and k-nearest neighbors classifiers under 5-fold cross-validation.

1.2 Findings

Class-wise F1 scores from our study are shown in Table 2. Both the random forest and KNN classifiers offered good performance for both datasets and all classes. We also examined the decision-tree classifiers to develop insights into the underlying logic of the classifiers (Figure 2).

Table 2: User classes classification F1 scores using attractor classes (width and depth) as features

Classifier	Climate Change	K-pop & BLM
Decision Tree	0.7022107	0.6812203
	0.8100659	0.7503566
	0.7709280	0.5570560
	0.9204545	0.6941992
Random Forest	**0.9177489**	0.9392151
	0.9726843	0.9594145
	0.9412149	0.9022602
	0.9923037	0.9303215
K-Nearest Neighbors	0.8893754	**0.9480612**
	0.9594382	**0.9675780**
	0.9378589	**0.9388073**
	0.9692860	**0.9498699**

Our findings offer several insights. The topological properties of attractors that people visit on belief landscapes are highly predictive of an individual's proclivity for belief change. The decision tree in Figure 3 illustrates that individuals positioned in shallow and wide belief clusters are more likely to change their beliefs often. Conversely, users situated in regions with densely converging opinions exhibit a lower propensity for belief change, suggesting that consensus has a stabilizing effect on beliefs. This finding corresponds to the prior observation that deeper attractors tend to be 'stickier' than shallow ones [6]. Moreover, we found these results to be similar in both landscapes, suggesting the generalizability of the model across different domains.

Our results also provide evidence that people who change their beliefs frequently tend to skirt around more densely populated cultural belief systems—that is, those that change their beliefs often may also be those that profess beliefs that co-occur relatively infrequently in a population. Additional analysis will be necessary to explore this finding further, but if correct, this may open new lines of inquiry into the belief dynamics of individuals.

REFERENCES

[1] Michael Bailey, Drew M. Johnston, Martin Koenen, Theresa Kuchler, Dominic Russel, and Johannes Stroebel. 2020. Social Networks Shape Beliefs and Behavior: Evidence from Social Distancing During the COVID-19 Pandemic. (12 2020). https://doi.org/10.3386/w28234 DOI: 10.3386/w28234.

[2] Janelle Benham. 2020. Best Practices for Journalistic Balance: Gatekeeping, Imbalance and the Fake News Era. *Journalism Practice* 14, 7 (8 8 2020), 791–811. https://doi.org/10.1080/17512786.2019.1658538

[3] Rex P. Bringula, Annaliza E. Catacutan-Bangit, Manuel B. Garcia, John Paul S. Gonzales, and Arlene Mae C. Valderama. 2022. "Who is gullible to political disinformation?" : predicting susceptibility of university students to fake news. *Journal of Information Technology Politics* 19, 2 (3 4 2022), 165–179. https://doi.org/10.1080/19331681.2021.1945988

[4] Rafael Evangelista and Fernanda Bruno. 2019. WhatsApp and political instability in Brazil: targeted messages and political radicalisation. *Internet Policy Review* 8, 4 (2019), 1–23. https://doi.org/10.14763/2019.4.1434 publisher: Berlin: Alexander von Humboldt Institute for Internet and Society.

[5] Joshua Introne. 2023. Measuring Belief Dynamics on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 17, 387–398.

[6] Shan Jiang, Miriam Metzger, Andrew Flanagin, and Christo Wilson. 2020. Modeling and Measuring Expressed (Dis)belief in (Mis)information. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (26 5 2020), 315–326. https://doi.org/10.1609/icwsm.v14i1.7302

[7] Hye Kyung Kim and Edson C. Tandoc. 2022. Consequences of Online Misinformation on COVID-19: Two Potential Pathways and Disparity by eHealth Literacy. *Frontiers in Psychology* 13 (2022). https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.783909 [Online; accessed 2024-03-03].

[8] Kerry Lee. 2004. Age, neuropsychological, and social cognitive measures as predictors of individual differences in susceptibility to the misinformation effect. *Applied Cognitive Psychology* 18, 8 (2004), 997–1019. https://doi.org/10.1002/acp.1075

[9] Jörg Reichardt and Stefan Bornholdt. 2006. Statistical mechanics of community detection. *Physical Review E* 74, 1 (18 7 2006), 016110. https://doi.org/10.1103/PhysRevE.74.016110 publisher: American Physical Society.

[10] Jaime E. Settle. 2018. *Frenemies: How Social Media Polarizes America.* Cambridge University Press.

[11] Tracy Jia Shen, Robert Cowell, Aditi Gupta, Thai Le, Amulya Yadav, and Dongwon Lee. 2019. How Gullible Are You? Predicting Susceptibility to Fake News. *Proceedings of the 10th ACM Conference on Web Science*, 287–288. https://doi.org/10.1145/3292522.3326055 [Online; accessed 2024-03-03].

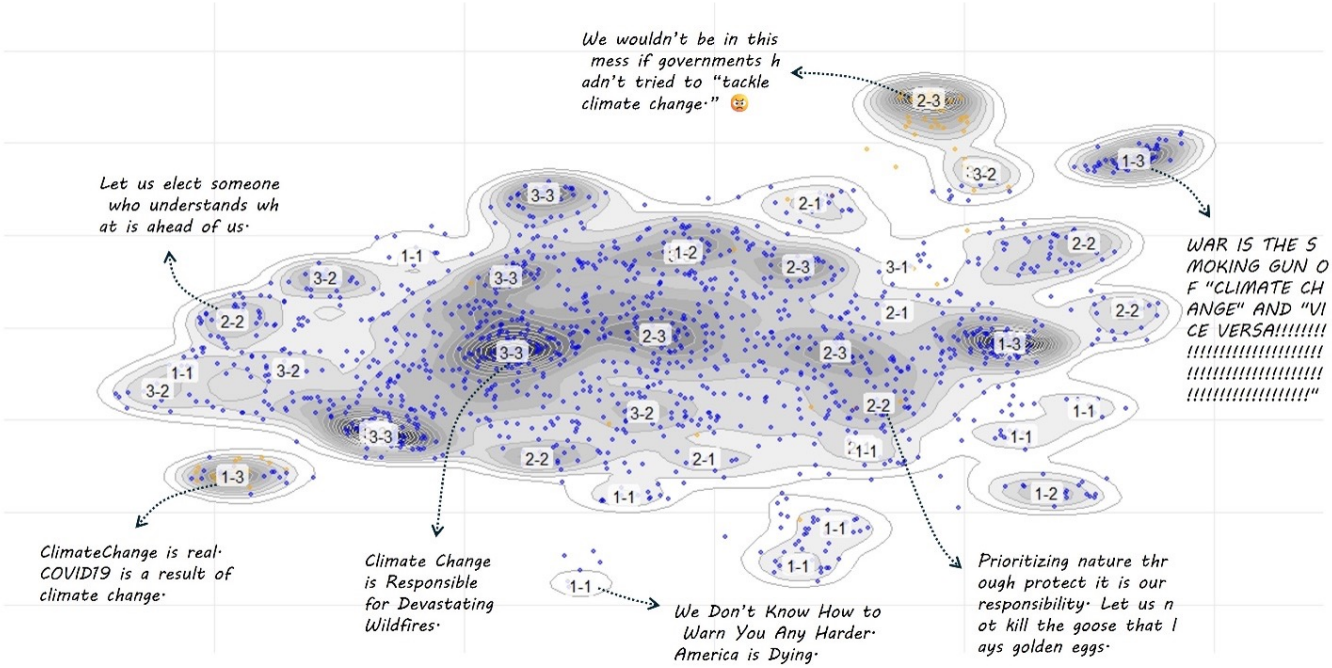


Figure 2: Climate change discussion belief landscape. The dots on the landscape represent a user at a time (only 0.1% are shown). Blue dots are believers, and orange dots are skeptics. The label numbers are attractor classes (9 classes in total) defined by dual quantiles (3 quantiles for each; '1' for low, '2' for medium, and '3' for high) of their width (number to the left) and depth (number to the right).

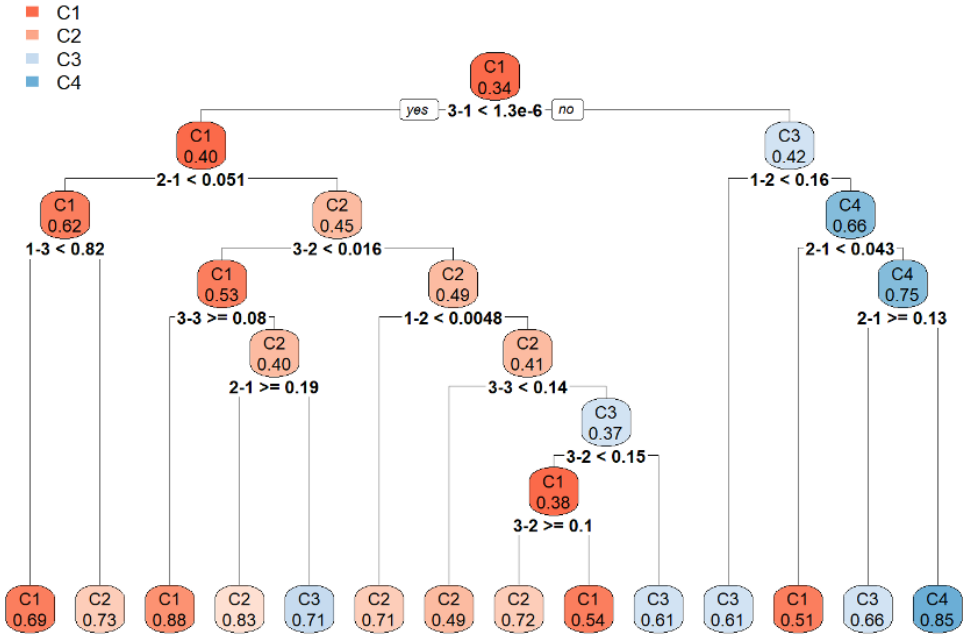


Figure 3: An example of decision tree visualization from climate change data

Instability in a Cultural Model of Belief Dynamics

Khashayar Khodadad Motarjemi
kkhodada@syr.edu
Syracuse University
Syracuse, New York, USA

Joshua Introne
jeintron@syr.edu
Syracuse University
Syracuse, New York, USA

ABSTRACT

We introduce and explore a novel model of belief dynamics that pits cultural forces against social forces and individual stochasticity. We hypothesize that a proper balance between these three forces leads to a point of criticality, wherein we observe ongoing dynamics of the formation and dissolution of belief clusters. Our simulations confirm this hypothesis, demonstrating that for every level of relative strength of social and cultural forces, there is a distinct threshold of stochasticity that triggers the continuous dynamics of coalition formation and dissolution.

CCS CONCEPTS

• **Applied computing** → **Psychology**.

KEYWORDS

Opinion Dynamics, Simulation, Complexity Theory

ACM Reference Format:
Khashayar Khodadad Motarjemi and Joshua Introne. 2024. Instability in a Cultural Model of Belief Dynamics. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

The study of belief dynamics examines how beliefs form, change, and spread within populations [6]. Claims about the movements of political parties, the emergence of conspiracy theories, and vaccine hesitancy all focus on belief dynamics. Discussions of polarization are particularly relevant in today's political landscape [1, 7, 9]. In an era where digital platforms significantly amplify the rapid dissemination and evolution of beliefs, a question of central importance is how the design and interaction of these platforms influence belief dynamics, and whether these influences enhance or undermine the collective intelligence of human societies.

Traditional models of opinion dynamics, such as those pioneered by [4] and later expanded in bounded confidence models like the Hegselmann-Krause model [10], have significantly advanced our understanding of how consensus or polarization may arise within societies. However, these models are largely theoretical and focus primarily on discrete opinions or single-issue dynamics. Beliefs in human societies tend to be interdependent, giving rise to culturally specific clusters of beliefs that co-occur within individuals [2, 3, 8].

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

A growing body of work has aimed to expand the analysis of belief dynamics to investigate the complex, interconnected, and continuous nature of belief systems [5, 6, 8]. Our study furthers this research by introducing a model of belief dynamics that accounts for multiple interacting beliefs and concentrates on the interplay between cultural and social forces, alongside individual stochastic behavior. Our model is motivated by several factors. First, digital technologies are potent mediators of both social interaction and cultural perceptions [e.g., 11], but we lack the tools needed to assess how their interplay modifies the complex dynamics of belief. Second, empirical validation of theoretical models of opinion dynamics is uncommon, which significantly limits their practical application. The model we propose is not only inspired by, but also aligns with, empirical strategies developed in [8], paving the way for future validation with existing social media data. Lastly, we aim to identify regions within the parameter space that prevent a collapse into a "frozen" belief state, such as polarization or consensus. We contend that these regions are crucial for collective intelligence, as populations in frozen belief states are less apt to assimilate new information.

We hypothesize that a proper balance between social and cultural forces leads to a point of criticality [1], wherein individual stochasticity drives the ongoing dynamics of the formation and dissolution of belief clusters. Our simulations indeed demonstrate that for every level of relative strength of social and cultural forces, there is a distinct threshold of stochasticity that triggers the continuous dynamics of coalition formation and dissolution. In the following, we present our model and initial findings.

1.1 Model and method

The Belief Landscape Framework (BLF) introduced in [8] envisions the space of possible beliefs as a landscape where coherent sets of beliefs create attractors that attract larger populations. This framework is based upon a few key assumptions:

(1) People gravitate towards coherent belief sets.
(2) The coherence of a belief set is determined by the number of individuals clustered around it.
(3) The perception of coherence increases as one's distance from that belief set diminishes.

In our framework, beliefs are considered coherent because they co-occur in the population, rather than for some intrinsic criterion of compatibility. Individuals gravitate towards more densely populated belief regions, taking cues from the concentration of cultural signals. However, the belief adoption process is also influenced by individuals' social networks. Factors such as personal inclinations introduce unpredictability, causing some people to be more or less influenced by their environment or connections.

Mathematically, we define an agent i's belief-state at time t as a vector of observed dispositions with respect to a set of discrete

propositions b_j. That is:

$$\mathbf{b}_{i,t} = b_j : b_j \in [0,1]\,,\ \text{j} = 1, ..., N$$

Where $\mathbf{B}_t$ represents the entire population's beliefs.

Cultural signaling in our model is akin to gravitational force. We approximate it with a law analogous to Newton's law of universal gravitation, where γ acts as the gravitational constant:

$$F_{c_{i,j}} = \frac{\gamma}{r_{i,j}^2} \cdot \hat{r}_{i,j}$$

Here, $r_{i,j}$ is the distance between agents i and j, and $\hat{r}_{i,j}$ denotes the direction from i to j. For simplicity, agents are assumed to have equal influence, but it is possible to model influencers as having more "mass." Considering the entire agent population, the total cultural force on agent i at time t is:

$$F_c\left(\mathbf{b}_{i,t}\right) = \sum_{j \neq i} F_{c_{i,j}}$$

Social influence on an agent is computed as a weighted mean of the belief-states of an agent's social ties. Given $\mathbf{A}$ as the tie strength matrix, the social forces acting on agent i are:

$$F_s\left(\mathbf{b}_{i,t}\right) = \sum_{j=1}^{M} \mathbf{A}_{i,j}\left(\mathbf{b}_{j,t} - \mathbf{b}_{i,t}\right)$$

Lastly, we introduce the concept of "energy" to account for each agent's randomness. Randomness is modeled as a Wiener process [8] represented as dW_i and the degree of randomness exhibited by an agent is governed by an energy parameter e_i. Integrating all these components, we represent the dynamics of a population's beliefs as:

$$\frac{d\mathbf{B}}{dt} = \rho F_s\left(\mathbf{B}_t\right) + (1-\rho)\, F_c\left(\mathbf{B}_t\right) + \sqrt{2\left(e \otimes 1_{1,N}\right)} \odot dW$$

Where ρ is a control parameter allowing us to vary the relative influence of cultural and social forces.

To test our hypothesis, we constructed an agent-based model governed by the above equation. First, we define convergence time as the number of time steps required for the mean Euclidean distance between agents in belief space to be less than a threshold t. Second, to evaluate the longitudinal clustering dynamics, we measure the variance in "roughness" of the belief space over time. To calculate roughness at any given time, we apply a kernel density estimate to the n-dimensional belief space to generate a scalar field $Z\left(x_1, x_2, \ldots, x_n\right)$. Roughness R is defined as the standard deviation of the magnitude of the gradient vector field ∇Z across the space:

$$R = \sigma\left(|\nabla Z|\right)$$

Where σ denotes the standard deviation computed over the magnitudes of the gradient vector field across the entire domain of Z. Longitudinal roughness is then just the variance of R over a fixed number of time steps.

For the experiments reported here, we conducted a parameter sweep over values of ρ in $[0, 1]$ and e in $[25, 230]$, with all agents set to the same value of e. In all cases, we used a social network of 100 agents arranged in a Newman-Watts small-world network [11] with $k = 4$ and $p = .25$, and agents resided in a ten-dimensional belief space. Fifteen trials were run at each parameter setting, and each measure of interest averaged over these trials unless otherwise stated. We set the convergence threshold t to 0.1 and evaluated longitudinal roughness over a period of 200 timesteps after the simulation has settled (10,000 timesteps).

1.2 Findings

Across all values of ρ, we find that all simulations converge to a single point of belief as long as e is small enough (Figure 1a); beyond this threshold, beliefs never appear to converge. Note that in the case of ρ=1, where cultural forces are absent, very little energy is necessary to prevent convergence. However, Figure 1b reveals another level of organization that appears beyond the boundary separating convergence from non-convergence. Here (the yellow band in Figure 1b), agents continually form "coalitions" that subsequently dissolve. Figure **??** examines a sample of the raw roughness values that underlie this result; what is notable is that there is no clear pattern in these dynamics nor does roughness vary in a predictable way across values of ρ. Roughness varies substantially, and this variance is apparent even when averaged over multiple trials.

In summary, our initial results offer support for our hypothesis that a critical point exists across all levels of ρ, leading to continuous dynamics in the clustering of beliefs. Additional analysis will be required to further understand these dynamics. However, these findings underscore the utility of the model and pave the way for "tuning" the belief dynamics of a specific system to target a particular dynamic regime.

REFERENCES

[1] Bill Bishop and Robert G Cushing. 2009. *The big sort: Why the clustering of like-minded America is tearing us apart.* Houghton Mifflin Harcourt.

[2] Mark J. Brandt and Willem W. A. Sleegers. 2021. Evaluating Belief System Networks as a Theory of Political Belief System Dynamics. *Personality and Social Psychology Review* 25, 2 (May 2021), 159–185. https://doi.org/10.1177/1088868321993751 Publisher: SAGE Publications Inc.

[3] Philip E Converse. 2006. The nature of belief systems in mass publics (1964). *Critical review* 18, 1-3 (2006), 1–74. Publisher: Taylor & Francis.

[4] Morris H. Degroot. 1974. Reaching a Consensus. *J. Amer. Statist. Assoc.* (March 1974). https://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10480137 Publisher: Taylor & Francis Group.

[5] Noah E. Friedkin, Anton V. Proskurnikov, Roberto Tempo, and Sergey E. Parsegov. 2016. Network science on belief system dynamics under logic constraints. *Science* 354, 6310 (Oct. 2016), 321–326. https://doi.org/10.1126/science.aag2624 Publisher: American Association for the Advancement of Science.

[6] Mirta Galesic, Henrik Olsson, Jonas Dalege, Tamara van der Does, and Daniel L. Stein. 2021. Integrating social and cognitive aspects of belief dynamics: towards a unifying framework. *Journal of The Royal Society Interface* 18, 176 (March 2021), 20200857. https://doi.org/10.1098/rsif.2020.0857 Publisher: Royal Society.

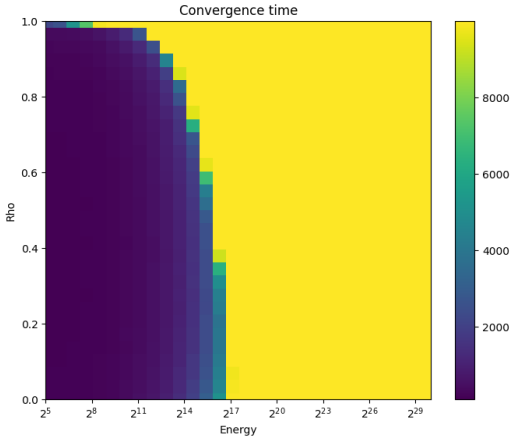
[7] Abigail Geiger. 2017. The Partisan Divide on Political Values Grows Even Wider. https://www.pewresearch.org/politics/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/

[8] Joshua Introne. 2023. Measuring Belief Dynamics on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 17 (June 2023), 387–398. https://doi.org/10.1609/icwsm.v17i1.22154

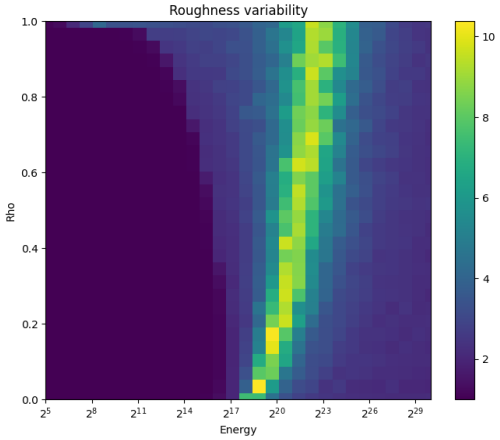
[9] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. 2012. Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly* 76, 3 (Jan. 2012), 405–431. https://doi.org/10.1093/poq/nfs038

[10] Hegselmann Rainer and Ulrich Krause. 2002. Opinion Dynamics and Bounded Confidence: Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation* 5, 3 (2002).

[11] Jaime E. Settle. 2018. *Frenemies: How Social Media Polarizes America.* Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108560573

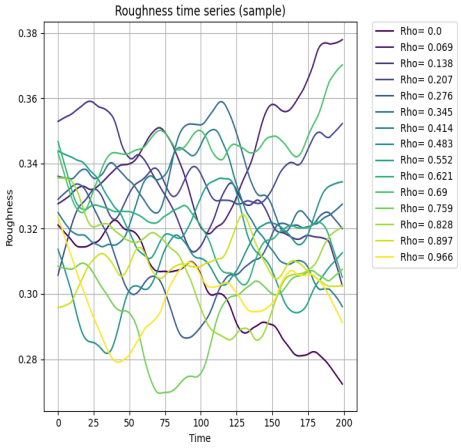


(a) Timesteps before all belief states collapse to the same vector; the yellow region never achieves convergence.

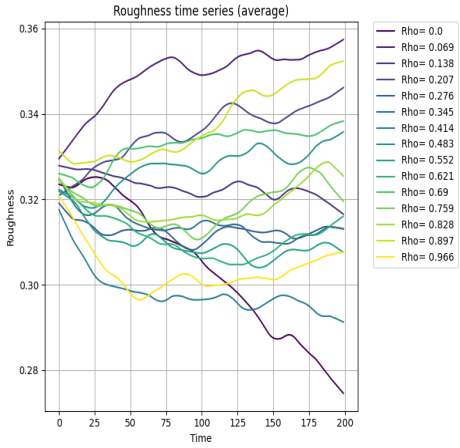


(b) Longitudinal roughness; the "ridge" reflects a critical point of continuous variation in the clustering of beliefs

Figure 1: Results of parameter sweep



(a) Longitudinal roughness for selected individual trials.



(b) Longitudinal roughness averaged over all trials.

Figure 2: Longitudinal roughness in the critical region of Figure 1b. Data in both plots smoothed via Loess smoothing.

An empirical study of the El Farol bar problem
Non-Archival Extended Abstract

EDGAR J. ANDRADE-LOTERO*, Universidad del Rosario, Colombia

ROBERT L. GOLDSTONE*, Indiana University, USA

ACM Reference Format:

Edgar J. Andrade-Lotero and Robert L. Goldstone. 2024. An empirical study of the El Farol bar problem Non-Archival Extended Abstract. In *Proceedings of ACM Collective Intelligence 2024 (CI2024).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Our investigation centers on the El Farol bar problem [2], serving as an elegant paradigm to assess coordination in groups that do not engage in explicit communication. The paradigm asks us to imagine a town with 100 adults, each desiring to attend the same bar every Thursday. The experience is enjoyable when the attendance is 60 or fewer individuals, but due to overcrowding, none of the attendees enjoys their experience if more than 60 people attend. The challenge lies in the fact that these 100 adults must collectively decide, without central authority or communication, whether to go to the bar. This problem has garnered significant theoretical and computational attention [4, 7–10, 13] and it permits possible solutions in the form of role specialization [5, 11], turn-taking [15], or both. Although there are studies that have approached very similar problems from an empirical perspective [1, 6, 11, 12, 14], a full-scale empirical examination of the El Farol bar problem, to our knowledge, does not exist. To address this gap, we have conducted an experimental study, establishing an empirical testbed for future theoretical analyses.

2 METHODS

2.1 Formalization and simulations

Given that there are different interpretations of the El Farol bar problem, it is important to exactly specify how we understand it. On each round t, each agent i makes a decision, a_t^i, which can be either 0 or 1 (0=do not go; 1=go), independently of and simultaneously to the others. We will denote by A_t the total number of visitors to the bar on round t. We will denote by N the total number of agents and μ the bar's capacity (or threshold) as a proportion of N. The payoff for each agent is defined as follows: It is 0 if $a_t^i = 0$; -1 if $a_t^i = 1$ & $A_t > \mu N$; and 1 if $a_t^i = 1$ & $A_t \leq \mu N$.

In addition to the experiment with human participants, we also implemented simulated agents which decided to go to the bar with a certain probability p. In one scenario, p was equal to μ (Random scenario). In another scenario, p was set to the empirically found proportions (using Maximum Likelihood modeling). Simulation parameters mirrored those of the corresponding human trial for comparison. Scenarios were run 10 times with different seeds to reduce noise.

*Both authors contributed equally to this research.

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 the author(s) 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.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

2.2 Materials and participants

We conducted two experiments. The first one included two-person groups involving 46 students from Universidad del Rosario (conducted in Spanish) where μ=50%. There were two between-groups treatments: the full and partial information conditions. The test was implemented in *nodeGame* [3]. The first screen presented the informed consent form followed by instructions and number of participants. On each round, the screen showed the bar's capacity and two buttons to go to the bar or not (see Figure 1, left panel). If the player did not make a decision within 30 seconds, the computer made the decision with equal probability. Afterwards, each player was shown their score for the round, their total score, the scores of the last 15 rounds, and, in the full information treatment, the attendance of each player during the last 15 rounds (see Figure 1, right panel). The game was repeated for 50 rounds. At the end, the computer calculated the payment for each player based on their score. Participants received their reward through electronic transfer.

The second experiment included from three- up to twelve-person groups involving 152 students from Indiana University (conducted in English). There where two between-groups treatments: the downwards condition, in which the bar's capacity was reduced two times (going from μ ≈75% to ≈50% to ≈25% each threshold 30 rounds); and the upwards condition. Both treatments received full information feedback. Participants received credit for their Psychology 101, 102, or 155 experiment participation requirement. Otherwise, the setup was the same for both experiments.

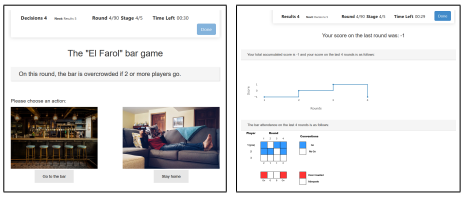


Fig. 1. The experimental task. Left panel: the player chooses whether to go to the bar or not. Right panel: the feedback screen, showing the score from the last round, the cumulative score, the scores from the last 15 rounds, and the attendance choices of each player over the last fifteen rounds. It also includes information on whether the bar was overcrowded on a given round.

3 RESULTS

Performance was measured in terms of average score *over the last 10 rounds*. For the two-player experiment, human participants had better performance than simulated agents in both Random and Maximum Likelihood scenarios (Figure 2, left panel). This shows that participants are basing their responses on more than just an overall probability of going to the bar. Also, there is no significant difference between treatments, full vs. partial information (second panel). The third panel shows the relationship between proportion of go rounds by average score. Each data point is one player. The large cluster in the area $(0.4, 0.6) \times (0.4, 0.65)$ contains players who took turns going to the bar on alternating rounds and obtained relatively good scores. All dyads lie along a slanted line with the player who goes to the bar more often earning a larger score. Moreover, the correlation of the proportion of go rounds between the two players in a dyad is ≈ -0.7. This means that one player complements the other player's drive to go. The same relationship between proportion of go rounds vs. average score with respect to dyads (right panel) shows that the higher the average rate of going, the lower the average score of the group. The last two panels indicate a classic social dilemma in which each person is incentivized to go to the bar often, but if both people do, then overcrowding results.

Turning to the multi-player experiment, we see that higher thresholds were associated with higher average scores (Figure 3, left panel). For thresholds ≈ 0.25, groups in the downwards treatment showed higher average scores than

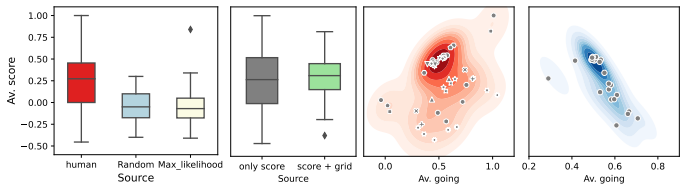


Fig. 2. The two-player experiment. Left panel: Distribution of av. scores for humans and simulated scenarios. Second panel: Distribution of av. scores for the two information treatments. Third panel: Two-variate kernel density estimate of the relationship between proportion of go rounds and av. score for players. Right panel: Two-variate kernel density estimate of go rounds and av. score for dyads.

the upwards treatment. No significant differences where found for thresholds ≈ 0.5 and ≈ 0.75. Humans performed worse than simulated Random agents for thresholds ≈ 0.25, but were better for ≈ 0.75 (second panel). For all thresholds, humans were better than simulated agents in the Maximum Likelihood scenario. The three panels in the right of Figure 3 show the relationship between proportion of go rounds and av. score, broken down by threshold. We can see "free riders" (i.e. players going every round) on the right side. Some of them obtained a high reward, but some got punished when other players also chose to go often. We stacked all pairs of players within a group-size and measured each player by their proportion of go rounds, producing two columns of data, which we used to measure the correlation of go rounds in each group-size, but we found that none of them was significant.

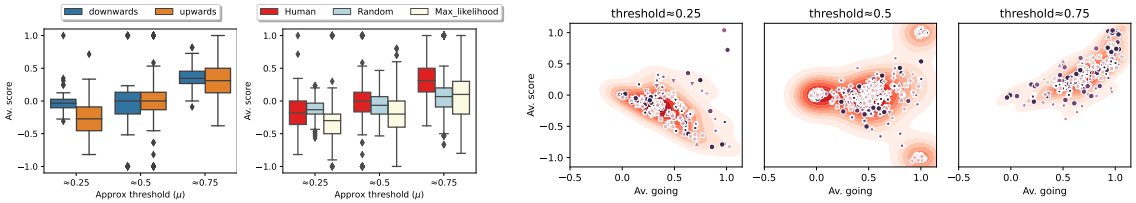


Fig. 3. The multi-player experiments. Left panel: Distribution of av. scores for the three thresholds coded by treatment. Second panel: Distribution of av. scores for the three thresholds coded by human or simulated scenario. Three panels on the right: Two-variate kernel density estimate of the relationship between proportion of go rounds and av. score for players (group is color- and shape-coded).

4 DISCUSSION AND FUTURE WORK

Our results indicate a social dilemma at the heart of the El Farol Game in which goals for the individual are not aligned with those for the group. Moreover, players' outcomes are systematically better than would be expected if they were randomly choosing to go to the bar with either the probability set by the bar's threshold or by their empirically obtained marginal probability. In the two-person groups, players complement each other's probability of going (if one players goes often then the other player will tend to go rarely) and many dyads show turn-taking from early on. The multi-player groups also showed some turn-taking but less robustly than the dyads. Groups with the downwards treatment performed better at lower thresholds than groups with the upwards treatment. For this paradigm, if you want a group to show more cooperative, prosocial behavior in a situation requiring considerable restraint (i.e. usually not going to the bar), it is better to give them a situation requiring less restraint first. Further exploration is imperative to unravel the intricate web of motivational forces that underlie behavior in this paradigm.

REFERENCES

[1] Edgar J. Andrade-Lotero and Robert L. Goldstone. 2021. Self-organized division of cognitive labor. *PLOS-ONE* 16, 7 (2021).

[2] W. Brian Arthur. 1994. Inductive Reasoning and Bounded Rationality. *The American Economic Review* 84, 2 (1994), 406–411.

[3] Stefano Balietti. 2017. nodeGame: Real-time, synchronous, online experiments in the browser. *Behavior Research Methods* 49, 5 (2017), 1696–1715.

[4] Ann M. Bell and William A. Sethares. 2001. Avoiding global congestion using decentralized adaptive agents. *IEEE Trans. Signal Process.* 49 (2001), 2873–2879.

[5] Ann M. Bell, William A. Sethares, and James A. Bucklew. 2003. Coordination failure as a source of congestion in information networks. *IEEE Trans. Signal Process.* 51 (2003), 875–885.

[6] Giulio Bottazzi and Giovanna Devetag. 2003. A laboratory experiment on the minority game. *Physica A: Statistical Mechanics and its Applications* 324, 1 (2003), 124–132. Proceedings of the International Econophysics Conference.

[7] Damien Challet, Matteo Marsili, and Gabriele Maria Ottino. 2004. Shedding light on El Farol. *Physica A: Statistical Mechanics and Its Applications* 332 (2004), 469–482.

[8] Shu-Heng Chen and Umberto Gostoli. 2017. Coordination in the El Farol Bar problem: The role of social preferences and social networks. *Journal of Economic Interaction and Coordination* 12, 1 (2017), 59–93.

[9] D.B Fogel, K Chellapilla, and P. J Angeline. 1999. Inductive reasoning and bounded rationality reconsidered. *IEEE transactions on evolutionary computation* 3, 2 (1999), 142–146.

[10] Reiner Franke. 2003. Reinforcement learning in the El Farol model. *Journal of Economic Behavior and Organization* 51, 3 (2003), 367–388.

[11] Robert L. Goldstone, Edgar J. Andrade-Lotero, Robert Hawkins, and Michael Roberts. 2023. The emergence of specialized roles within groups. *Topics in Cognitive Science* (2023).

[12] Robert X. D. Hawkins and Robert L. Goldstone. 2016. The Formation of Social Conventions in Real-Time Environments. *PLOS ONE* 11, 3 (03 2016), 1–14.

[13] Cristina Ponsiglione, Valentina Roma, Fabiola Zampella, and Giuseppe Zollo. 2015. The Fairness/Efficiency Issue Explored Through El Farol Bar Model. In *Scientific Methods for the Treatment of Uncertainty in Social Sciences*, J. Gil-Aluja J., A. Terceño-Gómez, J. Ferrer-Comalat, J. Merigó-Lindahl, and S. Linares-Mustarós (Eds.). Vol. 377. Springer.

[14] Michael E. Roberts and Robert L. Goldstone. 2011. Adaptive Group Coordination and Role Differentiation. *PLOS ONE* 6, 7 (2011), e22377.

[15] Peter Vanderschraaf and Brian Skyrms. 2003. Learning to Take Turns. *Erkenntnis* 59, 3 (2003), 311–347. https://doi.org/10.1023/a:1026046625024

Averaging, Communication, and Wisdom of Crowds

Non-Archival Extended Abstract

Ulrike Hahn
Birkbeck College, Univ. of London
London, U.K.
u.hahn@bbk.ac.uk

Leon Assaad
Ludwig Maximilians Universitaet
Munich, Germany
L.Assaad@campus.lmu.de

Jason W. Burton
Copenhagen Business School
Frederiksberg, Denmark
jb.digi@cbs.dk

ACM Reference Format:
Ulrike Hahn, Leon Assaad, and Jason W. Burton. 2024. Averaging, Communication, and Wisdom of Crowds: Non-Archival Extended Abstract. In *Proceedings of (ACM CI'24).* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

It has long been known that communication within a group may, in fact, serve to reduce wisdom of the crowd effects because of the correlation an exchange may induce [11]. In response, an empirical and modelling literature has sought to determine the implications of this for how best to structure communication and communication networks to maximise collective performance [1, 3, 9, 12, 13]. In this paper, we use an agent-based model to clarify different contexts of opinion/judgment heterogeneity and their implications for the benefits that will accrue through averaging and likely costs of communication.

2 MODELS AND SIMULATIONS

We used a recent agent-based modelling framework called NormAN (short for 'Normative Argument Exchange across Networks') [2] that was created to facilitate agent-based models that involve the exchange of individual arguments across a social network instead of modelling communication simply as opinion averaging [6] or as a contagion process [5]. In this it follows on from earlier models of argument exchange such as [14]; however, it uses Bayesian agents situated in a ground truth world so that accuracy can be investigated [2].

The model involves an underlying 'world' (represented by a Bayesian Belief network) that is used to specify the true state of the target claim at issue and to initialise the arguments that are available in principle. At the start of a run, agents are randomly assigned individual pieces of evidence (the number of pieces of initial evidence is a free parameter in the model). Agents then use their own copy of the underlying Bayes' net to update their belief in the target hypothesis given the evidence they have received. When communicating with their neighbours—the probability of communication at each time step is again a free parameter in the model— agents choose a piece of evidence from their memory according to one of several communication rules (we only use a random selection in the current simulations), and communicate that piece of evidence faithfully to their neighbour(s). Exchanging evidence as arguments pulls the Bayesian agents to the 'optimal posterior', that is, the posterior given all, in principle, available pieces of evidence in the simulated world for a given run. Crucially, optimal posteriors and initial mean belief may significantly diverge.

2.1 Divergence Between Optimal Posterior and Initial Mean Belief

To give an indication of the prevalence of substantial divergence, we used one of the pre-defined 'world models' contained in NormAN version 1.0 ("big net") to simulate evidence distributions across populations (for detailed description of argument generation in NormAN, see [2]). The generic "big net world" consists of a target hypothesis and nine possible pieces of evidence regarding the truth or falsity of that hypothesis.

We used the model to simulate worlds in which agents are initialised with a random draw of either 1, 3, 5, or 7 pieces of evidence each, out of the possible total of 9. We simulated 200 instances of each for a total of 800 evidence distributions. We then identified, for each run, the lowest degree of belief amongst the agents in the set, the highest degree of belief given the initial evidence, the mean belief in the population of agents given their initial evidence, and the optimal posterior. Results are shown in Figure 1. The main takeaway from these results is how likely it is that the optimal posterior is *more extreme* than the most extreme initial belief. This has the inevitable consequence that the collective mean is less extreme. As a consequence, averaging will inevitably pull a sizeable chunk of the population *away* from the best estimate.

The discrepancy arises, fundamentally, because the integration of individual pieces of evidence gives rise to discontinuous jumps and changes in agents' overall belief. This means, for one, that the belief dynamics of simulated societies that exchange arguments are fundamentally different to those that will emerge in models based on averaging a single opinion [for extensive discussion see 2, 14, 15]. In particular, beliefs may readily become more extreme (either for the population as a whole, or for different subgroups) than the initial estimates—a dynamic that is antithetical to the notion of averaging itself [see 2, 8, 14]. The fundamental reason averaging fails is because the total evidential value of three independent pieces of evidence in favour of a claim is cumulative: each piece of evidence *adds* weight. The combined evidential value is not the average evidential value of these three individual arguments.

By the same token, the initial mean *would be* the optimal posterior (or very close to it) *only* if arguments for and against were roughly present in equal number and equal weight. For matters of fact, such a distribution will very much be the exception not the rule. This follows directly from a Bayesian perspective [see 8]. However, it can readily be appreciated with an intuitive example as well: if it is, in fact, raining outside at present, it is extremely unlikely that there would be equal amounts of evidence speaking for and against the presence of rain. We consider evidence to be evidence precisely because it has some degree of lawful relationship

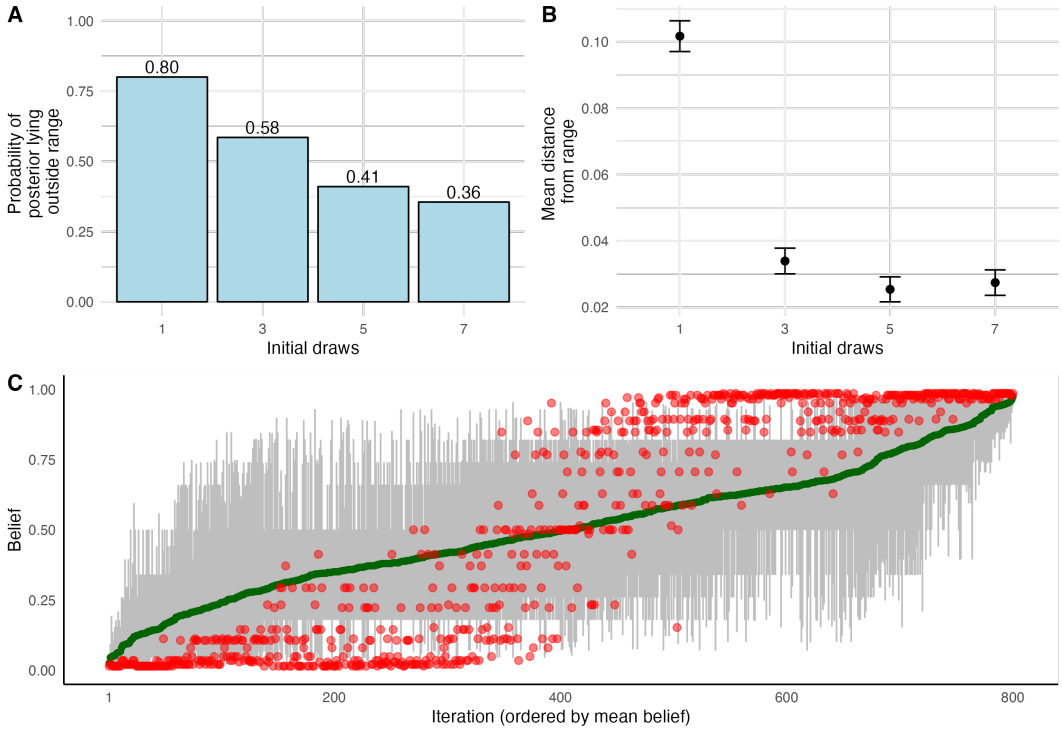


Figure 1: Results of simulations comparing optimal posteriors following argument exchange to agents' initial beliefs. A. The probability of the optimal posterior lying outside of the range of initial beliefs. Results are split by the number of initial draws parameter, which indicates whether the agents start with 1,3, 5, or 7 pieces of initial evidence (out of a possible 9). B. Mean distance of the optimal posterior from the range of beliefs following averaging (i.e., how far below the minimum belief or how far above the maximum belief the optimal posterior is if it is outside the range of initial beliefs). Results are again split by the number of initial draws parameter. C. All 800 runs (iterations) of our model, ordered by the initial mean belief, with each iteration represented as a grey vertical bar (the range of initial beliefs), a green point (the mean initial belief), and red point (optimal posterior following argument exchange).

to a hypothesis, so the true state of that hypothesis will, conversely, influence the (in principle) availability of evidence.

This means also that none of this hinges, fundamentally, on whether one is or is not adopting a Bayesian perspective, or modelling specifically Bayesian agents. It depends only on the way 'evidence' is related to the world [i.e., that a pregnancy test is considered evidence for pregnancy precisely (and only) because it is more likely to indicate 'pregnant' than 'not pregnant' when used by a pregnant woman] and how multiple pieces of evidence combine. A Bayesian perspective is simply a particular formalisation of these more fundamental points.

3 IMPLICATIONS FOR THE UTILITY OF AVERAGING, CROWD WISDOM, AND COMMUNICATION

Averaging does, in all cases of Fig. 1, improve accuracy with a 'wisdom of crowds effect' in as much as the squared error (to the optimal posterior) of the initial population mean is lower than the average individual accuracy in that individual population. This is because squared error, like other proper scoring rules such as the logarithmic scoring rule [4], penalises the long distances of the beliefs on the far side of the population mean more than it rewards short distances of the beliefs on the near side of the mean to the optimal value. Those gains, however, seem limited in light of the overall error.

To put all of this more generally, averaging boosts accuracy by eliminating variance. How useful that is depends on the source of that variance. Where it reflects random error, eliminating that variance will be extremely powerful and lead to estimates that are almost spookily accurate [7]. Where that variance reflects unique knowledge, by contrast, the effects will likely be far less profound. The discrepancy between the population average and the optimal posterior will, of course, decrease the more evidence individuals in the population have initially. Unfortunately, however, those will also be the circumstances in which the magnitude of the boost over average individual accuracy decreases.

This has implications for research on crowd wisdom and the benefits (or harms) of communication. Motivated by formal results [11], research has examined potential negative impacts of communication on crowd wisdom [10, 12, 13, 16]. The present results,

however, suggest that it would be a mistake to caution against communication as a threat to crowd wisdom in general.

The nature of the estimate, and with it the nature of the variance, matters fundamentally. Where fragmentation of evidence is the underlying source of variance, promoting the exchange of reasons seems quite likely to outweigh potential costs of dependence. Averaging fares poorly when agents' pre-deliberation beliefs are based on different pieces of specific evidence—especially if the body available evidence is unambiguously indicative of the hypothesis. These findings highlight the potential benefits of argument exchange on group accuracy and motivate further empirical and theoretical research into the merits and demerits of communication.

ACKNOWLEDGMENTS

The research reported in this paper was supported by the UK's Arts and Humanities Research Council grant AH/V003380/1, and a Mercator Fellowship to U.H. through Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 455912038. L.A. was supported by a Konrad-Adenauer Stiftung scholarship.

REFERENCES

[1] Abdullah Almaatouq, Alejandro Noriega-Campero, Abdulrahman Alotaibi, PM Krafft, Mehdi Moussaid, and Alex Pentland. 2020. Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences* 117, 21 (2020), 11379–11386.

[2] Leon Assaad, Rafael Fuchs, Ammar Jalalimanesh, Kirsty Phillips, Leon Schoeppl, and Ulrike Hahn. 2023. A Bayesian Agent-Based Framework for Argument Exchange Across Networks. *arXiv preprint arXiv:2311.09254* (2023).

[3] Jason W Burton, Abdullah Almaatouq, M Amin Rahimian, and Ulrike Hahn. 2024. Algorithmically mediating communication to enhance collective decision-making in online social networks. *Collective Intelligence* 3, 2 (2024), 26339137241241307.

[4] Arthur Carvalho. 2016. An Overview of Applications of Proper Scoring Rules. *Decision Analysis* 13, 4 (2016), 223 – 242.

[5] Damon Centola. 2018. *How behavior spreads: The science of complex contagions*. Vol. 3. Princeton University Press Princeton, NJ.

[6] Morris H DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical association* 69, 345 (1974), 118–121.

[7] Francis Galton. 1907. Vox populi (the wisdom of crowds). *Nature* 75, 7 (1907), 450–451.

[8] Ulrike Hahn. 2023. Individuals, collectives, and individuals in collectives: the in-eliminable role of dependence. *Perspectives on Psychological Science* (2023).

[9] Ulrike Hahn, Jens Ulrik Hansen, and Erik J Olsson. 2018. Truth tracking performance of social networks: how connectivity and clustering can make groups less competent. *Synthese* (2018), 1–31.

[10] Ulrike Hahn, Momme von Sydow, and Christoph Merdes. 2019. How Communication Can Make Voters Choose Less Well. *Topics in Cognitive Science* 11, 1 (2019), 194–206.

[11] Robin M Hogarth. 1978. A note on aggregating opinions. *Organizational Behavior and Human Performance* 21, 1 (1978), 40–46.

[12] Martin L Jönsson, Ulrike Hahn, and Erik J Olsson. 2015. The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition* 142 (2015), 191–204.

[13] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* 108, 22 (2011), 9020–9025.

[14] Michael Mäs and Andreas Flache. 2013. Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PloS one* 8, 11 (2013), e74516.

[15] Carlo Proietti and Davide Chiarella. 2023. The Role of Argument Strength and Informational Biases in Polarization and Bi-Polarization Effects. *Journal of Artificial Societies and Social Simulation* 26, 2 (2023).

[16] Kevin JS Zollman. 2010. The epistemic benefit of transient diversity. *Erkenntnis* 72, 1 (2010), 17–35.

The Epistemic Potential of Adversarial Cooperation

On Adversarial Cooperation

Almos C. Molnar

Brown University, almos_molnar@brown.edu

The literature on collective reasoning shows mixed results: Groups often outperform even the most competent individuals on some tasks. But, on occasion, groups perform even worse than the average individual. What determines group performance? I argue that the most effective driver of epistemic success in collectives is a self-organized group dynamic called "adversarial cooperation". Adversarial cooperation is defined as rival factions exchanging arguments and evidence with one another with the common aim of getting closer to the truth. Its unique identifier is an *a priori* division of cognitive labor based on complementary myside biases between those who develop and argue for a proposition ("conjecturers") and those who criticize and stress test it ("refuters"). This paper aims to 1) explain the theoretical underpinnings of this construct, 2) provide some of the key empirical evidence for it, and 3) outline directions for future experimental work.

Keywords and Phrases: collective reasoning, theory of adversarial cooperation, myside bias, division of cognitive labor, critical rationalism

1. INTRODUCTION

The literature on collective reasoning shows mixed results: Groups often outperform even the most competent individuals on some tasks. But, on occasion, groups perform even worse than the average individual [1] What determines group performance? In the past few decades, several important factors have been identified, including solution demonstrability and member competence [2], independent judgment aggregation [3], majority rule [4], and cognitive diversity [5]. I argue, however, that the most effective driver of epistemic success in collectives is a self-organized group dynamic that I call "adversarial cooperation". The aim of this paper is to explain this theoretical construct (Section 2), provide some of the key empirical evidence for it (Section 3), and outline directions for future experimental work (Section 4).

2. THE THEORY OF ADVERSARIAL COOPERATION (TAC)

When rival factions exchange arguments and evidence with one another with the common aim of getting closer to the truth, they are engaging in adversarial cooperation. The unique identifier of this collective dynamic is an *a priori division of cognitive labor* between those who develop and argue for a proposition ("conjecturers") and those who criticize and stress test it ("refuters"). This division of cognitive labor is not imposed top-down, but rather emerges dynamically: Members adopt these complementary, specialized roles based on their individual motivations [6]. It is such differential myside biases what ensure that the space of arguments both for and against a theory/proposition is better covered than how it otherwise would be.

Adversarial cooperation, as a theoretical construct, is predominantly inspired by the epistemology of Karl Popper, namely critical rationalism. For a neo-Popperian account of human reasoning, see [7]. In contrast to many influential epistemological frameworks, such as Bayesianism, critical rationalism regards the epistemic progress to be first and foremost a social affair rather than an intrapersonal one, best attained by the method of critical discussion (alternatively described as the dynamic of "conjectures-and-refutations") [8].

The *conjectures* component of this bipartite model refers to the fact that, first, theories or arguments have to be devised by way of guesswork. Popper likened this part of the epistemic process to myth-making and story-telling where there are no restrictions on the methods to be used or the possible sources of inspiration [9]. In this sense, critical rationalism emphasizes the roles that creativity and epistemic luck plays: Among the endless blind conjectures, one may happen to get it right (or more precisely, slightly righter) by sheer happenstance.

While conjecturing has a strong subjective component, the end products of the process are categorically objective: Theories or arguments have a linguistic form, and hence, they can be instantiated in public representations, such as utterances and writing. When they are, these human artifacts become subject to the same evolutionary selection pressures as any other animal artifacts found in the natural environment, such as nests, dams, and honey [10]. In the case of theories, this selection pressure comes primarily in the form of other people, and more broadly, from culture.

This is what the second part of the epistemic process, namely *refutations*, refer to. Cultural institutions that care about epistemic success, such as science and democracy, have evolved traditions of criticism, where it is considered normative to criticize and stress test ideas in

order to discover their true worth [8]. Such attempted refutations, whether empirical or theoretical, are primarily provided by people who have a personal stake in the domain of life to which the idea pertains and disagree with the theory in question. This is what ensures high quality control over the proliferation of conjectures. The outcome of this process is that those conjectures that survive attempted refutations preferentially spread in the natural environment (by way of further public representations) compared to refuted alternatives.

Therefore, the "objectivity of science is not a matter for the individual scientist but rather the social result of […] the friendly-hostile division of labour among scientists, of their co-operation and also of their competition" [11, p. 72). It is important to emphasize that within this collective dynamic there is a "methodological justification for individual scientists to be biased" one way or another, because it is only when theories are both tenaciously defended by some and tenaciously criticized by others that we can learn their real power [9, p. 93]. If there is not enough criticism, then epistemically unreliable ideas can linger on and cause problems. If there is not enough defense, then epistemically highly valuable ideas can die out before ever completely revealing their true merit. In each case, the tenacity needed is best attained by a certain level of personal bias, and even, affective involvement, on behalf of the disagreeing parties, rather than real or pretended dispassion or neutrality.

3. THE EMPIRICAL EVIDENCE

3.1. Myside bias supports better conjectures

In Molnar and Sloman [2024a], we tested TAC's prediction that complementary myside biases are collectively optimal for conjecturing (i.e. the first component of the epistemic process). In testing this prediction, we relied on prior research that shows that Democrats and Republicans in the US have divergent opinions about which social and economic issues are zero-sum. We hypothesized that merely framing a non-zero-sum problem to be solved in terms of one of these issues (e.g. immigration or corporate tax cuts) can lead members of each political group to qualitatively different solutions. We conjectured that prior beliefs about the problem being zero-sum will hinder problem-solving ability and that prior beliefs about the problem being non-zero sum will boost it.

Across four experiments, using two unique economic games that require a creative approach rather than mere numerical reasoning, we found empirical evidence for this phenomenon. We demonstrated that Democrats conjecture less effective solutions to a problem when it is framed in terms of corporate tax cuts, and more effective ones (e.g. win-win solutions) when the problem is framed in terms of pro-immigration policies, than when the formally identical problem is framed non-politically. Republicans exhibited the opposite pattern. Thus, we found differential conjecturing performance between the two political groups only in the two politicized problem frames, and in the opposing directions.

Crucially, we also showed that while political neutrals (i.e. politically non-affiliating and ideologically moderate individuals) perform equally well as Democrats and Republicans in the non-politicized problem frame, they perform worse than Democrats in the pro-immigration frame and worse than Republicans in the corporate tax cuts frame. Put differently, being biased towards an idea can often better facilitate the conjecturing of creative solutions in service of that idea than individual neutrality.

3.2. Myside bias supports better refutations

In Molnar and Sloman [2024b], we tested TAC's prediction that complementary myside biases are collectively optimal for refutations (i.e. the second part of epistemic process). For this, we use politicized versions of the Wason Selection Task and the Reduced Array Selection Task, to which the correct solutions require insight into the logic of scientific hypothesis testing, namely *modus tollens* [8].

We showed that both Democrats and Republicans are more likely to propose empirical tests that can only result in confirming evidence when they have to evaluate a public policy claim they favor, and more likely to propose empirical tests that can result in disconfirming evidence when they have to evaluate a public policy claim they disfavor, than when they have to evaluate a neutral public policy claim in the formally identical way. We also showed that the political adversaries of the policies are not just more likely to propose the normatively correct solution (i.e. the falsificationist response) on the tasks than political allies, but also more likely to provide the correct reasoning for their choice (i.e. modus tollens).

Once again, we found that political neutrals perform worse than the Republicans when they have to evaluate a policy claim towards which the Democrats are biased and worse than the Democrats when the have to evaluate a policy claim towards which the Republicans are biased. In other words, the space of relevant evidence for and against a hypothesis is collectively better covered in the presence of complementary myside biases than individual impartiality

4. OTHER EMPIRICAL PREDICTIONS AND FUTURE WORK

TAC makes several distinct predictions from previous scholarly approaches to explain collective reasoning performance:

- The truth (or truthlikeness) gradually emerges and is selected for as a natural product of adversarial cooperation, rather than being present at the outset. As such, the upper limit of group performance is not tied to its most competent member (*pace* [2])
- Most conjectures are wrong, and we should expose these and let them die instead of weighing all conjectures equally (*pace* "miracle of aggregation" or "wisdom of crowds" approaches, such as [3]).
- Democratic governance is epistemically superior to alternative political systems not because it relies on majority rule, but because it uniquely institutionalizes adversarial cooperation (*pace* Condorcet and other jury theorists; see [4]).
- It is not so much cognitive diversity, but sustained disagreement and its attempted rational resolution by way of exchanging arguments and evidence that has unique epistemic potential (*pace* [5]).

In other words, TAC predicts that group deliberation as a collective action mechanism outperforms (a) the most competent individuals, and (b) any other groups that rely on independent judgment aggregation or majority rule, provided that there is disagreement among group members (i.e. the competitive element), and group members share a common aim to find effective solutions (i.e. the cooperative element).

I am currently in the process of testing these predictions involving small groups of political partisans tasked with reasoning together. The political composition of the groups is manipulated so that there are politically homogenous groups (i.e. made up of either only Democrats or only Republicans), politically adversarial groups (i.e. made up of an equal number of Democrats and Republicans), and politically diverse groups (i.e. made up of members who are all ideologically unique). I also manipulate the collective action mechanism by which the groups have to decide. There are three conditions: Groups decide by 1) deliberation, 2) independent judgment aggregation, or 3) majority rule.

I predict that politically adversarial groups that decide via deliberation will outperform 1) the politically homogeneous groups and the politically diverse groups as well as 2) the other politically adversarial groups that decide via one of the other collective action mechanisms that does not involve deliberation (i.e. independent judgment aggregation or majority rule). I predict that this will hold true regardless of the type of reasoning tasks employed.

ACKNOWLEDGMENTS

The author would like to thank Steven Sloman for providing valuable comments and helpful suggestions. The work is supported by the Foundation of Individual Rights and Expression's Free Inquiry Grant Program under Grant No. 009491.

REFERENCES

[1] Norbert L. Kerr and R. Scott Tindale. 2004. Group performance and decision making. Annu. Rev. Psychol., 55 (February 2004), 623-655. https://doi.org/10.1146/annurev.psych.55.090902.142009

[2] Patrick R. Laughlin & Alan L. Ellis. 1986. Demonstrability and social combination processes on mathematical intellective tasks. Journal of Experimental Social Psychology, 22 (May 1986), 177–89. https://doi.org/10.1016/0022-1031(86)90022-3

[3] James Surowiecki. 2005. The wisdom of crowds. Anchor.

[4] Helene Landemore. 2012. Democratic reason: Politics, collective intelligence, and the rule of the many. Princeton University Press.

[5] Scott Page. 2007. The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition. Princeton University Press.

[6] Robert L. Goldstone, Edgar J. Andrade-Lotero, Robert D. Hawkins, and Michael E. Roberts. 2023. The Emergence of Specialized Roles Within Groups. Topics in Cognitive Science 16, 2 (February 2023), 257-281. https://doi.org/10.1111/tops.12644

[7] Almos C. Molna. 2024. Human Reasoning: A Neo-Popperian Approach. Manuscript submitted for publication.

[8] Karl Popper. 1963. Conjectures and refutations: The growth of scientific knowledge. Routledge.

[9] Karl Popper. 1994. The myth of the framework: In defence of science and rationality. Psychology Press.

[10] Karl Popper. 1972. Objective knowledge: An evolutionary approach. Oxford: Clarendon press.

[11] Karl Popper. 1984. In search of a better world: Lectures and essays from thirty years. Routledge.

[12] Almos C. Molnar and Steven A. Sloman. 2024a. Politicization Induces Zero-Sum Thinking in Problem Solving. Manuscript submitted for publication.

[13] Almos C. Molnar and Steven A. Sloman. 2024b. Public Policies Are Best Evaluated by Their Political Adversaries. Manuscript in preparation.

Nuanced Effect of Human Trust in AI on Human-AI Collaboration Performance

Danushka Bandara
dbandara@fairfield.edu
Fairfield University
Fairfield, Connecticut, USA

Robert Dillon
robert.dillon@student.fairfield.edu
Fairfield University
Fairfield, Connecticut, USA

Noor Khattak
noor.khattak@student.fairfield.edu
Fairfield University
Fairfield, Connecticut, USA

ABSTRACT

In this study, we investigate the impact of trust on collaborative performance in human-AI teams, focusing on the task of image authenticity judgment. Utilizing the CASIA dataset, we conducted a human subject experiment (n=18), assessing their trust levels in AI systems and their collaborative performance in image tampering detection tasks. Participants underwent trials where they evaluated images from the dataset while considering AI output decisions, varying in accuracy. We observed that trust significantly influences collaborative performance. When participants entrusted decision-making to AI predominantly, higher trust levels correlated with superior collaborative performance, surpassing expected performance by 247%. Conversely, in scenarios where participants relied more on human judgment, lower trust in AI led to heightened collaborative performance, exceeding expectations by 406.7%. Notably, when decision-making involved a combination of human and AI inputs, lower trust in AI was associated with a notable performance increase of 467%. These findings underscore the nuanced relationship between trust and performance in human-AI collaboration.

CCS CONCEPTS

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

KEYWORDS

Human-AI interaction, trust, performance

ACM Reference Format:
Danushka Bandara, Robert Dillon, and Noor Khattak. 2024. Nuanced Effect of Human Trust in AI on Human-AI Collaboration Performance. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Human trust in AI is a critical factor in determining the success of human-AI collaboration. It is important to understand when to trust or distrust AI to effectively leverage human expertise, particularly in situations where AI performance may be suboptimal. Additionally, human trust in AI is often overlooked, with a predominant focus on the machine's trustworthiness and performance. Recently, several objective methods of measuring human trust has been developed [1]. Making it a usable metric in Human Computer Interaction. Therefore it is important to understand the role that trust plays in decision making and performance.

We study the human AI system performance on the task of judging the authenticity of an image. Here we chose an existing doctored and non doctored image dataset named CASIA [2]. See an overview of the task presented in Figure 1, wherein we present the subjects with an image and obtain their response regarding whether they accept or reject the AI decision. The AI decision for the image was presented as: 'The AI system determined the above image is authentic/non-authentic. The 'AI' we used was simulated in a 'Wizard of Oz' method. Meaning the subjects were not aware of it being simulated.

The key findings of this study can be summarized as follows:

- Human AI collaboration does not always increase the performance of detecting image tampering. Especially when the AI performance is low.
- Human trust in AI improves the performance of the collaborative system only if the human mainly relies on the AI for decision-making.

2 METHODOLOGY

2.1 Images dataset

The CASIA dataset [2], short for the Chinese Academy of Sciences Institute of Automation (CASIA) dataset, is a collection of images commonly used in research related to computer vision, image processing, and machine learning. This dataset comprises a diverse range of images, including both manipulated (or "doctored") and authentic (non-doctored) images. These images are sourced from various sources and encompass a wide array of scenes, objects, and subjects. The CASIA dataset is often used in studies focusing on image manipulation detection, forensic analysis, and related fields.

2.2 Human subject experiment

In this human subject experiment, 18 participants aged 18-30 from a university in the northeastern United States underwent a series of image judgment tasks. Before the experiment, participants completed a pre-experiment questionnaire adapted from the Propensity to Trust Inventory [3] to assess their baseline trust in computers. Seated in front of a computer, participants were presented with consecutive images from the CASIA dataset and tasked with judging their authenticity, with 10 seconds of rest between trials. Using a "Wizard of Oz" methodology, participants were exposed to various simulated AI responses to the images, showcasing a range of AI behaviors and accuracy levels. Throughout the experiment, a trust measurement survey was administered to gauge subjective trust

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

levels in the AI system. Data collected included participants' image judgments, corresponding AI responses, and survey responses regarding trust. Analysis of the collected data aims to uncover the interplay between participants' baseline trust, observed AI responses, and subjective trust levels, shedding light on human-computer trust dynamics in AI-assisted tasks.

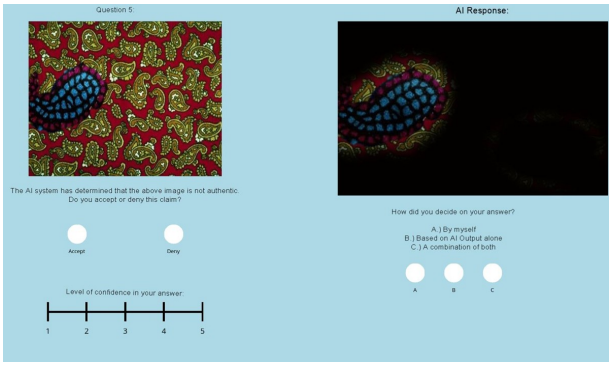


Figure 1: A testbed question which constitutes a single trial in our experiment.

3 DATA COLLECTION

The images were presented to the subjects using the Psychopy [4] platform. The subjects provided self-report responses after every image task to indicate the following,

- their acceptance/rejection of AI response
- their confidence in the acceptance/rejection decision
- whether the decision was made mainly by themselves/mainly using the AI response or a combination of both
- whether they trusted the AI during the current image task.

The subject responses were saved to a CSV file by the test bed for further analysis. The saved data also included whether the images presented were doctored/not doctored and whether the AI response presented to the participant was correct/incorrect.

3.1 Data analysis

We hypothesized that the system accuracy would be higher than the AI accuracy. However, as seen from the figure 2, this is not the case for every subject. Next, we wanted to delve more into the nuances of the performance differences. Particularly, we were interested in how the self-reported trust level of the participants affected the collaboration performance. Taking the group as a whole, we found that the average collaboration performance (accuracy) was 60% when the users reported mistrust of the AI, and 67% when they reported trust of the AI (p=0.087). Since this effect was not very significant, we explored the trust effect in more detail as described below.

Participants in this study did not make their decisions purely using the AI response, they could also use their own judgment or a combination of AI and their judgment to come to a decision. This was indicated in the self-report surveys administered during each trial. The next analysis is based on this aspect. We wanted to test how trust affected the performance when we stratified the

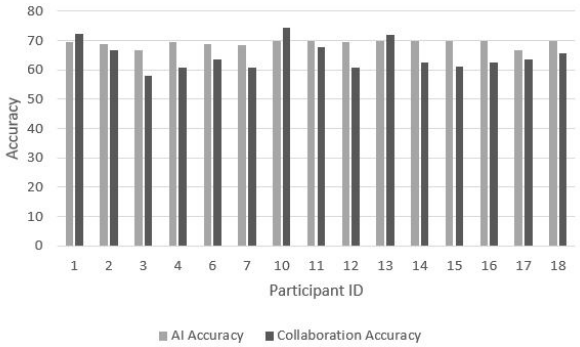


Figure 2: Variation of AI Accuracy and Collaboration Accuracy. The AI's accuracy is the number of correct responses provided by the AI. Collaboration accuracy is the number of correct responses provided by the system.

subject responses by the decision-making process of the subject. i.e. whether they made the decision mainly by themselves, mainly using AI, or a combination of both. We calculated the performance as the number of images that were correctly identified. The accuracy results for when the decision was made mainly by the participant are shown in table 1 and table 2. As seen in the tables, the collaboration accuracy depends on the AI accuracy, therefore doing a direct comparison between the collaboration performance between trust and mistrust conditions is not possible. To overcome this, we developed a model to obtain the expected collaboration accuracy.

Table 1: AI accuracy and collaboration accuracy for the mistrust condition when decision made by participant

Subject_ID	AI accuracy	Collaboration Accuracy
1	66.67	100
2	71.43	57.14
3	62.5	100
4	69.23	46.15
6	55	55
7	60	55
10	76.92	53.85
11	68.18	68.18
12	63.64	72.73
13	71.43	62.86
14	60.71	57.14
15	66.67	33.33
16	71.43	38.10
17	56.52	47.83
18	53.33	46.67

We found that the collaboration accuracy is linearly related to the AI accuracy (see figure 3). Using this linear model, we calculated the expected collaboration accuracy for each of the AI accuracy values. Then we checked how the average true collaborative accuracy compares to the expected collaborative accuracy. The results of this analysis are shown in figure 4.

Table 2: AI accuracy and collaboration accuracy for the trust condition when decision made by participant

Subject_ID	AI accuracy	Collaboration Accuracy
1	70.00	70.00
2	60.00	60.00
3	67.86	57.14
4	65.00	60.00
6	60.71	60.71
7	72.73	63.64
10	60.00	76.67
11	70.59	64.71
12	62.50	56.25
13	72.73	78.79
14	53.85	53.85
15	65.22	65.22
16	75.00	75.00
17	64.71	47.06
18	55.00	40.00

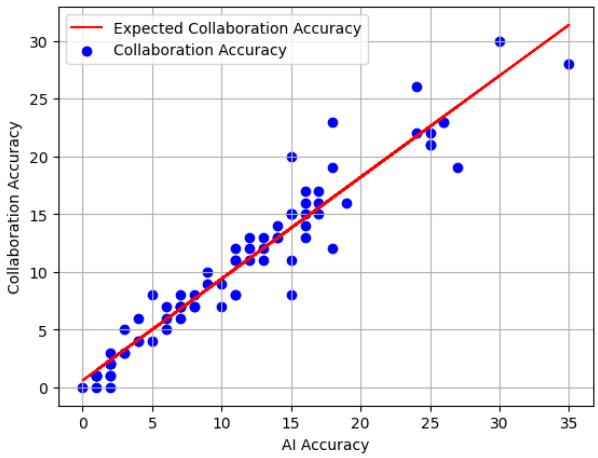


Figure 3: Linear regression model of expected collaboration accuracy based on AI accuracy.

4 DISCUSSION

We found that the average collaborative performance was higher than expected performance when they trusted the AI (247% higher) when the decision was made mainly using AI. However, when they made the decision mainly by themselves, the average performance was higher than expected (406.7% higher) when they did not trust the AI. In the case that the decision was made with a combination of AI and themselves, the performance was higher than expected (467% higher) when they did not trust the AI. This indicates that trust does have an effect on overall human-AI collaborative performance. When the trust level is lower, the performance is higher in cases where the humans make the decision themselves or with AI assistance. However, when they make decisions based mainly on AI, the higher trust levels result in higher performance. Our results show that there is a nuanced difference when it comes to the relationship

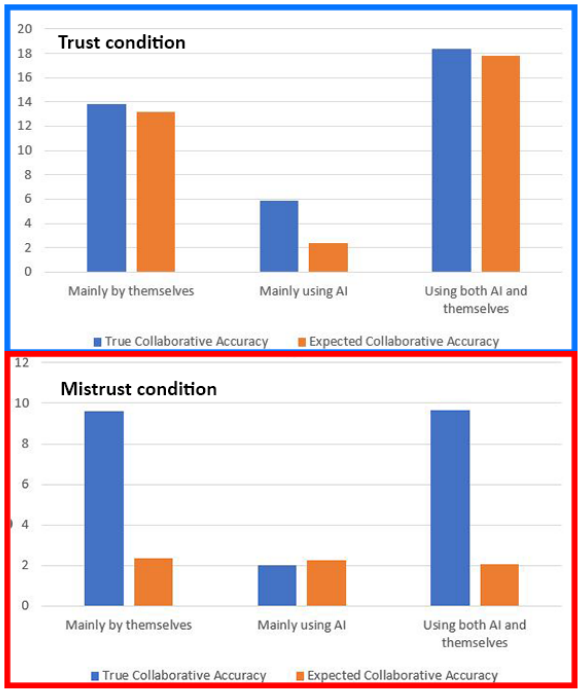


Figure 4: Comparison of true collaboration performance vs expected collaboration performance for the trust(and mistrust conditions

between trust and performance in human-AI teams. As decisions are made with more AI participation, higher trust tends to increase performance. When decisions are made with less AI participation, lower trust tends to increase performance. Of course our results are grounded in the context of image tampering as our experiment was limited to this task. Follow up studies could further explore the generalizability of our results. In essence, our findings underscore the significance of trust dynamics in shaping the performance of human-AI collaborative teams. The results indicate that the level of trust individuals place in AI systems significantly influences collaborative outcomes. Higher levels of trust in AI tend to elevate performance, particularly in scenarios where decision-making is predominantly AI-driven. Conversely, lower levels of trust in AI correlate with heightened performance when decision-making relies more on human judgment. Our study highlights the nuanced and nature of the relationship between trust and performance in human-AI collaboration, emphasizing the importance of understanding and managing trust dynamics to optimize collaborative outcomes.

REFERENCES

[1] D. Bandara and S. Sau. 2023. Are Scrutiny and Mistrust Related? An Eye-Tracking Study. In *International Conference on Human-Computer Interaction*. Springer Nature Switzerland, Cham, 539–545.

[2] Jing Dong, Wei Wang, and Tieniu Tan. 2013. CASIA Image Tampering Detection Evaluation Database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE. https://doi.org/10.1109/chinasip.2013.6625374

[3] Mark L. Frazier, Paul D. Johnson, and Stav Fainshmidt. 2013. Development and Validation of a Propensity to Trust Scale. *Journal of Trust Research* 3, 2 (2013), 76–97.

[4] Jonathan W. Peirce. 2007. PsychoPy—Psychophysics Software in Python. *Journal of Neuroscience Methods* 162, 1–2 (2007), 8–13.

AI-in-the-Loop Expert Labeling for LLM's Real-World Domain Adaptation

Bingsheng Yao
b.yao@northeastern.edu
Northeastern University
Boston, USA

Dakuo Wang*
d.wang@northeastern.edu
Northeastern University
Boston, USA

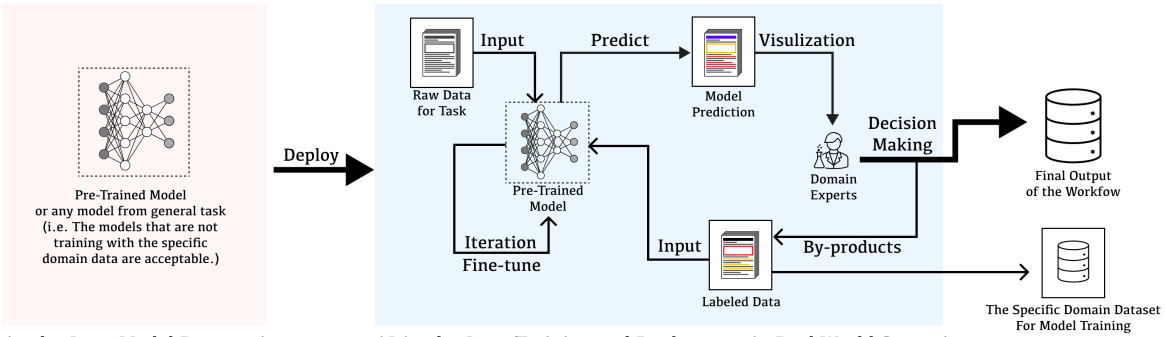


Figure 1: The proposed AI-In-The-Loop architecture for data annotation that centers on human experts' existing workflow.

ABSTRACT

The real-world success of pre-trained Large Language Models (LLMs) depends on how well they can adapt to the specific domain. For domain adaptation, LLMs either need to fine-tune with massive labeled data, or dynamically incorporate user feedback via Reinforcement Learning or Active Learning (AL) strategies. However, domain-expert-labeled large datasets are almost impossible in the real world, and data drifting happens (labeled data scarcity). We propose an AL-augmented LLM (AL-LLM) and an AI-in-the-loop (AITL) hosting architecture to handle the real-world challenges of labeled data scarcity and domain adaptation. AL-LLM has two AI modules (an Active In-Context Sampling (AICS) LLM, and an AL-based compact model) and a dynamic switching mechanism. The AITL architecture, differing from human-in-the-loop, centers on the human experts' existing workflow so that human experts do not feel themselves as data labeling slaves as they provide continuous human feedback to update the LLM in post-deployment.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**; **Human computer interaction (HCI)**.

*Corresponding author.

Permission to make digital or hard copies of part or all 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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

KEYWORDS

Expert Data Annotation, AI-in-the-Loop, Human-Centered AI Collaboration, Low-Resource, Large Language Model, Domain Adaptation

ACM Reference Format:
Bingsheng Yao and Dakuo Wang. 2024. AI-in-the-Loop Expert Labeling for LLM's Real-World Domain Adaptation. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 3 pages.

1 MOTIVATION AND BACKGROUND

The recent advances in Large Language Models (LLMs) have demonstrated profound world knowledge and exceptional task-solving capabilities including reasoning, arithmetic, and comprehension [2, 7, 10]. However, LLM's success has not yet been generalized into real-world business scenarios, mainly because there lacks an efficient training or robust hosting mechanism for LLMs to customize to the particular domain knowledge and adapt to the data drift. For instance, in clinical scenarios, the multi-modality of patient data, including imaging, time-series sensors, and textual descriptions are all analyzed by clinicians to facilitate their clinical decision-making process. For each disease and clinic (or even an individual doctor), their domain knowledge could vary a lot; thus, the LLM-based system has to be able to efficiently and dynamically adapt to domain contexts.

Traditional LLM domain adaptation methods typically require a huge amount of labeled data to fine-tune the model, or a small amount of labeled data to perform in-context learning (ICL) via prompt engineering [2]. However, for LLM fine-tuning, human experts' data labeling is hardly accessible at scale because of the high cost of hiring experts, their already overwhelmingly day jobs,

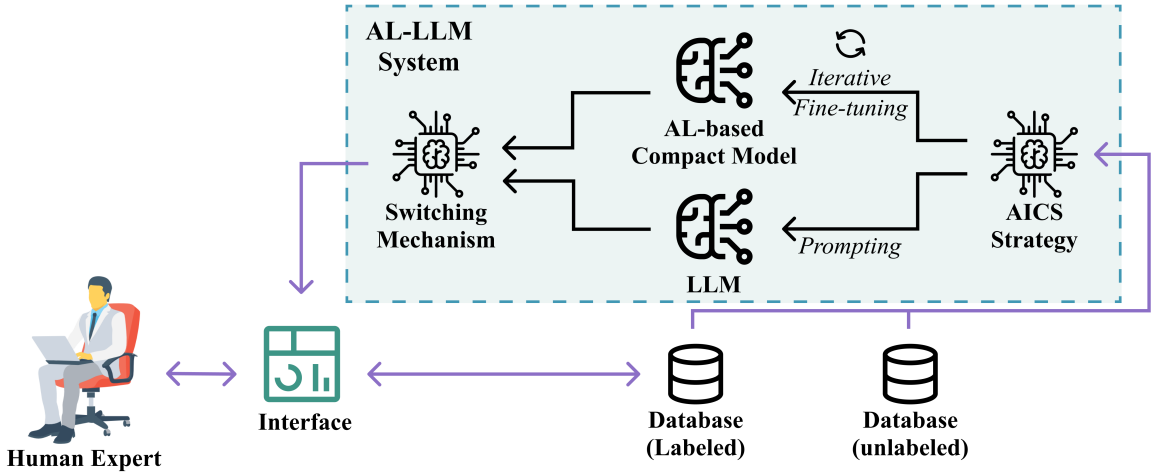


Figure 2: The proposed AL-LLM system overview. AL-LLM has two AI modules (an Active In-Context Sampling (AICS) LLM and an AL-based compact model) and a dynamic "switching" mechanism.

and their unwillingness to be the "data slave" for AI models [6, 11]. The ICL mechanism falls short as it tries to leverage the limited prompt input length to control the pre-trained LLM to focus on a highly specific domain (e.g., clinical). **How can we train and host a pre-trained LLM to perform well in any given new real-world domain (with limited or no labeled data)?**

In this work, we are inspired by Active Learning (AL) — a key technique in low-resource data annotation [8]. We propose to develop an AL-augmented LLM strategy (AL-LLM) and a novel AI-in-the-loop (AITL) system architecture to handle the limited labeled data, data drift, and low user acceptance challenges in real-world LLM deployment. **The AL-LLM algorithm** has two AI modules (an Active In-Context Sampling (AICS) LLM [12], and an AL-based compact model) and a dynamic switching mechanism to switch between these two modules in real-time. The two independent AI modules can continuously learn from human expert annotations at different stages, and the switching mechanism determines which module's suggestion to display at different stages. **The AITL architecture**, differing from the human-in-the-loop architecture, centers on the human experts' existing workflow so that humans do not feel themselves as data labeling slaves for the AI model. Human experts continuously provide critical data labeling as well as feedback towards model inference without realizing it — the whole human feedback flow is integrated into these human experts' existing workflow and existing user interface seamlessly, thus the data labeling task is invisible to them.

2 SYSTEM DESIGN

Recent research demonstrated that AL-based compact models can reliably outperform state-of-the-art LLMs (i.e., GPT-4 [1]) in real-world domain-specific tasks, but they may suffer from "cold-start" problems [5] — the early iterations of AL-based models' performance could be unstable and unsatisfactory [3]. Thus, we propose the AL-augmented LLM, a promising paradigm that leverages LLMs and AL-based compact models to complement each other.

2.1 AL-LLM Algorithm

Our AL-LLM has two AI modules: an LLM and an Active Learning (AL)-based compact model. Both modules will incorporate a novel Active In-Context Sampling (AICS) strategy, where the LLM module does not update model parameters, but the AL-based compact model will be iteratively fine-tuned on the data annotated by human experts.

Active In-Context Sampling (AICS). Recent exploration in optimizing LLMs' In-Context Learning (ICL) performance includes improving reliability and reducing sensitivities to prompt variations. For instance, fine-tuning the LLM with ICL format [9], and better selecting ICL examples [4, 13]. Yao et al. [12] proposed an In-Context Sampling (ICS) prompting mechanism that encompasses three steps: sample demonstration candidates, augment in-context learning (ICL) prompt inputs, and vote for the most reliable prediction. This approach aims to mitigate the limitations associated with prompt dependency and improve the overall robustness and reliability of AI-assisted annotations for LLMs. Nevertheless, this approach does not tune the model parameters and, as a result, falls short of resolving LLMs' lack of domain expertise. To bridge the gap, we propose a novel **Active In-Context Sampling strategy (AICS)** that supports both fine-tuning and prompting without finetuning. Traditional compact models can be fine-tuned with AICS whereas generic LLMs can either be fine-tuned or just prompting without fine-tuning. We will utilize a carefully designed AL sampling strategy that can efficiently and reliably select the most representative and potentially helpful data examples, and then construct multiple ICL prompt inputs for the same data, similar to the vanilla ICS mechanism. The two independent AI modules can continuously learn from human expert annotations at different stages. When it is early stage, the AICS LLM module performs better so it is chosen to provide prediction to the human experts, while the AL-based compact model is continuously fine-tuned; when passing the "cold start" stage, the AL-based compact model outperforms LLM, the system

switches; and lastly, when there is enough data, we fine-tune the LLM and reiterate the stages.

Switching Mechanism. Existing work illustrates that LLMs yield reliable and satisfactory results at early stages whereas the AL-based compact model can surpass LLMs in real-world tasks eventually [5]. At first, the LLM's prediction will be presented to the human expert, and the collected annotation will be used to train the AL model. When the AL model begins to outperform LLM, the mechanism will "switch" to present the AL model's prediction. Thus, the LLM's prediction can help overcome AL's "cold-start" problem and the AI-assisted tool can still benefit from continually improving performance. The "switch" mechanism efficiently and continually evaluates both modules in real-time.

2.2 AITL Architecture

The traditional training and deployment split is no longer valid given that pre-trained LLMs are already very powerful. Users, such as organizations, do not have enough resources to train LLMs. To better align LLMs with human experts' needs and preferences, we need to design an effective LLM hosting architecture that can provide a seamless and productive user experience, as shown in Fig.1. AITL architecture hides the hosting LLM behind the user interface that the domain experts are already using so that the end users are not even aware of its existence. When the LLM's performance is unsatisfactory, the model prediction is not shown to the domain expert, thus the users continuously perform their task independently as if they were doing today; only when the LLM's performance is good enough, it will show a subtle visualization on top of the existing UI to support expert's decision making. Any of the domain expert's implicit actions (e.g., ignore or approval) will be stored and automatically feedback into the model fine-tuning iteration.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan L Yuille, and Zongwei Zhou. 2022. Making your first choice: To address cold start problem in vision active learning. *arXiv preprint arXiv:2210.02442* (2022).

[4] Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified Demonstration Retriever for In-Context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4644–4668. https://doi.org/10.18653/v1/2023.acl-long.256

[5] Yuxuan Lu, Bingsheng Yao, Shao Zhang, Yun Wang, Peng Zhang, Tun Lu, Toby Jia-Jun Li, and Dakuo Wang. 2023. Human Still Wins over LLM: An Empirical Study of Active Learning on Domain-Specific Annotation Tasks. *arXiv preprint arXiv:2311.09825* (2023).

[6] Christoffer Bøgelund Rasmussen, Kristian Kirk, and Thomas B Moeslund. 2022. The challenge of data annotation in deep learning—a case study on whole plant corn silage. *Sensors* 22, 4 (2022), 1596.

[7] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207* (2021).

[8] Burr Settles. 2009. Active learning literature survey. (2009).

[9] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models Are Zero-Shot Learners. In *International Conference on Learning Representations*.

[10] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).

[11] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381.

[12] Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Sijia Liu, James Hendler, and Dakuo Wang. 2023. More Samples or More Prompt Inputs? Exploring Effective In-Context Sampling for LLM Few-Shot Prompt Engineering. *arXiv preprint arXiv:2311.09782* (2023).

[13] Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active Example Selection for In-Context Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9134–9148. https://doi.org/10.18653/v1/2022.emnlp-main.622

When Are Human-AI Groups Useful?

MICHELLE VACCARO, ABDULLAH ALMAATOUQ, AND THOMAS MALONE, MIT

1. INTRODUCTION

People increasingly work with artificial intelligence (AI) tools to accomplish goals in fields including medicine, finance, and law, as well as in daily activities such as traveling, shopping, and communicating. These human-AI systems possess tremendous potential given the complementary nature of humans and AI—the general intelligence of humans allows us to reason about diverse problems, and the computational power of AI systems allows them to accomplish specific tasks that people find difficult. In fact, a large body of work points to the potential of integrating AI into workflows to improve outcomes in multiple domains [Liu et al. 2023; Wamba-Taguimdje et al. 2020; Schwalbe et al. 2018]. On the other hand, a growing number of studies reveals that human-AI collaboration does not necessarily lead to better results when compared to the human or AI alone [Bansal et al. 2021; Buçinca et al. 2020; Lai et al. 2020; Zhang, Liao, and Bellamy 2020; Vaccaro and Waldo, 2019]. These seemingly contradictory results raise the question: When do humans and AI complement each other and when do they not?

2. METHODS

To address this question, we conducted a systematic literature review and meta-analysis of studies involving human-AI groups in which we compared the performance of the human-AI system to a baseline of the human alone or AI alone, whichever of the two performed better. We say that human-AI synergy exists when the human-AI group outperforms this baseline, akin to strong synergy in traditional groups of humans [Larson 2013].

We applied the following criteria to select studies that fit our research questions. First, the paper needs to present an original experiment that evaluates some instance in which a human and an AI system work together to perform a task. Second, it needs to report the mean and standard deviation of performance for (1) the human alone, (2) the AI system alone, and (3) the human-AI group, or enough information to calculate them from other quantities. Third, we required the paper to include the experimental design and the number of participants in each condition.

Given the interdisciplinary nature of human-AI interaction studies, we performed this search in multiple databases covering conferences and journals in the computer sciences, information sciences, and social sciences, as well as other fields. Through consultation with a library specialist in these fields, we decided to target the Association for Computing Machinery Digital Library (ACM DL), Association for Information Systems eLibrary (AISeL), and the Web of Science Core Collection (WoS) for our review. To focus on current forms of artificial intelligence, we limited the search to studies published between January 1, 2020 and June 30, 2023. To further ensure comprehensive coverage, we also conducted a forward and backward search on all studies we found that met our inclusion criteria.

3. RESULTS

Our initial literature search yielded 1,960 unique papers, and we identified 71 that met our inclusion criteria. These records reported the results of 101 unique experiments and contained 367 unique effect sizes. We synthesized these data in a three-level meta-analytic model that accounts for the dependency of effect sizes from the same experiment.

3.1 Overall Levels of Human-AI Synergy

We found that the human-AI groups performed significantly *worse*, on average, than either the human alone or AI alone (Fig. 1). The overall pooled effect was negative ($g = -0.21, p = 0.0094$, 95% confidence interval (CI) $[-0.37, -0.05]$) and small, according to common interpretations [Cohen 2013].

We also discovered substantial heterogeneity among the included effect sizes in our analysis ($I^2 = 97.7\%$). Most of the heterogeneity occurred within-experiments (73.7%), but there were also differences between-experiments (24.0%).

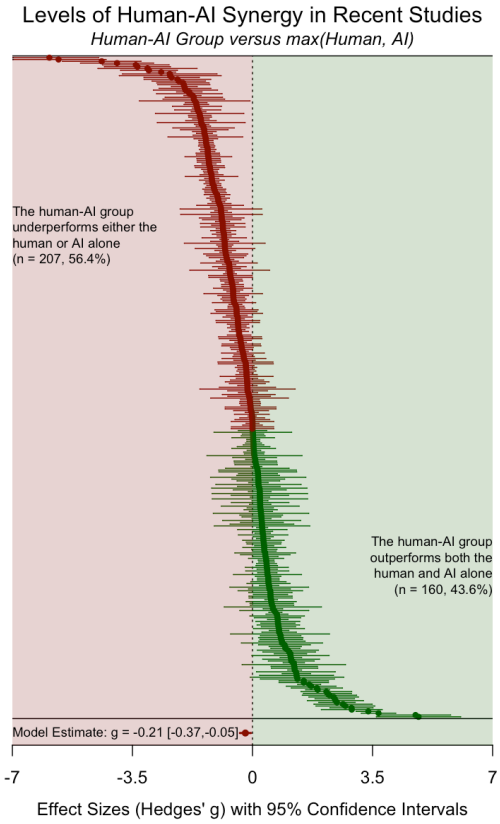


Fig. 1. Forest plots of all effect sizes ($k = 367$) included in the meta-analysis. The positions of the points on the x-axes represent the values of the effect sizes, and the bars indicate the 95% confidence interval for the effect sizes. The circle at the bottom of the graph represents the meta-analytic average effect size and confidence interval.

3.2 Moderators of Human-AI Synergy

Our moderator analysis identified characteristics of participants, tasks, and experiments that led to different levels of synergy in human-AI groups, and it helps explain sources of the heterogeneity in the main analysis (Fig. 2).

First, we found that the type of task significantly moderated synergy in human-AI groups ($p = 0.0094$). Among decision tasks—those in which participants decided between a finite set of options or provided some number value—the pooled effect size was significantly negative ($g = -0.25$, $95\%CI$ $[-0.42, -0.08]$), which indicates performance losses from combining humans and AI. In contrast, among creation tasks—those in which participants created some sort of open response content—the pooled effect size was positive ($g = 0.19, 95\%CI$ $[-0.09, 0.48]$). Even though the average performance gains for creation tasks was not significantly different from 0 (presumably because of the relatively small sample size of $n = 34$), the difference between levels of synergy for decision tasks versus creation

tasks was statistically significant. Relatedly, we found that the type of data involved in the task significantly moderated human-AI synergy ($p < 0.0001$).

We also found that the performance of the human and AI relative to each other impacted synergy in human-AI groups ($p < 0.0001$). When the AI alone outperformed the human, performance losses often occurred in the human-AI group relative to the AI alone ($g = -0.53, 95\%CI\ [-0.70, -0.35]$). On the other hand, when the human alone outperformed the AI, the human-AI group outperformed both the human and the AI alone with an average effect size of $g = 0.46$ ($95\%CI\ [0.28, 0.64]$).

Additionally, we found that the type of AI involved in the experiment moderated synergy in human-AI groups ($p < 0.0186$). The remaining moderators we investigated were not statistically significant (division of labor, explanation, confidence, study design, participant type).

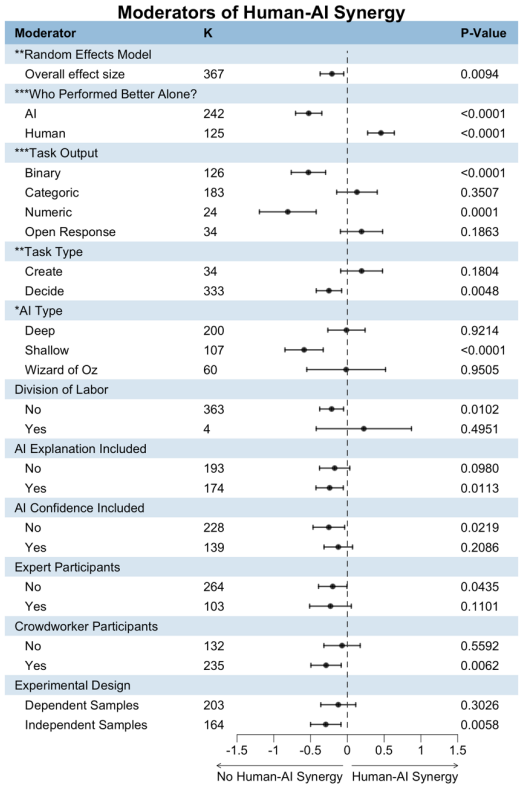


Fig. 2. Results from the three-level meta-regression models. Here, K is the number of included effect sizes for the moderator subgroup level, g is the estimated effect size with the corresponding 95% confidence interval for the moderator subgroup level, and the p-value corresponds to the estimated effect size for the moderator subgroup level. The stars in front of the moderator header indicate if there is a statistically significant difference between subgroups. $^{***}P < 0.001, {}^{**}P < 0.01, {}^{*}P < 0.05$.

Even though our main result suggests that—on average—combining humans and AI leads to performance losses, we do *not* think this means that combining humans and AI is a bad idea. On the contrary, we think it just means that there is still much work to be done on finding effective ways to integrate them, and our other results suggest promising ways to proceed. For example, given our result that, on average across our 300+ effect sizes, explanations provided by AI systems do not improve human-AI performance, we think researchers may wish to de-emphasize this line of inquiry. Rather, we encourage a shift in focus to the significant moderators we identified: the type of task performed and whether humans or AI alone are better at the task. We hope that this work can serve as a roadmap for researchers in selecting topics to study and for practitioners in designing and developing human-AI systems.

REFERENCES

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. DOI:http://dx.doi.org/10.1145/3411764.3445717

Zana Bucçinca, 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 (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. DOI:http://dx.doi.org/10.1145/3377325.3377498

Jacob Cohen. 2013. Statistical Power Analysis for the Behavioral Sciences. Academic press.

Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. DOI:http://dx.doi.org/10.1145/3313831.3376873

James R Larson Jr. 2013. In search of synergy in small group performance. Psychology Press.

Qi Liu, Ruihao Huang, Julie Hsieh, Hao Zhu, Mo Tiwari, Guansheng Liu, Daphney Jean, M Khair El-Zarrad, Tala Fakhouri, and Steven Berman. 2023. Landscape Analysis of the Application of Artificial Intelligence and Machine Learning in Regulatory Submissions for Drug Development from 2016 to 2021. *Clinical Pharmacology and Therapeutics* 113, 4 (2023), 771–774.

Nina Schwalbe and Brian Wahl. 2020. Artificial Intelligence and the Future of Global Health. *The Lancet* 395, 10236 (2020), 1579–1586.

Michelle Vaccaro and Jim Waldo. 2019. The Effects of Mixing Machine Learning and Human Judgment. *Communications of the ACM* 62, 11 (Oct 2019), 104–110. DOI:http://dx.doi.org/10.1145/3359338

Serge-Lopez Wamba-Taguimdje, Samuel Fosso Wamba, Jean Robert Kala Kamdjoug, and Chris Emmanuel Tchatchouang Wanko. 2020. Influence of Artificial Intelligence (AI) on Firm Performance: The Business Value of AI-based Transformation Projects. *Business Process Management Journal* 26, 7 (2020), 1893–1924.

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 (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. DOI:http://dx.doi.org/10.1145/3351095.3372852

Commitment Issues: Feedback, Commitment, and Performance in Algorithmically Managed Contexts

Allen S. Brown
Carnegie Mellon University
Pittsburgh, PA, USA
allenbrown@cmu.edu

Christopher R. Dishop
Carnegie Mellon University
Pittsburgh, PA, USA
cdishop@andrew.cmu.edu

Andrew Kuznetsov
Carnegie Mellon University
Pittsburgh, PA, USA
adkuznet@andrew.cmu.edu

Ping-Ya Chao
Carnegie Mellon University
Pittsburgh, PA, USA
pingyac@andrew.cmu.edu

Anita Williams Woolley
Carnegie Mellon University
Pittsburgh, PA, USA
awoolley@andrew.cmu.edu

ABSTRACT

The characteristics of non-routine, knowledge work that pose challenges to traditional management practices have proven similarly challenging to address with algorithmic performance management. However, as artificial intelligence technologies grow increasingly sophisticated, opportunities emerge to develop new approaches to managing algorithmically-mediated work that could enhance knowledge worker performance and motivation. We conducted a randomized, controlled experiment using a novel work simulation platform to examine the effect automated, real-time feedback has on worker commitment and performance under varying levels of task uncertainty. The results demonstrate that automated feedback can improve workers' commitment under high task uncertainty in online work environments and improve performance.

CCS CONCEPTS

• Applied computing → Law, social and behavioral sciences → Psychology • Human-centered computing → Human computer interaction (HCI) → Empirical studies in HCI

KEYWORDS

algorithmic performance management, feedback, commitment, human-agent teaming

1 Introduction

Over the past several decades, workforce demands have radically transformed from being dominated by routinized tasks to non-routine and knowledge-based work [1] that is increasingly gig-based [22], platform-mediated [7,15], and algorithmically managed [4,14]. Knowledge work involves the production, application, and/or distribution of knowledge,[8] with less tangible outputs than more traditional, manual work and has always posed significant challenges to conventional management practices. Productivity is more challenging to quantify in knowledge work, making productivity-based incentives ineffective as knowledge work motivation is primarily based on autonomy, mastery, and purpose [18]. The characteristics of knowledge work that make measuring and incentivizing performance more difficult also made early attempts to use artificial intelligence (AI) in performance management ineffective. Rooted in the assumptions of productivity-based management practices, these attempts were marked by increases in managerial surveillance and control, which exacerbated experiences of alienation, isolation, and distrust [5,6,20] and increased efforts at worker concealment [2,20]. Ultimately, such efforts were ineffective in supporting knowledge workers' internal motivation and commitment [12,15,19,23].

However, more sophisticated AI technologies offer the possibility of improving the conditions promoting internal motivation in knowledge work, which can produce a higher level of commitment [17]. For instance, most knowledge work is characterized by high task uncertainty, which hinders workers' "knowledge of results," a psychological state essential for internal work motivation [9]. At the same time, feedback plays a central role in developing workers' sense of mastery [16]. AI capabilities in digitally-mediated work contexts can enable real-time feedback across broadly distributed workforces [21] and could mitigate the detrimental effects of high task uncertainty to support the necessary conditions for knowledge workers' internal motivation and, consequently, their commitment and performance.

2 Method

To examine the effect of real-time feedback under varying conditions of task uncertainty, we conducted a randomized controlled experiment in the Collaborative Caring Virtual Testbed. [13] This platform simulates a caregiving environment with game-like qualities in which participants are hired to work as caregivers for a recipient with ongoing emergent problems and needs. The caregivers need to diagnose the issues that arise and use or develop various solutions to maintain the comfort and health of the recipient.

After completing a short training module that oriented them to the platform environment, participants worked a seven-minute simulated caregiving "shift," monitoring and attending to the patient's different needs. At the end of their shift, all participants completed short attention and manipulation checks before receiving a final, algorithm-based performance evaluation for their work. After receiving their evaluation score, participants completed a survey to capture additional measures before the experiment concluded.

2.1 Manipulations

Participants were randomly assigned to one of four conditions in a 2 (task uncertainty: high vs. low) x 2 (automated, real-time feedback: absent vs. present) between-subjects design, resulting in four treatment groups.

1. Low task uncertainty and real-time feedback (Low U + FB).
2. Low task uncertainty and no real-time feedback (Low U).
3. High task uncertainty and real-time feedback (High U + FB).
4. High task uncertainty and no real-time feedback (High U).

2.2 Measures

We captured participant perceptions using likert-type items measuring *Knowledge of Results* [9] as a key emergent state supporting internal motivation, and focused on *Affective Commitment [3]* as the main commitment indicator. We also measured *Task Uncertainty* [11] both to evaluate the success of the task uncertainty manipulation and to test its role as a moderator in the relationship of knowledge of results on commitment following their work shift. The Collaborative Caring Testbed also captured measures of participants' performance, which served both as the basis for the performance ratings individuals received after their shift was over and our dependent measure and ranged from "0: Fails to meet expectations" to "3: Exceeds expectations."

3 Results and Analysis

This study was designed to examine the effects of explicit, real-time feedback on participants' knowledge of results and, consequently, their performance and reported feelings of commitment to the work. Moreover, we expected that higher levels of task uncertainty would strengthen the effect of knowledge of results on commitment.

Manipulation checks confirmed that both experimental treatments successfully induced their intended effects. The indirect effect that automated, real-time feedback had on performance through knowledge of results was positive and significant (see Fig. 1). Additionally, results from a bootstrapping analysis with 10000 resamples according to PROCESS Model 15 [10] examining the effect on affective commitment returned a significant index of moderated mediation (see Fig. 2). Automated real-time feedback increases both performance and experienced commitment through improved knowledge of results. The effect on commitment is stronger and significant in environments with higher task uncertainty, and this effect persists even when controlling for an individual's quality of performance. Future work will capture the effect automated, real-time feedback has on measures of collective intelligence and worker persistence in simulated caregiving contexts.

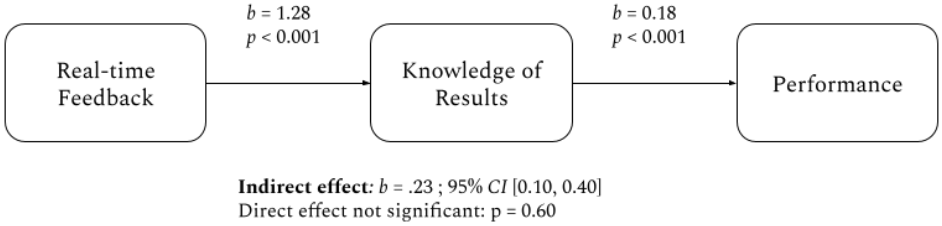


Figure 1. Effect of automated, real-time feedback on task performance through knowledge of results.

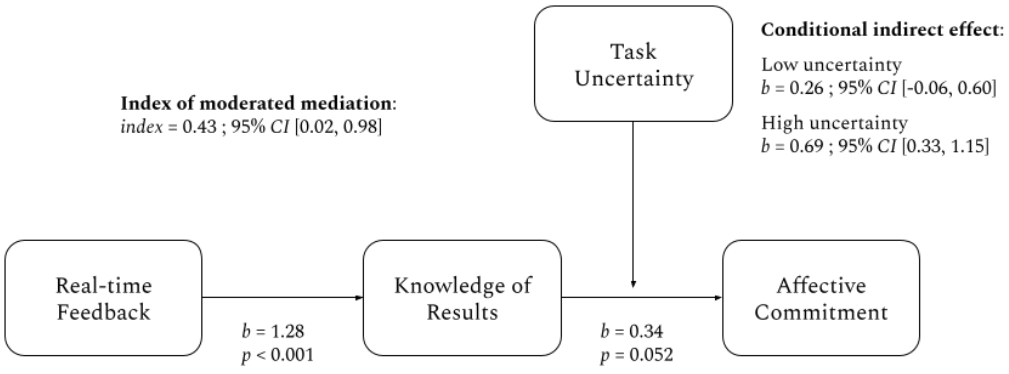


Figure 2. Moderated-mediation model for the effect of automated, real-time feedback on affective commitment.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant no. 2112633.

REFERENCES

[1] David H. Autor. 2015. Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives* 29, 3: 3–30.

[2] Ethan S. Bernstein. 2017. Making Transparency Transparent: The Evolution of Observation in Management Theory. *Academy of Management Annals* 11, 1: 217–266.

[3] Gary Blau. 2003. Testing for a four-dimensional structure of occupational commitment. *Journal of Occupational and Organizational Psychology* 76, 4: 469–488.

[4] Lindsey Cameron, Laura Lamers, Ulrich Leicht-Deobald, Christoph Lutz, Jeroen Meijerink, and Mareike Möhlmann. 2023. *Algorithmic management: Its implications for information systems research.* Communications of the Association for Information Systems.

[5] Jillian Chown. 2021. The unfolding of control mechanisms inside organizations: pathways of customization and transmutation. *Administrative Science Quarterly* 66, 3: 711–752.

[6] Corentin Curchod, Gerardo Patriotta, Laurie Cohen, and Nicolas Neysen. 2020. Working for an algorithm: power asymmetries and agency in online work settings. *Administrative Science Quarterly* 65, 3: 644–676.

[7] Gerald F. Davis. 2016. *The Vanishing American Corporation: Navigating the Hazards of a New Economy.* Berrett-Koehler Publishers.

[8] Peter F. Drucker. 1999. Knowledge-worker productivity: The biggest challenge. *California Management Review* 41, 2: 79–94.

[9] J. Richard Hackman and Gary R. Oldham. 1976. Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance* 16: 250–279.

[10] Andrew F. Hayes. 2013. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* Guilford Press, New York.

[11] Mark Keith, Haluk Demirkan, and Michael Goul. 2017. The role of task uncertainty in IT project team advice networks. *Decision Sciences* 48, 2: 207–247.

[12] Katherine C. Kellogg, Melissa A. Valentine, and Angéle Christin. 2020. Algorithms at work: the new contested terrain of control. *Academy of Management Annals* 14, 1: 366–410.

[13] Andrew Kuznetsov, Ping-Ya Chao, Christopher R Dishop, Allen S. Brown, and Anita Williams Woolley. 2024. The collaborative caring virtual testbed: A software platform for prototyping collective intelligence interventions for asynchronous care teams. In *Proceedings of the ACM Collective Intelligence Conference 2024,* Boston, MA (Demonstration).

[14] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: the impact of algorithmic and data-driven management on human workers. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 1603–1612.

[15] Mareike Möhlmann, Lior Zalmanson, Ola Henfridsson, and Robert Wayne Gregory. 2021. Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control. *MIS Quarterly* 45, 4.

[16] Elizabeth Wolfe Morrison and Robert J. Bies. 1991. Impression Management in the Feedback-Seeking Process: A Literaturereview and Research Agenda. *Academy of Management Review* 16, 3: 522–541.

[17] Sharon K. Parker and Gudela Grote. 2022. Automation, Algorithms, and Beyond: Why Work Design Matters More Than Ever in a Digital World. *Applied Psychology* 71, 4: 1171–1204.

[18] Gianpiero Petriglieri, Susan J. Ashford, and Amy Wrzesniewski. 2019. Agony and ecstasy in the gig economy: cultivating holding environments for precarious and personalized work identities. *Administrative Science Quarterly* 64, 1: 124–170.

[19] Sumita Raghuram, N. Sharon Hill, Jennifer L. Gibbs, and Likoebe M. Maruping. 2019. Virtual Work: Bridging Research Clusters. *Academy of Management Annals* 13, 1: 308–341.

[20] Hatim A. Rahman. 2021. The invisible cage: Workers' reactivity to opaque algorithmic evaluations. *Administrative Science Quarterly* 66, 4: 945–988.

[21] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, et al. 2019. Machine behaviour. *Nature* 568, 7753: 477–486.

[22] Gretchen M. Spreitzer, Lindsey Cameron, and Lyndon Garrett. 2017. Alternative work arrangements: two images of the new world of work. *Annual Review of Organizational Psychology and Organizational Behavior* 4, 1: 473–499.

[23] Angie Zhang, Alexander Boltz, Chun Wei Wang, and Min Kyung Lee. 2022. Algorithmic Management Reimagined For Workers and By Workers: Centering Worker Well-Being in Gig Work. *CHI Conference on Human Factors in Computing Systems*, ACM, 1–20.

Collective Adaptation of Human Behavior During a widespread Crises: a COVID-19 Case Study

Tamanna Urmi
urmi.t@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Christoph Riedl
c.riedl@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

David Lazer
D.lazer@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Mauricio Santillana
m.santillana@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Abstract

Humans have developed collective adaptation to protect them from crises many times in history during disease outbreaks, catastrophic natural events, and wars. In this paper, we dissect the nature and dynamics of the collective adaptation of protective behaviors during the COVID-19 pandemic to better understand how humans respond to the crises and to utilize this information for more powerful mathematical models. It could thereby lead to better understand the role of collective intelligence – how humans collectively utilize their perception of security, their knowledge, and their collaborative (or divisive) nature – to protect themselves from external threats. Infectious pathogens spreading in human populations have led to multiple disease outbreaks in human history that have had devastating consequences including deaths of large portions of communities. Despite being a crucial factor in determining the epidemic trajectory, human behaviors have largely been absent from epidemiological models which instead assume behavior remains constant over time.

We find that people responded dynamically to the state of the pandemic, increasing protective behaviors in reaction to a worsening state of the pandemic, and engaging in risk-exposing behaviors when the severity of disease transmission decreased. Over the first two years of the COVID-19 pandemic in the US, people experienced fatigue in adhering to protective behaviors. This was expressed by the decay of the proportion of people engaging in protective behaviors such as limiting their contact with people outside their household or wearing a mask. or example, the behavior of avoiding contact with other people decreased from 60% people adhering to it, at the beginning of the pandemic, to only 20% two years later, declining at a rate of 1.5% per month. The sensitivity of behaviors to situational awareness (mortality in their community or state) exhibited synchronous patterns for protective behaviors and asynchronous patterns for exposing ones. We find political affiliations modulate individuals' decay over time and also their sensitivity to changes in severity throughout the crisis. This study elucidates how collective adaptation forms and how it shifts during a crisis, and enables us to use behavior as precise parameters in mathematical models of the crisis.

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-XXXX-X/18/06

CCS Concepts

• **Applied computing** → *Decision analysis.*

Keywords

Adaptation, Health, Fatigue, Epidemiology, Behavior

ACM Reference Format:
Tamanna Urmi, Christoph Riedl, David Lazer, and Mauricio Santillana. 2024. Collective Adaptation of Human Behavior During a widespread Crises: a COVID-19 Case Study. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 3 pages.

1 Introduction

Humans have developed collective adaptation [6] to protect them from crises many times in history during disease outbreaks, catastrophic natural events, and wars. In this paper, we dissect the nature and dynamics of the collective adaptation of protective behaviors during the COVID-19 pandemic to better understand how humans respond to the crises and to utilize this information for more powerful mathematical models. It could thereby lead to better understand the role of collective intelligence [8] – how humans collectively utilize their perception of security, their knowledge, and their collaborative (or divisive) nature – to protect themselves from external threats.

Infectious pathogens spreading in human populations have led to multiple disease outbreaks in human history that have had devastating consequences including deaths of large portions of communities. The influenza outbreak in 1918 alone killed 33% of the world population, for example. From ancient plagues, influenza outbreaks, and smallpox, the severity of infectious disease outbreaks has been modulated by how humans have responded to them. As a result, it shaped population dynamics, societal structures, and everyday behaviors.

In the recent coronavirus disease 2019 (COVID-19) pandemic, public health authorities around the globe recognized the importance of behavioral change to mitigate disease transmission. This motivated behavioral interventions such as stay-at-home orders,

social distancing, hygiene practices, and mask-wearing. What role does changing human behavior play in how a public health crisis unfolds? In this study, we explore two key research questions:

- Do protective behaviors systematically relax as the crisis unfolds?
- Do protective behaviors spike during intense periods of the crisis?

We find that people responded dynamically to the state of the pandemic, increasing protective behaviors in reaction to a worsening state of the pandemic, and engaging in risk-exposing behaviors when the severity of disease transmission decreased.

Despite being a crucial factor in determining the epidemic trajectory, human behaviors have largely been absent from epidemiological models which instead assume behavior remains constant over time. Epidemiologists and public health officials have highlighted the importance of building comprehensive epidemiological models that integrate the dynamics of human behavior on many occasions in the past [2, 4, 5]. Previous mathematical modeling investigations demonstrated enhancement of models for smallpox and influenza by incorporating behavior parameters [7, 10]. Behavioral shifts alter the mixing pattern of humans, the duration of contact with pathogenic agents, and the probability of transmission by contact, thereby, impacting the overall transmission dynamics of the disease. Human behaviors served as explanations of certain transmission changes or as a representation of public sentiment during the COVID-19 pandemic but effective models incorporating these are yet to be developed [1, 9]. Behaviors have been incorporated in disease modeling in other scenarios. The smallpox transmission curve demonstrated high sensitivity to a shift in protective behaviors [10]. Multi-scale modeling of influenza transmission has shown that social contact pattern is a crucial factor for the size of the epidemic, and using simulated agent-based modeling Manfredi et. al. have shown the importance of adopting protective behaviors to contain the spread of influenza [11].

Traditional mechanistic Susceptible-Infected-Recovered (SIR) compartmental epidemiological models are fully specified by the number of people in the disease state and fixed transmission probability, but frequently, are behavior invariant [3]. An effective integrated model is yet to be developed. In this study, we characterize the trends and dynamics of risk-averting and risk-exposing behaviors during a public health crisis—the COVID-19 pandemic. We anticipate that it will enable the mathematical modeling community to retrospectively integrate them into enhanced epidemiological models, ultimately helping us better prepare for inevitable novel pandemics in the future.

2 Methods

Using data collected through a non-probability survey conducted among the adult population in the US from the early months of the COVID-19 pandemic (April 2020), we quantitatively characterize 15 distinct (risk-averting and risk-exposing) human behavior dynamics observed during the pandemic.

We investigate their interplay with the severity of the crisis, gauged by the death toll. As confirmed by Principal Component Analysis (PCA), we identify and group these behaviors into "protective" and "exposing" behaviors. Over 600,000 people participated in these surveys and their responses were weighted to compensate for the under-represented demographic groups. The survey waves are of irregular lengths and gaps. We used numerical interpolation to generate a monthly time series of behavior data from April 2020 to May 2022. By decomposing the time-series data into a linear trend and a deviation (anomaly) from the trend, we obtained behavior fluctuations. Cross-correlation and phase shift analyses were conducted on the behavior fluctuations and death counts to estimate people's thermostatic response to the death toll.

3 Results

Over the first two years of the COVID-19 pandemic in the US, people experienced fatigue in adhering to protective behaviors. This was expressed by the decay of the proportion of people engaging in protective behaviors such as limiting their contact with people outside their household or wearing a mask. In the first 26 months of the pandemic, risk-seeking behaviors have linearly increased and risk-avoiding behaviors have linearly decreased. For example, the behavior of avoiding contact with other people decreased from 60% people adhering to it, at the beginning of the pandemic, to only 20% two years later, declining at a rate of 1.5% per month (Fig. 1a). Conversely, exposing behaviors surged over time, for example, the use of mass transit more than doubled in two years. The sensitivity of behaviors to situational awareness (mortality in their community or state) exhibited synchronous patterns for protective behaviors and asynchronous patterns for exposing ones (Fig. 1b). The trends, dynamics, and cross-correlations exhibit similar patterns in different states but further analysis could elucidate important state-level variations.

We use self-reported political leaning and demographic information to explore heterogeneity in the population's response to the health emergency and characterize patterns of collective adaptation. We find political affiliations modulate individuals' decay over time (Fig. 1c) and also their sensitivity to changes in severity throughout the crisis (Fig. 1d). But, the paradigm that there are society is divided into adherent and non-adherent group is not accurate. The percentage of adherent among democrats are somewhat higher both for risk avoiding and risk seeking behaviors but the dynamics of change is the same across different political affiliation of states or individuals. Democrats tended to adhere to protective behaviors for longer and more intensely during increased risk (higher mortality), followed by independents and Republicans. Initially, widespread infections prompted significant protective responses, but they diminished over time. In contrast, increased mortality consistently elicited strong protective reactions throughout the crisis. (Fig. 1e). This suggests that while individuals generally become increasingly desensitized to the public health crisis, they continue to adapt their behavior based on reported spikes of the most severe negative outcomes (deaths).

4 Discussion

The irregularity of the survey wave frequency presented us with the challenge of producing a suitable time series for trend analysis with this data. We recognize that self-reported surveys of this nature could pose social desirability bias.

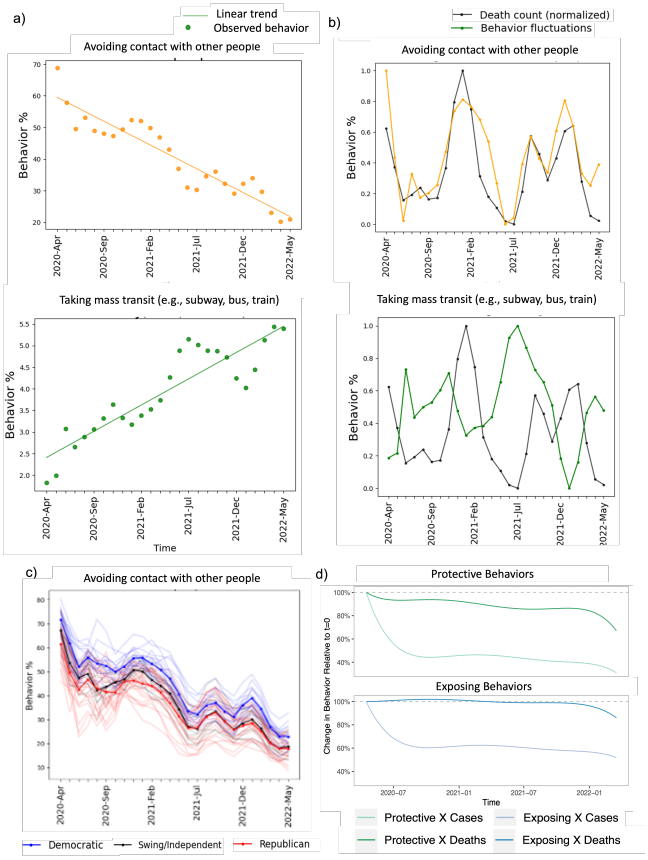


Figure 1: Collective adaptation during a pandemic. (a) Behavior trends for a risk-avoiding behavior "Avoiding contact with other people" and for a risk-exposing behavior "Taking mass transit." (b) The phase shift of the behavior anomalies/fluctuations with death count. (c) Aggregated behavior trend segregated by political leaning (d) Fatigue of preventive behavior adherence

Much of the modeling in this space assumes stability of behaviors and a steady collective adaptation. However, we should expect very substantial and predictable instability of behaviors, and, assuming that these behaviors actually affect transmission patterns, this will likely have significant implications for building spreading models. These analyses beg for understanding transmission dynamics of information regarding the state of the pandemic, how different people are differentially responsive, and how we might intervene in those communications to result in population responses that reduce mortality.

Generally, our research yields valuable insights into individual responses, revealing patterns of collective adaptation to complex phenomena threatening human well-being such as climate change, social fragmentation, and future pandemics. We believe that integrating human behavior dynamics will enhance the predictive accuracy of epidemiological models, enable our understanding of how societies adapt to life-threatening events and severe crises and also facilitate more effective policy-making strategies during such events.

References

[1] Jason Abaluck, Laura H. Kwong, Ashley Styczynski, Ashraful Haque, Md. Alamgir Kabir, Ellen Bates-Jefferys, Emily Crawford, Jade Benjamin-Chung, Shabib Raihan, Shadman Rahman, Salim Benhachmi, Neeti Zaman Bintee, Peter J. Winch, Maqsud Hossain, Hasan Mahmud Reza, Abdullah All Jaber, Shawkee Gulshan Momen, Aura Rahman, Faika Laz Banti, Tahrima Saiha Huq, Stephen P. Luby, and Ahmed Mushfiq Mobarak. 2021. Impact of community masking on COVID-19: A cluster-randomized trial in Bangladesh. *Science* 375, 6577 (Dec. 2021), eabi9069. https://doi.org/10.1126/science.abi9069 Publisher: American Association for the Advancement of Science.

[2] Chris Bauch, Alberto d'Onofrio, and Piero Manfredi. 2012. Behavioral Epidemiology of Infectious Diseases: An Overview. *Modeling the Interplay Between Human Behavior and the Spread of Infectious Diseases* (Nov. 2012), 1–19. https://doi.org/10.1007/978-1-4614-5474-8_1

[3] Jamie Bedson, Laura A Skrip, Danielle Pedi, Sharon Abramowitz, Simone Carter, Mohamed F Jalloh, Sebastian Funk, Nina Gobat, Tamara Giles-Vernick, Gerardo Chowell, et al. 2021. A review and agenda for integrated disease models including social and behavioural factors. *Nature human behaviour* 5, 7 (2021), 834–846.

[4] Nita Bharti. 2021. Linking human behaviors and infectious diseases. *Proceedings of the National Academy of Sciences* 118, 11 (March 2021), e2101345118. https://doi.org/10.1073/pnas.2101345118 Publisher: Proceedings of the National Academy of Sciences.

[5] Sebastian Funk, Marcel Salathé, and Vincent A. A. Jansen. 2010. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of The Royal Society Interface* 7, 50 (May 2010), 1247–1256. https://doi.org/10.1098/rsif.2010.0142 Publisher: Royal Society.

[6] Mirta Galesic, Daniel Barkoczi, Andrew M Berdahl, Dora Biro, Giuseppe Carbone, Ilaria Giannoccaro, Robert L Goldstone, Cleotilde Gonzalez, Anne Kandler, Albert B Kao, et al. 2023. Beyond collective intelligence: Collective adaptation. *Journal of the Royal Society interface* 20, 200 (2023), 20220736.

[7] Dongmin Guo, King C. Li, Timothy R. Peters, Beverly M. Snively, Katherine A. Poehling, and Xiaobo Zhou. 2015. Multi-scale modeling for the transmission of influenza and the evaluation of interventions toward it. *Scientific Reports* 5, 1 (March 2015), 8980. https://doi.org/10.1038/srep08980 Publisher: Nature Publishing Group.

[8] Christoph Riedl, Young Ji Kim, Pranav Gupta, Thomas W Malone, and Anita Williams Woolley. 2021. Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences* 118, 21 (2021), e2005737118.

[9] Yacob T. Tesfaldet and Nji T. Ndeh. 2022. Public face masks wearing during the COVID-19 pandemic: A comprehensive analysis is needed for potential implications. *Journal of Hazardous Materials Advances* 7 (Aug. 2022), 100125. https://doi.org/10.1016/j.hazadv.2022.100125

[10] S. Del Valle, H. Hethcote, J. M. Hyman, and C. Castillo-Chavez. 2005. Effects of behavioral changes in a smallpox attack model. *Mathematical Biosciences* 195, 2 (2005), 228–251. https://doi.org/10.1016/j.mbs.2005.03.006

[11] Sara Y. Del Valle, Susan M. Mniszewski, and James M. Hyman. [n. d.]. Modeling the Impact of Behavior Changes on the Spread of Pandemic Influenza. *Modeling the Interplay Between Human Behavior and the Spread of Infectious Diseases* ([n. d.]), 59. https://doi.org/10.1007/978-1-4614-5474-8_4 Publisher: Nature Publishing Group.

KnEvol: A Simulation for Examining the Role of Social Bias in Collective Intelligence

LaRoyce J. Covington
Joshua Introne
ljcoving@syr.edu
jeintron@syr.edu
Syracuse University
Syracuse, New York, USA

ABSTRACT

This paper introduces KnEvol, an agent-based model to examine the role of social bias in collective intelligence. We begin with a discussion of the factors that influence collective intelligence before describing the model and providing sample results of a parameter sweep. We conclude with a brief discussion of how KnEvol may be useful to researchers.

CCS CONCEPTS

• **Computing methodologies** → *Distributed artificial intelligence*;
• **Applied computing** → *Law, social and behavioral sciences.*

KEYWORDS

collective intelligence, homophily, reputation, specialization, complexity, agent-based model

ACM Reference Format:
LaRoyce J. Covington and Joshua Introne. 2024. KnEvol: A Simulation for Examining the Role of Social Bias in Collective Intelligence. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 3 pages.

1 BACKGROUND

Theories suggest that individuals are more likely to interact with and trust those who share similar characteristics, beliefs, and values [12]. Indeed, this *homophily bias* has been demonstrated to affect information behavior and processing at all levels [6]. Most relevantly, at the macro level, homophily has been demonstrated to promote polarization and segregation as individuals share information primarily with those like themselves [1]. This effect has practical implications as groups with diverse perspectives have been shown to outperform groups of high-ability individuals [9]. This suggests that homophily hinders group performance on complex tasks as such performance would benefit from a diversity of viewpoints and approaches.

Previous work has already tried to investigate task complexity and homophily bias in tandem. The meta-analysis in [3] found that the positive effects of demographic diversity are more pronounced when groups are dealing with complex tasks. Considering the role of homophily in fostering diversity within dispersed communities, this shows that further investigation into this phenomenon is warranted.

Reputation – the collective belief or opinion about an individual's character or standing within a community – has also been shown to influence trust, co-operation, and information sharing [13]. It is perhaps that reputation governs the extent to which an individual adopts the knowledge of others whereas homophily influences from whom an individual seeks information. A good reputation is hard earned, but quickly discarded once quality begins to slip [4].

Agent-based models (ABMs) have great potential in the study of collective intelligence. As [7] notes, ABMs provide "a generative, proof-by-construction approach to understanding social behavior." Indeed, [11] developed an ABM to investigate the role of homophily in the diffusion of social behaviors. Their model, based on [8] was used to find that homophily of all types promoted behavior diffusion. While behavior diffusion and collective intelligence are distinct phenomena, this demonstrates the utility of ABMs in the study of homophily and related social biases.

Although the individual influences of homophily and reputation on social influence have been acknowledged, their interaction and potential combined effects remain a subject of ongoing inquiry. [5], for example, found that the "most effective social environment for increasing the 'willingness' of obese individuals to adopt the behavior was one in which they interacted with others with similar health characteristics." Further, [5] found that the spread of health behaviors "were greatest in homophilous networks." This suggests that the combination of similarity and social reinforcement may have a stronger impact than either factor alone.

We are further drawn to the role of individual interests or specializations, which must intersect with homophily and reputation. Groups of generalists, for example, may lack the will to make headway on specific areas whereas groups of specialists may lack the proper interest overlap to get a clear picture. Both situations can cause the collective to be stuck in local optima within the solution space.

2 MODEL

Agent-based models have become common in the study of social behaviors [7, 11]. Keeping with this tradition, we developed an ABM named KnEvol (Knowledge Evolution) to investigate the dynamics of complexity and homophily within collective intelligence frameworks. In KnEvol, agents are embedded within a "universe" controlled by an NK random boolean network [10]. Each agent possesses a "goal," which represents a specific state in the universe.

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

This goal may constitute a subset of bits in the Random Boolean Network (RBN), with the number of bits in an agent's goal acting as a simulation parameter to regulate "specialization." Agents aim to accurately predict the states within the RBN that precede their goal. For example, if the RBN of the universe generates a sequence of states 4-12-18-23, and Agent A's goal state is 18, Agent A strives to make precise predictions for states 4 and 12, while disregarding state 23. We define the collection of all preceding states relevant to an agent's goal as their set of interests. An agent's fitness is determined by the proportion of interests they can predict accurately.

Agents are part of a fixed, undirected social network established at the simulation's start. Each agent's "brain" is another RBN, with N and K values matching those of the universe. Agents enhance their fitness by adjusting their RBN through either "thinking" or "communicating." "Thinking" involves random modifications to their brain akin to mutation in genetic algorithms [2], while "communicating" mirrors the crossover process, where agents adopt portions of a neighbor's brain based on homophily and reputation. In each update cycle of the simulation, an agent will engage in either communication or thinking, with the choice governed by a population-wide probability parameter. Any modifications to their brain are retained only if they enhance the agent's fitness.

In our model, homophily and reputation dictate which neighbors any given agent chooses to communicate with. The preference for a given communication partner is a multiplicative function of homophily, assessed at the Jaccard similarity between two agents' interests, and reputation, which is an agent's self reported fitness. The collective intelligence of the system is evaluated periodically, following a set number of update cycles, by having the population vote on predictions for all states in the universe's RBN. The collective intelligence is measured by the accuracy of these collective predictions.

This simulation framework enables the exploration of various interactions, specifically focusing on how social biases such as homophily and reputation, or individual specializations affect collective intelligence when adjusting the complexity of the universe (K).

3 EXPERIMENT

We instantiate a social network consisting of 100 agents with an average degree of 30 and use it for all trials. For this experiment we vary the NK model such that N = 8 and K varies. Further, we vary the agent's preference for others by manipulating a "minimum similarity" parameter. This dictates the minimum allowable value when one agent evaluates the similarity of another (i.e., if two agents have a similarity of N < min. similarity, the agent assesses similarity to be 'min similarity'. This has the effect of increasing homophily at low values and decreasing homophily at higher values.) Similarly, we vary the agent's propensity to adopt information from others by manipulating a "minimum fitness" value. This value interacts with the researcher-set overall probability of information adoption to adjust the degree an agent adopts information from a particular neighbor at a particular time. For each parameter combination we ran a simulation for 5,000 time steps, conducting a vote every 500 time steps. We ran 10 trials at each parameter set, using the same RBN for each at a given parameter set.

4 RESULTS AND DISCUSSION

In this initial analysis, we observe several performance patterns. Firstly, it is unsurprising that more complex problem domains (higher K) pose greater difficulty for agents. Moreover, specialization seems to have a direct impact. There may be an optimal level of specialization which does not directly vary with complexity. This optimal specialization level is most evident in lower complexity environments.

However, the dynamics of homophily and reputation interact with both problem complexity and specialization. Specifically, at lower specialization levels, reputation signals negatively affect collective intelligence, especially when combined with high levels of homophily. As agents become more specialized, this negative impact diminishes, although it remains evident in complex problem domains. Interestingly, at very high levels of specialization, reputation and homophily seem to lose their influence.

These initial findings, though exploratory, shed light on crucial aspects of our simulation and prompt specific hypotheses. We note that mutation, the agents' individual learning procedure, remains unaffected by manipulations. Our exploration focuses solely on parameters governing the crossover process. Additionally, while homophily remains constant throughout the simulation, increasing sensitivity to it reduces communication diversity over time. In contrast, reputation is dynamic, with agents who excel in configuring local networks becoming preferred sources of "sub-solutions."

This attention to reputation signals further reduces diversity over time. A possible explanation for our results lies in the combined reduction of diversity and homophily, ultimately impairing performance by limiting collective movement across the search space and potentially driving the population towards local optima rather than encouraging continued exploration. Whether this is due to the distribution of interests across the population, a mismatch between social signals and the problem space, or some other issue is currently unknown and warrants further study.

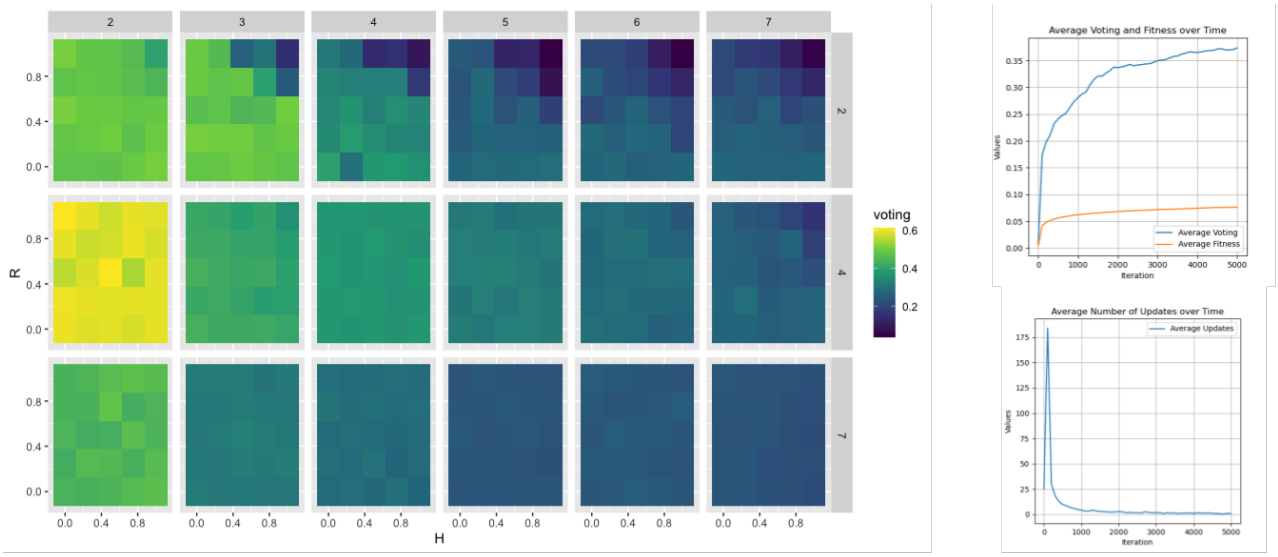


Figure 1: *Left:* A heatmap of collective voting performance at six values (2-7) of K and three specialization values (2, 4, 7); *Right-top:* Average voting (collective) and fitness (individual) for K = 4, homophily = reputation = 0.5, and specialization = 4; *Right-bottom:* The average number of updates over time for the same parameter set

REFERENCES

[1] Michele L. Barnes, John Lynham, Kolter Kalberg, and PingSun Leung. 2016. Social Networks and Environmental Outcomes. *Proceedings of the National Academy of Sciences* 113, 23 (June 2016), 6466–6471. https://doi.org/10.1073/pnas.1523245113

[2] David Beasley, David R. Bull, and Ralph Robert Martin. 1993. An Overview of Genetic Algorithms: Part 1, Fundamentals. *University computing* 15, 2 (1993), 56–69.

[3] Suzanne T. Bell, Anton J. Villado, Marc A. Lukasik, Larisa Belau, and Andrea L. Briggs. 2011. Getting Specific about Demographic Diversity Variable and Team Performance Relationships: A Meta-Analysis. *Journal of Management* 37, 3 (May 2011), 709–743. https://doi.org/10.1177/0149206310365001

[4] Silvia Bonaccio and Reeshad S. Dalal. 2006. Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences. *Organizational Behavior and Human Decision Processes* 101, 2 (Nov. 2006), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

[5] Damon Centola. 2011. An Experimental Study of Homophily in the Adoption of Health Behavior. *Science (New York, N.Y.)* 334, 6060 (Dec. 2011), 1269–1272. https://doi.org/10.1126/science.1207055

[6] Gokhan Ertug, Julia Brennecke, Balázs Kovács, and Tengjian Zou. 2022. What Does Homophily Do? A Review of the Consequences of Homophily. *Academy of Management Annals* 16, 1 (Jan. 2022), 38–69. https://doi.org/10.5465/annals.2020.0230

[7] Robert L. Goldstone and Marco A. Janssen. 2005. Computational Models of Collective Behavior. *Trends in Cognitive Sciences* 9, 9 (Sept. 2005), 424–430. https://doi.org/10.1016/j.tics.2005.07.009

[8] Mark Granovetter. 1978. Threshold Models of Collective Behavior. *Amer. J. Sociology* 83, 6 (1978), 1420–1443. jstor:2778111

[9] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[10] S. A. Kauffman and E. D. Weinberger. 1989. The NK Model of Rugged Fitness Landscapes and Its Application to Maturation of the Immune Response. *Journal of Theoretical Biology* 141, 2 (Nov. 1989), 211–245. https://doi.org/10.1016/s0022-5193(89)80019-0

[11] Qingjun Li, Haihua Hu, and Wei Yang. 2020. Homophily and Behavior Diffusion. *Social Behavior and Personality: an international journal* 48, 3 (March 2020), 1–10. https://doi.org/10.2224/sbp.8028

[12] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2003. Birds of a Feather: Homophily in Social Networks. https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.27.1.415. https://doi.org/10.1146/annurev.soc.27.1.415

[13] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The Value of Reputation on eBay: A Controlled Experiment. *Experimental Economics* 9, 2 (June 2006), 79–101. https://doi.org/10.1007/s10683-006-4309-2

Generative AI and Distributed Work: Evidence from Open Source Software

Manuel Hoffmann* Sam Boysel* Frank Nagle* Sida Peng† Kevin Xu‖

*Harvard Business School, Harvard University

†Microsoft Corporation

‖GitHub Inc.

This version: *March 15, 2024*

Please do not circulate or cite without permission!

Abstract: Artificial intelligence (AI) is at the beginning of a technological evolution that is slowly and steadily reaching all sectors of the economy. However, due to the novelty and breadth of AI, we still know relatively little about the labor impacts of AI, including both the effects on strategic human capital decisions as well as decentralized, often voluntary, contributions to public goods. Therefore, it imperative to understand how different work processes can be affected by this transformational technology. Using the context of open source software, we study the individual consequences of using a generative AI tool designed to help developers work more efficiently and effectively. We exploit a natural experiment arising from the deployment of GitHub Copilot, a generative programming AI released by the social coding platform in June 2022. We study the impact of generative AI for a whole year on millions of work activities by exploiting threshold eligibility requirements through a quasi-experimental regression discontinuity approach. We find that after having access to Copilot, top maintainers are able to code more and spend less time on project management tasks. Similar to other modern technologies, generative AI also induces maintainers to substitute away from high cost interactions (i.e. comments) towards low cost interactions (i.e. emoticons). Overall, a back-of-the-envelope calculation suggests that around thirty percent of the efficiencies arising from the programming AI are reallocated onto the platform.

JEL-Classification: H4, O3, J0

Keywords: Generative Artificial intelligence, Digital Work, Digital Economy, Open Source Software

Acknowledgement: The authors are grateful for financial and administrative support from GitHub and, in particular, from Peter Cihon. The authors are also indebted for comments by seminar participants at the research seminars at Harvard LISH in Cambridge, MA.

1 Introduction

Artificial intelligence (AI) is at the beginning of a technological evolution that is slowly and steadily reaching all sectors of the economy. However, due to the novelty and breadth of AI, we still know relatively little about the labor impacts of AI, including both the effects on strategic human capital decisions as well as decentralized, often voluntary, contributions to public goods. Therefore, it imperative to understand how different work processes can be affected by this transformational technology. Using the context of open source software, we study the individual consequences of using a generative AI tool designed to help developers work more efficiently and effectively.

We exploit aspects of the general access launch of Github Copilot to the broader public. In particular, a segment of developers who maintain the top OSS project on GitHub receive free Copilot access. Eligibility is determined by an internal ranking known to GitHub, and not the individual developers, allowing us to make credible claims about non-manipulation to establish a causal identification strategy to recover the effect of generative AI on developer activity. In the most basic version, maintainers with a ranking below 0 become eligible to receive free access to Copilot and those above do not. We use a panel of 175,990 distinct maintainers observed weekly from July 2022 through July 2023, which results in millions of maintainer-week observations for activity levels in public GitHub repositories and Copilot usage.

Within the data set of top maintainers, we find those with access to Copilot AI during the general access period are engaging in more coding activities while reducing time in activities required for project management. We also find that the quality of social interactions drops with the new technology. The developers with access are more likely to use low-quality, low-cost interaction methods such as emoticons, and less likely to use high-quality, high-cost means of communication via comments and text. Overall activities on the platform are dropping which indicates that the efficiency and time gains from coding for the individual maintainer are too a large proportion allocated outside of the platform.

Our results contribute to a growing literature on the productivity impacts of AI in important

ways. Early work in this area posited general productivity gains (Agrawal, Gans, and Goldfarb 2019 Corrado, Haskel, and Jona-Lasinio 2021, Raj and Seamans 2018), but that the gains may not be evenly distributed (Brynjolfsson, Rock, and Syverson 2018; Furman and Seamans 2019). Empirical work confirmed such predictions and found wide-ranging productivity benefits to using AI, at both the firm level (e.g., Czarnitzki, Fernández, and Rammer 2023) and the individual level (e.g., Fügener et al. 2022). Our work is consistent with this prior research but adds additional nuance to the labor augmenting technical change literature (Acemoglu 2003) as we provide one of the largest natural experiments of generative AI and it's impact on granular measured distributed work processes for over a one year time horizon "in the wild".

We complement the current literature that leverages IT and consultancy chat support AIs and measures them through experimentation (Brynjolfsson, Li, and Raymond 2023, Dell'Acqua et al. 2023) by investigating changes in work and social interaction processes over a year after the introduction of the programming LLM. While it has been recognized that some new technology may reduce social interactions, including social media and the internet in general (Allcott et al. 2020, Mosquera et al. 2020, Kraut et al. 1998), our findings highlight that a reduction in high quality social interactions may be accompanied by this novel technology as well. This finding has important implications for the strategic human capital literature (Wright, Coff, and Moliterno 2014) since it implies that managers could strategically leverage programs at work that heighten social interactions to accompany the proliferation of generative AI allowing for an offset of a potential drop in social interactions. In aggregate, our results build upon existing literature to answer questions related to how access to generative AI changes labor and social activities.

Beyond the identification of causal effects that generative AI has on decentralized work, our results suggest important implications for the future of OSS. OSS has become an increasingly critical part of the modern economy, such that 96% of corporate codebases include some OSS (Synopsys 2023). Further, recent studies estimate the value of OSS to be on the order of billions of dollars (Blind et al. 2021, Robbins et al. 2021) or trillions of dollars when accounting for usage (Hoffmann, Nagle, and Zhou 2024). Additionally, firm usage of, and contribution to, OSS has

important implications for firm productivity (Nagle 2018, Nagle 2019) and entrepreneurial activity (Wright, Nagle, and Greenstein 2023). However, despite the importance of OSS, many critical projects are under-resourced (Nagle et al. 2020b) as numerous firms free-ride on the efforts of others without giving back (Lifshitz-Assaf and Nagle 2021) leaving volunteer maintainers burnt out and overwhelmed (Raman et al. 2020). As our results show, generative AI may offer a solution to help address these concerns and allow top maintainers to more easily contribute to the common good by solving issues faster. Prior research has shown that OSS developers and maintainers generally contribute to OSS because it gives them a creative outlet and they do not want to spend their time on managerial tasks like security and documentation (Nagle et al. 2020a). AI-powered tools may make it easier to quickly address such managerial tasks, so developers can spend time in a manner they prefer, while still ensuring the security, stability, and usability of OSS.

The Echo Chambers of Complexity: Exploring the Impact of Design Complexity on Groupthink and Innovation- An Experimental Study

Soumyakant Padhee
soumyakant.padhee@lr.edu
College of Business & Economics,
Lenoir-Rhyne University
Hickory, North Carolina, USA

Babak Heydari
b.heydari@northeastern.edu
Dept. of Mechanical & Industrial Engineering and
Network Science Institute,
Northeastern University
Boston, USA

Abstract

The interplay between individual exploration and group reliance (exploitation) is crucial in collective problem-solving, impacting outcomes in complex search landscapes. Existing research highlights the advantage of increased individual exploration as complexity rises, yet it is ambiguous whether reliance on collective intelligence diminishes in tandem, a critical aspect for understanding behavioral and theoretical mechanisms in group dynamics. In this paper, we explore how heightened complexity may lead agents to favor exploiting group knowledge over individual exploration, potentially amplifying 'free riding' tendencies. We tested this in a pre-registered human-subject experiment, where participants emulating recombination search evaluated a search task's complexity before indicating their preference for group collaboration and reliance on collective intelligence. Our findings unveil an exploration social dilemma that intensifies with task complexity. This dilemma manifests as an increasing tendency among individuals to rely on existing knowledge as complexity rises, which paradoxically undermines collective intelligence. However, we also discover that this trend towards inefficiency can be mitigated by the presence of individuals who typically underestimate task complexity and inherently fulfill the team's exploratory needs.

CCS Concepts

• **Human-centered computing** → *Social networks.*

Keywords

Collective problem-solving, Task Complexity, Groupthink, and Innovation

ACM Reference Format:
Soumyakant Padhee and Babak Heydari. 2024. The Echo Chambers of Complexity: Exploring the Impact of Design Complexity on Groupthink and Innovation- An Experimental Study. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 4 pages.

Unpublished working draft. Not for distribution.

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

1 Introduction

Collective problem solving, viewed as a sequential search process, involves agents blending individual exploration and leveraging group knowledge to devise new solutions. The interplay between independent exploration and reliance on others (exploitation) has been a focal point in collective intelligence, organizational design, and cognitive science literature. The ideal balance between exploration and exploitation within a group depends on numerous factors, primarily the complexity of the search landscape, governed by the interconnectedness of decision variables. Existing research posits that as complexity of the landscape escalates, there is a group advantage in agents relying more on individual exploration, thereby evading the pitfall of sub-optimal local peaks that could stymie the innovation process [6], [9]. However, the group advantage doesn't necessarily translate into aligned individual behavior and it remains ambiguous whether agents also tend to decrease their reliance on the group's collective intelligence as complexity increases [1, 2]. This understanding is crucial for examining the correlations between theoretical and behavioral mechanisms [10] and identifying the conditions that facilitate the advantages of collective intelligence.

In this paper, we hypothesize that with increased complexity, agents might incline more towards minimizing risk and cognitive load by exploiting group knowledge rather than embracing the uncertainty of individual exploration. In essence, the lure of capitalizing on the efforts of others without corresponding contribution may amplify [3, 4]. If this hypothesis proves correct, it essentially models a social dilemma, where individual incentives conflict with collective good, resulting in an inefficient equilibrium. In order to experimentally probe this hypothesis, we conducted a pre-registered experiment and implemented it in a lab setting, encompassing tasks of variable complexities. The experiment was structured in two stages, designed to assess how innovators, acting as participants in team configurations, gauge the complexity of a search task in the first stage and subsequently reveal their inclinations towards relying on the collective knowledge in the second stage.

In our experimental set-up, innovating agents, as grouped participants, are tasked to delight a preset virtual agent in stage-one by offering various alternative solutions in the form of re-combination of factors (Figure 1). To which the virtual agent responds with its delightness score computed by the values of interdependent factors that constitute the solution. We chose this setup to eliminate the effect of prior expertise or individual biases towards the subject

Unpublished working draft. Not for distribution.

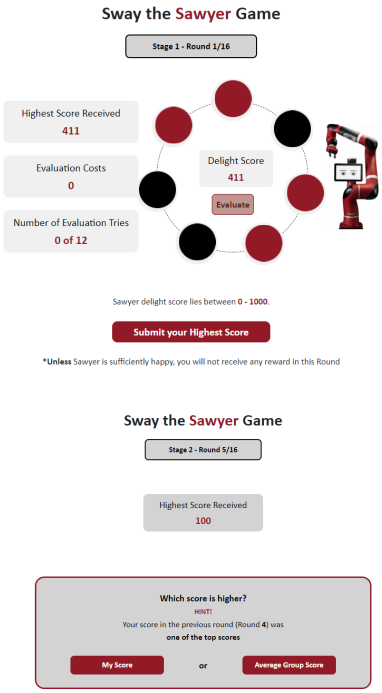


Figure 1: Snapshots of the game showing (top) stage-1 Task, (bottom) stage-2 highest score guessing with nudging feedback based on performance in previous round.

matter, while being able to tune the level of complexity. The computational NK model was utilized to create the search landscapes, offering a robust platform for dissecting the dynamics of exploration and exploitation under different levels of complexity. With varying levels of complexity treatments corresponding to the level of interdependencies between solution factors and participants make multiple trials by offering various recombined solutions to get to a higher delightness score from virtual agent. This gives them a taste of the complexity of search landscape and exposes them to a social choice dilemma in the next stage. Stage-two involves participants guessing the highest delightness score from among their own scores and average group scores, and upon a successful guess, they are incentivized on the condition that their own score is also better than a certain cut-off score. While multiple factors may contribute to dependence on a group's collective intelligence [8], we examine the conditions under which such behavior increases with complexity and test several related hypotheses to find explanation for the mechanism. We conducted in-lab experiments on 57 subjects, divided into teams of 3. Participants were recruited from the students of Master's in Engineering Management at Northeastern University all with engineering background and typically between 2-3 years of work experience before joining the program.

Our first finding, a validation of the main hypothesis, shows that given the choice of weakly dominant incentives, the majority of participants choose to exploit — with increased reliance on group score — rather than exploring individually; and once controlled for other factors, this tendency increases with complexity, resulting in

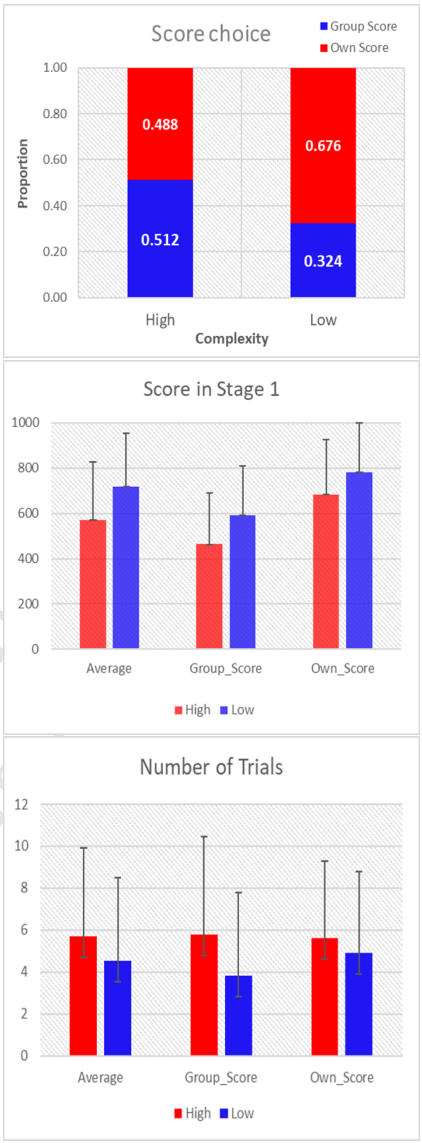


Figure 2: Results showing (top) proportion of group score versus own score choices, (middle) average score and (bottom) number of trails by participants in Stage 1 under high and low complexity treatment conditions.

inefficient outcomes as described earlier (Figure 1). However, we see substantial, systematic heterogeneity in the behavior of agents, motivating the second research question. To understand this heterogeneity and its effect on the overall group dynamic, we introduced a metric of self-reliance (ratio of games they chose to explore in the second stage) for each agent based on their behavior in all the rounds of the game and studied its impact on their performance under varying complexity treatment [5]. Analysis confirmed notable heterogeneity in self-reliance among participants, demonstrated by a distribution shift under different complexity scenarios and performance (Figure 2). A high level of self-reliance, indicating

Unpublished working paper. Not for distribution.

a preference for exploration over exploitation, is associated with lower average performance at the individual level, especially at high complexity. Paradoxically, these individuals collectively contribute to an improvement of group performance, at the expense of their average individual performance.

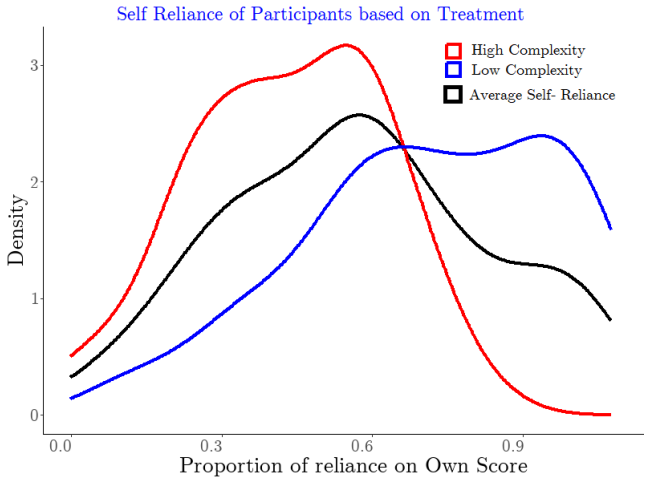


Figure 3: Distribution of reliance on own score at varying complexity.

What explains this heterogeneity? To unravel this, we categorized individuals based on their average discrepancies in assessing task complexity (as declared at the conclusion of stage one in each round). Despite a broad range of agent behaviors, we identify two principal types: those who consistently overestimate task complexity (often labeling both complex and some simple tasks as "high" complexity) and those who underestimate it (frequently classifying both simple and some complex tasks as "low" complexity). We term these groups "Complexity Overestimators" and "Complexity Underestimators," respectively. Our findings suggest that this dichotomy in complexity perception significantly accounts for the observed variability in agent behaviors, particularly in terms of their decision-making self-reliance [7]. Complexity Overestimators, who habitually perceive tasks as more challenging than they are, tend to lean more on existing team knowledge, increasing their reliance on the group as task complexity escalates. Complexity Underestimators, conversely, display a stronger inclination towards self-reliance, intensifying this trait as the complexity of tasks increases Moreover, individuals with lower self-reliance base their decisions on their perception of task complexity, while those with higher self-reliance are influenced more by the actual complexity [3]. These findings (Table 1) highlight the nuanced interplay between individual and group score, perceived and actual complexity, and the impact of individual traits in collective problem-solving.

In summary, our findings unveil an exploration social dilemma that intensifies with task complexity. This dilemma manifests as an increasing tendency among individuals to rely on existing knowledge as complexity rises, which paradoxically undermines collective intelligence. However, we also discover that this trend towards inefficiency can be mitigated by the presence of individuals who typically underestimate task complexity and inherently fulfill the team's exploratory needs. Despite earning lower individual rewards on average, these agents are pivotal in enhancing the group's collective intelligence. This revelation holds profound implications for policy and governance, emphasizing the need to balance the composition of team members based on their attitudes towards complexity perception. Furthermore, it underscores the importance of providing support for Complexity Underestimators, particularly in environments characterized by high complexity.

Unpublished working draft. Not for distribution.

	(M27)	(M37)	(M47)	(M57)	(M67)	(M77)	(M87)	(M97)
Intercept	**-1.362** ***	**-1.494** ***	**-2.395** ***	**-1.473** ***	**-0.9261** ***	**-1.833** *	**-2.395** ***	**-0.6404** ***
	-2.46E-13	(0.1578)	(0.5919)	(0.3923)	(0.2072)	(0.7585)	(0.5919)	(0.1517)
Self Reliance	**2.724** ***	**2.733** ***	**4.206** ***	1.749		2.594	**4.206** ***	
	-6.86E-13	(0.01955)	(0.9231)	(1.486)		(2.195)	(0.9231)	
Complexity (Low)		0.2274	1.693	0.1162	0.1162	1.83	1.693	0.2274
		(0.263)	(0.8991)	(0.2754)	(0.2754)	(0.9673)	(0.8991)	(0.263)
Player_Score				0.0007157	0.0007157	-0.0009302		
				-0.0003745	-0.0003745	-0.001259		
Num Trails				0.003586	0.003586	-0.02461		
				(0.02616)	(0.02616)	(0.1228)		
Performer - Low						-2.259	**-1.876** ***	**-0.2784** ***
						(1.496)	(0.07154)	(0.001815)
Self Reliance: Low Complexity			-2.348			-2.739	-2.348	
			(1.364)			(1.447)	(1.364)	
Self Reliance: Player_Score						0.002687		
						-0.001917		
Self Reliance: Num Trails						0.04824		
						(0.1746)		
nobs	912	912	912	912	912	912	912	912
logLik	-573.5	-572.2	-567.1	-570.2	-570.2	-564.2	-567.1	-572.2
AIC	1261	1260	1252	1260	1260	1254	1252	1260
BIC	1535	1540	1536	1549	1549	1558	1536	1540
nobs.1	912	912	912	912	912	912	912	912

Standard errors are heteroskedasticity robust. *** p 0.001; ** p 0.01; * p 0.05.

Table 1: "Correct Prediction" estimated by self-reliance and Treatment of complexity

Unpublished working draft. Not for distribution.

References

[1] Daniel Barkoczi and Mirta Galesic. 2016. Social learning strategies modify the effect of network structure on group performance. *Nature Communications* 7, 1 (oct 2016), 1–8. https://doi.org/10.1038/ncomms13109 arXiv:1606.00753

[2] Chelsea M Campbell, Eduardo J Izquierdo, and Robert L Goldstone. 2022. Partial copying and the role of diversity in social learning performance. *Collective Intelligence* 1, 1 (aug 2022), 263391372210818. https://dl.acm.org/doi/10.1177/26339137221081849

[3] Ketika Garg, Paul E Smaldino, and Christopher T Kello. 2024. Evolution of explorative and exploitative search strategies in collective foraging. *Collective Intelligence* 3, 1 (jan 2024), 1–16. https://doi.org/10.1177/26339137241228858

[4] Naomi Ehrich Leonard and Simon A Levin. 2022. Collective intelligence as a public good. *Collective Intelligence* 1, 1 (aug 2022), 263391372210832. https://doi.org/10.1177/26339137221083293

[5] Ioanna Lykourentzou, Federica Lucia Vinella, Faez Ahmed, Costas Papastathis, Konstantinos Papangelis, Vassilis-Javed Khan, and Judith Masthoff. 2022. Self-organization in online collaborative work settings. *Collective Intelligence* 1, 1 (aug 2022), 263391372210780. https://doi.org/10.1177/26339137221078005

[6] J.G. March. 1991. Exploration and exploitation in organizational learning. *Organization Science* 2, 1 (1991), 71 – 87. arXiv:z0009

[7] Niccolò Pescetelli, Anna Katharina Hauperich, and Nick Yeung. 2021. Confidence, advice seeking and changes of mind in decision making. *Cognition* 215 (oct 2021). https://doi.org/10.1016/j.cognition.2021.104810

[8] Christoph Riedl, Young Ji Kim, Pranav Gupta, Thomas W. Malone, and Anita Williams Woolley. 2021. Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences of the United States of America* 118, 21 (may 2021), e2005737118. https://doi.org/10.1073/pnas.2005737118

[9] Jan W. Rivkin. 2000. Imitation of complex strategies. *Management Science* 46, 6 (2000), 824–844. https://doi.org/10.1287/mnsc.46.6.824.11940

[10] Vince J. Straub, Milena Tsvetkova, and Taha Yasseri. 2020. The cost of coordination can exceed the benefit of collaboration in performing complex tasks. *Collective Intelligence* 2, 2 (apr 2020), 263391372311569. https://doi.org/10.1177/26339137231156912 arXiv:2009.11038

Large Language Models (LLMs) for Communication Disorders: Collective Intelligence from Speech Language and Hearing Science

YAO DU, University of Southern California, USA
BINGSHENG YAO, Northeastern University, USA
DAKUO WANG*, Northeastern University, USA

The rise of artificial intelligence and machine learning (ML) approaches offers novel support and enhancement to collaborative work in healthcare, including diagnostic decision-making and clinical services for speech-language and hearing disorders. This paper reviews existing ML research in speech, language, and hearing science, and then discusses ways to build collective intelligence using Large Language Models (LLMs) for understanding voice, speech, and cognitive communication in humans.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; **Collaborative and social computing theory, concepts and paradigms**.

Additional Key Words and Phrases: speech language hearing science, human-centered artificial intelligence, large language models

ACM Reference Format:
Yao Du, Bingsheng Yao, and Dakuo Wang. 2018. Large Language Models (LLMs) for Communication Disorders: Collective Intelligence from Speech Language and Hearing Science. In *Proceedings of ACM Collective Intelligence (CI '24).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Speech language and hearing science is an interdisciplinary field that connects speech science, language science, and hearing science to investigate human communication across the lifespan. Speech-language pathologists (SLPs) and audiologists are professionals who screen, assess, evaluate, treat, and monitor pediatric, adult, and geriatric clients with communication disorders across educational settings (e.g., K-12 classrooms), medical settings (e.g., hospitals, skilled nursing facilities, clinics), home health, and virtual telehealth settings. Therefore, multimodal data streams can be stored in educational learning management systems (LMSs), electronic medical records (EMRs), and telehealth platforms. To better understand how these disorders impact individuals' ability to communicate, the American Speech-Language-Hearing Association (ASHA) provided several key areas of SLPs' professional scope of practice, also known as the "Big 9" scope of practice [1, 19]. Individuals with communication disorders may receive clinical diagnosis and intervention services across different biophysiological levels (e.g., voice, swallowing), sensory and perceptual levels (e.g., hearing, articulation, fluency), and linguistic-social-cognitive levels (e.g., receptive and expressive language, social and cognitive communication). Despite the increased adoption and use of web, mobile, and assistive technology tools for speech, language, and hearing services[19], there is a growing need and knowledge gap in applying machine learning (ML) approaches such as Large Language Models (LLMs) to multimodal clinical data to improve service delivery[7, 8, 17, 18].

*Corresponding author.

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 the author(s) 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.
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

1.1 Biophysiological Level: Voice Disorders

Voice data are easy, inexpensive, and noninvasive to capture as accurate biomarkers with different modalities, including but not limited to videomics (video endoscopy analysis), (2) radiomics (radiology images), (3) acoustic data (spectrograms), and (4) clinical data [7]. Voice disorders can be classified as either (1) organic disorders due to physical changes in structural mechanism (e.g., nodules) or neurogenic due to impairments in the nervous system (e.g., tremors), or (2) functional disorders due to voice use (e.g., vocal fatigue) [5]. Voice disorders are typically evaluated via various dimensions (e.g., voice quality, pitch, and loudness) through screening and assessment of respiration, phonation (how the vocal folds and breath are used to make sound), resonance (the voice quality that results from the balance of sound vibrations in the throat, mouth, and nose), and vocal range and flexibility (pitch and range, loudness, endurance). Prior studies have explored applying ML algorithms to a variety of multimodal data (e.g., audio samples, video, and clinical notes) for improving screening, detection, management, and prognosis of voice disorders [13, 23], laryngeal cancer [7], and gender-affirming voice care [6]. Researchers also reported challenges in identifying appropriate audio tasks, acoustic features, processing algorithms, and additional factors in recording environments [11]. Since the majority of the research studies are primarily trained on "single-institution databases with limited data sets and a single data modality" [7] and lack a collective voice databank to curate a shared database across institutions, therefore, ongoing efforts to streamline the proper voice data pipeline have been initiated in the past few years by clinicians [9].

1.2 Sensory & Perceptual Level: Speech Disorders

Speech is the behavior of forming and sequencing the sounds of human communication using different articulators such as the mouth, lips, and tongue. Individuals with speech sound disorders may have two possible causes: organic vs. functional [4]. Organic speech sound disorders are caused by motor/neurological causes (e.g., dysarthria, apraxia), structural deficits (e.g., cleft lip/palate), or sensory/perceptual causes (e.g., hearing loss); functional speech sound disorders have unknown etiologies and can result in either articulation disorders (error in producing individual speech sounds) or phonological disorders (errors in rule-based errors that affect more than one sound). Differing from voice data, since speech data are typically captured in the form of an audio waveform, audio analysis requires transforming raw waveform into spectrogram for analysis based on time and frequency [25]. Most ML studies on speech disorders have focused on developing databases for specific speech disorders (e.g., dysarthria), patient populations (e.g., amyotrophic lateral sclerosis), or use cases in areas such as bilingual speech perception [24], speech sound disorders [12, 22], and individuals with neurodegenerative conditions [20, 21].

1.3 Linguistic, Social, and Cognitive Level: Cognitive-Communication Disorders

Language is a system of symbols that can be broken down into expressive language (language production and expression) and receptive language (language comprehension and understanding) in different spoken, gestural, and written forms. Language assessment and intervention can focus on language content (semantics or vocabulary), form (morphology, syntax), and function (pragmatics or use). Different types of language disorders may have different manifestations in pediatric and adult populations (e.g., childhood spoken language disorders, aphasia) [2, 3], and the process of collecting language samples is a common practice for patients across the lifespan. These linguistic data gathered via clinical activities such as language assessments, narrative storytelling, and conversations, can become important databases for LLM researchers. In addition, language disorders may also lead to social communication disorders and cognitive communication disorders (e.g., traumatic brain injury or TBI, and dementia), which are also important ASHA Big 9

areas. Specifically, cognitive aspects of communication include attention, memory, problem-solving, organization, etc. Intervention requires not only the use of complex linguistic skills but also high-level cognitive processing and additional cognitive load. In recent years, tools such as ChatGPT and voice assistants have been increasingly utilized in TBI and dementia research for detection and prediction [15, 16] and supporting memory and social relationships [10, 14].

2 DISCUSSION

2.1 Building Collective Intelligence Data Pipeline for Speech, Language, and Voice Data

Recent technological advances enabled different ML approaches such as LLMs with opportunities to serve as diagnostic and intervention supports to generate materials, simulate human-like communication, and serve as an intelligent speech-language and voice assistant for the therapist [8, 10, 14]. Despite varying degrees of implementation of LLMs for different multimodal data, recent studies have predominately focused on using MLs for detecting or classifying disease severity, making repeated clinical decisions to identify ongoing symptoms as diseases progress, and simulating the progress of particular degenerative conditions. It is particularly worth noting that many of these ML approaches did not happen during synchronous in-person or online patient encounters, which leads to unlimited potential for multimodal data to be gathered, curated, and analyzed via asynchronous communication and hybrid service delivery (e.g., in-person clinician visit and telehealth ML support). In the future, with immersive technology tools (e.g., mobile health, virtual reality) for direct and indirect patient-provider communication and interaction, more ML approaches can be directly implemented and integrated into clinical simulations both prior to and after direct patient care. However, to ensure patient and clinician interactions can provide the best data input and output pipeline, it is highly recommended that both AI and clinical researchers partner with human computer interaction designers and researchers, who can ensure any interfaces that are utilized for patient and clinician data collection will have inclusive and accessible audiovisual, tactile, and conversational interface design.

2.2 Creating A Collective Intelligence Research Community Using LLMs for Speech-Language and Hearing Science

To work on multimodal data, researchers need to integrate theoretical frameworks and evidence-based assessment and intervention procedures from clinicians while attending to critical system design and engineering principles to ensure the feasibility and implementation of ML. It is critical to acknowledge that research in this area is contributed by diverse scholarly communities of researchers, such as (1) physical science researchers in auditory neuroscience, image and signal processing, and perception and cognition who investigate the biological and physical processes of human communication, (2) social science researchers in psychology and linguistics who investigate the psychological and linguistic process of communication, and (3) clinical practitioners such as speech-language pathologists and audiologists who diagnose and treat individuals with speech, language, and hearing disorders. Specifically, working on language disorders may require communication with language science researchers who have already established systematic clinical practice procedures and research methodologies implementing both linguistic and statistical approaches for data analysis. Partnerships with language scientists can expedite AI researchers' progress toward developing theory-driven models and algorithms in LLM research. Creating such a collaborative research community is the first step towards building a truly translational science research community. As HCI and AI researchers, a collaborative framework Ai-RPD (AI, Research, Practice, and Design) could support team-building initiatives that can contribute to the collective intelligence from different scientific communities using ML approaches such as LLMs.

REFERENCES

[1] American Speech-Language-Hearing Association. 1993. Definitions of communication disorders and variations. https://www.asha.org/policy/rp1993-00208/

[2] American Speech-Language-Hearing Association. n.d.. Aphasia. https://www.asha.org/practice-portal/clinical-topics/aphasia/

[3] American Speech-Language-Hearing Association. n.d.. Childhood Spoken Language Disorders. https://www.asha.org/public/speech/disorders/preschool-language-disorders/

[4] American Speech-Language-Hearing Association. n.d.. Speech Sound Disorders-Articulation and Phonology. https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology/

[5] American Speech-Language-Hearing Association. n.d.. Voice Disorders. https://www.asha.org/practice-portal/clinical-topics/voice-disorders/

[6] Yael Bensoussan, Jeremy Pinto, Matthew Crowson, Patrick R Walden, Frank Rudzicz, and Michael Johns III. 2021. Deep learning for voice gender identification: proof-of-concept for gender-affirming voice care. *The Laryngoscope* 131, 5 (2021), E1611–E1615.

[7] Yael Bensoussan, Erik B Vanstrum, Michael M Johns III, and Anaïs Rameau. 2023. Artificial intelligence and laryngeal cancer: from screening to prognosis: a state of the art review. *Otolaryngology–Head and Neck Surgery* 168, 3 (2023), 319–329.

[8] Yao Du and Felix Juefei-Xu. 2023. Generative AI for Therapy? Opportunities and Barriers for ChatGPT in Speech-Language Therapy. https://openreview.net/forum?id=cRZSr6Tpr1S

[9] Emily Evangelista, Rohan Kale, Desiree McCutcheon, Anais Rameau, Alexander Gelbard, Maria Powell, Michael Johns, Anthony Law, Phillip Song, Matthew Naunheim, et al. 2023. Current Practices in Voice Data Collection and Limitations to Voice AI Research: A National Survey. *The Laryngoscope* (2023).

[10] Clément Gakuba, Charlene Le Barbey, Alexandre Sar, Gregory Bonnet, Damiano Cerasuolo, Mikhael Giabicani, and Jean-Denis Moyer. 2024. Evaluation of ChatGPT in predicting 6-month outcomes after traumatic brain injury. *Critical care medicine* (2024), 10–1097.

[11] Rijul Gupta, Dhanshree R Gunjawate, Duy Duong Nguyen, Craig Jin, and Catherine Madill. 2024. Voice disorder recognition using machine learning: a scoping review protocol. *BMJ open* 14, 2 (2024), e076998.

[12] Sarah E Gutz, Marc F Maffei, and Jordan R Green. 2023. Feedback From Automatic Speech Recognition to Elicit Clear Speech in Healthy Speakers. *American Journal of Speech-Language Pathology* 32, 6 (2023), 2940–2959.

[13] Sarika Hegde, Surendra Shetty, Smitha Rai, and Thejaswi Dodderi. 2019. A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice* 33, 6 (2019), 947–e11.

[14] Gahangir Hossain, Znaida Simpson Pomare, and Gayle Prybutok. 2024. ChatGPT: A Companion for Dementia Care. In *2024 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1–6.

[15] Eli Kurtz, Youxiang Zhu, Tiffany Driesse, Bang Tran, John A Batsis, Robert M Roth, and Xiaohui Liang. 2023. Early Detection of Cognitive Decline Using Voice Assistant Commands. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[16] Xiaohui Liang, John A Batsis, Youxiang Zhu, Tiffany M Driesse, Robert M Roth, David Kotz, and Brian MacWhinney. 2022. Evaluating voice-assistant commands for dementia detection. *Computer Speech & Language* 72 (2022), 101297.

[17] Yi Liu, Yuekang Li, Gelei Deng, Felix Juefei-Xu, Yao Du, Cen Zhang, Chengwei Liu, Yeting Li, Lei Ma, and Yang Liu. 2023. ASTER: Automatic speech recognition system accessibility testing for stutterers. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 510–521.

[18] Qisheng Lu and Shaomei Wu. 2024. Towards Fair and Inclusive Speech Recognition for Stuttering: Community-led Chinese Stuttered Speech Dataset Creation and Benchmarking. (2024).

[19] Kathryn Lubniewski, Lori Price, Yao Du, and Nikita Soni. 2023. Using Multi-Stakeholder Perspectives to Enhance Integration of Mobile Technology for Students with Communication Needs. *ISTES BOOKS* (2023), 175–230.

[20] Robert L MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A Ladewig, Jimmy Tobin, Michael P Brenner, Philip C Nelson, et al. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia.. In *Interspeech*, Vol. 2021. 4833–4837.

[21] Katrin Tomanek, Katie Seaver, Pan-Pan Jiang, Richard Cave, Lauren Harrell, and Jordan R Green. 2023. An analysis of degenerating speech due to progressive dysarthria on ASR performance. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[22] Subhashini Venugopalan, Jimmy Tobin, Samuel J Yang, Katie Seaver, Richard JN Cave, Pan-Pan Jiang, Neil Zeghidour, Rus Heywood, Jordan Green, and Michael P Brenner. 2023. Speech Intelligibility Classifiers from 550k Disordered Speech Samples. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[23] Laura Verde, Nadia Brancati, Giuseppe De Pietro, Maria Frucci, and Giovanna Sannino. 2021. A deep learning approach for voice disorder detection for smart connected living environments. *ACM Transactions on Internet Technology (TOIT)* 22, 1 (2021), 1–16.

[24] Zilong Xie, Rachel Reetzke, and Bharath Chandrasekaran. 2019. Machine learning approaches to analyze speech-evoked neurophysiological responses. *Journal of Speech, Language, and Hearing Research* 62, 3 (2019), 587–601.

[25] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. 2023. Audio diffusion model for speech synthesis: A survey on text to speech and speech enhancement in generative ai. *arXiv preprint arXiv:2303.13336* (2023).

Orchestrate: An AI-Powered Cross-Device System to Enhance Group Cohesion and Collective Intelligence at Workshops

Jude Rayan
jrayan@ucsd.edu
The Design Lab, University of California San Diego
La Jolla, California, USA

Jinmao Wang
jiw117@ucsd.edu
The Design Lab, University of California San Diego
La Jolla, California, USA

Yifan Gong
y5gong@ucsd.edu
The Design Lab, University of California San Diego
La Jolla, California, USA

Nan Jiang
n3jiang@ucsd.edu
The Design Lab, University of California San Diego
La Jolla, California, USA

Steven P. Dow
spdow@ucsd.edu
The Design Lab, University of California San Diego
La Jolla, California, USA

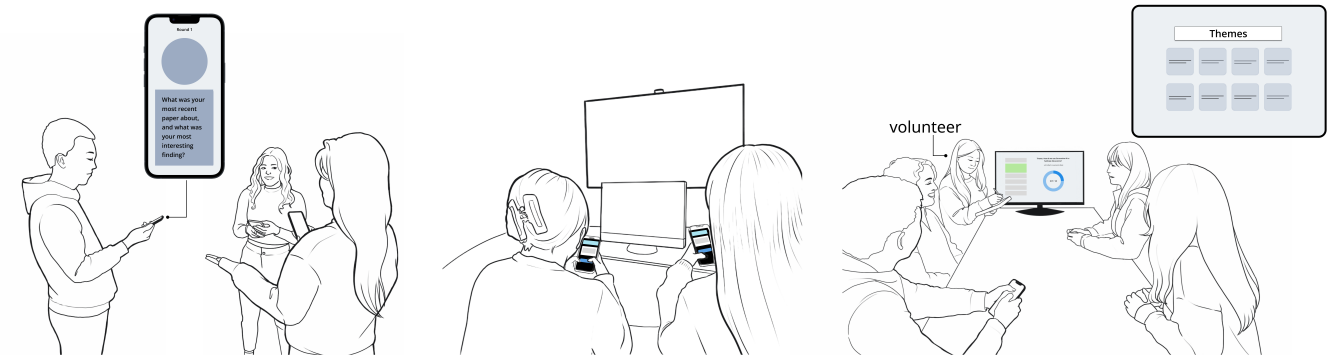


Figure 1: (left) During the Mingle stage, workshop participants get personalized questions to introduce themselves; (middle) During the Ideate stage, participants use their mobile phones to see personalized inspiration questions for ideation and submitting their ideas; (right) During the Synthesize stage, all the submitted ideas will be clustered as themes. The large room display will show these themes and the tables will discuss each of these themes. Participants will be assigned to their theme discussion, based on the content of their submitted ideas. Each table will also have a volunteer to take down notes. After this stage, they will move onto the Reflect stage where participants from other groups can provide feedback.

ABSTRACT

An important ingredient for collective intelligence is creating space for exchange and exposure of new perspectives between individuals. Workshops have the potential to facilitate CI, but real-world constraints can impede communities from building cohesion, exchanging perspectives, and capturing and integrating new insights into existing knowledge representations. We propose a cross-device system to facilitate workshops by coordinating information across personal and shared displays and by leveraging Generative AI to create conversation cues based on participant interests and ongoing conversation.

KEYWORDS

Generative AI, Workshops, Social Computing

ACM Reference Format:
Jude Rayan, Jinmao Wang, Yifan Gong, Nan Jiang, and Steven P. Dow. 2024. Orchestrate: An AI-Powered Cross-Device System to Enhance Group Cohesion and Collective Intelligence at Workshops. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 3 pages.

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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI '24, June 26–29, 2024, Boston, MA
© 2024 Copyright held by the owner/author(s).

1 INTRODUCTION

Collective Intelligence (CI) thrives when there are opportunities to share diverse perspectives [9], to create group cohesion [3], and to capture new knowledge on top of existing knowledge [5].

The citizen science game FoldIt, for instance, captures individual decisions and actions to get insights on protein designs that could help prevent or treat important diseases [7]. Online voting systems, like conversational swarm intelligence, explore different strategies for reaching consensus [15]. Workshops, on the other hand, are a more common and flexibly approach for enabling CI; workshop participants interact fluidly, share ideas, and chart out agendas for future collaborations [8].

Whether virtual or in-person, the workshop format can pose a number of difficulties. The constrained time frame for workshops — whether it's several hours or days — places a burden on organizers and limits opportunities for exploration and exchange of information [16]. For instance, participants from different backgrounds can struggle to establish common ground [14] or to connect with others who may provide new insights, due to a lack of time for direct interaction. Even with effective methods to enable mingling among participants [18, 22], the challenge of initiating conversations persists. Also since many in-person workshops use paper and other analog resources, it creates a challenge around preserving new knowledge in a digital form that can be viewed and built upon later [2]. Online or virtual workshops can ease the burden of travel and thus enable more people to participate, including underrepresented or under-resourced groups [4]. However, remote communication technology (connectivity issues, the learning curve for new tools, etc) [14]. While the virtual format makes it easier to digitally capture insights discussed at a workshop, social interactions often suffer [2, 18]. Hybrid workshops that attempt to blend in-person and online interactions are difficult to manage and often treat the remote people like second-class participants [17, 20].

The research community has offered solutions to increase the momentum and productivity of workshops. For example, the FreeForm Templates project explored how to computationally model "design data" captured on a collaborative digital whiteboard to augment later thinking exercises [11]. Capturing design data across a community enables the creation of interactive visualizations, such as stakeholder maps [10]. To bridge the communication gap between participants joining remotely and in-person during conferences, researchers explored various approaches to telepresence to enhance togetherness such as movable cameras [12], robots [21], and virtual avatars [24].

Recent advances in technology could help unlock the full potential of CI-oriented workshops. For one, most in-person workshop attendees have mobile devices with internet connectivity that could be used to deliver personalized information; large and relatively inexpensive displays can be positioned in workshops to convey shared information. Researchers have also explored the concept of proxemics surrounding devices allowing unique interaction opportunities based on their proximity to the user [1, 6, 23]. Meanwhile, advances in Generative AI (Gen AI) have made it possible to quickly interpret and generate responses to long form text inputs. Given the challenges and opportunities, our research explores: How can we combine advances in Gen AI and multi-display environments to create more effective workshops? We propose Orchestrate, a workshop system that addresses challenges pertaining to organizing in-person workshops and leverages Gen AI to enhance team cohesion, awareness, and collective intelligence.

2 SCENARIO

Imagine the 2025 Collective Intelligence Conference hosting a series of workshops dedicated to fostering awareness regarding the United Nations Sustainable Development Goals (SDGs) [13]. The participants are a diverse cohort, including experts in their respective fields, as well as enthusiastic novices who are looking forward to meeting new people. Their goals include networking and knowledge exchange, with a unified desire to gain new insights and innovate solutions to sustainable development.

2.1 Workshop Setup and Logistics

The co-located workshop will include 30-50 participants and organizers within the same space with room for standing discussions and tables for group discussions. A large common display visible to all participants will be used for topic introductions and instructions. Participants will scan a QR code with their mobile devices where a browser interface will guide them to meet and ideate with other attendees. Each table will have a discussion facilitator and medium-size shared display for group-level information. Gen AI will be weaved into the workshop by providing conversation cues during mingling and ideation, and by helping to organize ideas and synthesize group proposals.

Before the workshop starts, participants will be invited to complete a preliminary survey to assess their background, research interests, and particular engagement with and interest in the UN SDGs. The data collected from this survey will provide valuable insights into the participants' expertise and will enable Orchestrate to dynamically form groups and cues for effective discussions.

2.2 Stages of the Workshop

The stages described are tailored for a specific type of workshop, but the various components and time allocated to each stage can be modified as required to suit the workshop's objectives and constraints.

- Mingle (5-min per round for 5 rounds): In order to facilitate group members getting acquainted with one another and fostering team cohesion, Orchestrate will make recommendations on how to group people over the course of several rounds and provide them personalized cues to support socializing/mingling. The group assignments and cues will be displayed on their mobile phone. Members can find their groups in the standing area and participate accordingly.
- Ideate (8-min per round for 3 rounds): This stage enables participants to share their ideas and engage in discussions in smaller groups (2-3 ppl). With the help of personalized cues aimed at different dimensions of the problem space, the goal is to support divergence across the entire workshop by engaging everyone in parallel ideation. Members can request more inspiration cues through the same interface. Orchestrate will cluster ideas based on topic similarity and then also recommend which participants will continue developing each cluster during the Synthesize stage.
- Synthesize (45 min to 1 hour): Orchestrate will synthesize all the submitted ideas and cluster them as themes with the help of Gen AI. Additionally, members will now be assigned to each of those themes as a bigger group (7-8 ppl). Members

will then sit down at assigned large tables and review all the ideas generated for a particular topic. Each table will have a large display and an assigned student volunteer. In this stage, Orchestrate will pick each member to talk about their best idea or the theme behind some of their submitted ideas to ensure turn-taking. Each member will be able to view their list of submitted ideas on their mobile phone. After each idea introduction, the other members will be given a set time to share their thoughts or feedback. This process repeats till every member gets to introduce their ideas or theme of ideas. Orchestrate will also be transcribing this conversation. The role of the student volunteer is to take down notes from the discussions to make sure the transcription did not miss out on any important information. Near the end of the timeframe, Orchestrate can generate a fleshed-out proposal for the group to collaboratively edit. The final proposal will be a synthesis of top ideas and the discussion transcripts/notes written by the student volunteer

- Reflect: Orchestrate could continue to support interactions post-workshop by asking participants from other groups to give feedback on aspects of proposals and providing suggestions for collaborative action.

3 DISCUSSION

Orchestrate is aimed at accomplishing a set of goals to facilitate collective intelligence processes during workshops. Firstly, the system must successfully aid group cohesion but also needs to carefully strike a balance between creating connections and cues with tech, and distracting people from having genuine conversations (we don't want people staring at their displays the whole time). Qualitative analysis of conversation transcript data during the workshop such as communication sequences [19] can help us study discussion dynamics and group cohesion. Secondly, Orchestrate must also support the workshop organizer's specific plans and objectives, while remaining flexible to the situation on the ground. Some groups might want to talk for longer and might have to end discussions abruptly to stay aligned in time with the other groups. System log data will give insights into the time spent across different activities and the influence of Gen AI. Finally, Orchestrate must capture the knowledge generated during these discussions to facilitate reflection and to enable future groups to build upon it. We aim to prototype Orchestrate to get insights that could accelerate CI for a variety of workshop formats.

REFERENCES

[1] Till Ballendat, Nicolai Marquardt, and Saul Greenberg. 2010. Proxemic interaction: designing for a proximity and orientation-aware environment. In *ACM International Conference on Interactive Tabletops and Surfaces*. 121–130.

[2] Zoe M Becerra, Nadia Fereydooni, Andrew L Kun, Angus McKerral, Andreas Riener, Clemens Schartmüller, Bruce N Walker, and Philipp Wintersberger. 2021. Interactive workshops in a pandemic: the real benefits of virtual spaces. *IEEE Pervasive Computing* 20, 1 (2021), 35–39.

[3] Prerna Chikersal, Maria Tomprou, Young Ji Kim, Anita Williams Woolley, and Laura Dabbish. 2017. Deep structures of collaboration: Physiological correlates of collective intelligence and group satisfaction. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 873–888.

[4] Cesar O Estien, Eli B Myron, Callie A Oldfield, Ajisha Alwin, and Ecological Society of America Student Section. 2021. Virtual scientific conferences: benefits and how to support underrepresented students. *The Bulletin of the Ecological Society of America* 102, 2 (2021), e01859.

[5] Dawn G Gregg. 2010. Designing for collective intelligence. *Commun. ACM* 53, 4 (2010), 134–138.

[6] Jens Emil Grønbæk, Mille Skovhus Knudsen, Kenton O'Hara, Peter Gall Krogh, Jo Vermeulen, and Marianne Graves Petersen. 2020. Proxemics beyond proximity: Designing for flexible social interaction through cross-device interaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[7] Firas Khatib, Seth Cooper, Michael D Tyka, Kefan Xu, Ilya Makedon, Zoran Popović, David Baker, and Foldit Players. 2011. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences* 108, 47 (2011), 18949–18953.

[8] Evangelos Kosmidis, Panagiota Syropoulou, Stavros Tekes, Philipp Schneider, Eleftherios Spyromitros-Xioufis, Marina Riga, Polychronis Charitidis, Anastasia Moumtzidou, Symeon Papadopoulos, Stefanos Vrochidis, et al. 2018. hackAIR: Towards raising awareness about air quality in Europe by developing a collective online platform. *ISPRS International Journal of Geo-Information* 7, 5 (2018), 187.

[9] Jan Marco Leimeister. 2010. Collective intelligence. *Business & Information Systems Engineering* 2 (2010), 245–248.

[10] Stephen MacNeil, Zijian Ding, Kexin Quan, Ziheng Huang, Kenneth Chen, and Steven P Dow. 2021. ProbMap: Automatically constructing design galleries through feature extraction and semantic clustering. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*. 134–136.

[11] Stephen MacNeil, Ziheng Huang, Kenneth Chen, Zijian Ding, Alexander Yu, Kendall Nakai, and Steven P Dow. 2023. Freeform Templates: Combining Freeform Curation with Structured Templates. In *Proceedings of the 15th Conference on Creativity and Cognition*. 478–488.

[12] Hideyuki Nakanishi, Yuki Murakami, and Kei Kato. 2009. Movable cameras enhance social telepresence in media spaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 433–442.

[13] United Nations. n.d.. The 17 Goals | Sustainable Development. https://sdgs.un.org/goals Accessed on May 24, 2024.

[14] Eslam Nofal. 2023. Participatory Design Workshops: Interdisciplinary Encounters within a Collaborative Digital Heritage Project. *Heritage* 6, 3 (2023), 2752–2766.

[15] Louis Rosenberg, Gregg Willcox, Hans Schumann, and Ganesh Mani. 2024. Conversational Swarm Intelligence amplifies the accuracy of networked groupwise deliberations. In *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 0086–0091.

[16] Jonathan D Salter, Cam Campbell, Murray Journeay, and Stephen RJ Sheppard. 2009. The digital workshop: Exploring the use of interactive and immersive visualisation tools in participatory planning. *Journal of environmental management* 90, 6 (2009), 2090–2101.

[17] Shervin Shirmohammadi, Shun-Yun Hu, Wei Tsang Ooi, Gregor Schiele, and Arno Wacker. 2012. Mixing virtual and physical participation: The future of conference attendance?. In *2012 IEEE International Workshop on Haptic Audio Visual Environments and Games (HAVE 2012) Proceedings*. IEEE, 150–155.

[18] Jaeyoon Song, Christoph Riedl, and Thomas W Malone. 2021. Online mingling: supporting ad hoc, private conversations at virtual conferences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–10.

[19] Evelyn TS Tan, Katja Rogers, Lennart E Nacke, Anders Drachen, and Alex Wade. 2022. Communication sequences indicate team cohesion: A mixed-methods study of Ad Hoc league of legends teams. *Proceedings of the ACM on Human-Computer Interaction* 6, CHI PLAY (2022), 1–27.

[20] Anthony Tang, Carman Neustaedter, and Saul Greenberg. 2007. Videoarms: embodiments for mixed presence groupware. In *People and Computers XX—Engage: Proceedings of HCI 2006*. Springer, 85–102.

[21] Katherine M Tsui, Munjal Desai, Holly A Yanco, and Chris Uhlik. 2011. Exploring use cases for telepresence robots. In *Proceedings of the 6th international conference on Human-robot interaction*. 11–18.

[22] Gustavo Umbelino, Matin Yarmand, Samuel Blake, Vivian Ta, Amy Luo, and Steven P Dow. 2021. Prototeams: Supporting team dating in co-located settings. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27.

[23] Julie R Williamson, Joseph O'Hagan, John Alexis Guerra-Gomez, John H Williamson, Pablo Cesar, and David A Shamma. 2022. Digital proxemics: Designing social and collaborative interaction in virtual environments. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–12.

[24] Kevin Yu, Gleb Gorbachev, Ulrich Eck, Frieder Pankratz, Nassir Navab, and Daniel Roth. 2021. Avatars for teleconsultation: Effects of avatar embodiment techniques on user perception in 3d asymmetric telepresence. *IEEE Transactions on Visualization and Computer Graphics* 27, 11 (2021), 4129–4139.

Collective Intelligence in Agile Teams: Reviewing Facilitated Processes

Stephen J. Thorpe

Project Management Research Office, Software Engineering Research Centre, Department of Computer Science and Software Engineering, Auckland University of Technology, Stephen.Thorpe @ aut.ac.nz

ABSTRACT

Agile presents facilitated ways of working that have fast become amongst the most popular approaches to team collaboration, especially in the software engineering field. In an agile context the work of the facilitator is often undertaken by the Scrum Master role. These facilitators help groups as a process guide rather than being highly involved in the content or decision-making. They draw on both a knowledgebase of tools, methods and techniques, as well as the person-centered qualities of the interpersonal relationships between the facilitator and group. A number of studies have explored the detection and presence of *Collective Intelligence* (CI) in teams, however, there remained an opportunity to explore the work of facilitators and the processes they use with groups to shift them away from traditional control models to that of mutual learning. CI offers a significant opportunity to further enhance agile practices. Through a systematic literature review this paper explored the intersection between facilitated agile approaches and collective intelligence within the context of software delivery. It appeared that despite the focus on team collaboration there is scant literature on CI within the software engineering literature explored. The review of agile teamwork showed little evidence of CI being understood. This paper sheds light on the potential of CI in agile that may be invaluable for organizations aiming to enhance their agility and management capabilities.

CCS CONCEPTS

• Software and its engineering • Software organization and properties • Human-centered computing • Collaborative and social computing • Collaborative and social computing theory, concepts and paradigms.

Keywords: Agile, Large-Scaled Agile Methods, Collective Intelligence, Software Delivery.

ACM Reference Format:

1 INTRODUCTION

Collective Intelligence (CI) can be considered a crucial factor in the success of teams and organizations, as it influences their problem-solving capabilities and overall performance (Woolley, et al., 2015). As we discover more about CI and the ways that a team's collective intelligence potential may be reached, we find that CI may not only be about the factors relating to the structure or the characteristics of the individuals in a team, but more about what is going on between them. For example, studies have demonstrated that teams consistently perform better than the best individuals when solving analytical tasks, both in performance and speed (Carey & Laughlin, 2012; Laughlin et al., 2006). Similarly, teams also have been found to outperform a set of an equivalent number of independent workers. Even though each member of a group may put in less effort on average, the gains of collaboration have usually dominated such effort losses (Mao et al, 2016). Facilitators such as Hunter and Thorpe (2020) have discussed approaches that facilitators may consider in generating presence and accessing collective intelligence with the groups and teams they work with.

1.1 Agile is still eating the world

What's somewhat unique about software delivery is that most of the tasks, belonging to the development of software, can be considered complex tasks (de Campos Candria & Araujo, 2022) and in need of a higher level of skills at both the individual and team level. While many have discussed the global phenomenon of agile ways of working - particularly in technology development and delivery (see Denning 2018; 2019; 2024), little research has explored collective intelligence with facilitated ways of working at the team level, nor the relationship between the facilitator's role or the process they use. At scale, agile has been found to run into challenges when dealing with complexity, complex systems, addressing the dilemmas between the needs of many stakeholders, large code bases, and alignment in distributed teams (Nowroozi, et al., 2020; Vaipulu, et al., 2023). This represents a gap in our knowledge and presented an opportunity to explore the literature at the intersection between facilitated agile approaches and collective intelligence within the context of software delivery. The following research question was formulated: *What is known about CI in facilitated agile approaches within the context of software development and delivery?* The aim of the review was to investigate and evaluate the current state of knowledge and whether there was evidence of CI being understood within the literature on agile teamwork.

2 RESEARCH METHOD

Following Kitchenham's (2004) procedures for conducting a systematic literature review (SLR), the review sought relevant peer-reviewed literature in the ACM Digital Library and IEEE Xplore published in the last 5 years.

3 RESULTS

From the literature a census of 14 papers were identified that appeared to address the question of CI in a software context. To summarize, agile concepts identified in the review are mapped against some of the prominent CI concepts in Table 1.

Table 1: Summary of concepts in the literature

CI Concept: Equality in Conversational Turn-Taking
Agile Concepts from the Review
A distributed set of flattened structures (Mora & Velandia, 2020)
Expectations about behavior (Zhang et al., 2020)
Trust - the intensity with which links were maintained (Zhang et al., 2020)
Self-organization and re-structuring (Akhrif et al., 2019)
Signals from contribution patterns were predictive of success (Horvát et al., 2023)
Coalition formation, power dynamics, reciprocity, and performance were linked to collaboration formation agency (Lykourentzou et al., 2022)
Aligned values & shared agreements (Rashid et al., 2023)
The facilitation of team norms an important role in enabling team performance (Stray et al. 2016),
Knowledge exchange between team members was linked with higher performance (Ryś & Gloor, 2023)
Communication with team members was also statistically significant (Ryś & Gloor, 2023)
CI Concept: Social Sensitivity of Group Members
Agile Concepts from the Review
Feelings and perceptions impact the way relationships perform (Ferreira et al. 2019)
The burden of communication did not outweigh the benefits as member size increased (Murić et al., 2019)
The importance of psychological safety and the role of reflection in changing norms (Stray et al., 2016)
Agile processes reinforced acceptable behaviors (Stray et al., 2016)
A team's interpersonal relationships moderate the effect of team diversity (Diegmann & Rosenkranz, 2016)
The intersection between an agile mindset & digital emotional intelligence lay in self-awareness (Ryś & Gloor, 2023)
Psychological empowerment, communication and collaboration, and respect (Ryś & Gloor, 2023)
Empathy essential for self and others, to evaluate emotions and to utilize them effectively (Ryś & Gloor, 2023)
CI Concept: Team Diversity
Agile Concepts from the Review
Reputation - Expectations about behavior (Zhang et al., 2020)
More collaborators found to be beneficial (Murić et al., 2019)
Cognitive diversity highly important (Ochoa-Zambrano 2021)
Team cognitive style and cognitive style heterogeneity (Diegmann & Rosenkranz, 2016)
The diversity of professional background, nationality, and previous experience an important factor influencing collaboration (Ryś & Gloor, 2023)

4 DISCUSSION

What may be understood from the review is that facilitated agile processes along with its underpinning principles may provide both a mindset for psychological empowerment as well as processes that lead to proactive behaviors, such as, collaborative communication, power sharing, turn taking, and the sharing of knowledge and information. These practices also align well with promoting sensitivity to others, resilient relationships, and valuing diversity as a source of creative input to creative team problem solving.

Agile methods rely heavily on teamwork, team composition, communication, and interpersonal relationships, and yet software engineering research has traditionally tended to view the creation of technology and information systems as more important than the particular management frameworks or activities leading to its creation and delivery. Team research has been seen merely as adjacent (e.g., Kozlowski & Bell, 2003; Kozlowski & Ilgen, 2006). Some research has tended to apply theories of organizational psychology on IT use, adoption, and acceptance, rather than on its creation. This has been somewhat unfortunate as the reality over several decades has been that IT projects are not so much technological as sociological in nature (DeMarco & Lister, 1987, Denning 2018; 2019; 2024). Coordination and communication between various stakeholders are known to be necessary for successful implementation (Gallivan & Keil, 2003, Ko et al., 2005), and creating a shared understanding between stakeholders has been well established as a major driver for information system development success (Gallivan & Keil, 2003, Tan, 1994). While the individual and organization-wide level of effects on software development have been a focus, team-level effects are covered less so in the literature, and it can be argued that results have been contradictory (Diegmann & Rosenkranz, 2016).

5 CONCLUSION

Given the importance of a team's collaboration process in CI, this review has explored software engineering literature to identify what relationship agile ways of working may have, or might, offer in relationship to CI. 14 peer-reviewed articles from the IEEE Xplore and ACM Digital Databases were reviewed, however, there still remains an opportunity to explore the intersection between facilitated agile approaches and collective intelligence within the context of software delivery.

The aim of the review was to investigate and evaluate the current state of knowledge and whether there was evidence of CI being understood within the literature on agile teamwork. It would be difficult to argue that this has been achieved in this particular review. However, while the review may not have identified the determinants of how CI may function in agile teams, it has identified some similarities in agile approaches that appear to align well. This review of predominantly software engineering literature on agile teamwork showed little evidence of CI being understood. The opportunity remains to broaden the search and to seek further insights into facilitated agile processes and their relationship with CI.

REFERENCES

Akhrif, O., El bouzekri el idrissi, Y., & Hmina, N. (2019, October). Enabling smart collaboration with Smart University Services. In *Proceedings of the 3rd International Conference on Computer Science and Application Engineering* (pp. 1-6).

de Campos Candria, D., & de Araujo, R. M. (2022, May). Crowdsourcing software development-a possible path? In *Proceedings of the XVIII Brazilian Symposium on Information Systems.* (pp. 1-8).

Carey, H. R., & Laughlin, P. R. (2012). Groups perform better than the best individuals on letters-to-numbers problems: Effects of induced strategies. *Group Processes and Intergroup Relations, 15*(2), 231-242.

Denning, S. (2018). Why Agile is eating the world. *Forbes*, January 2.

Denning, S. (2019). Why Agile's Future Is Bright. *Forbes*, August 25.

Denning, S. (2024). Why The World's Most Valuable Firms Are So Agile. *Forbes*, February 26.

Demarco, T. & Lister, T. (1987) *Peopleware: Productive Projects and Teams*, Dorset House.

Diegmann, P., & Rosenkranz, C. (2016). Improving software development efficiency–how diversity and collective intelligence shape agile team efficiency. (2016). *International Research Workshop on IT Project Management 2016*. Paper 7. http://aisel.aisnet.org/irwitpm2016/7

Ferreira, T. M., Costella, F. L., Zanetti, A. B., da Silva, S. E., Zanatta, A. L., & De Marchi, A. C. B. (2019, May). CrowdRec: A prototype recommendation system for crowdsourcing platforms using Google Venture Design: Google Venture Design Sprint. In *Proceedings of the XV Brazilian Symposium on Information Systems* (pp. 1-8).

Gallivan, M. J. & Keil, M. (2003). The user-developer communication process: A critical case study, *Information Systems Journal, 13*(1), 37-68.

Horvát, E. Á., Dambanemuya, H. K., Uparna, J., & Uzzi, B. (2023, April). Hidden indicators of collective intelligence in crowdfunding. In *Proceedings of the ACM Web Conference 2023* (pp. 3806-3815).

Hunter, D., & Thorpe, S. J. (2020). The Gift of Presence in Groups: An unfolding story of transformative learning. In *Social Ecology and education: transforming worldviews and practices* (pp. 89-100). Routledge.

Kitchenham, B. A. (2004). Procedures for Undertaking Systematic Reviews, *Joint Technical Report*, Computer Science Department, Keele University (TR/SE-0401) and National ICT Australia Ltd. (0400011T.1).

Ko, D. G., Kirsch, L. J. and King, W. R. (2005). Antecedents of knowledge transfer from consultants to clients in enterprise system implementations, *MIS Quarterly, 29*(1), 59-85.

Kozlowski, S. W. and Bell, B. S. (2003). *Work Groups and Teams in Organizations*, Wiley, London.

Kozlowski, S. W. and Ilgen, D. R. (2006). Enhancing the Effectiveness of Work Groups and Teams, *Psychological Science in the Public Interest, 7*(3), 77-124.

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: effects of group size. *Journal of Personality and social Psychology, 90*(4), 644.

Lykourentzou, I., Vinella, F. L., Ahmed, F., Papastathis, C., Papangelis, K., Khan, V. J., & Masthoff, J. (2022). Self-organization in online collaborative work settings. *Collective Intelligence, 1*(1), 26339137221078005.

Mao, A., Mason, W., Suri, S., & Watts, D. J. (2016). An experimental study of team size and performance on a complex task. *PLOS One, 11*(4), e0153048.

Mora, J. A. H., & Velandia, M. C. C. (2020, June). Proposal for a conceptual decision-making model based on collective intelligence for SMEs of the photovoltaic solar energy sector in Colombia. In 2020 *47th IEEE Photovoltaic Specialists Conference (PVSC)* (pp. 0559-0563). IEEE.

Murić, G., Abeliuk, A., Lerman, K., & Ferrara, E. (2019). Collaboration drives individual productivity. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1-24.

Nowroozi, A., Teymoori, P., Ramezanifarkhani, T., Besharati, M. R., & Izadi, M. (2020). A crisis situations decision-making systems software development process with rescue experiences. *IEEE Access, 8*, 59599-59617.

Ochoa-Zambrano, J. (2021, June). How collective intelligence can gear agility with sustainability. In *Proceedings of the International Conference on Agile Software Development*, (pp. 69-77). Cham: Springer International Publishing.

Rashid, M. T., Wei, N., & Wang, D. (2023). A survey on social-physical sensing: An emerging sensing paradigm that explores the collective intelligence of humans and machines. *Collective Intelligence, 2*(2), 26339137231170825.

Ryś, M., & Gloor, P. A. (2023). There is truth in legends: Lessons for team performance from hackathons. *Computer, 56*(8), 38-47.

Stray, V., Fægri, T. E., & Moe, N. B. (2016). Exploring norms in agile software teams. In Product-Focused Software Process Improvement. In *Proceedings of the 17th International Conference, PROFES 2016*, Trondheim, Norway, November 22-24, 2016, (pp. 458-467). Springer International Publishing.

Tan, M. (1994). Establishing mutual understanding in systems design: An empirical study, *Journal of Management Information Systems, 10*(4), 159-182.

Vaipulu, T., Lal, R., & Thorpe, S. (2023). A systematic review of complexity in agile software project delivery. In *Proceedings of the Australasian Conference on Information Systems* (ACIS 2023), https://aisel.aisnet.org/acis2023/82.

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective intelligence in teams and organizations. *Handbook of Collective Intelligence*, 143-168.

Zhang, J., Liao, S., Tang, L., Liu, Y., Huang, S., & Wang, X. (2020, December). Agile C2 organizational decision allocation and pattern evolution methods. In 2020 *IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)* (Vol. 9, pp. 357-370). IEEE.

Harnessing Partisan Motives to Solve the Misinformation Problem

Non-Archival Extended Abstract

Cameron Martel†
Sloan School of Management
Massachusetts Institute of Technology
Cambridge MA USA
cmartel@mit.edu

Jennifer Allen†
Sloan School of Management
Massachusetts Institute of Technology
Cambridge MA USA
jnallen@mit.edu

Gordon Pennycook
Department of Psychology
Cornell University
Ithaca NY USA
grp58@cornell.edu

David G. Rand
Sloan School of Management
Massachusetts Institute of Technology
Cambridge MA USA
rand.dave@gmail.com

ABSTRACT

Partisan motives are often conceptualized as fundamentally in opposition to accuracy-directed motives. Rather than being opposed, however, it may be that political and accuracy motivations simply operate independently – in which case partisan motives may not necessarily interfere with truth discernment. Here, we evaluate this possibility in the context of crowd evaluations of (mis)information. To empirically assess this, we carried out a survey study and analyzed field data from Twitter's crowdsourced fact-checking platform Community Notes. These data show that more politically motivated individuals are more active community fact-checking participants, helping sustain overall contribution levels. Furthermore, more politically motivated participants engage in more politically biased flagging yet exhibit the same or better flagging discernment as less politically motivated participants. Our results challenge the notion that partisan motives inherently undermine the ability (and willingness) to evaluate truth – instead political motivation may actually be the key to motivating the provision of high quantity and quality fact-checks.

CCS CONCEPTS

• Collaborative and social computing • misinformation • social media

KEYWORDS

Community Notes, partisanship, accuracy motives

† Shared first- and corresponding authorship

Permission to make digital or hard copies of part or all 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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CI'24, June 2024, Boston, Massachusetts USA
© 2024 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06…$15.00

ACM Reference format:

Cameron Martel, Jennifer Allen, Gordon Pennycook, and David G. Rand. 2024. Harnessing Partisan Motives to Solve the Misinformation Problem: Non-Archival Extended Abstract: Insert Subtitle Here. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27-28, Boston, MA, 4 pages.*

1 Introduction

Many scholars conceptualize directional partisan motives as fundamentally in tension with accuracy- or truth-seeking motives [1-5]. This lens is commonly applied to the topic of misinformation, where it is argued that increasing polarization has created a "post-truth" world where people actively disregard accuracy in pursuit of partisan goals [6-10].

Concerns about partisan motives are particularly salient when considering "wisdom of crowds" approaches for identifying misinformation – a burgeoning strategy for keeping up with the rapid pace of false news online. Although crowds of laypeople have been shown to effectively identify misinformation at scale with high accuracy [11-15], partisan motives may undermine such efforts – particularly in a voluntary system in which users can choose what to fact-check. Directionally motivated partisans may exploit misinformation-reporting mechanisms to target content that is simply counter-partisan, rather than inaccurate. Recent evidence indeed shows that participants in community fact-checking systems are more likely to flag counter-partisan posts and evaluate co-partisan fact-checks as high quality [16-17]. Other research has posited that these political biases might increase flagging of reliable and neutral sources [18] and could decrease reporting of false news that users agree with politically.

Rather than being diametrically opposed, however, it may be that these motives simply operate independently. This

alternative perspective resonates with recent findings that engaging in reasoning neither amplifies nor reduces partisan bias – but instead increases truth discernment for both politically concordant and discordant information [19-20]. If partisan motives and accuracy motives are independent, then – rather than partisan motives necessarily reducing truth discernment – these two motivations may interact in more complicated ways [21-22]. Here we ask whether partisan motives undermine or work in tandem with accuracy motives in the domain of crowdsourced identification of misinformation.

2 Methods and Results

2.1 Partisan motives predict more biased, but more discerning, misinformation flagging

We empirically assess whether misinformation flagging is a context where partisan and accuracy motives work in opposition or in tandem. We first test our predictions with a preregistered survey study using Americans recruited from the survey firm Lucid [25] (n=2,380), quota-matched to the national representativeness. Participants were told that they would be shown many social media posts about a variety of topics. They were then asked to help social media companies identify misinformation by identifying the posts they believed to be false. If they believed a post to be false, they were instructed to flag it and write an explanation about why. For posts they believed to be true or not misleading, they were asked to just continue scrolling to the next post. Participants were taken to a simulated newsfeed where they could scroll through 80 posts; 20 posts were entertainment items (items that were either true or had no verifiable claim of fact), and the remaining 60 posts were news headlines balanced on both veracity and whether they were more favorable to Democrats or Republicans (as determined via pretesting, [26] for details).

Participants were more likely to flag false, politically discordant posts than any other type of headline, followed by posts that were false and politically concordant (b=-0.234, p<0.001). Flag rates were substantially lower for entertainment (b=-1.688, p<0.001), concordant true (b=-1.798, p<0.001), and even discordant true (b=-1.580, p<0.001) headlines relative to discordant false posts.

We also assessed the impact of political motivation by measuring (at the study onset) out-party animosity, issue polarization, and political knowledge. We conducted principal component analysis on these measures and found that all three constructs had high weights in the first principal component. We therefore consider this first principal component a continuous measure of aggregate political motivation across these dimensions.

Next, we evaluated the relationship between political motivation and overall flagging rates. Across all participants who started the main headline flagging task, 30.7% flagged no headlines, underscoring how even in a survey task environment, motivating the provision of fact-checks is of paramount concern. We found that more politically motivated participants contributed a greater number of overall flags (average marginal effect (AME) from quasi-Poisson model: b=1.493, p<.001). To illustrate this, a median split on political motivation shows that low politically motivated participants (n=1,189) wrote 5,305 flags, while high politically motivated participants (n=1,191) contributed 10,259 flags. These findings are qualitatively similar when independently assessing our political motivation measures – greater out-party animosity, issue polarization, and political knowledge are each associated with increased overall flagging.

We find that more politically motivated participants showed greater bias in flagging. That is, such participants flagged more discordant than concordant posts overall (ordinary least squares (OLS) regression with analytic weighting by flag count: b=0.034, p<0.001). Critically, however, more politically motivated individuals were also more discerning in their flagging – they flagged fewer true and non-political posts as false (OLS regression with analytic weighting by flag count: b=-0.054, p<0.001). We further examine the specific types of posts highly politically motivated participants flag in particular. Namely, more politically motivated individuals specifically flag more false, discordant posts than all other headline types (average marginal contrasts from quasi-Poisson model for political motivation (baseline = false, discordant): false, concordant: b=0.289, p<0.001; true, discordant: b=0.825, p<0.001; true, concordant: b=0.955, p<0.001; non-political: b=0.965, p<0.001).

In sum, we find that (a) participants were much more likely to flag false headlines than true headlines, (b) participants were more likely to flag discordant than concordant headlines, and (c) more politically motivated participants flagged headlines at about twice the rate of less politically motivated participants and exhibited greater bias and discernment in their flagging. Overall, less politically motivated users wrote 3,265 flags on false headlines and 2,040 flags on true and entertainment headlines, while more politically extreme users contributed 7,896 flags on false headlines and 2,363 flags on true headlines. Thus, not only is the true positive rate greater for more politically motivated participants (77% versus 61.5%) but excluding flags from more politically motivated users would decrease the number of total flags by about two-thirds.

2.2 Empirical testing in the field: Twitter's Community Notes

To evaluate whether our findings generalize to an actual online news evaluation system, we assess evidence from Twitter's crowdsourced fact-checking program Community Notes (formerly Birdwatch). Twitter users admitted to Community Notes are able to classify tweets as 'Misinformed

or potentially misleading' and write a free response 'note' explanation as to why. We assessed a dataset of notes from the first six months of Community Notes. We focused on instances where notes classified the original tweet as potentially misleading. In total, our dataset consisted of 4,442 notes classifying 3,151 unique tweets as potentially misleading.

Of 4,025 users who applied and were accepted as Community Notes users, only 1,046 contributed notes – such that 74% of users did not write a single note. Given this low participation rate, we next consider the role of partisan motivation as a predictor of actual fact-check provisioning. We assess political extremity as the absolute value of users' partisanship score, as inferred by the accounts users follow (following [27]). We also find that more politically extreme users write more community notes (AME for quasi-Poisson model: b=0.451, p<.001).

However, are these notes actually high quality – or rather, do more politically motivated Community Notes users simply classify all counter-partisan content as misleading? To evaluate this question, we first sent a subset of 461 unique tweets to be evaluated on veracity by two professional fact-checkers, as well as to be evaluated by Amazon Mechanical Turk workers on partisan favorability. Given these classifications, we first find, promisingly, that 79.4% of notes agreed with our professional fact-checkers and were written about tweets classified as potentially misleading. In particular, we find that the substantial majority of these notes were written about false, discordant tweets. However, we also find a notable proportion of notes written on true, discordant tweets (16.2%). How does political extremity relate to the types of notes users contribute? As predicted, we find that more politically extreme users write a greater proportion of notes on politically discordant tweets (b=0.051, p=0.010). That is, politically extreme users are more biased in their community fact-checking. However, we do not find evidence that more politically extreme users are more or less discerning in their fact-checking – we do not find that political extremity is associated with substantially more or less agreement with professional fact-checker evaluations (b=-0.021, p=0.132).

When assessing note writing by tweet category, we indeed find that more politically extreme users are especially more likely to write notes on false, discordant tweets relative to false, concordant (average marginal contrast from quasi-Poisson model from political extremity (baseline = false, discordant): b=0.275, p<.001) and true, concordant (average marginal contrast: b=0.287, p<.001) tweets. However, we find mixed evidence that political extremity is more associated with false, discordant note writing than true, discordant note writing (average marginal contrast: b=0.226, SE=0.054, z(5,772)=4.15, p<.001, and OLS model interaction is significant; but overall quasi-Poisson model interaction is not significant), such that discernment is relatively similar across political extremity levels.

These field data from Twitter demonstrate that more politically extreme users contribute more to an online community fact-checking system, where the majority of users do not participate at all. Despite these users exhibiting greater bias in fact-checking by preferentially flagging counter-partisan content, more politically motivated users maintain the same overall level of flagging discernment. Practically, consider the contributions of low versus high politically extreme users (median split) on the subset of tweets we had evaluated for veracity and partisan favorability. Less politically extreme users wrote 314 notes on tweets classified as false and 83 notes on tweets classified as true, while more politically extreme users contributed 632 notes on false tweets and 169 notes on true tweets. While the true positive rate is nearly equivalent (79.1% versus 78.9%), excluding notes from more politically extreme users would result in only about one-third as many correct notes contributed. Politically motivated users are integral to the sustained usability of crowdsourced fact-checking systems – and can promote a high quantity of evaluations without reducing flagging discernment.

3 Discussion

Widely held theories posit that political polarization undermines high quality information sharing [6]. Coupled with robust evidence of partisan selective fact-checking [16], a major concern looming over the potential of crowdsourced fact-checking is that individuals will flag too much – inappropriately flagging true but counter-partisan content. The current work suggests that the much bigger hurdle is insufficient flagging – most people are not motivated enough by accuracy concerns alone to fact-check anything. Rather than undermine the system, partisan motives may help address this challenge by promoting people to flag content that is both misleading and counter-partisan. In sum, we argue that crowdsourcing can successfully work to identify misinformation at scale because of – rather than in spite of – partisan motivations.

REFERENCES

[1] Taber, C. S., & Lodge, M. Motivated skepticism in the evaluation of political beliefs. *Am. J. Pol. Sci.* **50**, 755-769 (2006).

[2] Kunda, Z. The case for motivated reasoning. *Psychol. Bull.* **108**, 480-498 (1990).

[3] Druckman, J. N. The politics of motivation. *Crit. Rev.* **24**, 199-216 (2012).

[4] Druckman, J. N., & McGrath, M. C. The evidence for motivated reasoning in climate change preference formation. *Nat. Clim. Change* **9**, 111-119 (2019).

[5] Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. Motivated numeracy and enlightened self-government. *Behav. Public Policy* **1**, 54-86 (2017).

[6] Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *Am. Polit. Sci. Rev.* **115**, 999-1015 (2021).

[7] Peterson, E., & Iyengar, S. Partisan gaps in political information and information-seeking behavior: motivated reasoning or cheerleading? *Am. J. Pol. Sci.* **65**, 133-147 (2021).

[8] Van Bavel, J. J., & Pereira, A. The partisan brain: An identity-based model of political belief. *Trends Cogn.* **22**, 213-224 (2018).
[9] Rathje, S., Roozenbeek, J., Van Bavel, J. J., & van der Linden, S. Accuracy and social motivations shape judgments of (mis)information. *Nat. Hum. Behav.* 1-12 (2023).
[10] Marie, A., & Petersen, M. B. Motivations to affiliate with audiences drive the sharing of partisan (mis)information on social media. *OSF Preprints* (2023).
[11] Martel, C., Allen, J., Pennycook, G., & Rand, D. G. Crowds can effectively identify misinformation at scale. *Perspect. Psychol. Sci.* 17456916231190388 (2022).
[12] Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. Scaling up fact-checking using the wisdom of crowds. *Sci. Adv.* **7**, eabf4393 (2021).
[13] Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *J. Online Trust Safety* **1** (2021).
[14] Resnick, P., Alfayez, A., Im, J., & Gilbert, E. Informed crowds can effectively identify misinformation. *arXiv preprint arXiv:2108.07898* (2021).
[15] Arechar, A. A., Allen, J., Berinsky, A. J. *et al.* Understanding and combatting misinformation across 16 countries on six continents. *Nat Hum Behav* (2023).
[16] Allen, J., Martel, C., & Rand, D. G. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *Proc. 2022 CHI Conf. Hum. Factors Comput. Syst.* 1-19 (2022).
[17] Yasseri, T., & Menczer, F. Can crowdsourcing rescue the social marketplace of ideas? *arXiv preprint* (2021). https://arxiv.org/abs/2104.13754
[18] Coscia, M., & Rossi, L. Distortions of political bias in crowdsourced misinformation flagging. *J. R. Soc. Interface* **17**, 20200020 (2020).
[19] Bago, B., Rand, D. G., & Pennycook, G. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *J. Exp. Psychol. Gen.* **149**, 1608-1613 (2020).
[20] Pennycook, G., & Rand, D. G. The psychology of fake news. *Trends Cogn.* **25**, 388-402 (2021).
[21] Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. The wisdom of polarized crowds. *Nat. Hum. Behav.* **3**, 329-336 (2019).
[22] Duncan, M. Selective rating: partisan bias in crowdsourced news rating systems. *J. Inf. Technol. Politics* **19**, 360-375 (2022).
[23] Anderson, E. T., & Simester, D. I. Reviews without a purchase: Low ratings, loyal customers, and deception. *J. Mark. Res.* **51**, 249-269 (2014).
[24] Goes, P. B., Guo, C., & Lin, M. Do incentive hierarchies induce user effort? Evidence from an online knowledge exchange. *Inf. Syst. Res.* **27**, 497-516 (2016).
[25] Coppock, A., & McClellan, O. A. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Res. Polit.* **6(1)**, 2053168018822174 (2019).
[26] Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychol.* **7(1)**, 25293 (2021).
[27] Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychol. Sci.* **26**, 10, 1531–1542 (2015).

Short title: COOPERATING WHEN COMPETING IN INNOVATION CONTESTS

Full title: COOPERATING WHEN COMPETING: HOW AND WHEN DO COMPETITIVE REWARD STRUCTURES IN INNOVATION CONTESTS AFFECT COOPERATIVE BEHAVIOR

Submission type: Non-Archival Extended Abstract

Christoph Grimpe

Copenhagen Business School, Department of Strategy and Innovation, cg.si@cbs.dk

Marion Poetz

Copenhagen Business School, Department of Strategy and Innovation, mp.si@cbs.dk

Nathan Rietzler

Northeastern University, D'Amore-McKim School of Business, n.rietzler@northeastern.edu

Keywords: innovation contests, cooperation, competition, social interdependence theory, configurational approach

EXTENDED ABSTRACT

Innovation contests have gained a prominent role in firms' efforts to access external knowledge that they can integrate into their innovation process (e.g., Boudreau et al. 2011, Körpeoğlu and Cho 2018, Piezunka and Dahlander 2015, Terwiesch and Xu 2008). Based on the notion of broadcast search to reach more distant sources of innovation (Afuah and Tucci 2012), problems are posted to potential problem solvers who compete for the best solution which is then rewarded by the problem holder with a prize (Liu et al. 2014). But many contest platforms also allow individual problem solvers to cooperate. They can exchange knowledge by posting their ideas, commenting on others' ideas, and incorporating ideas and comments from others into their own solution before they submit. By allowing contestants to cooperate, platforms seek to reap the benefits of both cooperation and competition, that is, to enable knowledge recombination to produce superior solutions (Girotra et al. 2010, Singh and Fleming 2010) and to increase the effort that individuals exert (Boudreau et al. 2011, Körpeoğlu and Cho 2018). However, cooperating by exchanging ideas and knowledge in innovation contests with others creates a public good dilemma (Cabrera and Cabrera 2002), as group members can benefit from the shared resources without depleting them, irrespective of whether they contributed to their production (Olson 1965), and hence free-ride (Sweeney 1973). Moreover, exchanging ideas and knowledge is associated with cost, even though modest, and may imply individuals lose their competitive edge (Cabrera and Cabrera 2002).

While innovation contests assign individuals to a competitive reward structure, social interdependence theory of cooperation and competition would lead us to believe that cooperative reward structures are important to motivate individuals to exchange ideas and knowledge (Deutsch 1949, Tjosvold 1989). Social interdependence exists as the group of individuals is linked by competitive social interaction in which the individuals' goal achievements are negatively correlated because only one or a few can obtain the prize (Beersma et al. 2003, Johnson 2003, Johnson and Johnson 1989). However, the simple notion that cooperative rewards promote knowledge exchange and mutually supportive behavior in groups while competitive rewards lead individuals to keep valuable information proprietary and to harm the progress of others (Beersma et al. 2003) seems incorrect in light of widespread evidence that individuals do cooperate in innovation contests (Riedl and Seidel 2018). For example, contestants heavily use the discussion board on Kaggle.com, an online contest platform for data scientists, statisticians, and machine learning experts, on which they can, but do not have to share their coding scripts, allowing others to read, discuss, vote for, and integrate them into their own submission. Prior research suggests that individuals may be motivated to cooperate because they seek to gain higher status or reputation in a crowd of problem solvers, because the contest may be about finding a pro-social, not-for-profit solution, because learning from each other may increase the winning chances, or because individuals derive pleasure and fun from engaging in the contest (e.g., Füller et al. 2014, Jin et al. 2021, Kathan et al. 2015, Majchrzak and Malhotra 2020). These insights indicate that the relationship between competitive reward structures and cooperation might be influenced by a set of interrelated contest and solver attributes. But it is precisely these interrelations that complicate the identification of the effect of competition on cooperation in innovation contests. In that sense, understanding the conditions under which competitive reward structures lead individuals to cooperate (or not) is critical for understanding the boundary conditions of social interdependence theory and its theoretical predictions (Kistruck et al. 2016).

Hence, in this paper, we ask how and when competitive reward structures in innovation contests influence the likelihood that individuals will choose to cooperate. To answer this question, we extend social interdependence theory by integrating insights from the literature on coopetition. Defined as a relationship between two or more actors that simultaneously engage in cooperation and competition (Bengtsson and Kock 2014), coopetition has been predominantly viewed as an arrangement on the firm level that enables exchanging knowledge and other types of resources to contribute to an organization's innovation performance (Cassiman et al. 2009, Gnyawali and Park 2011, Ritala and Hurmelinna-Laukkanen 2013). But coopetition is also a pervasive phenomenon on other levels. For example, organizational units in multiunit organizations typically compete for internal resources as well as status and rewards and, at the same time, seek to learn from each other and

exchange knowledge (Tsai 2002). On the individual level, employees compete for promotion, but still work together in day-to-day business (DeVaro 2006, Baumann et al. 2019). Despite the relevance of innovation contests for firms' innovation performance (Jeppesen and Lakhani 2010), prior literature has yet to view cooperation in innovation contests as coopetition on the individual level, offering an opportunity for theorizing in extension of social interdependence theory. In particular, innovation contests typically involve one-time interactions between problem solvers instead of repeated opportunities to interact that firm employees have.

Our research proceeds in two steps. First, we study the role of different competitive reward structures (with no, low, and high degrees of competition) on the likelihood of individuals to cooperate in an innovation contest. Competition typically results from several attributes, including evaluation, reward, and a win-lose aspect that is characteristic to competitive situations (Amabile et al. 1996; Brown et al. 1998, Murayama and Elliot 2012). We focus on the structural competition that an individual is subject to in a given situation and manipulate monetary reward structures. Consistent with social interdependence theory, we argue that more competitive reward structures (higher levels of the prize money) lead individuals to refrain from cooperating in innovation contests. Second, since individuals may have all kinds of reasons to cooperate irrespective of competition, we are interested in how different attribute configurations in an innovation contest regarding the problem solvers' own attributes, the attributes of other contest participants, and the attributes of the contest itself influence the likelihood that contestants will choose to cooperate, given different competitive reward structures. Configurational approaches facilitate insights into the equifinality of different combinations of attributes or characteristics (Fiss 2007), i.e., we seek to understand which configurations would lead individuals to cooperate under non-competitive, lowly competitive, and highly competitive reward structures of the innovation contest.

Empirically, we conduct two studies. In Study 1, we implement an experimental vignette study with 405 participants to examine the effect of different competitive reward structures on individuals' likelihood to cooperate in innovation contests. Following Collins and Smith (2006), we measure cooperation as an individual's decision to exchange ideas and knowledge with other contest participants. Our results show a negative effect of competitive reward structures on cooperation that is more pronounced for individuals' likelihood to post their ideas and provide comments than to incorporate ideas and knowledge from others in their own ideas. While more than half of the participants are inclined to cooperate when there is no competition at all, counterintuitively, about one-third of the participants continue to cooperate even in the face of high competition.

In Study 2, we apply a configurational approach to examine the complex decision-making of individuals in a contest, with the goal of better understanding why so many individuals choose to cooperate even under competitive conditions. We use a multi-step approach to identify and categorize potentially relevant drivers of cooperation under competition by means of a literature review and interviews with real-world participants in innovation contests that enable cooperation among contestants. The outcome of this attribute exploration feeds into an adaptive choice-based conjoint (ACBC) study with 562 study participants. Despite competing for the prize, we identify several configurations of attributes that increase the likelihood of cooperation in innovation contests. In each competitive reward structure, surprisingly even in the absence of competition, cooperation is more likely when contests are configured to reduce fears of misappropriation. Moreover, in each competitive reward structure, cooperation is also more likely for configurations that include a broader and more diverse set of similarly important attributes, dominated by either the expectation of extrinsic rewards or the pursuit of participants' intrinsic motivations. While configurations focusing on expectations of extrinsic rewards emphasize the existence of rewards for sharing ideas and knowledge with others or the ability to benefit extensively from the contributions of others, configurations focusing on participants' desire to pursue their intrinsic motivations emphasize fun and learning experiences that a contest creates or their personal interest in the topic. The two motivational configurations vary with the degrees of competition, with intrinsic and extrinsic attributes blended in the two competitive structures and clearly separated in the non-competitive structure. In this sense, while competition generally discourages contestants from cooperating, we can identify several equifinal attribute configurations that increase the likelihood of cooperation and explain the per se paradoxical observation of widespread cooperation in innovation contests.

We contribute to the extant literature in at least three ways. First, understanding when individuals outside organizational boundaries – who do not form teams but constitute a group linked only by competitive social interaction – engage in coopetition given different configurations of innovation contests extends social interdependence theory (Johnson and Johnson 1989, Johnson 2003, Beersma et al. 2003) and, more broadly, advances research on the management and organization of innovation (Argyres and Silverman 2004, Dahlander and Frederiksen 2012). Social interdependence theory typically focuses on an organizational context, in which team members employed by the organization are assigned to either competitive or cooperative reward structures (Deutsch 1949, Tjosvold 1989). While this suggests that organizations must choose between reaping the benefits from either competition (e.g., Boudreau et al. 2011) or cooperation (e.g., Singh and Fleming 2010), our research shows that specific configurations of innovation contests can harness both competition and cooperation. By explicating configurations of contest and solver attributes, our research explains the widespread occurrence of cooperation under competition in a context outside of organizational boundaries in which individuals typically interact with each other just once.

Second, we advance literature on innovation contests (e.g., Boudreau et al. 2011, Jeppesen and Lakhani 2010) and knowledge sharing in online communities (e.g., Jarvenpaa and Majchrzak 2010, Wasko and Faraj 2005) by systematically identifying the attributes that influence cooperation under competition. Despite the theoretical interest in and practical relevance of innovation contests as an emerging organizational form (e.g., Afuah and Tucci 2012, Riedl and Seidel 2018), the decision of individuals to cooperate in such contests has remained undertheorized. Previous literature has predominantly studied attributes influencing this decision in isolation, such as reciprocity considerations as drivers of knowledge sharing in online communities (e.g., Kathan et al.

2015), or on different levels of analysis, such as the relationship between competitive or cooperative rewards and knowledge sharing in teams (e.g., Baer et al. 2010, Beersma et al. 2003). Our research advances a configurational perspective, highlighting the interrelationship between contest and solver attributes that lead to a number of equifinal configurations increasing an individual's likelihood of cooperation under competition.

Third, we advance a model of individual-level coopetition in innovation contests that is informed by social interdependence theory and rooted in a configurational analysis. While prior coopetition research has typically focused on the firm level (Cassiman et al. 2009, Gnyawali and Park 2011, Ritala and Hurmelinna-Laukkanen 2013) or the unit and individual level within the boundaries of the organization (Tsai 2002, DeVaro 2006, Baumann et al. 2019), our research fills a void by extending coopetition research to a context outside organizational boundaries. In turn, this extension allows to view phenomena related to open innovation more generally (Chesbrough 2003), such as broadcast and distant search for innovation (Afuah and Tucci 2012, Jeppesen and Lakhani 2010), through a coopetition lens.

REFERENCES

Afuah, A., C.L. Tucci. 2012. Crowdsourcing As a Solution to Distant Search. *Academy of Management Review* **37**(3) 355-375.

Amabile, T.M., R. Conti, H. Coon, J. Lazenby, M. Herron. 1996. Assessing the Work Environment for Creativity. *Academy of Management Journal* **39**(5) 1154-1184.

Argyres, N.S., B.S. Silverman. 2004. R&D, organization structure, and the development of corporate technological knowledge. *Strategic Management Journal* **25**(89) 929-958.

Baer, M., R.T.A. Leenders, G.R. Oldham, A.K. Vadera. 2010. Win or lose the battle for creativity: The power and perils of intergroup competition. *Academy of Management Journal* **53**(4) 827-845.

Baumann, O., J.P. Eggers, N. Stieglitz. 2019. Colleagues and Competitors: How Internal Social Comparisons Shape Organizational Search and Adaptation. *Administrative Science Quarterly* **64**(2) 275-309.

Beersma, B., J.R. Hollenbeck, S.E. Humphrey, H. Moon, D.E. Conlon, D.R. Ilgen. 2003. Cooperation, Competition, and Team Performance: Toward a Contingency Approach. *Academy of Management Journal* **46**(5) 572-590.

Bengtsson, M., S. Kock. 2014. Coopetition—Quo vadis? Past accomplishments and future challenges. *Industrial Marketing Management* **43**(2) 180-188.

Boudreau, K.J., N. Lacetera, K.R. Lakhani. 2011. Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis. *Management Science* **57**(5) 843-863.

Brown, S.P., W.L. Cron, J.W. Slocum, Jr. 1998. Effects of trait competitiveness and perceived intraorganizational competition on salesperson goal setting and performance. *Journal of Marketing* **62**(4) 88-98.

Cabrera, A., E.F. Cabrera. 2002. Knowledge-Sharing Dilemmas. *Organization Studies* **23**(5) 687-710.

Cassiman, B., M.C. Di Guardo, G. Valentini. 2009. Organising R&D Projects to Profit From Innovation: Insights From Co-opetition. *Long Range Planning* **42**(2) 216-233.

Chesbrough, H.W. 2003. *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business Press, Cambridge, MA.

Collins, C.J., K.G. Smith. 2006. Knowledge Exchange and Combination: The Role of Human Resource Practices in the Performance of High-Technology Firms. *Academy of Management Journal* **49**(3) 544-560.

Dahlander, L., L. Frederiksen. 2012. The Core and Cosmopolitans: A Relational View of Innovation in User Communities. *Organization Science* **23**(4) 988-1007.

Deutsch, M. 1949. A theory of co-operation and competition. *Human Relations* **2**(2) 129-152.

DeVaro, J. 2006. Strategic promotion tournaments and worker performance. *Strategic Management Journal* **27**(8) 721-740.

Fiss, P.C. 2007. A set-theoretic approach to organizational configurations. *Academy of Management Review* **32** 1180-1192.

Füller, J., K. Hutter, J. Hautz, K. Matzler. 2014. User Roles and Contributions in Innovation-Contest Communities. *Journal of Management Information Systems* **31**(1) 273-308.

Girotra, K., C. Terwiesch, K.T. Ulrich. 2010. Idea Generation and the Quality of the Best Idea. *Management Science* **56**(4) 591-605.

Gnyawali, D.R., B.-J. Park. 2011. Co-opetition between giants: Collaboration with competitors for technological innovation. *Research Policy* **40**(5) 650-663.

Jarvenpaa, S.L., A. Majchrzak. 2010. Research commentary—vigilant interaction in knowledge collaboration: challenges of online user participation under ambivalence. *Information Systems Research* **21**(4) 773-784.

Jeppesen, L.B., K.R. Lakhani. 2010. Marginality and problem-solving effectiveness in broadcast search. *Organization Science* **21**(5) 1016-1033.

Jin, Y., H.C.B. Lee, S. Ba, J. Stallaert. 2021. Winning by Learning? Effect of Knowledge Sharing in Crowdsourcing Contests. *Information Systems Research*.

Johnson, D.W. 2003. Social interdependence: interrelationships among theory, research, and practice. *American Psychologist* **58**(11) 934.

Johnson, D.W., R.T. Johnson. 1989. *Cooperation and competition: Theory and research*. Interaction Book Company.

Kathan, W., K. Hutter, J. Füller, J. Hautz. 2015. Reciprocity vs. free-riding in innovation contest communities. *Creat. Innov. Manag.* **24**(3) 537-549.

Kistruck, G.M., J. Robert B. Lount, B.R. Smith, J. Brian J. Bergman, T.W. Moss. 2016. Cooperation vs. Competition: Alternative Goal Structures for Motivating Groups in a Resource Scarce Environment. *Academy of Management Journal* **59**(4) 1174-1198.

Körpeoğlu, E., S.-H. Cho. 2018. Incentives in Contests with Heterogeneous Solvers. *Management Science* **64**(6) 2709-2715.

Liu, T.X., J. Yang, L.A. Adamic, Y. Chen. 2014. Crowdsourcing with All-Pay Auctions: A Field Experiment on Taskcn. *Management Science* **60**(8) 2020-2037.

Majchrzak, A., A. Malhotra. 2020. *Unleashing the Crowd*. Springer.

Murayama, K., A.J. Elliot. 2012. The competition–performance relation: A meta-analytic review and test of the opposing processes model of competition and performance. *Psychological Bulletin* **138**(6) 1035.

Olson, M. 1965. *The logic of collective action. Public goods and the theory of groups*. Harvard University Press, Cambrige, MA.

Piezunka, H., L. Dahlander. 2015. Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. *Academy of Management Journal* **58**(3) 856-880.

Riedl, C., V.P. Seidel. 2018. Learning from mixed signals in online innovation communities. *Organization Science* **29**(6) 1010-1032.

Ritala, P., P. Hurmelinna-Laukkanen. 2013. Incremental and Radical Innovation in Coopetition—The Role of Absorptive Capacity and Appropriability. *Journal of Product Innovation Management* **30**(1) 154-169.

Singh, J., L. Fleming. 2010. Lone inventors as sources of breakthroughs: Myth or reality? *Management Science* **56**(1) 41-56.

Sweeney, J.W. 1973. An experimental investigation of the free-rider problem. *Social Science Research* **2**(3) 277-292.

Terwiesch, C., Y. Xu. 2008. Innovation Contests, Open Innovation, and Multiagent Problem Solving. *Management Science* **54**(9) 1529-1543.

Tjosvold, D. 1989. Interdependence and Power Between Managers and Employees: A Study of the Leader Relationship. *Journal of Management* **15**(1) 49-62.

Tsai, W. 2002. Social Structure of "Coopetition" Within a Multiunit Organization: Coordination, Competition, and Intraorganizational Knowledge Sharing. *Organization Science* **13**(2) 179-190.

Wasko, M.M., S. Faraj. 2005. Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly* 35-57.