Artificial Swarming shown to Amplify Accuracy of Group Decisions in Subjective Judgement Tasks

Gregg Willcox[1], Louis Rosenberg[1], David Askay[2], Lynn Metcalf[2], Erick Harris[2], and Colin Domnauer[3]

[1] Unanimous AI, San Francisco CA 94115, USA
[2] California Polytechnic State University, San Luis Obispo CA 93401, USA
[3]University of California Berkeley, Berkeley CA 94702

Abstract. New technologies enable distributed human teams to form real-time systems modeled after natural swarms. Often referred to as Artificial Swarm Intelligence (ASI) or simply "human swarming", these real-time systems have been shown to amplify group intelligence across a wide range of tasks, from handicapping sports to forecasting financial markets. While most prior research has studied human swarms with 20 to 100 members, the present study explores the ability of ASI to amplify accuracy in small teams of 3 to 6 members. The present study also explores if conducting multiple swarms and aggregating by taking a "vote of swarms" can further amplify the accuracy. A large set of 66 small teams were engaged in this study. Each team was given a standard subjective judgement test. Participants took the test both as individuals and real-time swarms. The average individual scored 69% correct, while the average swarm scored 84% correct ($p<0.001$). In addition, aggregation of multiple swarms revealed additional amplifications of accuracy. For example, by randomly selecting sets of 3 swarms and aggregating by plurality vote, average accuracy increased to 91% ($p<0.001$). These results suggest that when small teams make subjective judgements as real-time swarms, they can be significantly more accurate than individual members, and that their accuracy can be further amplified by aggregating the output across small sets of swarms.

Keywords: Swarm Intelligence, Collective Intelligence, Artificial Intelligence, Human Swarming, Wisdom of Crowds, Artificial Swarm Intelligence, ASI

1 Introduction

In the natural world, Swarm Intelligence (SI) enables social organisms to rapidly aggregate their collective insights and reach optimal decisions by forming real-time closed-loop systems that converge in synchrony. Swarm Intelligence has been deeply studied across many social species, from schools of fish and flocks of birds, to swarms of honey bees. In recent years, advances in networking technology and artificial intelligence have enabled human groups to form similar systems online, moderated by AI algorithms. Often referred to as Artificial Swarm Intelligence (ASI) or simply "human swarming," this novel approach has been shown to significantly increase the accuracy of group decisions across a wide variety of tasks, from handicapping sporting events to forecasting financial markets [1 - 7].

While ASI has been found to significantly amplify decision-making accuracy in human groups comprised of 20 to 100 members, few studies have investigated the use of swarming among small groups on the order of 3 to 6 members. This is important to scholarship and practice as many important decisions are made by small teams.

2 Enabling "Human Swarms"

Unlike birds, bees and fish, humans have not evolved the natural ability to form real-time swarms, as we lack the innate mechanisms used by other species to form closed-loop systems. Schooling fish detect vibrations in the water around them. Flocking birds detect high-speed motions propagating through the group formation. Swarming bees generate complex body vibrations called a "waggle dance" that encode

Future of Information and Communication Conference (FICC), San Francisco, 2019.

assessment information. To enable networked human groups to form similar closed-loop systems, a cloud-based platform called “swarm.ai” was developed, as shown below in Figure 1. It enables distributed human groups, connected over the internet, to make collective predictions, decisions, and assessments by working together as closed-loop swarms.

When using the swarm.ai platform, networked human teams answer questions by collaboratively moving a graphical pointer to select from a set of answer options. Each participant provides input by manipulating a graphical magnet with a touchscreen or mouse. By adjusting the position and orientation of their magnet with respect to the moving puck, participants express their intent, not as a discrete vote, but a stream of vectors that varies freely over time. Because all members adjust their individual intent continuously in real-time, the swarm explores the decision-space, not based on the input of any single member, but based on the emergent dynamics of the full system. The complex behavioral interactions among the full population are processed by swarming algorithms in real-time, empowering the unified system to converge on solutions that maximizes the collective confidence and conviction of the group.

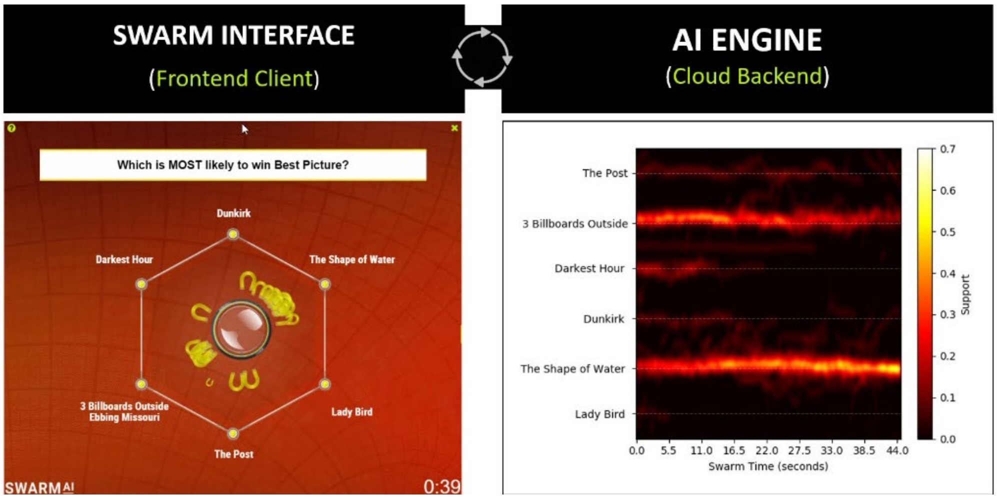


Fig. 1. Architecture of the **swarm.ai** platform with graphical client and cloud-based AI engine

It is important to note that participants not only vary the direction of their intent but also modulate the magnitude of their intent by adjusting the distance between their magnets and the pointer, which is commonly represented as a graphical puck. Because the graphical puck is in continuous motion across the decision-space, users need to continually move their magnets so that they stay close to the puck’s rim. This is significant, for it requires that all participants, regardless of group size or composition, to be engaged continuously throughout the decision process, evaluating and re-evaluating their intent in real-time. If a participant stops adjusting their magnet with respect to the changing position of the puck, the distance grows and the participant’s influence on the group’s decision wanes.

Thus, like bees vibrating their bodies to express sentiment in a biological swarm, or neurons firing to express conviction levels within a biological neural-network, the participants in an artificial swarm must continuously update and express their changing preferences during the decision process or lose their influence over the collective outcome. This is generally referred to as a “leaky integrator” structure and common to both swarm-based and neuron-based systems. In addition, intelligence algorithms monitor the behaviors of swarm members in real-time, inferring their relative conviction based on their actions and interactions over time. This reveals a range of behavioral characteristics within the population and weights their contributions accordingly.

3 Accuracy Study

To assess the ability of ASI to amplify the accuracy of team decisions in subjective judgement tasks, a large study was conducted across a set of 66 working groups (i.e. business teams), each of 3 to 6 members. Each of these teams were already engaged in a long-term project together and had already established a working relationship among themselves. In total, 330 subjects participated in this study.

All were college students in business, communication studies, and engineering courses, for which the team project was a significant component.

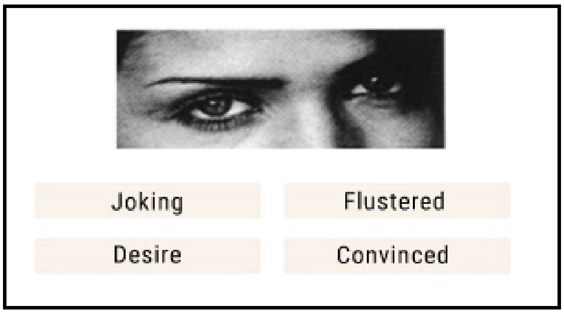


Fig.2. Sample Question from Standard RME Test

To rigorously measure accuracy in a standardized subjective judgement task, a widely used instrument was employed known as the "Reading the Mind in the Eyes" or RME test [8]. The test includes 35 questions, each of which provides a facial image cropped so that only a narrow region around the eyes is shown. A set of four options are provided that describe the emotion expressed by the person in the image, requiring participants to assess the emotional state based only on the eyes. An example question from a standard RME test is shown above in Figure 2. Four options are provided, only one of which accurately represents the emotion of the depicted individual.

Prior studies have shown that the RME test is a reliable instrument with strong internal consistency and test-retest stability [9]. The performance in this subjective judgement task has been shown to indicate the Social Sensitivity of the test taker and is generally used for that purpose. [10-12]. To test whether real-time swarming enabled small working groups to amplify their performance in the RME task, a two-stage process was employed. First, each of the 330 study participants were administered a 35-question test individually through an online survey. To limit bias and knowledge of correct answers, individual scores were not disclosed. In the second stage, each of the 66 teams was administered the RME test through an online swarming platform. This enabled each team to converge on each subjective judgement by working together as a real-time system, moderated by swarming algorithms. Teams were discouraged from communicating with each other verbally during the assessment, instead relying only on the closed-loop interaction afforded by the platform.

For each of the 35 subjective judgements in the test, the platform displayed one of the 35 facial images to all members of each team, along with the four potential assessments of that image. Each team was allotted up to 60-seconds to coverage upon an answer as a real-time swarm. Figure 3 below is a snapshot of a team member's screen during a real-time swarm response. The magnets in the image represents the pull of each teammate at one instant in time. It should be noted that to discourage conformity, participants did not see the magnets during the actual swarming session.

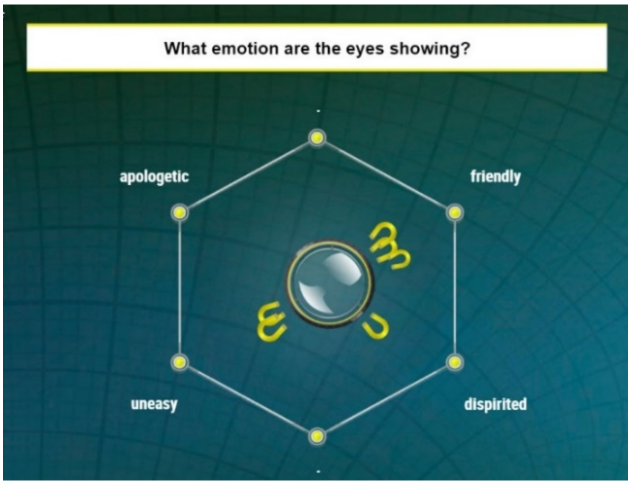

What emotion are the eyes showing?

Fig. 3. Swarming Group responding to RME question.

4 Data and Analysis

The RME test was administered to 330 individuals across 66 teams and produced four unique datasets. First, we received fully completed individual assessments from 283 participants (86% response rate) totaling over 9,000 item responses. These responses were used to calculate individual RME scores for each participant. Second, these same responses were aggregated by team to generate a "plurality vote" RME score for each question. This was calculated by assessing the most popular answer among the team for each question. For questions where the vote was split evenly across multiple answers (i.e. there was no plurality winner) a "deadlock" was determined and classified as incorrect. This provided a dataset of over 2,000 plurality vote responses. Third, a swarm RME score for each team was calculated from the responses collected through the online swarming platform. This provided a dataset of over 2,000 swarm-based responses. For questions where the swarm could not converge upon an answer within the 60 second time limit, a "deadlock" was determined and classified as incorrect.

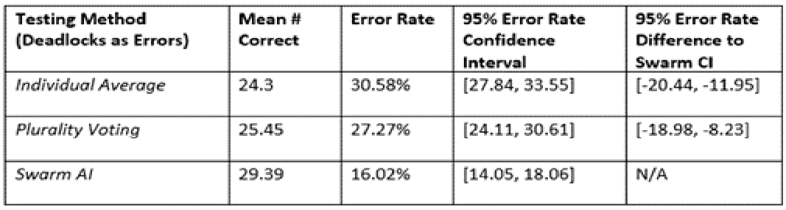
Finally, a "vote of swarms" RME score was generated by selecting random grouping of swarms for each question from the set of 66 teams and determining the final decision by plurality vote across the grouping. This was performed using a bootstrapping technique across a range of groupings of size S=3 to S=10 and repeated 1000 times for each size. For example, for S=3, random groupings of three swarms were selected from the dataset and an RME score was generated based on a plurality vote across those 3 swarms. This process was repeated 1000 times for groupings of 3 swarms.

5 Results

Across the set of 330 subjects, each participating in one of 66 teams, a comparison was performed among four conditions:

1. **Individuals** - participants taking RME test alone
2. **Votes** - teams taking RME test by plurality vote
3. **Swarms** - teams taking RME test as real-time systems
4. **Votes of Swarms** – taking plurality vote among swarms

Mean scores and error rates for RME were calculated for the individual, plurality, and swarm generated scores. As shown in the table of Figure 4 below, the average individual RME score was 24.3, which corresponds to an error rate of 30.6%.



Testing Method (Deadlocks as Errors)	Mean # Correct	Error Rate	95% Error Rate Confidence Interval	95% Error Rate Difference to Swarm CI
Individual Average	24.3	30.58%	[27.84, 33.55]	[-20.44, -11.95]
Plurality Voting	25.45	27.27%	[24.11, 30.61]	[-18.98, -8.23]
Swarm AI	29.39	16.02%	[14.05, 18.06]	N/A

Fig. 4. Error Rates and Confidence Intervals

The average of each team's plurality RME score was 25.45, which corresponds to an average error rate of 27.3%. When enabling the teams to work together as a swarm, the average RME score increased to 29.4, which corresponds to an average error rate of 16.0%. In other words, by working together as a swarm, the 66 teams, on average, reduced their error rates by 41%. This demonstrates that working as a swarm can significantly increase accuracy in subjective judgement tasks as compared to both individual performance and team performance by plurality vote.

To assess statistical significance, a bootstrap analysis of the error rate for each method was performed across 10,000 trials. The 95% confidence intervals and p-values were calculated for the difference between individual RME, plurality RME, and swarm RME scores. The results show that the swarm significantly outperforms both individual ($\mu_{difference} = 14.6\%$ error, $p < 0.001$) and plurality scores ($\mu_{difference} = 11.3\%$ error, $p < 0.001$). The bootstrapped error comparison is shown below in Figure 5.

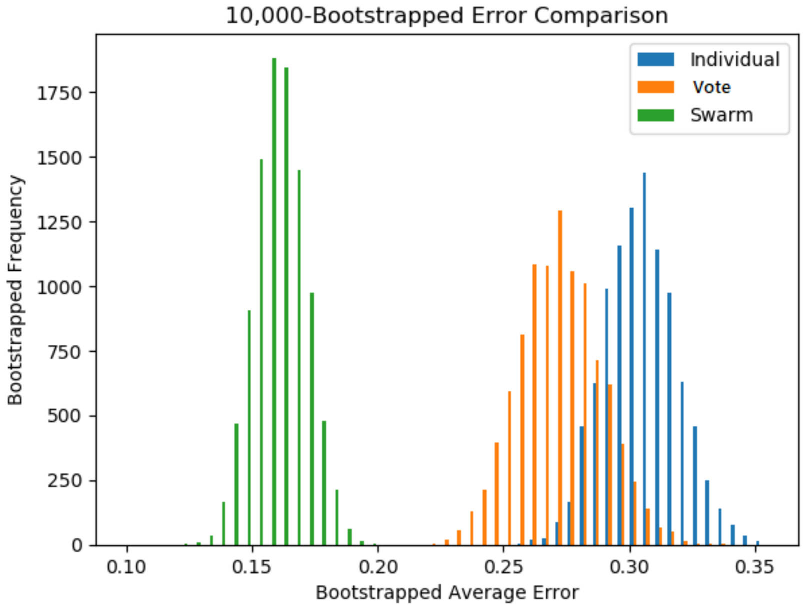
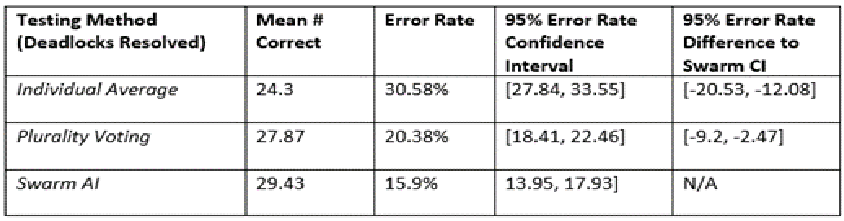


Fig. 5. Bootstrapped Average Error Rate

With respect to deadlocks, a comparison was made between the rate of deadlocks determined by plurality vote as compared to the rate of deadlocks reached by swarms. Across the 66 working groups, plurality voting resulted in deadlocks in 14% of questions. Across those same groups, when working together as swarms, the rate of deadlocks dropped substantially to 0.6% of questions. This is a significant improvement, reducing the need for further steps to resolve undecided groups.

In addition, an analysis was performed assuming deadlocked votes were resolved by giving partial credit for tied answers that include a correct response: half credit for a two-way tie, third credit for a three-way tie, etc. To balance this, deadlocked swarms were given the chance to resolve immediately following a deadlock in another 60-second swarm, with the answer chosen in this second round selected as the final answer. There were no swarms that deadlocked twice in a row.

As shown in the table of Figure 6 below, when deadlocks were resolved using partial credit, plurality vote averaged an RME score of 27.9 (an error rate of 20.4%). When enabling the swarms to work together as real-time systems and resolve their deadlocks in a follow-up swarm, the swarm RME score increased to 29.4 (an error rate of 15.9%). In other words, even when giving partial credit for deadlocks in group responses determined by plurality vote, the swarm outperformed.



Testing Method (Deadlocks Resolved)	Mean # Correct	Error Rate	95% Error Rate Confidence Interval	95% Error Rate Difference to Swarm CI
Individual Average	24.3	30.58%	[27.84, 33.55]	[-20.53, -12.08]
Plurality Voting	27.87	20.38%	[18.41, 22.46]	[-9.2, -2.47]
Swarm AI	29.43	15.9%	13.95, 17.93]	N/A

Fig. 6. Error Rates and Confidence Intervals with Deadlocks Resolved.

To assess statistical significance, a bootstrap analysis of the error rate for each method was again performed across 10,000 trials. We find that the swarm outperforms both the plurality vote (µdifference = 4.5% error, p < .001) and individuals (µdifference = 14.7% error, p < .001). The bootstrapping of the error rate confidence intervals is shown below in Figure 7.

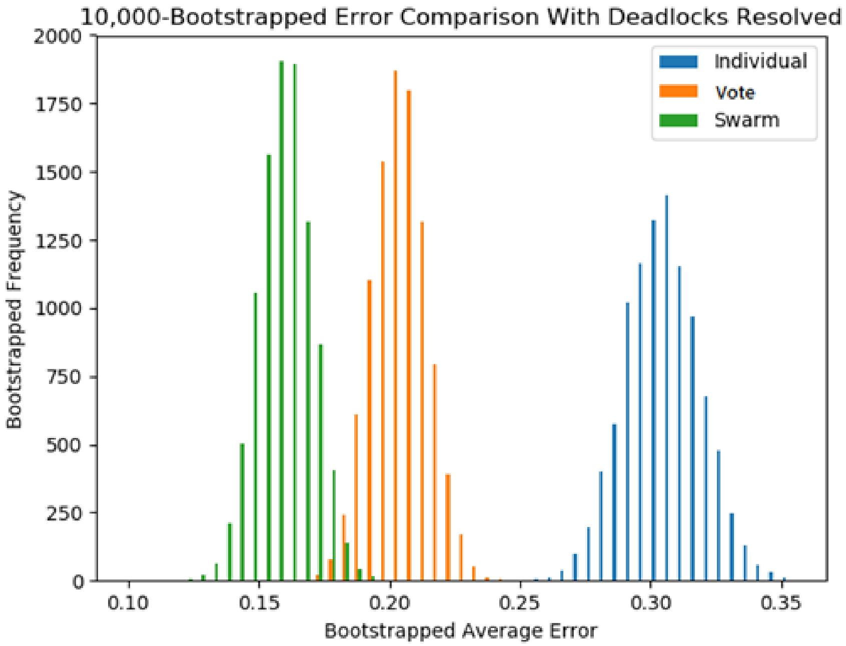


Fig. 7. Bootstrapped Average Error Rate

In addition to comparing against the average individual, the swarm can be compared against all individuals. On average, swarms are in the 92nd percentile of individuals, indicating that an average swarm scores better than 92.2% of individuals taking the test alone. The histogram of user performance and average swarm performance is shown below in Figure 8.

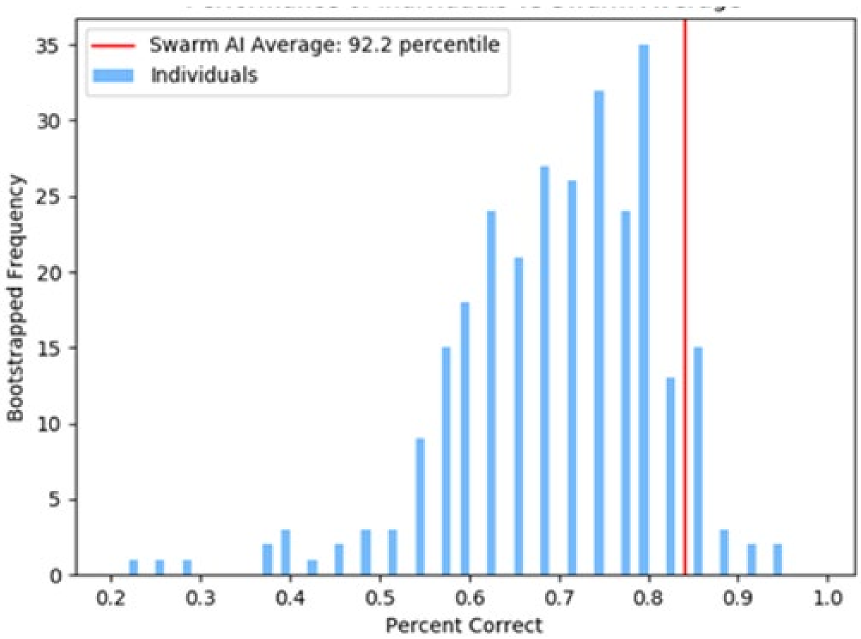


Fig. 8. Bootstrapped Average Error Rate

Finally, we explored the aggregation of swarm responses by plurality vote to assess if the accuracy on the subjective judgement RME test could be further amplified as compared to individual swarm responses. This process, referred to herein as "aggregations of swarms," was conducted by bootstrapping the average error rate of aggregations of swarms across a range of aggregation sizes from S=3 to S=9, with 1000 iterations of randomly selected aggregations performed for each aggregation size. The results are shown below in Figure 9. The single swarm case (S=1) is bootstrapped and shown to depict how error rate decreases as the number of swarms aggregated increases.

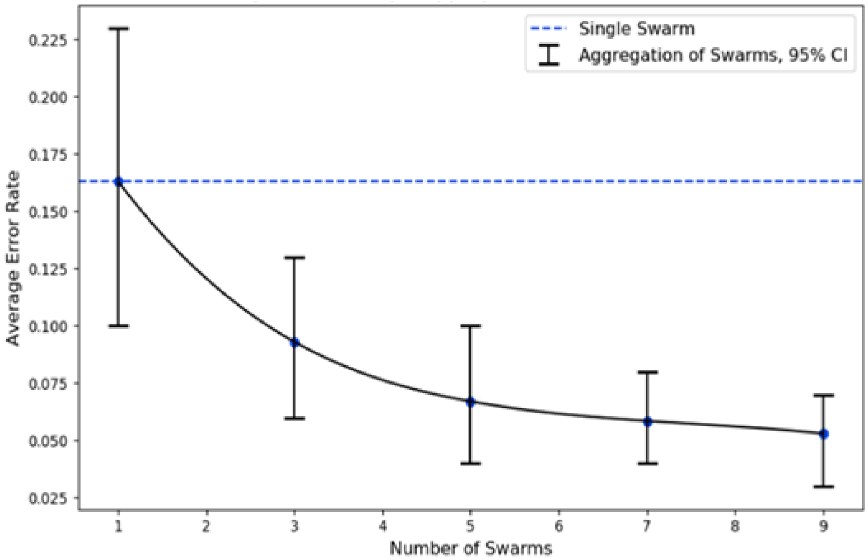


Fig. 9. Accuracy when Swarms are Aggregated by Plurality Vote

We find that increasing the number of swarms aggregated decreases the error rate. In addition, the variation in performance decreases as the number of swarms aggregated increases. Not only do votes of swarms become more accurate as more swarms are aggregated, but they also become more consistent. The aggregation of as few as three swarms significantly outperforms single swarms (µdifference = 6.9% error, p = .007) and the aggregation of five swarms significantly outperforms the aggregation of three swarms (µdifference = 2.3% error, p < .044).

The bootstrapped average error histogram created for individuals, swarms, and aggregations of three swarms is shown below in Figure 10. We find that an aggregation of three swarms outperforms individuals by an average error of 21.3%. So, by working together in swarms, and then aggregating three swarms together, the average error is reduced by 70% as compared to individual performance.

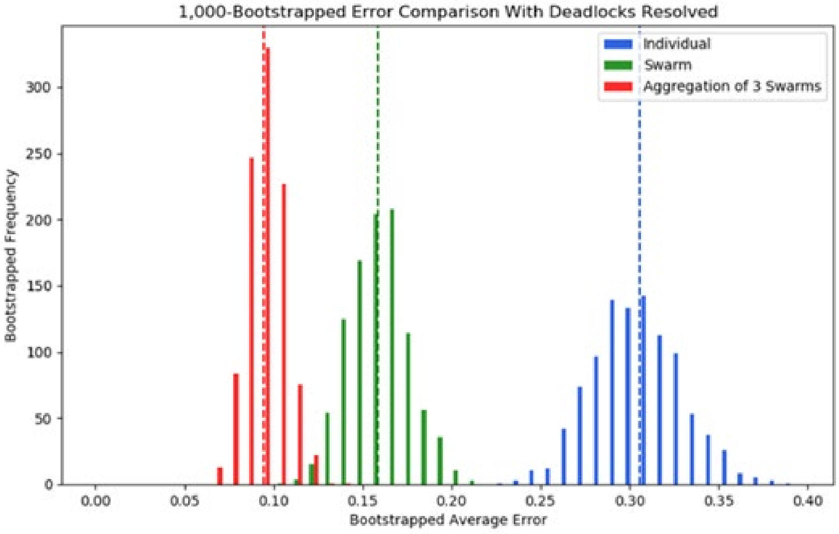


Fig. 10. Individual vs Swarm vs Aggregation of Swarms

6 Conclusions

The results of this study suggest that small "human swarms" are significantly more accurate than individuals in subjective judgement tasks. As demonstrated across 66 working groups, each of 3 to 6 members, subjective judgement accuracy increased from 69% correct to 84% correct when participants worked together as real-time swarms. This corresponds to a reduction in error rate by 41%. In addition, the results of this study suggest that small human swarms are significantly more accurate than those same groups reaching subjective judgements by plurality vote, which demonstrated 73% accuracy. The probability that the swarm outperformed the individuals and the group vote by chance was very low ($p < 0.001$).

In addition, results of this study suggest that by aggregating the output from multiple human swarms, we can further increase accuracy on subjective judgment tasks. A range of aggregation sizes were explored from S=3 to S=9. Even when aggregating only three swarms at a time (S=3), a significant increase in accuracy was observed, boosting performance from 84% correct for single swarms to 91% for aggregations. In other words, by having small human groups perform subjective judgement tasks as swarms, and then aggregating small sets of swarms, individual performance was increased from 69% accuracy (50th percentile) to 91% accuracy (98th percentile). These are a very significant results and suggests that real-time swarming may be a powerful method for boosting team performance, even among small teams of only 3 to 6 members.

References

1. Rosenberg, L.B., "Human Swarms, a real-time method for collective intelligence." Proceedings of the European Conference on Artificial Life 2015, pp. 658-659
2. Rosenberg, L. "Artificial Swarm Intelligence vs Human Experts," Neural Networks (IJCNN), 2016 International Joint Conference on. IEEE.
3. Rosenberg, L. Baltaxe, D and Pescetelli, N. "Crowds vs Swarms, a Comparison of Intelligence," IEEE 2016 Swarm/Human Blended Intelligence (SHBI), Cleveland, OH, 2016, pp. 1-4.
4. Baltaxe, D, Rosenberg, L., and N. Pescetelli, "Amplifying Prediction Accuracy using Human Swarms", Collective Intelligence 2017. New York, NY; 2017.
5. Rosenberg, L, Pescetelli, N, and Willcox, G. "Human Swarms Amplify Accuracy in Financial Predictions," Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), IEEE Annual, New York, NY, 2017.
6. Rosenberg, L, Willcox, G., Halabi, S., Lungren, M, Baltaxe, D. and Lyons, M. "Artificial Swarm Intelligence employed to Amplify Diagnostic Accuracy in Radiology," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2018 (Nov 2-4).
7. Rosenberg, L. and Willcox, G. "Artificial Swarms find Social Optima: (Late Breaking Report)," 2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), Boston, MA, 2018, pp. 174-178.
8. S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, I. Plumb, The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. J. Child Psychol. Psychiatry 42, 241 (2001).
9. Vellante M, Baron-Cohen S, Melis M, Marrone M, Petretto DR, Masala C, Preti A: The "reading the mind in the eyes" test: systematic review of psychometric properties and a validation study in Italy. Cogn Neuropsychiatry. 2012, 18: 326-354.
10. Fiske ST, Taylor SE: Social cognition: From brains to culture. 2013, London, UK: SAGE Publications Limited
11. Kunda Z: Social cognition: Making sense of people. 1999, Cambridge, MA: The MIT Press
12. Frith CD, Singer T (2008) The role of social cognition in decision making. Philos Trans R Soc Lond B Biol Sci 363:3875–3886