Enhancing Decision-Making in Human-Machine Teams

Akashdeep Nijjar, Ittipat Promnorakid, Riccardo Poli, Caterina Cinel
School of Computer Science and Electronic Engineering,
University of Essex, Colchester, UK
Emails: {an22851,ip22667,rpoli,ccinel }@essex.ac.uk

Thomas Reed
Defence Science and Technology Laboratory, Porton Down, UK
Email: treed@mail.dstl.gov.uk

Christopher Baker
School of Computer Science and Mathematics,
Liverpool John Moores University, Liverpool, UK
Email: c.m.baker@ljmu.ac.uk

Stephen Hinton, Stephen Fairclough
School of Psychology,
Liverpool John Moores University, Liverpool, UK
Emails: {S.F.Hinton,S.Fairclough}@ljmu.ac.uk

***Abstract*—One of the key advantages of groups is their ability to pool resources and share information effectively, enabling coordinated efforts that result in emergent behaviors and enhanced performance. Here we present preliminary data from two experiments that are part of a larger study aimed at developing and testing a prototype system of a "superorganism" designed to augment decision-making within human-AI teams. One key aspect of the system, is the collaboration between humans and AI agents (AIs). In each team, humans and AIs make decisions as peers, and the optimality of each team decision depends, among other things, on mutual trust and personality, of both the human and AIs. We manipulated team composition and AIs' personality and investigated how these, and human trust in AI, affect individual and team performance. Our preliminary results suggest complex dynamics. There was a clear advantage for groups, compared to individual decision-making, as expected. However, even though humans trusted the "humanized" AI more than the non-humanized AI, performance did not necessarily benefit from it.**

***Index Terms*—Human-machine teams, decision making, personality, brain-computer interface**

I. Introduction

One major advantage of groups is their ability to combine resources and exchange information, enabling coordinated efforts that function as "superorganisms". Integrating AIs in teams can further enhance performance by complementing human strengths and compensating for weaknesses [1]. Trust -— both among team members and toward AI agents —- is a crucial factor for effective human-AI collaboration. Research has found that affect and competence are key indicators of acceptance and trust of a team member, including when this is an AI [2]. Users typically exhibit unrealistically high expectations and trust toward AIs; these, however, significantly drop after encountering errors and are then very difficult to rebuild [3]. Hence, it is essential to lead users towards a more realistic attitude toward AI. This may be achieved through "humanization" of the AI system. Anthropomorphic characteristics, such as AI "personality" traits, play an important role in the human-AI relationship [4]. Perceiving AI devices as social actors makes users act toward them in a similar fashion to human interactions [5] allowing an understanding of the AI behavior. Studies find that humanization increases acceptance [6] and likeability [7].

When it comes to aggregating individual decisions within groups, traditional approaches often involve averaging individual scores or relying on majority voting. However, more advanced methods go beyond these by weighting individual responses according to confidence – a metric that is typically well-aligned with decision accuracy [8]. This confidence-weighted aggregation has been shown to enhance collective performance, compared to when more traditional methods are used [9]. Additionally, confidence plays a role in team dynamics. When individual effort is unobservable, teams often suffer from free-riding which can be alleviated by overconfident individuals. The latter are perceived by team members to exert more effort, motivating them to do the same, thereby enhancing overall team performance [10].

Also well known to influence decision-making and team performance is personality [11]. For instance, individuals high in extraversion may prefer making decisions intuitively and spontaneously, whereas those high in conscientiousness may prefer to make decisions after careful reflection [11]. A positive attitude towards group collaboration is also found in extroverted individuals, whereas introverted individuals prefer to work independently [12]. Attitudes to risk and confidence levels can also affect performance [13].

In this study, we address two aspects of human-AI collaboration: if human's personality as well as an AI's confidence and personality affect the individuals' and teams' performance, and how performance is affected by team composition (i.e., different AI agents joining a human agent) and trust in the AI(s). To do so, we performed two experiments in which human-AI teams of either two (one human, one AI) or three (one human, two AIs) were engaged in a strategic decision-

making task based on a pandemic scenario as described in section II. In Experiment 1, participants played the pandemic game with either an ”overconfident” or ”correctly” confident AI, and we investigated how this affects individual and team performance. In Experiment 2, we further manipulated the AIs’ personality by introducing a humanized AI, and we tested performances and trust when playing with just one - either correctly confident or humanized- or both AIs. Across both experiments, we investigated decision strategies and human personalities complementary to each AI.

II. Methodology

A. Experiments

In the two experiments [1], participants were asked to allocate limited resources to contain the spread of a pandemic across different cities, with varying levels of severity, and under conditions of uncertainty. In each experiment, participants completed eight blocks, where each block consisted of eight sequences of decisions. For each sequence, participants were given a total of 30 resource units to allocate. A sequence contained a variable number of infected cities — randomly selected to range between 3 and 10 — and was presented through a fictional map indicating the cities where the pandemic is spreading (Figure 1(a)). After the participant’s decision for a location, other cities became visible on the same map, one after the other. For each city, participants received information about the severity of the outbreak (indicated by the shade of red of the circle corresponding to each city) and updates on the pandemic evolution of affected cities where previous allocation decisions were made (in Figure 1(a), this is shown by a black arrows pointing upwards or downwards). With each new city, participants had to choose (with a mouse click) between two resource amount options (visible at the top of Figure 1(a)). Participants made their decisions with the understanding that selecting the largest options would increase the chances of containing the pandemic; on the other hand, they also knew that those resources were limited (30 in total for each sequence of cities) and did not know in advance how many cities would appear in each sequence. After each allocation decision, participants reported their confidence in their choice, on a scale from 0 to 100%. When they were presented with a new map (i.e., a new sequence of cities) they resources were re-set to 30. In both experiments, the participants played individually during the first two blocks and were joined by one or two AIs in the remaining six blocks. In Experiment 1 (N=20; 10 male, 10 female aged 32 ± 12 years), participants played with one AI, either an ”overconfident” AI (OAI), or a ”correctly” confident AI (CAI). Both AIs performed identically, the only difference being the level of confidence they reported (more detail below in Sections II-C and II-D).

In Experiment 2 (N=20; 12 male, 8 female aged 39 ± 23 years, none of the participants took part in Experiment 1), a humanized AI (HAI) was introduced, while the second AI was the CAI (the same as in Experiment 1). The HAI mimics the decision-making behavior of the average human participant in the first experiment (more detail below in Sections II-C). Participants were not told about the AI’s varying characteristics. They played with both the HAI and CAI in half of the experiment, and one of them in the other half. In the blocks where decisions were made in a team, resources were shared

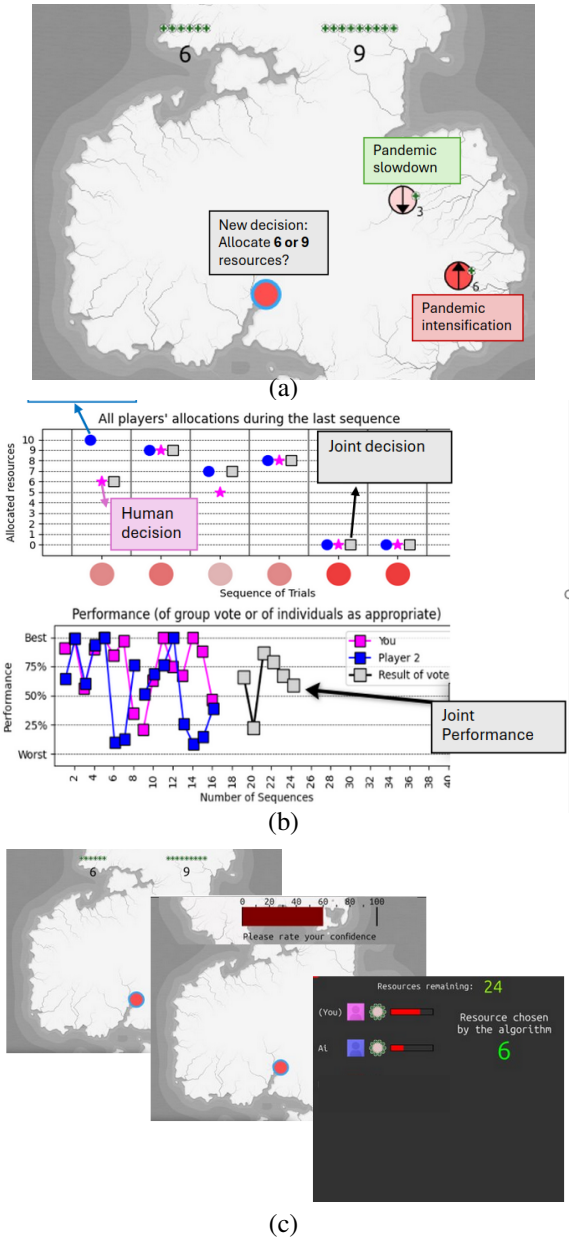


Fig. 1. a) Example of a trial showing a map with cities affected the pandemic. b) At the end of each sequence participants were given feedback: a panel showing both the team members’ resource choices (blue circles and pink stars) and final team decision (gray square) for each city (top panel); a second panel showing the performance of each player (pink and blue squares) and combined performance (gray square) of current and previous sequences in the game (bottom panel). c) Example of sequence of displays, showing, from left to right, resource options, confidence report and team’s choices and confidence

[1]Both protocols received favorable opinion from the UK Ministry of Defence (MoD)’s Research Ethics Committee in July 2024, and were performed in accordance with relevant guidelines and regulations. The tasks for both experiments were designed after consultation with the MoD.

between the participant and the AI(s). For each city, each team member made a resource allocation choice individually, and then the final team decision was the individual choice associated with the highest level of confidence. When three players were in the team (one human and two AIs), and therefore one of the two resource options was chosen by at least two of them, confidence ratings were averaged for that option. If players were equally confident, the decision was made at random. Once the team decision was determined, and before a new city appeared on the map, a new display appeared illustrating the resource choice and confidence of each team member for the most recent decision, as well as the final team decision. The display also showed the remaining amount of resources (see Figure 1(c), rightmost panel). At the end of each sequence of cities, participants were shown a display illustrating the amount of resources each player chose and the resulting team choice, as well as on the resulting optimality of team performance (Figure 1(b)).

B. Evaluation of performance

We assessed both individual and team-level performance via the *optimality rate*, defined as the percentage of optimal decisions within a sequence of trials. A decision is considered optimal if it leads to the lowest possible damage caused by the pandemic at the end of the sequence. Throughout the game, participants tended to adopt different resource allocation strategies, such as "matching" or "overallocating." The *matching* strategy involves selecting the resource option closest to the severity of the city, while *overallocation* means choosing the highest possible resource option. Although overallocation results in a faster reduction in severity—making it the optimal choice in shorter sequences— in longer sequences, this approach can lead to waste of resources. Thus, the most effective approach is *careful overallocation*—a strategy that balances the benefits of higher resource use with the need to plan long-term.

C. AI Players

In Experiment 1, participants were paired with either CAI or OAI. These AIs acted the same, the only difference between them being their reported confidence, as described in section II-D. The AI's decisions were derived from a Neural Network with two hidden layers with the following inputs: 1) the pandemic severity in the current city, 2) the number of cities to which allocations were already made in the sequence, 3) the remaining resources and 4) the resource choices for the current city. The network was trained on 64 sequences with approximately 360 decisions. The AIs were purposely trained on limited data in order to introduce a weakness, and with the goal of, firstly, preventing human participants relying too heavily on the AI and, secondly, fostering trust and likability. The CAI/OAI follow the overallocation strategy (see Section II-B), resulting in best performance with short- and medium-length sequences. The CAI and OAI performances are both approximately 65%, with the difference between them being in their level of confidence (see Section II-D). The adoption of a specific strategy aims to allow the human participant to understand the AI and its strengths/weaknesses and confidence over time. Participants are expected to become sensitive (and motivated) in reaction to the AI being overconfident or appropriately so.

In Experiment 2 participants were paired with either an CAI or HAI. The HAI was based on Linear Discriminant Analysis using the same input features as for CAI/OAI with the exception that the remaining resources are replaced by the current block. The model was trained on data from all 20 participants in Experiment 1 (5694 decisions in total). Thus, it adopts a decision strategy that mimics the behavior of those participants, initially starting with the matching strategy and gradually moving closer to the optimal strategy (Figure 5(b)). The similarity between the HAI and the human participants was expected to lead to higher trust and acceptance.

D. AI confidence

Human confidence is self-reported by participants after each decision on a scale ranging from 0 (no confidence at all) to 100 (full confidence). Confidence is known to be associated with the probability that a decision is correct. So, we set the confidence level of the AIs primarily based on the optimality of the choices. Specifically, the CAI confidence was given by the following equation:

$$\mathrm{conf}_{\mathrm{CAI}} = (0.5 \times o_1 + 0.4 \times o_2 + 0.2 \times d) \times e^{-0.0625 \times r}$$

where o_1 is the expected sequence performance given the current choice, o_2 is a binary indicator of whether the current choice is the optimal option, d is the difficulty of the choice, and r is an artificial *response time* obtained by randomly sampling from the distribution of human response times. The OAI was simply 10% more confident than the CAI, so:

$$\mathrm{conf}_{\mathrm{OAI}} = \mathrm{conf}_{\mathrm{CAI}} + 0.1$$

For HAI, the confidence for optimal (HAI_o) and suboptimal decisions (HAI_s) are given by:

$$\mathrm{conf}_{\mathrm{HAI_o}} = hc_o \qquad \mathrm{conf}_{\mathrm{HAI_s}} = hc_s - 0.1,$$

where hc_o and hc_s are randomly sampled from the distributions of human confidences (from Experiment 1) for optimal and suboptimal decisions, respectively.

E. Evaluation of personality

Personality was evaluated using the 10-item Big Five Inventory [14]. Additionally, to measure impulsiveness, risk affinity and ambiguity tolerance, we administered the 15-item Barrat Impulsiveness Scale [15], the Balloon Analog Risk Task [16] and the Multiple Stimulus Types Ambiguity Tolerance Scale-II [17]. We investigated (human) personality differences between higher- and lower-performing teams, based on which AI was part of the human-AI team (OAI, CAI, or HAI). For each AI, a Multilayer Perceptron (MLP) model was trained to distinguish between the higher/lower-performing groups using Leave-One-Out Cross-Validation (excluding one participant/team). The network had a single hidden layer with two neurons, using

the logistic activation function. The BIG 5 personality features, as well as risk affinity and ambiguity tolerance were used as input features. Input data were scaled prior to classification.

III. Results

A. Team performance and team composition

In Experiment 1, the advantage of teams over individuals was confirmed ($p<0.001$) (Figure 2(a)). On average, human participants' and the CAI/OAI's individual decisions (the CAI and OAI were designed to make the same decisions, only their confidence changed, as described in Sections II-C and II-D) align with the optimal decision in approximately 60% and 65% of cases, respectively. Consequently, team performance in human-AI collaborations improved by 20% and 15% compared to individual performances. However, team performance was unaffected by the type of AI. Interestingly, participants' confidence tended to align with the AI's confidence: as shown in Figure 2(b), participants playing with the OAI (designed to be 10% more confident than the CAI) gradually exhibited higher confidence. However, individual human performance was unaffected by the AI's confidence and not correlated to the reported confidence.

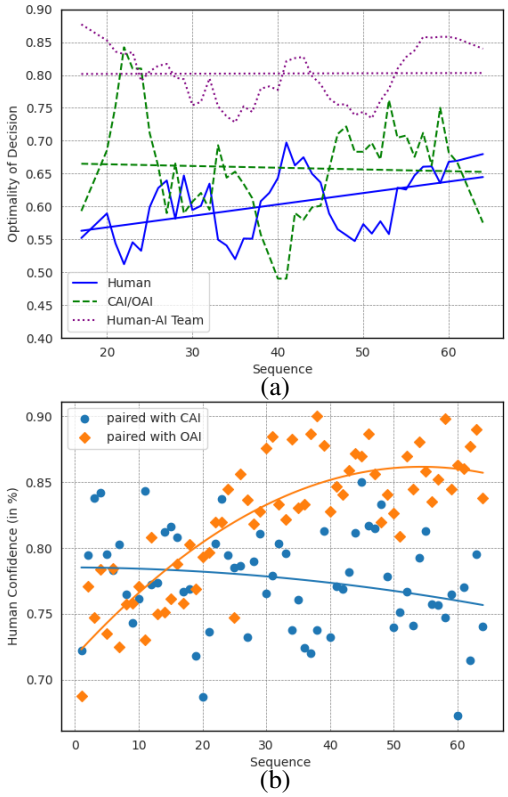


Fig. 2. Experiment 1: (a) Individual and human-AI team performance (N=19) (CAI and OAI choices were identical); (b) Participants' confidence when paired with the CAI (blue dots) and the OAI (orange dots).

In Experiment 2, participants played with one AI in one half of the experiment, and both AIs in the other half. When playing with one AI only, this was either the CAI or the HAI. As expected, when playing with the CAI, participants exhibited similar behavior to Experiment 1, with participants' performance being worse than the AI's performance at first and then gradually converging towards it, and team performance being superior to individual (both human and CAI) performance (see Figure 3(a)). Team performance was superior when playing with the CAI compared to when playing with the HAI (see Figure 3). This suggests that the combination of choosing an overallocation strategy with a correlated confidence in the CAI results in error correction (even though human confidence is unaffected by performance) and, ultimately, in a team performance that is superior to both the AI and human performance (see Figure 3(a)). The same effect was not observed in the human-HAI team, where team performance was better than human performance, but worse than the individual HAI performance, this, likely due to both the humans' and HAIs' confidence being unaffected by the optimality of their decisions. In Experiment 2, we also compared performance of one-AI vs. two-AIs teams, shown in Figure 4. The one-AI condition included teams paired with either the CAI or an HAI. The CAI/HAI performance reflects the average performance of both AIs (Figure 4(a)). In both conditions, individual human performance increased similarly over time. Overall, team performance in both conditions was generally superior to individual performance, though with some exceptions in the two-AI condition ((Figure 4(b)), where the HAI surpasses team performance in the final sequences.

B. Team Performance and personality

As described in Section II-E, we trained an MLP to examine whether human personality traits can predict team performance levels (high or low) across different AI-type teams.

For the OAI, higher-performing teams were those with joint performance scores above 80% ($N = 4$), while lower-performing teams scored below 75% ($N = 4$). Two teams performing too close to the mean were excluded. The MLP after Leave-One Out cross-validation (see II-E) achieved high vs. low performance classification accuracy of 75%. The most influential feature in the model was agreeableness, with the highest negative weights. Looking at the signs and magnitudes of other weights suggested that *teams composed of individuals who are disagreeable, introverted, emotionally stable, and less open to new experiences — but also conscientious, impulsive, tolerant of ambiguity, and comfortable with risk — tend to achieve the highest performance when working with OAI.*

For the CAI, higher-performing teams were those with joint performance scores above 80% ($N = 7$), while lower-performing teams scored below 65% ($N = 7$). Data from both experiments was included here, again teams performing close to the average were excluded. The MLP model after Leave-One Out crosss-validation achieved 71% classification accuracy. The most influential trait was neuroticism, with high negative weights. Looking also at the other weights indicated that *individuals who are emotionally stable, introverted, less open to new experiences — but also agreeable, impulsive, tolerant of ambiguity, and comfortable with risk— are the best team mates for CAI.*F

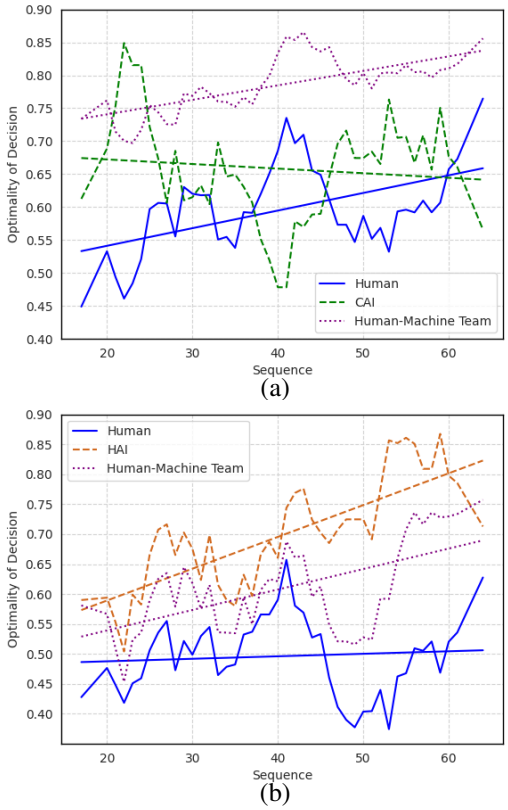


Fig. 3. Experiment 2: individual and team performance when playing with either the CAI (a) or the HAI (b).

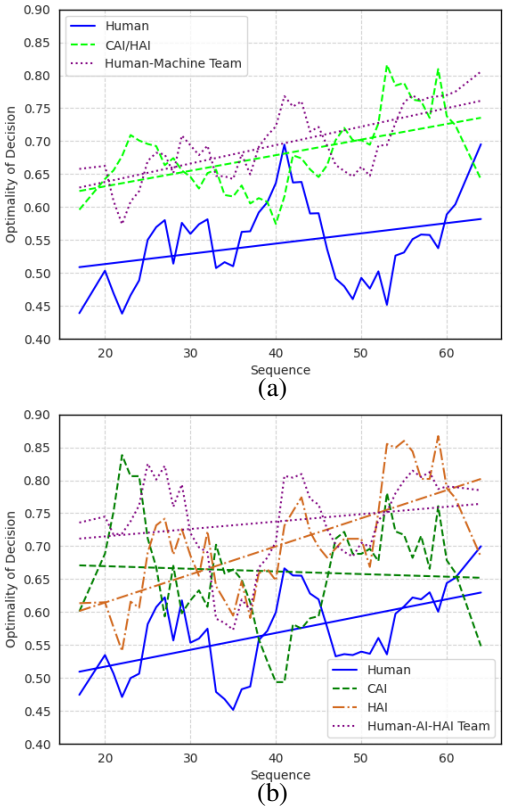


Fig. 4. Experiment 2: Individual and team performance when playing with one (a) vs. two (b) AIs. In (a), the AI's performance (in green) is the average of the CAI and HAI.

For the HAI, higher- and lower-performing teams were those with joint performance scores above and below 64%, respectively ($N = 5$ in each group). Here, the MLP model achieved 100% classification accuracy using just two features: agreeableness (with a positive weight) and ambiguity tolerance (weighted negatively), suggesting that *individuals who are agreeable but less tolerant of ambiguity performed best when paired with the HAI*.

C. Decision strategies and trust in AI

In Experiment 2, participants were also asked to rate their trust in both the CAI and HAI, on a scale from 0 - 100 %, directly after playing with both of them. As expected, trust in the HAI was significantly higher than trust in the CAI (medians 75 % and 65 % respectively, $p < 0.05$). Higher trust in the HAI might be related to its consistent, gradually improving performance, in contrast to the CAI whose performance is more context-dependent: because of its overallocation strategy, it performs better on sequences with fewer cities, which results in inferior performance compared to the participants in longer sequences (see for example sequences 38 to 45 in 3). Moreover, the HAI adopts a decision strategy that mirrors human behavior, which is likely fostering participant's trust even though team performance is on average inferior than when paired with the CAI. In early interactions with the CAI, participants tend to allocate fewer resources, adhering closely to a matching strategy. However, under the influence of the CAI, they gradually increase their resource allocation, moving toward the optimal strategy (Figure 5(a)). Interestingly, when human participants interact with the HAI, they do not exhibit the same shift in strategy. Without being exposed to the CAI's overallocation strategy, they consistently apply the matching strategy throughout the game (Figure 5(b)). Possibly, the HAI's human-like behavior reinforces this approach, subtly affirming the participants' initial strategy rather than challenging it.

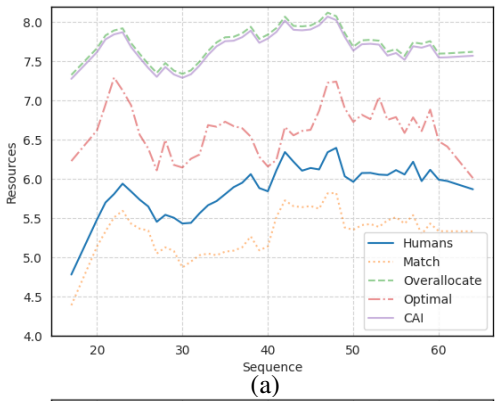
IV. Discussion and Conclusions

With the right ingredients, groups can act as "superorganisms" with enhanced sensing capabilities and intelligence. In this study, our objective was to identify such ingredients for human-AI teams, specifically focusing on decision-making under uncertainty, where there are no right or wrong choices, but only choices with different degrees of optimality. Teams included either one human and one AI (Experiments 1 and 2) or one human and two AIs (Experiment 2).

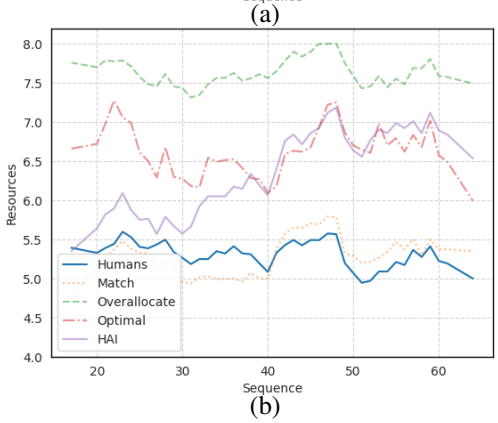
The findings confirmed that human-AI teams generally outperform both the human and the AI,[2] although there were exceptions — particularly when teams included two AIs.

Specifically, when the AIs in the team employed overallocation strategies (CAI and OAI) and their confidence correlated with decision optimality, *team performance surpassed that of*

[2]The resulting superorganism could thereby be considered both *superhuman* and *super-AI* [18].



(a)



(b)

Fig. 5. Experiment 2: a) Possible and applied strategy by participants paired with CAI (a) and when paired with HAI (b).

individual members, whether human or AI. However, when we introduced *humanized AIs* (HAIs), designed to mimic human behavior, we found that while they positively influenced human trust in the AI, this *did not lead to improved team or individual performance*. This is likely due to a low confidence-vs-optimality correlation in human participants.

These findings confirm the *critical role of metacognition* in team decision-making -— in particular, the ability to accurately judge the quality of own decisions (confidence), *for both humans and AI teammates*.

Our neural networks trained to distinguish between high- and low-performing teams based on the human player's personality and AI type uncovered some *important preliminary patterns to guide human-AI team co-selection*. Notably, both OAI and CAI were complemented by emotionally stable, introverted, and risk and ambiguity-tolerant individuals. Conversely, conscientiousness was a key positive trait to work with the OAI but was neutral for the CAI. Also, agreeableness was *negatively* associated with performance when working with the OAI, but *positively* associated with performance when teaming with the CAI, as well as with the HAI.

Naturally, more experimental data will be needed to corroborate and extend the findings reported in this article. Future work will need to focus on larger teams, deeper analyses of the influence of personality on the dynamics and performance of human-machines teams, strategies to train participants for improved metacognition, and the use of behavioral, physiological and neural data to improve calibration of human confidence.

Acknowledgment

This work was supported by the UK Defence Science and Technology Laboratory.

References

[1] A. H. DeCostanza, A. R. Marathe, A. Bohannon, A. W. Evans, E. T. Palazzolo, J. S. Metcalfe, and K. McDowell, "Enhancing human-agent teaming with individualized, adaptive technologies: A discussion of critical scientific questions," *US Army Research Laboratory Aberdeen Proving Ground United States*, 2018.

[2] P. Gupta, T. N. Nguyen, C. Gonzalez, and A. W. Woolley, "Fostering collective intelligence in human–ai collaboration: laying the groundwork for cohumain," *Topics in cognitive science*, 2023.

[3] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International journal of human-computer studies*, vol. 58, no. 6, pp. 697–718, 2003.

[4] C. Pelau, D.-C. Dabija, and I. Ene, "What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry," *Computers in Human Behavior*, vol. 122, p. 106855, 2021.

[5] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of social issues*, vol. 56, no. 1, pp. 81–103, 2000.

[6] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, "In the shades of the uncanny valley: An experimental study of human–chatbot interaction," *Future Generation Computer Systems*, vol. 92, pp. 539–548, 2019.

[7] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, pp. 313–323, 2013.

[8] R. G. Stephens, C. Semmler, and J. D. Sauer, "The effect of the proportion of mismatching trials and task orientation on the confidence–accuracy relationship in unfamiliar face matching.," *Journal of Experimental Psychology: Applied*, vol. 23, no. 3, p. 336, 2017.

[9] S. L. Sporer, S. Penrod, D. Read, and B. Cutler, "Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies.," *Psychological Bulletin*, vol. 118, no. 3, p. 315, 1995.

[10] S. Gervais and I. Goldstein, "Overconfidence and team coordination," *Available at SSRN 470787*, 2003.

[11] M. N. Riaz, M. A. Riaz, and N. Batool, "Personality types as predictors of decision making styles.," *Journal of Behavioural Sciences*, vol. 22, no. 2, 2012.

[12] W. R. Forrester and A. Tashchian, "Effects of personality on attitudes toward academic group work.," *American Journal of Business Education*, vol. 3, no. 3, pp. 39–46, 2010.

[13] T. Woodman, S. Akehurst, L. Hardy, and S. Beattie, "Self-confidence and performance: A little self-doubt helps," *Psychology of sport and exercise*, vol. 11, no. 6, pp. 467–470, 2010.

[14] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[15] J. H. Patton, M. S. Stanford, and E. S. Barratt, "Factor structure of the barratt impulsiveness scale," *Journal of clinical psychology*, vol. 51, no. 6, pp. 768–774, 1995.

[16] C. W. Lejuez, J. P. Read, C. W. Kahler, J. B. Richards, S. E. Ramsey, G. L. Stuart, D. R. Strong, and R. A. Brown, "Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart).," *Journal of Experimental Psychology: Applied*, vol. 8, no. 2, p. 75, 2002.

[17] D. L. McLain, "Evidence of the properties of an ambiguity tolerance measure: The multiple stimulus types ambiguity tolerance scale–ii (mstat–ii)," *Psychological reports*, vol. 105, no. 3, pp. 975–988, 2009.

[18] R. Poli, "Super-human and super-AI cognitive augmentation of human and human-AI teams assisted by brain computer interfaces," in *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '23, (New York, NY, USA), p. 3, ACM, 2023.