Humans in the Group, Computers in the Coop. Comparison of individual and collective improvement in cognitive tasks in adjunct AI settings

Chiara Natali[1,2][0000−1111−2222−3333], Luca Marconi[1][2222−−3333−4444−5555], Caterina Fregosi[1][1111−2222−3333−4444], and Federico Cabitza[1,3][0000−1111−2222−3333]

[1] Università degli Studi di Milano-Bicocca, Department of Computer Science, Systems and Communication, Milan, Italy chiara.natali@unimib.it
[2] University of Applied Sciences and Arts of Southern Switzerland, Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland
[3] IRCCS Ospedale Galeazzi-Sant'Ambrogio, Milan, Italy

Abstract. This paper explores how different human-AI collaboration protocols (HAI-CPs) influence performance in individual human+AI 'centaur' settings as well as in 'computer in the group' settings. Establishing effective HAI-CPs is essential to advancing Human Work Interaction Design (HWID), especially as industries move from the question of AI adoption to the challenge of integrating AI sustainably and effectively into collaborative workflows. Based on an experiment with logic puzzles supported by a simulated GPT-like AI, this study assesses AI's impact on decision-making, particularly regarding automation bias (the degree to which AI misleads users) and algorithmic aversion (the rejection of accurate AI recommendations). Our findings reveal that AI significantly enhances the cognitive performance of both centaurs and groups ('Bigae'), acting as a performance leveler by improving outcomes for lower performers more than high performers. Centaurs and Bigae's reported perceptions and observed error rates suggest that collective intelligence leads to less reliance on AI in group settings, relegating the machine to a more peripheral (adjunct) role in the decision-making process. While collectives are still improved by adding AI systems as additional average-performing teammates, deliberation with peers still appears to be the most powerful booster of human performance, and collective intelligence still outperforms individual intelligence, even when this is supported by AI.

Keywords: Human-AI Interaction · Group Decision-Making · Hybrid Intelligence · Technology Dominance · Kasparov's Law · Frictional AI

1 Introduction

Recent research has consistently confirmed the benefits of Artificial Intelligence (AI) in individual decision-making, where AI-based support is leveraged for cognition- and knowledge-intensive tasks in single-user settings [22]. This body of work has strengthened the legitimacy of the 'centaur' model (human-AI hybrid intelligence) in Human-AI Interaction (HAII) [19].

The centaur model symbolized a synergistic collaboration in which the complementary strengths of human and machine intelligence are integrated to achieve superior outcomes. The model gained prominence through the experiences of Chess Grandmaster Garry Kasparov, whose widely publicized defeat against IBM's chess-playing system *Deep Blue* in 1997 marked a pivotal moment in the evolving narrative of human versus machine intelligence. Rather than framing this as a zero-sum competition, Kasparov envisioned a new paradigm for collaboration.

This vision led to his creation of *Advanced Chess*, a format where human players enhance their decision-making by partnering with AI. It was in this context that the term 'centaur' emerged, describing the *Advanced Chess* player who blends AI's computational efficiency with their own intuition and creativity.

This term has been previously used in workplace studies, with Dell'Acqua et al. [19] defining as 'centaurs' those workers who strategically delegate tasks between AI and themselves, leveraging the strengths of each.

While the centaur model has demonstrated success in individual settings, achieving effective and sustainable integration into broader work environments presents an ongoing challenge.

Extending this concept to group deliberations, where decisions are made collectively, introduces added complexity in the interactions between humans, AI, and group members in collaborative settings [23]. This complexity, however, presents opportunities for richer outcomes [14] as it can be mitigated through collective sense-making, where team members jointly interpret and understand the AI system, enhancing group cohesion and performance [23].

This paper contributes to this nascent strand of research by exploring how different human-AI collaboration protocols (HAI-CPs) [8] influence group performance in hybrid human-AI decision-making settings, particularly in the context of solving logic puzzles collaboratively.

Understanding and establishing optimal HAI-CPs is crucial for advancing Human Work Interaction Design (HWID). As industries shift their focus from adopting AI to integrating it effectively into collaborative workflows and complex analytic tasks [19], it becomes clear that the quality of human-AI interactions is as critical—if not more so—than the quality of AI itself for decision-making and performance. This requires a detailed investigation of the dynamics that characterize human-machine interactions [34], resonating with Malone's call for a deeper understanding of how human intelligence and computational power can together address complex cognitive challenges effectively [27].

Integrating AI into collaborative decision-making must be approached as a matter of sustainability, not only due to the environmental costs of AI but also in

consideration of its potential long-term impacts on human skills, responsibility, and workforce inclusion.

This broader view on sustainability aligns with HWID's commitment to fostering the UN Sustainable Development Goal of 'Decent Work' (SDG 8) by prioritizing skill-preserving and meaningful work environments [15]. This includes addressing "cognitive work environment problems", defined by Sandblad et al. as "when properties of the work environment hinder the workers to use their skills efficiently" [36]. Tackling the complex cognitive challenges in technology-intensive workplaces requires careful design and are thus less frequently addressed than physical and psychosocial factors, with Bouzekri et al. [5] admitting in 2023 that (until then) "little HWID research has so far addressed [...] social dimensions of sustainability".

In this study, we use the mythological and historical terms 'Centaurs' and 'Bigae' to metaphorically describe three different HAI-CPs, as detailed Section 3. The centaur (1), a mythological hybrid, represents a system where humans and AI collaborate as a single unit, blending human intuition with AI's computational power. The Roman two-horse chariots, named 'Bigae', represent instead the group collaboration setting, further categorized into Group-first Bigae (2) and AI-first Bigae (3). Historically, Bigae are driven by charioteers named 'Aurigae': here, we use this term to refer to the individuals within the group making their autonomous decision during the pre-group deliberation phase.

This study is also a first of its kind in assessing *technology dominance* [1] in group deliberation settings. This refers to the impact of AI on decision-making in terms of automation bias (the extent AI misleads users or groups) and algorithmic aversion (the extent humans reject AI recommendations) [7].

By examining these phenomena, we aim to shed light on how AI shapes decision-making in collaborative environments, offering insights that are critical for understanding and optimizing Human-AI interaction in workplace settings for their long-term sustainability.

2 The Adjunction Perspective to the Computer in the Group

The integration of AI into decision-making processes offers a dual perspective: AI acting as an agent that can either collaborate with or supplant human users, or AI serving as a powerful tool enhancing human capabilities, positioned as "supertools and active appliances" that amplify human capabilities to achieve masterful outcomes, rather than as teammates or collaborators [39]. Further emphasizing group dynamics, Malone suggests shifting from placing humans in the loop to integrating AI systems within the human group [27], a sentiment driven further by Shneiderman's mantra, "Humans in the group; computers in the loop" [38]

Elaborating on these proposals, we propose the concept of *adjunct* AI [12], likening it to a computer "in the *coop*", that is, relegated to an ancillary and peripheral position and function, especially so for decision-support systems based

on Large Language Models (LLM) that have recently been equated to "stochastic parrots" by Bender et al. [4]. This isolation "in the coop", stems from reflections on the potentially adverse effects of AI on individual cognition, including automation bias, complacency, and algorithm aversion [25].

Adjunct AI can be considered an instance of Frictional AI [31], in that it ensures that decision support systems stimulate user cognitive activation, countering the motivational dip observed when humans collaborate with high-productivity AI [18]. When adjunct, AI acts as a post-consultation or second opinion mechanism which requires initial human decision-making. This aims at mitigating both anchoring bias [14, 2] and possibly groupthink phenomena [3].Anchoring bias, where initial AI recommendations heavily influence subsequent deliberations, risks constraining decision-makers' perspectives [14]. For example, it has been observed that many groups often commence their deliberative process by discussing topics related to AI recommendations [14]. While this direct initiation of discussions based on AI recommendations might anchor decisions, the diversity of opinions within a group can serve as a counterbalance, diluting the anchoring effect through exposure to varying viewpoints [14]. Still, the absence of individual pre-deliberation decisions may lead to a homogenization of thought, known as *groupthink*, where dissent and innovation are stifled [3]. This dynamic interplay suggests that the structure of interaction with AI—whether initiating discussion or following individual deliberation—significantly influences the cognitive biases at play within group decision-making processes. By encouraging initial independent decision-making, followed by AI consultation, Adjunct AI can facilitate a balance between leveraging AI insights and preserving individual and collective critical thinking.

Previous research underscores the advantage of group decision-making where AI contributes to group discussion (Bigae) over human-AI dyads (Centaurs) [10]. Rather than serving as oracles [30], AI systems should be catalysts for collective discourse, leveraging human transactive memory and social skills to mitigate the biases associated with AI-dominated decision-making, or agential AI [10]. The impact of AI in decision-making is difficult to understate, as Decision Support Systems posited as "structuring the cognitive process latent to the decision-making" even when providing access to the necessary information sources [37]. Our approach advocates for positioning AI as an enabler of enhanced cognitive engagement and collective intelligence, shifting its role from a dominant agent to a supportive, consultative entity [12]. This reimagined interaction seeks to amplify the benefits of human-AI collaboration, leveraging the unique strengths of both while mitigating known cognitive biases and fostering a more dynamic, thoughtful, and hybrid [20] decision-making environment.

3 Methods

3.1 Participants

Forty-six Master's students from a 'Communications and Technology' degree participated, forming 22 human-AI pairs, or 'Centaurs', and six teams, or 'Bigae'

(i.e., chariots), who act as a collective, expressing a single decision. We chose this term to represent the enhanced collective dynamics through human-AI teamwork as sort of 'super minds', in Malone's terminology [27]. Bigae are composed of four participants named 'Aurigae' (i.e., charioteers), who are individual decision-makers in the initial step. We further categorized Bigae in two different HAI-CPs based on the timing of AI advice: Group-first teams discussed potential solutions before AI consultation, while AI-first teams received AI input before deliberation (see Figure 3). These two HAI-CPs have been previously studied in the literature focusing on identifying and studying cognitive biases only in individual decision-making processes arising from interaction with AI support systems (see, for example, [6] and [8]). Finally, Centaurs are human-AI pairs where an individual works directly with the AI support. All three participants configurations are illustrated intuitively in Figure 2.

The composition of these groups, albeit random, ensured gender diversity with a majority of women (on average, 2.71 women and 1.28 men per four-person group), possibly aligning with Woolley et al.'s observations on ideal group composition for team performance [42] albeit the significance of gender for team cooperation has been debated in the HWID literature [44]. The variation in experimental protocol according to the assigned HAI-CP are detailed in Section 3.3 and in Figure 3.

3.2 Materials

Interface. We employed a simulated 'Wizard-of-Oz' [16] GPT-like interface developed via *Figma*, as shown in Figure 1. The system appeared to respond to images of pattern reasoning tasks or transcriptions of word logic puzzles, to which it provided both answers and accompanying rationales.

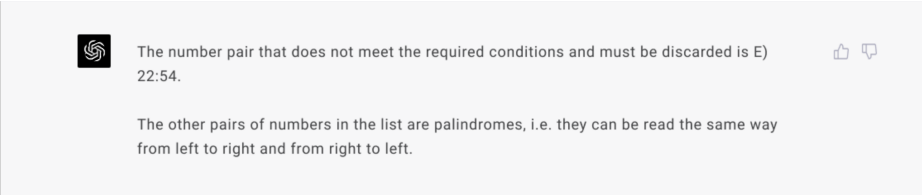


Fig. 1: Simulated interface shown to the participants, with an example of an erroneous AI suggestion in response to the puzzle *"Deduce the pair of numbers to discard from those proposed below: A) 11:22; B) 01:11; C) 15:51; D) 16:61; E) 22:54"* (screenshot translated from Italian).

To assess group and individual performance in terms of accuracy and examine the influence of AI, even when providing incorrect answers, we designed the 19 responses of the simulated AI so that would be 13 correct and 6 incorrect, resulting in an accuracy rate of 68.42%.

This accuracy rate closely aligns with, but exceeds, the 62.5% accuracy rate set by Chen et al. [13] in their study on human-AI decision-making with explanations. The authors chose this rate to oversample incorrect AI predictions, where oversampling incorrect predictions enabled deeper investigation into participant reliance on AI.

Furthermore, selecting a clearly less-than-perfect accuracy rate mirrors realistic scenarios reflecting the limitations of LLMs. Given the well-documented struggles of LLMs with logical tasks [35], the chosen accuracy rate could be considered generous, especially considering the period of experimentation, which occurred before the release of more advanced GPT systems like *ChatGPT-4o*. This design choice aimed to present the AI as a realistic tool, with strengths and weaknesses, to provide participants with an authentic experience of interacting with current AI limitations.

To minimize the risk of negative priming due to inaccurate answers and foster user trust in the AI system, we designed the initial sequence of responses to include 9 consecutive correct answers. The first 2 answers displayed highly relevant and compelling explanations, while lower-quality explanations were deliberately included in answers 3, 5, 7 and 9. Wrong answers were introduced later at questions 10, 12, 14, 16, 18 and 19, with all other responses being correct. This design choice allowed us to investigate the level of user trust in an AI that makes mistakes, as well as to verify potential automation biases.

This strategy aimed to create an initial positive impression of the AI as a reliable, though not infallible, tool. This positive priming could subsequently influence user willingness to accept AI suggestions even when incorrect.

While alternative response patterns might foster a more dynamic trust-building experience, the use of an initial sequence of correct answers as a trust-building strategy has been documented in the literature on interactive XAI systems [32, 33]. This approach was chosen for its effectiveness in enabling cross-comparison of trust formation among different groups, allowing us to study automation bias in a controlled and consistent manner.

Pre- and Post-test Assessments. We also conducted both pre- and post-test assessments of perceived trust and utility regarding AI interaction, with questionnaires drawing from validated scales such as the Artificial Intelligence Anxiety Scale [41] and the Technology Acceptance Model [17]. These questionnaires included questions on trust in AI and the perceived benefits of AI support and allowed us a comprehensive view of participants' attitudes towards AI. We aggregated the TAM-related questions to construct a UTILITY score, which were associated with a .85 Cronbach's Alpha.

3.3 Procedure

In this study, we investigated the effects of AI on decision-making accuracy and cognitive biases in both individual and group contexts. The experiment involved administering 19 logic puzzles, including numerical and alphanumeric

logic, deductive reasoning, graphic interpretation, and anagrams from the website *Youmath*[4]. The 19 items were selected and extracted from higher-level quizzes so that the task would be challenging for the subjects, thus evoking the need to include the AI's opinion and support. The rationale behind choosing these types of puzzles was to simulate tasks that individuals commonly face alone but could benefit from collaborative problem-solving and AI support. While the puzzles are generally solved individually, our aim was to examine how transitioning from individual to collective, AI-augmented problem-solving could reveal shifts in cognitive reliance and performance. This setup allowed us to analyze whether group deliberation could mitigate the influence of automation bias or algorithmic aversion typically observed in solitary decision-making.

Fig. 2: On the left, a representation of the Centaur configuration (human - upper body, AI system - lower body). On the centre, the AI-first Biga configuration (four humans - the 'Aurigae', AI system - the horse). On the right, the Group-first Biga (Aurigae first deliberate together: only then will they take the reins of the AI system).

Having introduced in Section 3.1 the general concept of the different collaboration protocols used in our study, we now provide a more detailed description of each of specific configuration within the experiment procedure. In our experiment, we tested three different HAI-CPs to better understand their influence on decision-making outcomes:

- **Group-First Bigae:** In this configuration, group members (Aurigae) discussed potential solutions within their groups before consulting the AI system. This protocol aimed to observe the natural flow of group deliberation and the influence of AI when introduced after an initial group decision was made. This allowed us to measure how AI input could alter or refine group consensus.
- **AI-First Bigae:** Conversely, in this protocol, group members (Aurigae) first received the AI's suggestion before engaging in group discussions. The goal here was to understand how AI could prime discussions, potentially influencing group dynamics by anchoring initial deliberations to the AI's suggestion.
- **Centaurs:** In this individual-focused collaboration model, each participant worked directly with the AI system, aiming to combine human decision-making skills with AI's computational power. This setting allowed us to

[4] https://www.youmath.it

isolate the impact of AI on individual performance without the complexities introduced by group interactions.

These protocols were chosen to explore the dynamics of human-AI collaboration beyond its outcomes in terms of accuracy, emphasizing the process of how decisions are made within different collaborative settings.

The experimental phase took place in-person, within university spaces. It began with an initial questionnaire (via *Google Forms*) to gauge the participants' predispositions towards teamwork and AI interaction, followed by six collaborative sessions (plus one centaur-specific session), and concluded with a post-test questionnaire. Each session lasted approximately 1 hour and 30 minutes, including the initial and final questionnaires.

Phase 1: Initial Questionnaire. The initial questionnaire involved completing 18 questions on Google Forms to assess the predisposition to interaction with AI systems and, for aurigae participants, to teamwork. The questions can be divided into two subcategories: a section of 3 questions on the perceived usefulness of using artificial intelligence; a section of 9 questions on the predisposition to group work. Users were asked to indicate their degree of agreement along a six-point Likert scale-like semantic differential where 1 corresponded to (totally disagree) and 6 to (totally agree), to mitigate central tendency bias.

Phase 2: Collaborative Sessions. Following the initial questionnaire, participants were then sequentially shown 19 logic puzzles with four answer options ('Case presentation' in Figure 3) and asked to provide anonymous individual answers via LimeSurvey[5], including their confidence level for each response on a 4-value ordinal scale, where 1 corresponded to the minimum confidence and 4 to the maximum confidence (Step 1, 'First Individual Answer' and 'Confidence Collection', 60").

The collaborative session protocol then varied by HAI-CP, as illustrated in Figure 3. Specifically:

- **Group-First Bigae**: Participants discussed possible solutions together and were asked to provide a first group decision, denoted as the 'auriga' decision (Step 2, '1° Group Deliberation', 60"). After that, participants consulted the AI advice and together developed the final 'biga' (group+AI) answer (Step 3, 'FHD collection', 60"). Both responses were manually marked by the moderators.
- **AI-First Bigae:** Participants consulted the AI before having the opportunity to talk to each other, and after two minutes of group discussion were asked to provide a final 'biga' answer (Step 2, 'FHD collection', 120"). This response was manually marked by the moderators.

[5] http://limesurvey.org

– **Centaurs:** The participant consulted the AI system and provided their final answer (Step 2, 'FHD collection', 120"). This response was marked independently by the participants on the LimeSurvey platform.

In all configurations, the total time allocated for each logic puzzle was 180".

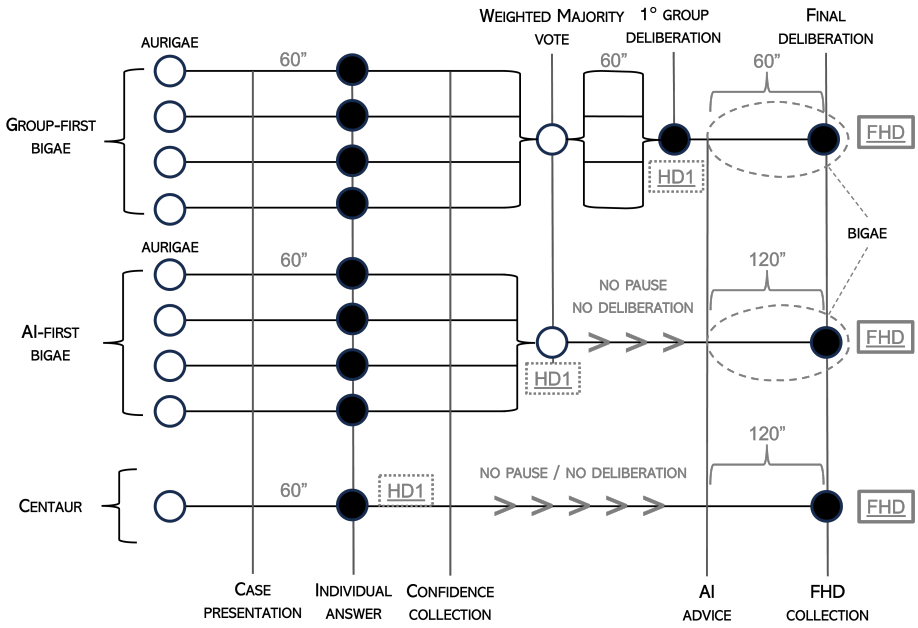


Fig. 3: Group deliberation scheme for Group-First Bigae, AI-First Bigae and Centaurs.

Phase 3: Final Questionnaire. A final questionnaire assessed participants' attitudes towards AI post-experiment, mirroring the initial questionnaire but with varied question order and phrasing to ensure participants would not mindlessly replicate the answers provided before the experiment.

3.4 Reliance Patterns

Reliance patterns have been proposed as a critical framework for evaluating the influence of Artificial Intelligence within AI-enhanced decision-making processes [7]. These patterns are rooted in the theory of technology dominance [1] and, as outlined in Table 1, they describe how humans interact and rely on AI systems during the decision-making process through a sequence of decision stages: an initial, unassisted human decision (HD1), the subsequent AI advice (AI), and the final AI-supported human decision (FHD). By analyzing the correctness (and/or

incorrectness) of decisions at each stage, reliance patterns offers a structured method to quantify the impact of AI on human decision-making, that allows to assess the extent to which individuals might adjust their initial judgments after considering AI input. For example, AI advice enhances decision-making by encouraging the rejection of initially incorrect human decisions (HD1) in favor of correct AI advice, or conversely, AI support may detrimentally lead individuals to discard their correct initial decisions for incorrect final decisions (FHD) based on erroneous AI recommendations. We elaborated these reliance

Table 1: Definition of Automation Bias and Conservatism Bias as reliance patterns between human decision makers and their AI system. In the first three columns, 0 denotes an incorrect decision, and 1 a correct decision.

Human judgment (HD1)	**AI support** (AI)	**Final decision** (FHD)	**Reliance pattern**
1	0	0	Detrimental Over-reliance **(Automation Bias)**
0	1	0	Detrimental Self-reliance **(Conservatism Bias)**

patterns via the open-access tool *Human - AI Interaction Assessment*[6] to assess the levels of automation bias and conservatism bias displayed by the responders. The collection of HD1, AI and FHD is illustrated in Figure 3. Notably, for Group-First Aurigae, HD1 refers to the first, pre-AI Group (Bigae) Deliberation. For AI-First Aurigae, instead, HD1 was the weighted average of individual, pre-AI responses, according to the responder's confidence. Weighted averages were considered in place of HD1 in order to isolate the effect of group deliberation in a cross-HAI-CP comparison (Group First vs. AI-First) as well an intra-HAI-CP comparison (Weighted Average vs. Group Deliberations in Group-First Bigae).

4 Results and Discussion

4.1 Centaurs vs. Bigae: *Kasparov's Law* in Action

Initial human group performance was 57.89% for AI-first Bigae and 50% for Group-first Bigae, compared to the AI performance of 68.42%. To guide our analysis, we built on the study by Cabitza et al. [9] on radiological double reading, which explored Kasparov's law [24]. This law posits that 'Weak *[sic]* human + machine + better process is superior to a strong computer alone.' In their study, the researchers found that groups of humans who initially perform worse than AI can significantly outperform it when their judgments are aggregated via majority voting. We observed a similar phenomenon in our study, as both Group-first

[6] https://haiiassessment.pythonanywhere.com, last accessed 29/12/2024

and AI-first HAI-CPs showed marked improvements in group performance after deliberation.

Final group performance reached 78.95% for AI-first Bigae and 75% for Group-first Bigae, representing gains of +21.06% and +25%, respectively, compared to their initial performance. Notably, these final performances also exceeded AI performance, though not statistically significant, with AI-first Bigae outperforming AI by +10.53% ($t(56) = 1.6$, $p = .109$) and Group-first Bigae by +6.58% ($t(75) = 1$, $p = .300$). In contrast, Centaurs, who did not aggregate their judgments with others, achieved a final performance of 69.14%, comparable to AI albeit marginally higher (+0.72%), and this difference not statistically significant ($t(417) = 0.3$, $p = .756$). As these findings suggest alignment with Kasparov's law, further studies are needed to robustly validate this trend with a larger sample to further support the idea that a 'better process' emerges through collaborative human efforts, rather than relying solely on individual human-AI pairings.

Moreover, we observed that the AI acted as a performance leveler, as shown graphically in Figure 4: global performance variance decreased and the average performance of the low-performers increased. This means that the low-performers (at pre-support step), both individuals and teams, improved more than high-performers at the post-support step. This phenomenon was analytically confirmed through a negative (-0.55) and statistically significant Pearson Correlation Coefficient ($p=0.002$), as well as by the angular coefficient of the regression line (-0.43).

As for Automation Bias, it was lower for Aurigae as expected, while for Conservatism Bias we did not observe any statistically significant difference, though Aurigae had higher variance (Figure 5). Despite these differences, both Aurigae and Centaurs reported increased trust in AI, with Aurigae showing a significantly higher trust level post-test ($p=0.028$, effect size=0.31).

4.2 Deep Dive into Group Deliberation: AI-First vs. Group-First

We found no statistically significant difference in teamwork predisposition between the Aurigae in Group-first and AI-first Bigae ($p=0.2258$, effect size=0.23), suggesting that our findings on group performance were not inadvertently skewed by this variable.

While in AI-first Bigae we could not decouple the collective intelligence effect (that is the power of discussion) from the AI influence, this was possible in the Group-first Bigae. In fact, the first group deliberation of AI-first Bigae necessarily followed exposure to AI advice and could as such be influenced by it. Instead, the initial group decision of the Group-first Bigae was the result of the collective deliberation of the members preceding any exposure to AI advice.

We observed no significant difference in the perceived utility of AI between the AI-first and Group-first HAI-CPs ($p=0.2114$, effect size=0.17), although Group-first Aurigae experienced lower Automation Bias and higher Conservatism Bias (Figure 5), thus relying less on AI.

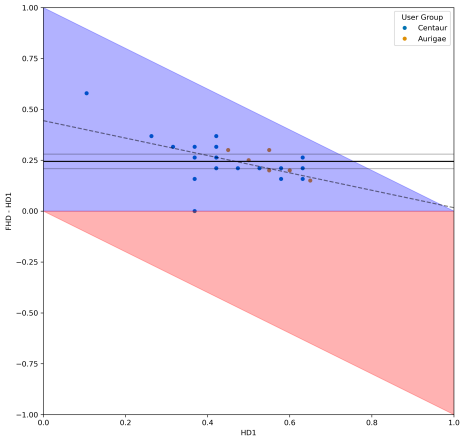
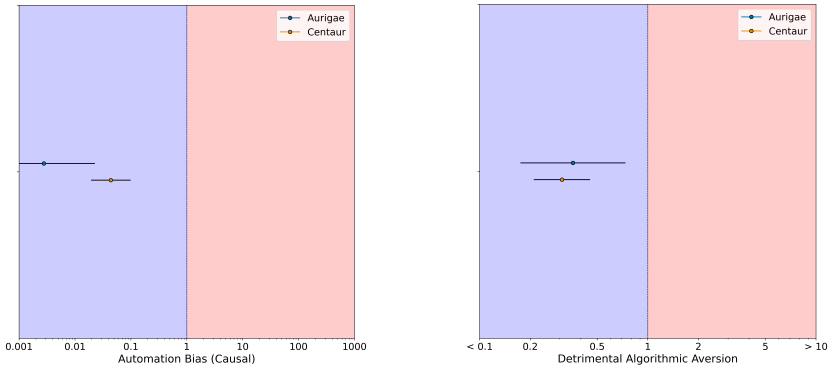


Fig. 4: Benefit Diagram for Aurigae and Centaurs, showing an increase in performance for both types of users.



(a) **(b)**

Fig. 5: (a) Automation Bias Diagram for Aurigae and Centaurs, showing a significantly lower automation bias for Aurigae and (b) Conservatism Bias Diagram for Aurigae and Centaurs, showing similar levels for Centaurs and Aurigae.

For AI-first groups, the average of 8.3 errors at Weighted Average became 3.7 errors after exposure to AI advice (FHD), a 58% decrease. Group-first Bigae exhibited a lower basal Weighted Average accuracy, averaging at 9.5 HD1. However, their transition from Weighted Majority (38 errors) to Group Deliberation (32 errors) marked a 16% error reduction, while the final error count post-AI exposure dropped by 41% compared to Group Deliberation (and 50% relative to Weighted Majority). This provides a more granular description of the difference in final accuracy observed in Section 4.1 which was 3.95% lower for Group-First Bigae.

Our conjecture is that, within the Group-first setting, discussions that precede AI consultation risk entrenching groupthink phenomena and marginalizing AI input due to a pronounced algorithmic aversion. Specifically, Group-first configurations face the challenge of integrating AI advice within a constrained timeframe — having only a minute to incorporate AI perspectives following earlier group deliberation. Conversely, AI-first groups benefit from two minutes of deliberation time in light of the AI advice, allowing for a comprehensive evaluation of AI suggestions alongside individual initial decisions. This structure may curb groupthink as well as encouraging a more critical engagement with AI input, where discussion is initiated by critical reflections on AI's contributions [14]. While group deliberation undeniably influences and enhances outcomes, the integration of AI within these discussions emerges as powerful a determinant of group performance. In light of the critical role of the initial private decision [3], we advocate for a model where humans remain at the core of the group, with AI serving as an adjunct 'in the coop', acting as a trigger to group deliberation that also allows for its suggestions to be doubted and discussed by the human team. This approach balances AI's role in stimulating dialogue and ensuring that human intelligence remains at the forefront of collective decision-making.

5 Conclusion

Our findings reveal AI's potential in augmenting group intelligence and facilitating productive deliberation, suggesting the importance of integrating AI into collaborative settings through sustainable, human-centred approaches. This study on HAI-CPs aligns with HWID's goal to foster working environments that prioritize human skill, intuition, and well-being, in line with the Sustainable Development Goal of "decent work for all" (SDG 8) [15]. By tailoring effective HAI-CPs for specific work environments, we can support more resilient teams where individuals benefit from AI support without experiencing dependency or skill erosion.

The various HAI-CPs studied in this work can contribute to the identification and understanding of new human-AI interaction patterns in collaborative work environments. Specifically, these protocols illustrate how different approaches to AI integration—whether AI-first, human-first, or adjunct—can shape user experience in ways that either enhance or inhibit effective collaboration. This insight complements HWID's emphasis on the careful design of interaction models that

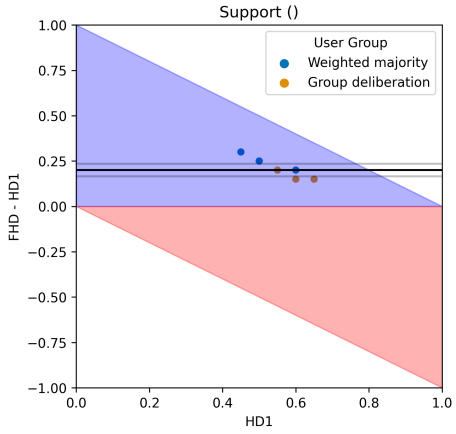


Fig. 6: Benefit Diagram for Weighted Majority vs Group Deliberation Aurigae, showing the higher basal performance for group deliberation and clear improvement for both WM and GD following exposure to AI advice.

are adaptive to the specific needs of users and their work environments [21, 40, 29], and further develops the research on the validity of *Kasparov's Law* [24, 10].

Sustainability, as understood by Kuhlman and Farrington's, extends beyond environmental preservation to include other essential resources as legacies for future generations [26, 5]. In this context, recent concerns about over-reliance on AI highlight the importance of safeguarding human skill and intuition as integral to sustainability at large. Specifically, we have shown how the use of AI in an auxiliary role has been shown to enhance the ability to maintain consistent performance among participants, particularly benefiting those with initially lower performance levels by promoting a more equitable distribution of cognitive load within the group.

However, we acknowledge certain limitations in our study's methodology and scope. Specifically, the formation of groups did not account for individual performance levels or internal leadership dynamics. In addition, our use of logic tasks does not adequately reflect current capabilities of LLMs [43], so we chose to simulate these ones as a proxy to explore AI-enabled 'hybrid intelligence' in a non-professional setting, preferring it to notion-driven tasks such as trivia. Another potential limitation of this study is related to the choice of task types. Although our study focused on logic puzzles typically used for individual cognitive training, we acknowledge that different types of tasks could potentially have influenced the outcomes. Tasks that require more explicit collaboration, such as problem-solving that benefits from diverse skill sets or tasks involving creativity, might yield different group dynamics and AI reliance patterns.

Future research should then investigate the differences between AI contributions that are limited to simple 'advice and explanation' (as in our study) and more complex, interactive scenarios facilitated by advanced LLMs.

The study was designed with a high degree of control, including specific tasks (logic puzzles), a fixed sequence of correct and incorrect AI responses, and predefined collaboration protocols. These choices were necessary to isolate variables like automation bias and trust-building. While the study's controlled design allows for rigorous exploration of HAI-CP dynamics, its generalizability to real-world work environments is inherently limited by the specificity of tasks and experimental conditions.

Despite these constraints, the findings offer a foundational understanding of HAI-CP dynamics, providing hypotheses for how similar patterns might manifest in professional settings. Future research should test these hypotheses in diverse and realistic contexts, incorporating varied task types, AI capabilities, and organizational constraints to validate and refine their applicability.

Ultimately, our study contributes to broader research on AI's potential to enhance group intelligence and stimulate human discussion, thus ensuring that AI does not merely act as *deus ex crowd* for challenging tasks [28]. Instead, the adjunction approach exemplifies what has been termed *frictional AI* [31, 11] by intentionally incorporating elements of resistance - in this instance, positioning AI in an ancillary role. This 'friction' encourages deeper engagement, critical thinking, and more comprehensive evaluation among decision-makers, ultimately enhancing the decision-making process.

This leads to the provocative statement of the title: *Humans in the group, computers in the coop*, as to motivate and call for further research into sustainable Human-AI interaction configurations where AI is moved to an ancillary role. Such configurations, or HAI-CPs, hold the potential to reinforce and elevate human intelligence in workplace settings, advancing both performance and collaboration.

Acknowledgements

C. Natali gratefully acknowledges the PhD grant awarded by the Fondazione Fratelli Confalonieri, which has been instrumental in facilitating her research pursuits. Chiara Natali gratefully acknowledges the financial support provided by the Federal Commission for Scholarships for Foreign Students in the form of the Swiss Government Excellence Scholarship (ESKAS No. 2024.0002) for the academic year 2024-25.

F. Cabitza and C. Fregosi acknowledge funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision making. CUP: H53D23008090001 funded by the European Union - Next Generation EU.

References

1. Arnold, V., Sutton, S.G., et al.: The theory of technology dominance: Understanding the impact of intelligent decision aids on decision maker's judgments. Advances in accounting behavioral research **1**(3), 175–194 (1998)

2. Bach, A.K.P., Nørgaard, T.M., Brok, J.C., Van Berkel, N.: "if i had all the time in the world": Ophthalmologists' perceptions of anchoring bias mitigation in clinical ai support. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2023)
3. Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., Frith, C.: What failure in collective decision-making tells us about metacognition. Philosophical Transactions of the Royal Society B: Biological Sciences **367**(1594), 1350–1365 (2012)
4. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of ACM FA. pp. 610–623 (2021)
5. Bouzekri, E., Barricelli, B.R., Clemmensen, T., Hertzum, M., Masoodian, M.: Sustainable human-work interaction designs. In: IFIP Conference on Human-Computer Interaction. pp. 674–679. Springer (2023)
6. Buçinca, Z., Malaya, M.B., Gajos, K.Z.: 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), 1–21 (2021)
7. Cabitza, F., Campagner, A., Angius, R., Natali, C., Reverberi, C.: Ai shall have no dominion: on how to measure technology dominance in ai-supported human decision-making. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–20 (2023)
8. Cabitza, F., Campagner, A., Ronzio, L.e.a.: Rams, hounds and white boxes: Investigating human–ai collaboration protocols in medical diagnosis. Artificial Intelligence in Medicine **138**, 102506 (2023)
9. Cabitza, F., Campagner, A., Sconfienza, L.M.: Studying human-ai collaboration protocols: the case of the kasparov's law in radiological double reading. Health information science and systems **9**, 1–20 (2021)
10. Cabitza, F., Campagner, A., Simone, C.: The need to move away from agential-ai: Empirical investigations, useful concepts and open issues. International Journal of human-computer studies **155**, 102696 (2021)
11. Cabitza, F., Natali, C., Famiglini, L., Campagner, A., Caccavella, V., Gallazzi, E.: Never tell me the odds: Investigating pro-hoc explanations in medical decision making. Artificial Intelligence in Medicine **150**, 102819 (2024)
12. Cabitza, F., Natali, C., et al.: Open, multiple, adjunct. decision support at the time of relational ai. FAIA **354** (2022)
13. Chen, V., Liao, Q.V., Wortman Vaughan, J., Bansal, G.: Understanding the role of human intuition on reliance in human-ai decision-making with explanations. Proceedings of the ACM on Human-computer Interaction **7**(CSCW2), 1–32 (2023)
14. Chiang, C.W., Lu, Z., Li, Z., Yin, M.: Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–18 (2023)
15. Clemmensen, T., Clemmensen, T.: Hwid policymaking. Human Work Interaction Design: A Platform for Theory and Action pp. 237–266 (2021)
16. Dahlbäck, N., Jönsson, A., Ahrenberg, L.: Wizard of oz studies: why and how. In: Proceedings of the 1st international conference on Intelligent user interfaces. pp. 193–200 (1993)
17. Davis, F.D., et al.: Technology acceptance model: Tam. Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption pp. 205–219 (1989)
18. Dell'Acqua, F., Kogut, B., Perkowski, P.: Super mario meets ai: Experimental effects of automation and skills on team performance and coordination. Review of Economics and Statistics pp. 1–47 (2023)

19. Dell'Acqua, F., McFowland III, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., Lakhani, K.R.: Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper (24-013) (2023)
20. Dellermann, D., Ebel, P., Söllner, M., Leimeister, J.M.: Hybrid intelligence. Business & Information Systems Engineering **61**, 637–643 (2019)
21. Gonçalves, F., Campos, P., Clemmensen, T.: Human work interaction design: An overview. In: Abdelnour Nocera, J., Barricelli, B.R., Lopes, A., Campos, P., Clemmensen, T. (eds.) Human Work Interaction Design. Work Analysis and Interaction Design Methods for Pervasive and Smart Workplaces. pp. 3–19. Springer International Publishing, Cham (2015)
22. Jarrahi, M.H.: Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. Business horizons **61**(4), 577–586 (2018)
23. Karvonen, H., Heikkilä, E., Wahlström, M.: Artificial intelligence awareness in work environments. In: Human Work Interaction Design. Designing Engaging Automation: 5th IFIP WG 13.6 Working Conference, HWID 2018, Espoo, Finland, August 20-21, 2018, Revised Selected Papers 5. pp. 175–185. Springer (2019)
24. Kasparov, G.: Deep thinking: where machine intelligence ends and human creativity begins. Hachette UK (2017)
25. Kliegr, T., Bahník, Š., Fürnkranz, J.: A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. Artificial Intelligence **295**, 103458 (2021)
26. Kuhlman, T., Farrington, J.: What is sustainability? Sustainability **2**(11), 3436–3448 (2010)
27. Malone, T.W.: Superminds: The surprising power of people and computers thinking together. Little, Brown Spark (2018)
28. Malone, T.W., Bernstein, M.S.: Handbook of collective intelligence. MIT press (2022)
29. Miao, X., Hou, W.j.: Research on the integration of human-computer interaction and cognitive neuroscience. In: Bhutkar, G., Barricelli, B.R., Xiangang, Q., Clemmensen, T., Gonçalves, F., Abdelnour-Nocera, J., Lopes, A., Lyu, F., Zhou, R., Hou, W. (eds.) Human Work Interaction Design. Artificial Intelligence and Designing for a Positive Work Experience in a Low Desire Society. pp. 66–82. Springer International Publishing, Cham (2022)
30. Miller, R.A., Masarie, F. E., J.: The demise of the "greek oracle" model for medical diagnostic systems. Methods of information in medicine **29**(01), 1–2 (1990)
31. Natali, C., et al.: Per aspera ad astra, or flourishing via friction: Stimulating cognitive activation by design through frictional decision support systems. In: CEUR WORKSHOP PROCEEDINGS. vol. 3481, pp. 15–19 (2023)
32. Nourani, M., King, J., Ragan, E.: The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 8, pp. 112–121 (2020)
33. Nourani, M., Roy, C., Block, J.E., Honeycutt, D.R., Rahman, T., Ragan, E.D., Gogate, V.: On the importance of user backgrounds and impressions: Lessons learned from interactive ai applications. ACM Transactions on Interactive Intelligent Systems **12**(4), 1–29 (2022)
34. Peng, A., Nushi, B., Kiciman, E., Inkpen, K., Kamar, E.: Investigations of performance and bias in human-ai teamwork in hiring. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 12089–12097 (2022)

35. Plevris, V., Papazafeiropoulos, G., Jiménez Rios, A.: Chatbots put to the test in math and logic problems: A comparison and assessment of chatgpt-3.5, chatgpt-4, and google bard. AI **4**(4), 949–969 (2023). https://doi.org/10.3390/ai4040048, https://www.mdpi.com/2673-2688/4/4/48
36. Sandblad, B., Gulliksen, J., Åborg, C., Boivie, I., Persson, J., Göransson, B., Kavathatzopoulos, I., Blomkvist, S., Cajander, Å.: Work environment and computer systems development. Behaviour and Information Technology **22**(6), 375–387 (2003)
37. Seymoens, T., Ongenae, F., Jacobs, A., Verstichel, S., Ackaert, A.: A methodology to involve domain experts and machine learning techniques in the design of human-centered algorithms. In: Human Work Interaction Design. Designing Engaging Automation: 5th IFIP WG 13.6 Working Conference, HWID 2018, Espoo, Finland, August 20-21, 2018, Revised Selected Papers 5. pp. 200–214. Springer (2019)
38. Shneiderman, B.: Human-centered artificial intelligence: Three fresh ideas. AIS Transactions on Human-Computer Interaction **12**(3), 109–124 (2020)
39. Shneiderman, B.: Human-centered ai. Issues in Science and Technology **37**(2), 56–61 (2021)
40. Vicente, K.J.: Cognitive work analysis: Toward safe, productive, and healthy computer-based work. CRC press (1999)
41. Wang, Y.Y., Wang, Y.S.: Development and validation of an artificial intelligence anxiety scale: An initial application in predicting motivated learning behavior. Interactive Learning Environments **30**(4), 619–634 (2022)
42. Woolley, A., Malone, T., et al.: What makes a team smarter? more women. Harvard business review **89**(6), 32–33 (2011)
43. Zhang, H., Huang, J., Li, Z.e.a.: Improved logical reasoning of language models via differentiable symbolic programming. arXiv preprint arXiv:2305.03742 (2023)
44. Zhi, X., Zhou, R.: The influence of automation and culture on human cooperation. In: IFIP Working Conference on Human Work Interaction Design. pp. 123–140. Springer (2021)