Time added on: the impact of multiple AI teammates on referee decision-making

Maia Low[1], Jolene Cox[1], Brandon King[1], Scott McLean[1], Chris Baber[2] & Paul Salmon[1]

[1]Centre for Human Factors and Systems Science, University of the Sunshine Coast, Australia
[2]Schools of Computer Science, University of Birmingham, UK

SUMMARY

There is a critical need to understand how the increasing deployment of AI technologies within Human-Autonomy Teams (HATs) impacts performance in different contexts. We investigated the effect of HAT composition (Human-AI dyad versus Human-AI-AI triad) on football referee decision-making performance for foul or no foul decisions in a series of English Premier League match excerpts. The findings demonstrated that decisions took longer in the human-AI-AI triad condition but decision accuracy and confidence were not impacted by HAT composition.

KEYWORDS

Artificial intelligence, Teamwork, Human autonomy teams, Sports officiating

Introduction

Teams and teamwork are changing, with an increasing use of Artificial Intelligence (AI) and Human-Autonomy Teams (HATs) across society. HATs comprise humans and intelligent autonomous agents working interdependently toward a common goal, where the autonomous agent is a computer entity with a partial or high degree of self-governance in decision making, adaptation, and communication (O'Neill et al., 2022). Though there is an extensive knowledge base surrounding human-human teams, there are key knowledge gaps regarding optimal HAT functioning, and how AI technologies can be designed to optimise HAT performance. Critically, though studies examining HATs are emerging, they have largely focused on simplistic human-AI dyads and few studies have explored larger HATs comprising multiple AI team members. This is despite the fact that HATs with multiple AI team members are emerging across society.

AI technologies are increasingly being deployed in sports officiating teams (e.g., football Video Assistant Referee (VAR), tennis Hawkeye, and baseball automated ball-strike system). As AI continues to advance, the addition of further technologies in sports officiating teams is likely. Given limitations in previous HATs research, the impact of having multiple AI team members on team performance is not clear. This paper describes the findings from an experimental study that aimed to investigate how HAT composition influences referee decision accuracy, speed, and confidence in football VAR officiating scenarios. Specifically, we aimed to compare performance in a human-AI dyad versus a human-AI-AI triad VAR team.

Methods

The study used a within-subjects experimental design with three dependent variables (decision accuracy, decision-making speed, and decision confidence) and one independent variable (HAT composition: human-AI dyad and human-AI-AI triad). In the human-AI dyad condition, participants received a decision support recommendation from one AI teammate, and in the human-

AI-AI triad condition, participants received decision-support recommendations from two AI teammates. A Wizard of Oz design was used to simulate AI team members, where participants interacted with a computer system that appeared autonomous but was pre-programmed by the research team. Participants were told that they were interacting with two prototype versions of a referee decision support system currently in the beta testing phase, labelled REFINE (Referee Evaluation & Foul Interpretation using Neural Engineering). REFINE[1] provided advice from one AI agent (human-AI dyad condition) and REFINE[2] provided advice from two AI agents (human-AI-AI triad condition). In the present study the REFINE[2] recommendations were always in agreement with one another.

52 participants took part in the study and were asked to make a foul or no foul decision for in-game excerpts taken from English Premier League games (10 in each condition). A set of candidate excerpts were reviewed prior to the study by the research team and those where there was agreement on the correct decision outcome were selected for inclusion in the study. During the study, the order of video presentation was randomised for each participant, and the assignment of team composition condition was counterbalanced randomly across participants. The effect of HAT composition on decision-making performance was examined using a repeated measures multivariate analysis of variance (MANOVA). This frequentist analysis was complemented with separate Bayesian paired-samples *t*-tests to indicate the extent of evidence (Bayes Factor [BF]) in favour of the null (no difference between HAT composition conditions, BF_{01}) or the alternative (difference between conditions, BF_{10}).

Results

Findings from the MANOVA indicated a significant effect of HAT composition, $p = .025$. Follow-up univariate analysis indicated that only decision-making speed significantly differed between the human-AI dyad ($M = 26.08$, $SE = 1.11$) and human-AI-AI triad ($M = 31.19$, $SE = 1.89$) conditions, $p = .002$. The Bayesian analysis indicated strong evidence for this finding ($BF_{10} = 16.06$). Decision accuracy and decision confidence did not differ between the two conditions, $ps \geq .46$.

Table 1: Summary of Results

	Dyad ***M* (*SE*)**	**Triad** ***M* (*SE*)**	***F*(1,51)**	***p***	**BF**
Decision accuracy (prop correct)	0.72 (0.02)	0.74 (0.02)	0.46	.499	$BF_{01} = 5.31$ (moderate evidence)
Decision-making speed (seconds)	26.08 (1.11)	31.19 (1.89)	10.76	.002	$BF_{10} = 16.06$ (strong evidence)
Decision confidence	4.06 (0.09)	4.02 (0.07)	0.54	.464	$BF_{01} = 5.12$ (moderate evidence)

Discussion

The findings suggest that the presence of multiple AI team members in football VAR teams may slow down the referee decision-making process. This aligns with previous research in other areas that has suggested that the presence of a second AI increases participant decision-making time (Lu et al., 2024). A potential explanation for our findings is that participants took additional time as they considered each AI recommendation independently; however, further investigation is recommended.

Given the need for rapid decision making in sports officiating, the findings suggest that adding multiple AI team members to sports officiating teams may be problematic. The findings also contribute to the knowledge base around optimal HATs design generally and suggest careful

consideration should be given to HAT composition where multiple AI team members are possible, particularly in safety-critical industries where timely decision making is important (e.g., healthcare, aviation, defence). Further work exploring the impact of multiple AI team members on teamwork and team performance in different contexts is therefore encouraged.

References

Lu, Z., Wang, D., & Yin, M. (2024). Does more advice help? The effects of second opinions in AI-assisted decision making. Proceedings of the ACM on Human-Computer Interaction, 8, 1–31. https://doi.org/10.1145/3653708

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors, 64*(5), 904–948. https://doi.org/10.1177/0018720820960865

Rietz, F., Sutherland, A., Bensch, S., Wermter, S., & Hellström, T. (2021). WoZ4U: An opensource wizard-of-oz interface for easy, efficient and robust HRI experiments. *Frontiers in Robotics and AI, 8*, 668057. https://doi.org/10.3389/frobt.2021.668057