Ex(plainable) Machina: how social-implicit XAI affects complex human-robot teaming tasks

Marco Matarese[1-2], Francesca Cocchella[3], Francesco Rea[2], Alessandra Sciutti[3]

[1] University of Genoa, DIBRIS department
[2] Italian Institute of Technology, RBCS unit
[3] Italian Institute of Technology, CONTACT unit

Introduction

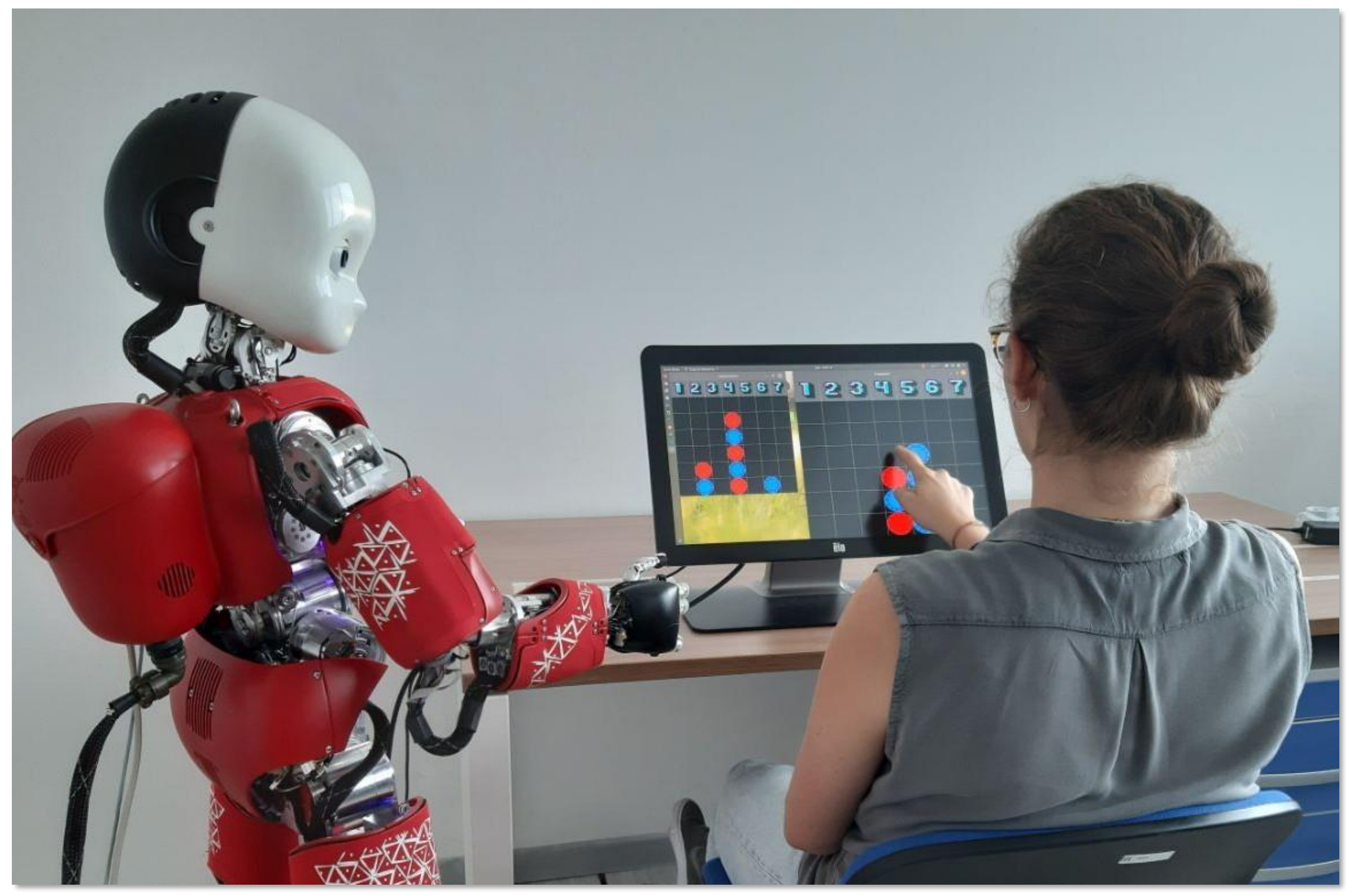
The HRI context is particularly suitable for a social and **user-centered XAI** [1] because people easily adapt their interaction habits to robots [2], and we expect the robots have long-term and personalized interactions with us [3].
However, we know little about the effects of personalized XAI in **social HRI** contexts [4]. In this work, we compare two explanation approaches in a **collaborative HRI decision-making task**: we called them *classical* (CF) and *shared experience*-based (SE).

Research question

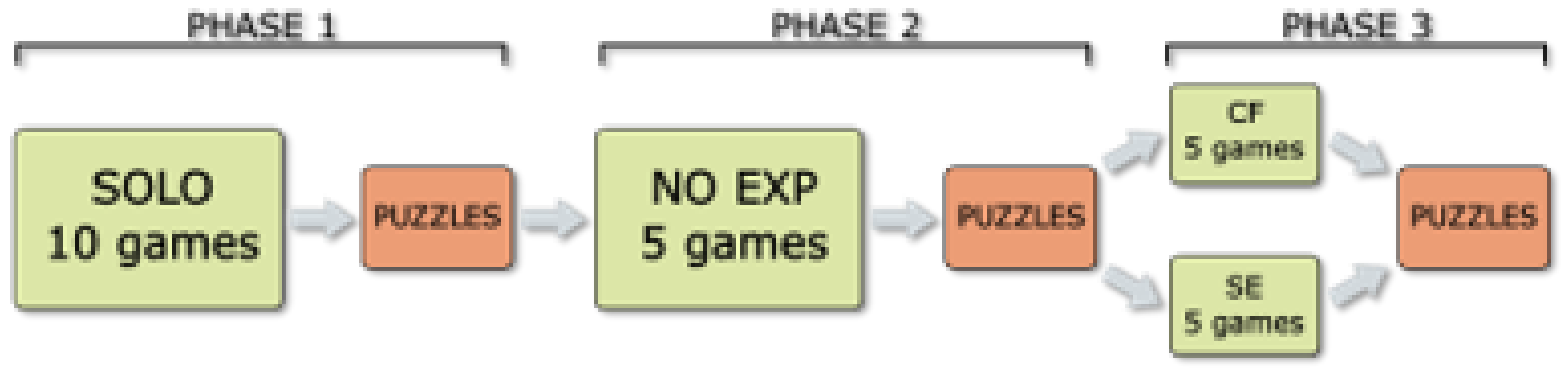
Are explanations based on shared experience more effective than classical ones during human-robot collaborative decision-making tasks?

Methodology

iCub and the participants played the **Connect 4** game against the COM. We had **three phases** which corresponded to the experimental conditions.



1. **SOLO**: participants played alone.
2. **NO EXP**: iCub participated, but it produced only suggestions.
3. **EXP**: iCub participated and it produced also explanations; with half of the participants, it used CF explanations, for the other half the SE ones.

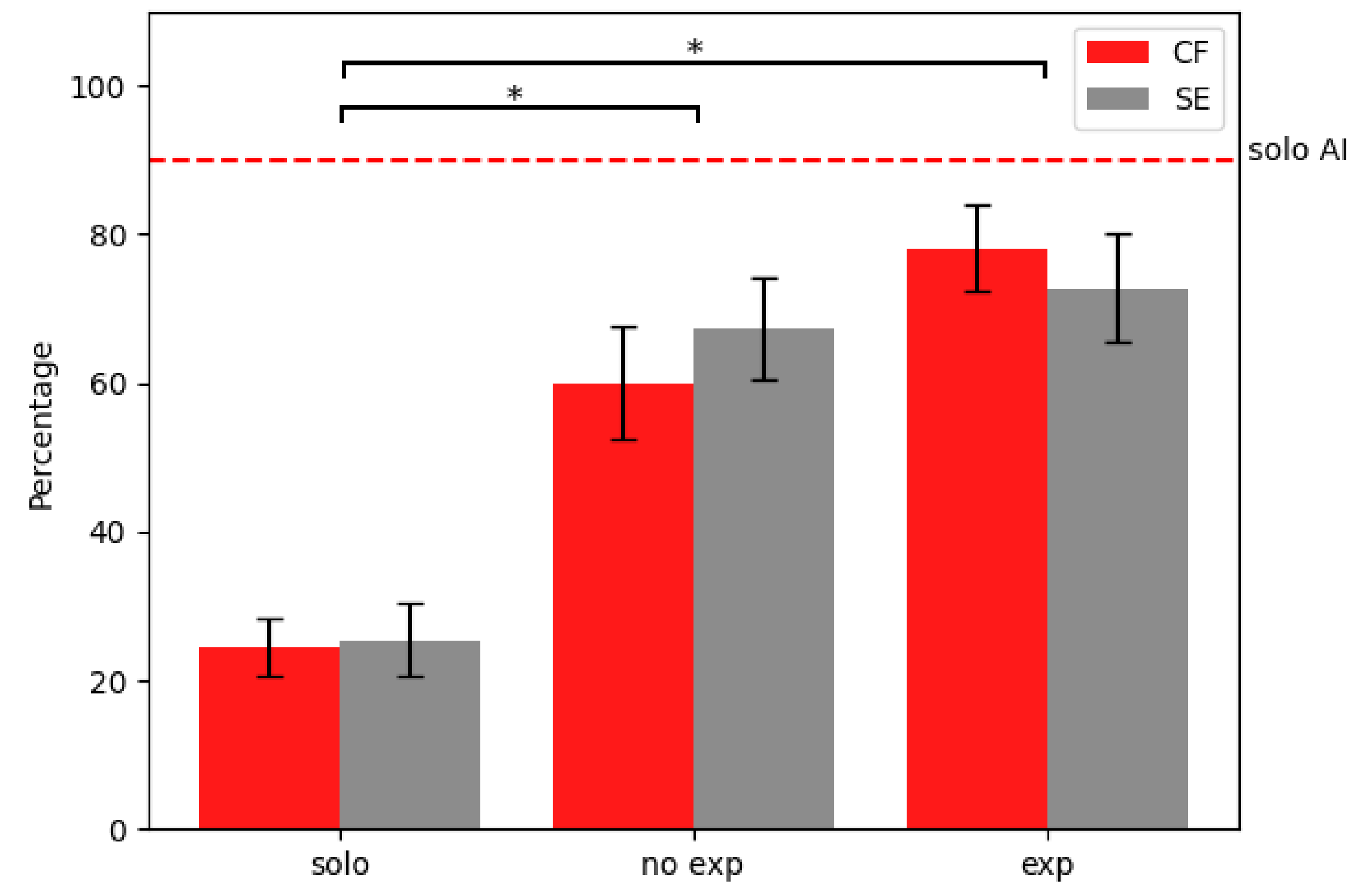


We compared participants' response to types of counterfactual explanations.

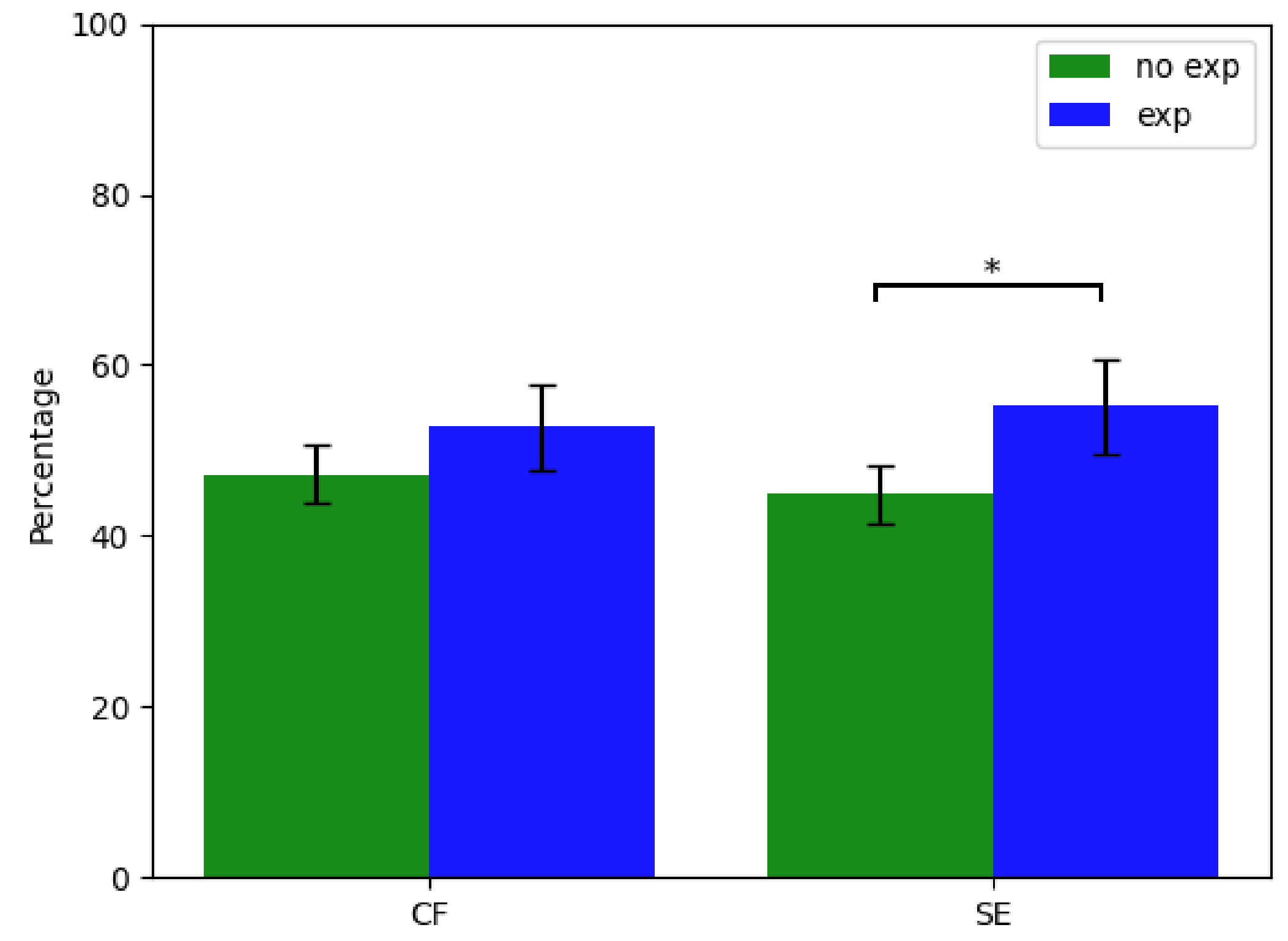
- **CF explanations**: more precise, but hardly previously encountered by participants.
- **SE explanations**: less precise, but taken from the previous games.

Results

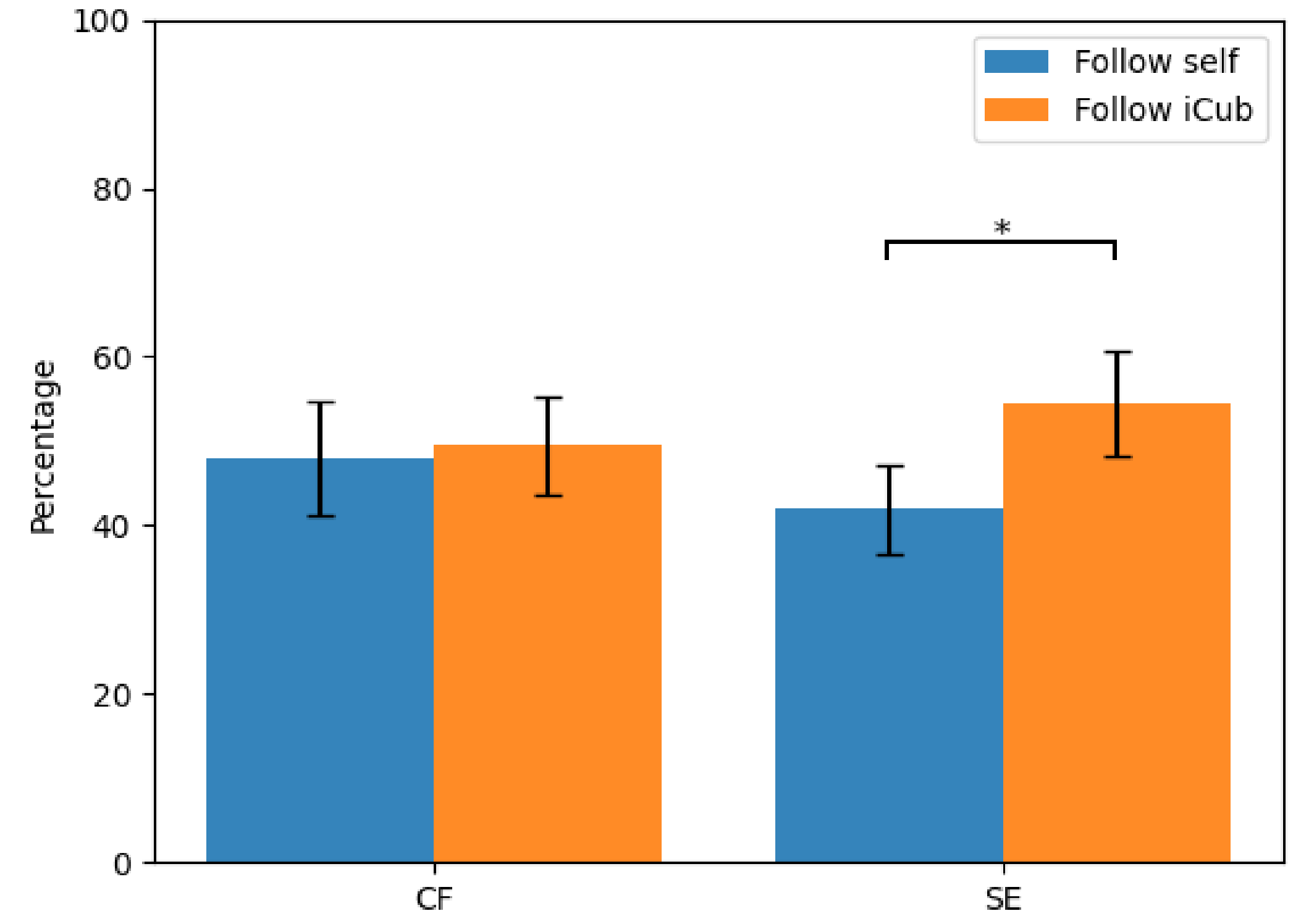
Performance against the COM



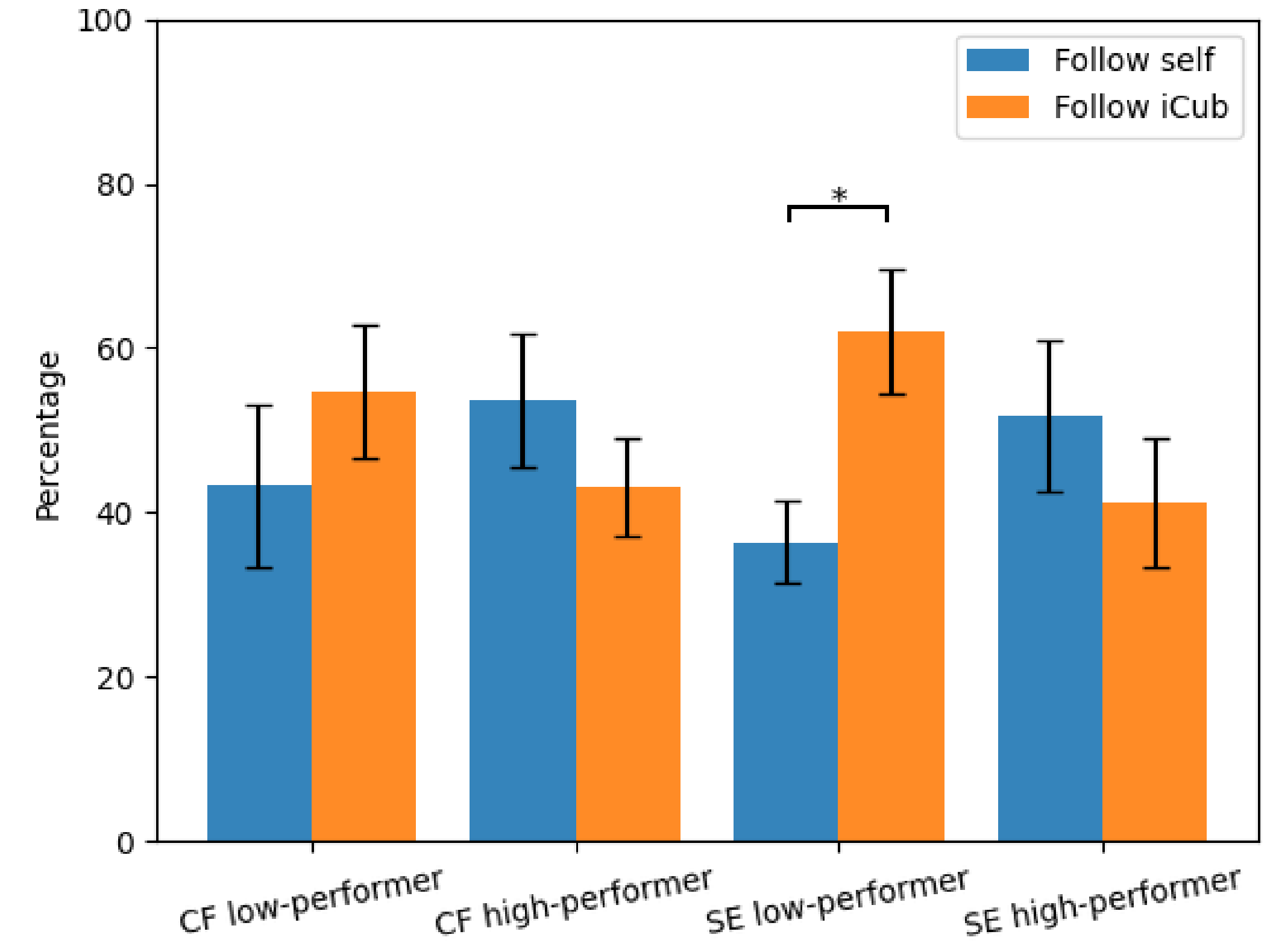
Participants learning



Robot's persuasiveness w.r.t. the explanations type



Robot's persuasiveness w.r.t. participants' performance



Conclusion

- SE explanations led to higher persuasiveness than CF ones.
- The two explanation strategies maintained comparable team performance.
- Low-performer participants followed the robot more than high-performed ones: this highlights the potential issues for letting non-expert users interact with expert robots.

References

[1] T. Miller. *Explanation in artificial intelligence: Insights from the social sciences*. Artificial Intelligence, 2019.
[2] M. Ahmad, O. Mubin, and J. Orlando. *A systematic review of adaptivity in human-robot interaction*. Multimodal Technologies and Interaction, 1(3), 2017.
[3] M. Matarese, F. Rea, and A. Sciutti. A user-centred framework for explainable artificial intelligence in human-robot interaction. ArXiv preprint arXiv:2109.12912, 2021.
[4] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling. *Explainable agents and robots: Results from a systematic literature review*. AAMAS '19, 2019.

marco.matarese@iit.it