

# Cooperation in hybrid human-AI populations: Can artificial agents reshape collective behaviour?

*Laboratory name:* Econophysics Lab

*Internship location:* Econophysics Lab, CFM, 23 rue de l'Université, Paris

*Supervision:* Michael Benzaquen

*Email:* michael.benzaquen@cfm.com

## Project at a Glance

As artificial agents become participants in social and economic interactions, even a small fraction of them may alter collective behaviour. This internship will investigate how cooperation emerges in populations containing both humans and artificial agents. Using empirical data from social-dilemma experiments together with simple statistical-physics and evolutionary-game models, we will ask when artificial agents stabilise cooperation, when they disrupt it, and which mechanisms control the transition between these regimes.



Figure 1: Illustration of a hybrid population of humans and artificial agents interacting through a social network. Image generated by ChatGPT.

## Project Description

Cooperation is a collective phenomenon: individually advantageous behaviour can produce poor outcomes for the group, while reciprocity, imitation, social structure and information sharing can stabilise very different collective regimes. In our previous work, we showed that restricting imitation to local stabilise can qualitatively reshape the dynamics of cooperation and stabilise cooperative states [1]. We also studied how individual learning, social learning and cultural diversity can coexist in stable populations, and how diversity itself can improve collective information processing [2].

The increasing presence of autonomous artificial agents adds a new ingredient to these problems. Rather than studying AI agents in isolation, the interesting question is how their presence changes the dynamics of an otherwise human population. Recent experimental and theoretical work shows that artificial agents can alter cooperation through delegation, reciprocity and strategic interaction, both in dyadic games and in public-goods settings [3, 4, 5, 6].

The project will begin from empirical human behaviour rather than from arbitrary behavioural rules. Existing public datasets from repeated social-dilemma experiments can be used to estimate how cooperation depends on previous actions, payoffs, group behaviour and individual heterogeneity. These empirical response rules will then serve as the baseline for a minimal model of a mixed population in which a fraction of individuals is replaced by artificial agents following controlled strategies. The student will investigate how the density, strategy and network position of artificial agents affect macroscopic cooperation. Of particular interest will be whether small fractions of agents can trigger collective transitions, whether local placement matters more than global density,

and whether simple strategies such as reciprocity, imitation or conditional cooperation have qualitatively different effects. Let us stress that the aim is not to have large language models repeatedly play a Prisoner’s Dilemma and then label them “cooperative” or “selfish”. Nor is it to anthropomorphise synthetic agents. Large language models could eventually be used as one implementation of behavioural policies, but only after the mechanisms are understood with simpler and interpretable models.

The broader objective is to understand hybrid human-AI populations as a new class of collective system. If artificial agents become increasingly common participants in negotiation, recommendation, coordination or shared-resource problems, even weak local interactions may alter the collective state of the system. The internship will seek to identify the minimal mechanisms through which such changes occur.

## References

- [1] P. Bousseyroux, G. Zérah, and M. Benzaquen, “Group-level imitation may stabilize cooperation,” *Journal of Statistical Mechanics: Theory and Experiment*, 123407 (2025).
- [2] B. de Courson, L. Fitouchi, J.-P. Bouchaud, and M. Benzaquen, “Cultural diversity and wisdom of crowds are mutually beneficial and evolutionarily stable,” *Scientific Reports* **11**, 16566 (2021).
- [3] J. W. Crandall, M. Oudah, Tennom, F. Ishowo-Oloko, S. Abdallah, J.-F. Bonnefon, M. Cebrian, A. Shariff, M. A. Goodrich, and I. Rahwan, “Cooperating with machines,” *Nature Communications* **9**, 233 (2018).
- [4] E. Fernández Domingos, I. Terrucha, R. Suchon, J. Grujić, *et al.*, “Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma,” *Scientific Reports* **12**, 8492 (2022).
- [5] A. Hintze and C. Adami, “Promoting cooperation in the public goods game using artificial intelligent agents,” *npj Complexity* **3**, 3 (2026).
- [6] I. Barreda-Tarrazona, A. Jaramillo-Gutiérrez, M. Pavan, and G. Sabater-Grande, “Exploiting the machine: Human cooperation with artificial agents in prisoner’s dilemma,” *Journal of Economic Psychology* **116**, 102925 (2026).