Enhancing Human Cooperation through AI-Assisted Decision-Making in Competitive Groups

Name(s) of PhD Advisor(s): Jean Claude Dreher
Host Laboratory: Institut des Sciences Cognitives, Lyon, CNRS

Abstract
Modern societies are increasingly characterized by persistent competition, which may contribute to social fragmentation and a decline in interpersonal trust. In this context, the concept of coopetition, which is cooperation among competing individuals or groups, has been shown to generate greater collective and societal benefits than purely competitive strategies. This project investigates how AI-based decision-support systems can foster cooperation in situations involving conflicting interests. We propose a series of large-scale, real-time online experiments comparing human-only groups with human–AI collaborative groups. Participants will engage in repeated social dilemma tasks involving both within-group and between-group interactions. We hypothesize that: (a) in the absence of AI support, participants will exhibit higher cooperation rates and more tolerant strategies within groups than between groups, where competitive and exploitative behaviors are expected to be more prevalent; (b) AI-assisted decision support will increase cooperation rates in both within-group and between-group contexts, with stronger effects emerging under structured, repeated interactions. To test these hypotheses, we will implement two levels of AI assistance designed as decision-support systems rather than autonomous agents. These systems will provide context-sensitive recommendations aimed at promoting cooperative strategies while preserving human autonomy and decision-making control. The first level focuses on within-group coordination, while the second targets intergroup interactions, enabling the study of how AI can facilitate cooperation across social boundaries. Together, these experiments aim to demonstrate how human–AI collaboration can promote cooperation in competitive environments and under conditions of conflicting interests. By examining behavioral adaptation over time, this project will provide insights into the mechanisms through which AI can support more cooperative and socially beneficial decision-making.

Short Description of the Hosting Research Group / Lab
The Dreher Lab is an interdisciplinary research group at the Institut des Sciences Cognitives (CNRS) dedicated to advancing knowledge in the neural bases of decision making, computational social neuroscience and human–AI collaboration (https://dreherteam.wixsite.com/neuroeconomics). The lab integrates expertise across several complementary domains: Multi-Agent Systems and Reinforcement Learning: Developing and analyzing AI agents that interact with humans or other agents in complex social environments, with a focus on ethically aligned and cognitively informed AI systems. Large-Scale Human Experiments: Conducting online and laboratory-based studies to investigate behavior in social dilemmas, collective decision-making, and cooperative tasks, supported by established experimental infrastructures. Human–Computer Interaction and Ethics: Exploring trust, transparency, interpretability, and the long-term cognitive and behavioral impact of AI-assisted decision-making.

Why This Lab is a Suitable Environment
Methodological Support: The lab has extensive experience in designing multi-agent environments and conducting large-scale behavioral experiments, providing a strong technical and experimental foundation for this project.
Interdisciplinary Expertise: With backgrounds spanning computational social science, cognitive psychology, social neuroscience, and machine learning, the lab offers an ideal environment for investigating both the technical and human dimensions of AI-assisted collaboration.
Active Research Community: The lab is embedded in a network of international collaborations focused on the societal impact of AI, offering opportunities for interdisciplinary exchange, workshops, and mentorship. JC Dreher is the PI of the PEPR COMCOMBBrr 'Computational modeling of collaboration and mentalization for benevolent bots in social networks' (https://www.pepr-ensemble.fr/en/projects/#COMCOMBR). This Phd proposal is highly complementary to this COMCOMBBrr project and the Phd student will closely work with the post-doc recruited in Dreher's lab on the COMCOMBBr project.

Description of the PhD proposal

Context (and Scenarios)

Cooperation in society not only creates reciprocity norms within communities but also establishes safety nets during crises such as natural disasters or economic downturns (Ramadan et al., 2026; Luengo-Oroz et al., 2020). Although cooperation increases collective welfare, a recent United Nations report indicates that societal willingness to cooperate is declining, which has led to increased distrust and social fragmentation (United Nations, 2025). In a fragmented world, where people are divided into categories such as ethnicity, nationality and compete against one another, whichcooperation becomes more and more challenging.

Problem and Objectives

Recent studies have shown that not only cooperation enhances welfare within partnerships, but that this can also occur between competitors (Brandenburger & Nalebuff, 2011; Jooss et al., 2023; Yao et al., 2023). This phenomenon, often referred to as "co-opetition," highlights how competitors can generate greater collective benefits when they move beyond purely self-interested behavior and collaborate for broader societal gain. Conversely, when individuals or groups act solely in their own interest and engage in constant competition, conflicts arise that reduce overall societal welfare. Artificial intelligence (AI) models have been shown to increase cooperation among individuals, as well as within human–AI hybrid systems (Ateeq et al., 2024; Shao et al., 2026; Crandall et al., 2018). Furthermore, recent studies show that AI-assisted interactions can enhance prosocial behaviors in real world settings(Sharma et al., 2023), and influence human decision-making process (Thakkar et al., 2026). However, little is known about **whether AI can promote cooperation between competing groups** and thereby enhance overall societal benefit. This project aims to investigate how AI-generated suggestions to humans can increase collective welfare, both experimentally and using computational models.

A key challenge lies in human trust and motivation: individuals may prioritize personal interests or distrust AI recommendations, thereby limiting the effectiveness of AI in promoting cooperation. Even when AI provides optimal suggestions based on collective welfare, humans may disregard such advice due to perceived conflicts with their own interests. This project will therefore examine how AI-to-human suggestions can foster cooperation in competitive environments, comparing their effectiveness in enhancing welfare both within groups and between conflicting groups.

Brief Overview of the State of the Art

Recent large meta-analyses have shown that, on average, human–AI combinations perform worse than the best-performing human or AI alone (Vaccaro et al., 2024; Burton et al., 2024). Furthermore, the effectiveness of human–AI collaboration depends on the type of task: decision-making tasks are often associated with performance losses, whereas creative tasks tend to show performance gains compared to humans or AI alone.

In this PhD project, we aim to examine how AI-to-human suggestions can improve not only within-group outcomes (i.e., coordination between individual and collective interests) but also between-group outcomes (i.e., coordination between competing groups). We plan to develop computational models of collaboration mediated by intelligent agents and to conduct controlled experiments to measure the effectiveness of AI assistance in resolving "within group" personal competition but also to resolve "between groups" conflict, so as to channel competitive efforts towards an overall general good.

We will evaluate the extent to which humans adhere to AI recommendations and design interventions to mitigate mistrust toward AI systems. By integrating insights from game theory, group decision-making, computational social science/economics, and cognitive psychology, this research aims to investigate how AI systems can enhance human cooperation and improve overall societal welfare.

Research Questions

The central question driving this PhD project is how AI-to-human suggestions can increase societal welfare by encouraging cooperation under competitive conditions. We seek to answer 3 specific questions regarding AI-human assistance to promote cooperation when humans are making a group decision. **(1) Which AI system designs might effectively improve overall societal welfare?** Since competition often reduces potential collective benefits, the first question focuses on algorithm design.

The goal is to develop AI systems that promote cooperation while individuals face decisions involving conflict between personal and collective interests. Based on interaction history, AI can infer individual strategies and provide tailored recommendations. For example, AI could act as a group-level (within-group) or global (between-group) advisor, offering suggestions that emphasize collective benefits. Identifying which design most effectively maximizes societal welfare will provide practical guidelines for deploying AI systems to support cooperation in fragmented real-world contexts. **(2) Can AI suggestions improve cooperation and overall societal benefit in human-human group decision-making contexts?** At first glance, AI acting as a group-oriented advisor may enhance welfare when individuals face conflicts between self-interest and collective benefit. However, human tendencies toward self-preservation and potential mistrust of AI may weaken this effect—even when AI provides accurate recommendations. Understanding how the degree of human adherence to AI advice influences cooperation dynamics and group outcomes is therefore central to this research. **(3) How can we reduce mistrust and improve human–AI interaction?** To optimize human–AI interactions, this project will develop or adapt indicators that measure the degree of human adherence to AI recommendations. It will also explore ways to highlight the potential benefits individuals overlook when ignoring AI advice. By designing consistent and effective interventions, this research aims to identify strategies that strengthen trust in AI and enhance cooperative behavior over repeated interactions.

Theoretical Foundations

Our main hypothesis is that individuals rely on active inference, encompassing action, perception, and learning processes grounded in Bayesian principles, to navigate social dilemmas. Specifically, individuals continuously observe their counterparts' behavior, update their beliefs over latent strategy types, and select actions based on the expected utility of inferred future outcomes. Decisions to cooperate are further modulated by social affiliation, with in-group members typically afforded greater trust and tolerance. Our lab has developed Bayesian models of group decision-making in social dilemmas (Park et al., 2019; Khalvati et al., 2019; Philippe et al., 2024), which we will extend in two directions. First, we will fit these models to participants' choices to recover individual-level latent parameters governing belief updating, social preferences, and risk attitudes. Second, and critically for this project, we will use the same generative framework to define the AI decision-support agent, so that the AI's inferences about counterparts and the participant's own inferences are formally commensurable. This shared formalism allows us to quantify, on a per-trial basis, the divergence between the participant's posterior beliefs and the AI's, and to test whether AI recommendations exert influence by correcting biased posteriors, by expanding the strategic horizon (longer planning depth than humans typically deploy), or by re-weighting the social objective (shifting the implicit utility function toward collective welfare). Distinguishing among these mechanisms is a core theoretical contribution of the project.

Approach and methods

Phase 1: Large-scale behavioral experiment of the iterated prisoner's dilemma game (PDG) with group-based design: The objective of this experiment is to examine how participants interact with multiple in-group and out-group counterparts in repeated PDG settings, and to evaluate to what extent the pursuit of self-interest undermines collective welfare. We will recruit 400 adult participants, divided into 10 online sessions (40 participants per session). Each participant will interact with four peers, over 10 repeated PDG rounds of dyadic interaction per peer. The experiment will follow a within-subjects design with two main conditions: in-group versus out-group interactions. In each session, participants will be divided into two groups. Among the four peers with which each participant interacts, two will be in-group members and two will be out-group members. The task is based on a variant of the classic iterated PDG, widely used to study conflicts between individual and collective interests. Participants will repeatedly decide whether to cooperate or compete with each peer. Importantly, they will not be informed of the total number of rounds, to reduce end-game effects. In total, each participant will complete 40 PDG rounds. We will compare performance across conditions by measuring the cooperation rate (both players cooperate) and exploitation rate in PDG (e.g., is there a difference in cooperation/exploitation rate between in and out group members?).

In addition to the behavioral analyses, we will develop computational models of collaboration at the individual level based on active inference based on Bayesian principles. These models will capture how participants form the expectation about their counterparts' strategies, update these beliefs over repeated interactions, and make decisions under uncertainty in social dilemma contexts. By estimating latent

cognitive variables such as expectations, learning variables (eg. prediction errors), and sensitivity to social outcomes, the models will provide a formal account of individual decision-making processes during cooperation and competition. In a second step, these computational models will be extended to incorporate mediation by intelligent agents. Specifically, AI-generated recommendations will be integrated into the modeling framework as an additional source of information influencing belief updating and action selection. This approach will allow us to quantitatively assess how AI interventions alter individual learning dynamics, strategic adaptation, and ultimately group-level cooperation. Together, this modeling framework will bridge individual cognition and AI-mediated interaction, providing a computational account of collaboration in complex social environments.
We hypothesize that within-group interactions will yield higher cooperation rates and lower exploitation rates. We expect participants to adopt more tolerant strategies (e.g., continued cooperation despite prior defection) in within-group contexts. The findings are expected to demonstrate how in-group favoritism strengthens cohesion within groups while increasing fragmentation between groups. This reflects real-world dynamics in which individuals face not only conflicts between personal and collective interests, but also intergroup competition. The results will inform the second phase of the PhD project, which focuses on designing AI systems to promote cooperation both within and between groups.

Phase 2:Intervention to increase cooperation with AI-assistance in human-human group decision making: To increase cooperation within and between groups, we propose implementing AI-based interventions in which AI acts as a *decision-support agent* trained on behavioral data to provide recommendations that maximize societal welfare (Levine et al., 2022).
We propose two intervention strategies.

Strategy 1: AI Group Advisor (within-group cooperation). The goal of this intervention is to enhance within-group cooperation by introducing an AI decision-support agent that issues recommendations aligned with collective group payoff. Participants engage in the same iterated PDG used in Phase 1 but, before each decision, receive an AI-generated suggestion together with a brief rationale. Final action selection remains under participant control (human-in-the-loop).
The AI is designed as a Bayesian decision-support agent that mirrors, in computational form, the same inference process we attribute to human participants. For each counterpart a participant faces, the AI maintains and continuously updates a belief about that counterpart's likely strategy, drawn from a small set of canonical strategies known to characterize human behavior in iterated PDG settings (such as Tit-for-Tat, Generous Tit-for-Tat, Win-Stay-Lose-Shift, and consistent cooperators or defectors) (Axelrod & Hamilton, 1981). As the interaction unfolds, the AI revises these beliefs in light of each new observed action, allowing it to track how a counterpart is behaving and to anticipate their next move. These beliefs are then combined with an explicit social objective that defines what the AI is trying to optimize. In Strategy 1, the objective is the average payoff across members of the participant's in-group, so that recommendations promote choices serving the group as a whole rather than narrow self-interest. The AI looks a few interactions ahead rather than only at the immediate round, which is important because cooperation in iterated dilemmas often pays off over time even when it is locally costly. For each upcoming decision, the AI selects the action that, given its current beliefs about the counterpart, is expected to yield the best group-level outcome.
Rather than starting from uninformative assumptions, the AI is calibrated using behavioral data collected during Phase 1. The distribution of strategies observed in baseline (no-AI) play, estimated separately for in-group and out-group interactions, serves as the AI's starting expectations when it is deployed in Phase 2. This ensures that the agent's inferences are ecologically grounded in how this population actually behaves, rather than in idealized assumptions, and gives the system realistic priors from the very first round of interaction.
Each participant interacts with four counterparts (two in-group, two out-group) in interleaved rounds. The AI tracks each counterpart independently, building a separate behavioral profile for each one. This allows the AI to give recommendations that are tailored to the specific partner being faced at any given moment, and to provide rationales the participant can interpret in terms of that particular counterpart's history. The system is also designed to remain responsive to change: if a counterpart's strategy drifts over the course of the interaction, the AI gradually places more weight on recent behavior than on older observations, keeping its predictions current.
A reasonable concern is that, because the AI uses the same kind of Bayesian reasoning we attribute

to participants, its recommendations might simply mirror what a thoughtful participant would already conclude. We expect three concrete sources of added value. First, prior research consistently shows that humans plan over shorter horizons than is optimal in iterated games, so the AI's longer look-ahead alone produces meaningfully different recommendations. Second, participants' implicit objectives tend to weight personal payoff more heavily than collective welfare, whereas the AI's objective is explicitly group-oriented. Third, by comparing AI recommendations against a model fit to each participant's own choices, we can decompose the AI's contribution into three interpretable components: improved beliefs about counterparts, longer planning, and a more collectively oriented objective. This decomposition is itself a methodological contribution and directly informs RQ1 by clarifying which feature of AI assistance most effectively promotes cooperation.
For each recommendation, the AI displays a brief rationale: its best current guess about the counterpart's strategy, the action it predicts the counterpart will take next, and the expected outcome for the participant and for the group under each possible choice. This makes the AI's reasoning transparent and human-readable, supports informed acceptance or rejection of the suggestion, and provides the basis for the trust-calibration analyses developed under RQ3.

Strategy 2: AI World Advisor (between-group cooperation). The goal of this intervention is to enhance cooperation between groups in the context of probable intergroup conflict. The experimental setup will be similar to Strategy 1, except that the AI will provide recommendations aimed at maximizing global welfare across both groups, rather than within-group benefit. We hypothesize that AI-assisted interventions will increase cooperation rates in both within-group and between-group contexts compared to conditions without AI support. These findings will highlight the potential of AI to mediate conflicts at multiple levels thereby fostering cooperation and contributing to a more cohesive society. Potential risks include over-reliance on AI recommendations, reduced individual autonomy, and algorithm aversion. This project will explicitly measure perceived autonomy, trust, and reliance to ensure that AI systems support rather than override human decision-making.

Evaluation of contributions. By systematically observing how different AI assistance strategies affect within or between groups' decision-making when facing conflicting interests, the PhD project aims to make several key contributions. It will furnish quantitative evidence on whether conflicts between humans can be improved by AI that is generating suggestion based on societal benefit.The key methodological advances of this PhD proposal include (1) an AI system that tunes into human behavior and behaves as a mediator to mitigate conflicts. This will allow us to test computational models of collaboration mediated by intelligent agents with controlled experiments made in groups of huma,s; (2) the use of machine learning tools for behavioral pattern analysis, which will be used in both model simulation and data analysis. Finally, our pre-trained AI system should increase cooperation in group decision-making. Theoretically, the study will employ a Bayesian model to account for how decisions about cooperation and competition develop over time under AI influence, thereby shedding light on how human-AI suggestion can improve essential societal functions in the face of a fragmented society.

Nature of digital collaboration
The form of collaboration examined in this PhD project centers on group coordination, cooperative problem-solving, and conflict resolution, with AI acting as a decision-support agent that guides group judgments toward cooperation. The experiments will be conducted online in real time using a platform that enables synchronous interaction among participants. Within this environment, AI will provide recommendations regarding whether to cooperate or compete, thereby potentially reducing competitive behavior and promoting collective benefit. Importantly, participants retain full autonomy: they may accept, reject, or override AI suggestions, resulting in a human-in-the-loop framework.
Each session, lasting approximately 90 mn, will involve medium-sized groups (around 40 participants) engaging in repeated interactions. This design allows us to examine the longitudinal effects of AI suggestions on cooperative behavior. The platform is fully remote, enabling participation from geographically diverse populations via a browser-based interface. It is also designed to be scalable, allowing future studies to include several hundred participants simultaneously. In addition, the system supports dynamic updates and integrated real-time data collection. The key features of the system include transparency mechanisms, which explain AI recommendations, and feedback loops, enabling participants to reflect on both individual and collective outcomes.

Together, these elements approximate real-world conditions of modern distributed teams and digital communities. By embedding AI systems that provide recommendations at different levels (e.g., group-level vs. global-level), this project directly addresses the core concern of PC3: whether repeated human collaboration can be improved through AI support, and whether AI-mediated interaction can enhance the quality and efficiency of collective decision-making in society.

Contribution to digital collaboration: Expected result and impact

This PhD project aims to advance understanding of the interaction between AI-assisted collaboration and human group decision-making, contributing to the broader field of hybrid intelligence and digital teamwork. At the experimental level, we expect the following outcomes: (1) Increased cooperation: AI suggestions will lead to higher cooperation rates, even in competitive environments, thereby improving overall societal welfare. (2) Sustained cooperative behavior: Through repeated interaction, performance feedback, and explanatory mechanisms, participants are expected to develop a longer-term willingness to cooperate. (3) Improved collective decision-making: Group decisions are expected to shift from competitive dynamics toward more coordinated and mutually beneficial outcomes.

At the theoretical level, this project integrates computational social science approaches to group decision-making with cognitive psychology perspectives on decision-making under conflict. In particular, it will develop Bayesian computational models that capture how individuals form beliefs about others, update these beliefs through repeated interactions, and make cooperation decisions under uncertainty. These models will then be extended to incorporate AI-mediated recommendations as an additional source of information influencing belief updating and action selection. By embedding such cognitively grounded mechanisms into cooperative multi-agent frameworks, the project aims to bridge the gap between AI optimization and human cooperative behavior.

Methodologically, the project will contribute by (1) Developing metrics to capture long-term cooperation dynamics. (2) Providing tools to measure human adherence to AI recommendations. (3) Potentially delivering an open-source experimental framework for studying AI-mediated cooperation.

From an applied perspective, the findings will generate practical design guidelines for AI systems that support group decision-making. These insights can inform policymakers and organizations seeking to balance individual and collective interests in AI-assisted environments. Aligned with PC3, this project demonstrates how AI interventions can enhance group decision-making and increase societal welfare.

Positioning in the eNSEMBLE program

Within the eNSEMBLE program, this project directly contributes to PC3, which focuses on how AI systems can promote cooperation while addressing conflicts between individual and collective interests. This project directly contributes to the development of computational models of collaboration mediated by intelligent agents by designing and empirically testing AI-based decision-support systems in group settings which can further be used in field studies. The proposed AI system functions based on a computational model that optimizes how individuals coordinate, adapt, and make decisions under conditions of cooperation and conflict. By integrating behavioral data from repeated social dilemma interactions, the model captures dynamic of decision processes, strategic adaptation, and responses to AI-generated recommendations. This allows for the simulation and prediction of collaborative behavior in multi-agent environments, where both human participants and AI systems interact. Furthermore, the project combines these computational approaches with large-scale online experiments, enabling the validation of model predictions against observed human behavior. In doing so, it bridges theoretical modeling and empirical research, providing a robust framework for understanding and improving AI-mediated collaboration in complex social systems.

The program's emphasis on hybrid socio-technical systems is directly reflected in this project's integration of Bayesian inference models and AI-assisted decision-making. The research contributes to eNSEMBLE's goal of designing AI systems that strategically promote cooperation.

Methodologically, the project aligns with the program's focus on large-scale online experimentation, using repeated, digitally mediated interactions with real-time adaptation. This provides both practical relevance and an opportunity to address ethical considerations related to AI transparency and human autonomy.

Finally, the project embodies eNSEMBLE's interdisciplinary approach by combining computational social science, cognitive psychology, and AI design. Its findings will also contribute to broader program

themes such as trust, transparency, and governance, helping to inform how AI can be effectively integrated into collaborative, team-based, and community-driven environments.

References

Ateeq A, Milhem M, Alzoraiki M, Dawwas MIF, Ali SA and Yahia Al Astal A (2024) The impact of AI as a mediator on effective communication: enhancing interaction in the digital age. *Front. Hum. Dyn.* 6:1467384.

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. *Science*, 211, 1390-1396.

Brandenburger, A. M., & Nalebuff, B. J. (2011). Co-opetition: A revolution mindset that combines competition and cooperation: The game theory strategy that's changing the game. *Currency/Doubleday.*

Burton, J.W., Lopez-Lopez, E., Hechtlinger, S. et al. How large language models can reshape collective intelligence*. Nat Hum Behav* 8, 1643–1655 (2024).

Crandall, J.W., Oudah, M., Tennom et al. Cooperating with machines*. Nat Commun* 9, 233 (2018).

Jooss, S. (2023). Beyond competing for talent: an integrative framework for coopetition in talent management in SMEs*. International Journal of Contemporary Hospitality Management*, 35(8), 2691–2707.

Levine, D. K., Modica, S., & Rustichini, A. (2026). Cooperating through leaders. *The Economic Journal*, ueag027.

Luengo-Oroz, M., Hoffmann Pham, K., Bullock, J. et al. Artificial intelligence cooperation to support the global response to COVID-19. *Nat Mach Intell* 2, 295–297 (2020).

Ramadan, E., Abdalla, S., Al Mamari, W., & Al Hosani, N. (2026). Building resilient communities: The impact of social capital on disaster recovery in Oman. *Community Development*, 57(2), 181–201.

Shao, E., Wang, Y., Qian, Y. et al. SciSciGPT: advancing human–AI collaboration in the science of science*. Nat Comput Sci* 6, 301–315 (2026).

Sharma, A., Lin, I.W., Miner, A.S. et al. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat Mach Intell* 5, 46–57 (2023).

Thakkar, N., Yuksekgonul, M., Silberg, J. et al. A large-scale randomized study of large language model feedback in peer review*. Nat Mach Intell* 8, 326–336 (2026).

United Nations. (2025). World Social Report 2025: A new policy consensus to accelerate social progress. Department of Economic and Social Affairs.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis*. Nature Human Behaviour*, 1-11.

Yao, G., Zhao, H., Hu, Y., & Zheng, X. (2023). Exploring knowledge sharing and hiding on employees' creative behaviors: A coopetition perspective. *Journal of Innovation & Knowledge*, 8(4), 100447.