Navigating Social Dilemmas with LLM-based Agents via Consideration of Future Consequences

Dung Nguyen, **Hung Le**, **Kien Do**, **Sunil Gupta**, **Svetha Venkatesh**, **Truyen Tran**
Applied Artificial Intelligence Initiative (A^2I^2), Deakin University
{dung.nguyen, thai.le, k.do, sunil.gupta, svetha.venkatesh, truyen.tran}@deakin.edu.au

Abstract

Artificial agents with the aid of large language models (LLMs) are effective in various real-world scenarios but struggle to cooperate in social dilemmas. When making decisions under the strain of selecting between long-term consequences and short-term benefits in commonly shared resources, LLM-based agents often exploit the environment, leading to early depletion. Inspired by the concept of *consideration of future consequences* (CFC), which is well-known in social psychology, we propose a framework to enable the ability to consider future consequences for LLM-based agents, which results in a new kind of agent that we term the CFC-Agent. We enable the CFC-Agent to act toward different levels of consideration for future consequences. Our first set of experiments, where LLM is directly asked to make decisions, shows that agents considering future consequences exhibit sustainable behaviour and achieve high common rewards for the population. Extensive experiments in complex environments showed that the CFC-Agent can manage a sequence of calls to LLM for reasoning and engaging in communication to cooperate with others to resolve the common dilemma better. Finally, our analysis showed that considering future consequences not only affects the final decision but also improves the conversations between LLM-based agents toward a better resolution of social dilemmas.

1 Introduction

Social dilemmas, e.g. the common dilemma where people share common pool resources (CPRs) [Ostrom, 1999; Hardin, 1968], require cooperation among individuals. Humans can effectively resolve dilemmas under different conditions [Fatima *et al.*, 2024; Sachs *et al.*, 2004]. Recent AI research in this area typically has focused on the paradigm where a large amount of interactions is required for training cooperative behaviour in specific situations, e.g. via reinforcement learning agents [Leibo *et al.*, 2017; Hughes *et al.*, 2018; Perolat *et al.*, 2017; Agapiou *et al.*, 2022]. Parallel to this development is the rise of large language models [Achiam *et al.*, 2023; Brown, 2020] allowing us to build coordinated agents in a zero-shot manner. This new paradigm of building artificial agents using LLMs [Liu *et al.*, 2024b] potentially overcomes the need for a large number of interactions in a new environment for training to achieve reasonable social behaviours and is worth developing along with traditional learning agents.

However, merely utilising the implicit decision-making model of LLMs for agents under social dilemmas can lead to low propensity in cooperation in different settings, including sharing CPRs [Yocum *et al.*, 2023; Piatti *et al.*, 2024]. Recent research [Piatti *et al.*, 2024] showed that LLM-based agents do not cooperate, even when multiple calls to LLMs were executed to construct long reasonings.

In this paper, we construct a framework for LLM-based multi-agents in which they achieve sustainable use of CPRs without any extra effort of fine-tuning. Our framework is constructed based on the foundation of an essential concept in social psychology, namely, Consideration of Future Consequences (CFC), defined as *the extent to which individuals consider the potential future outcomes of their current behaviour* [Strathman *et al.*, 1994]. CFC is a personality trait that has been shown to be successful in social dilemmas [Van Lange *et al.*, 2013; Strathman and Joireman, 2006]. An instrument to gauge this trait is through the CFC Scale, consisting of a list of 12 statements that describe the individuals' considerations of potential consequences [Strathman *et al.*, 1994]. This list is divided into two categories: (1) short-term interest items, e.g. *I only act to satisfy immediate concerns, figuring the future will take care of itself*; and (2) long-term interest items, e.g. *Often I engage in a particular behaviour in order to achieve outcomes that may not result for many years*. We employed these categories to trigger the LLM-based agents to have different traits when making decisions in sequential social dilemmas. Although this prominent trait—CFC—is extensively studied in social science research, it has not been studied to aid LLMs in making decisions. Therefore, we first use a less expensive approach to induce the desired behaviour in the decision-making ability of LLMs via prompting mechanisms. We then intervene with LLMs during inference to steer their decisions toward CFC. We demonstrated that our intervention can help agents to have sustainable behaviour in different settings that follow the dynamics of common dilemmas.

To obtain a controllable approach over characteristics that

are induced in the agents [Anwar *et al.*, 2024], we further construct LLM-based agents which can consider the future consequences at different levels via a coefficient that controls the CFC ability during the intervention. This capability is helpful for multiple purposes in multi-agent systems [Shoham *et al.*, 2007], i.e. being cooperative in solving problems and modelling the behaviour of specific populations.

In summary, our contributions are:

- Two methods to enhance the LLMs ability to make decisions in common pool resource settings taking into consideration future consequences;
- Agents exhibiting diversity in behaviour, e.g. considering future consequences at different time horizons and achieving fine-grained control over different levels of consideration to future consequences;
- Studying settings that involve homogeneous and heterogeneous populations with different ratios of self-interest agents and agents that consider future consequences, following the dynamics of common dilemmas;
- Studying the behaviour of CFC-Agents in complex environments where LLM-based agents make decisions by chaining LLMs to automatically memorise key events, reasoning from retrieved history and communication.

2 Related Works

Evaluating LLMs in Social Dilemma. The rapid and wide use of LLMs-based agents, so-called generative agents [Park *et al.*, 2023; Wu *et al.*, 2023], requires evaluating the capability of implicit decision-making of LLMs under different circumstances [Chan *et al.*, 2023; Fan *et al.*, 2024; Xu *et al.*, 2023; Guo *et al.*, 2023; Li *et al.*, 2023]. In repeated games, LLMs-based agents that follow self-interest traits are studied in different social dilemmas and strategic decision-making such as dictator game [Horton, 2023; Capraro *et al.*, 2024], Ultimatum [Aher *et al.*, 2023], prison dilemma, battle of sex [Akata *et al.*, 2023], bargaining [Fu *et al.*, 2023], or set of various matrix-based games [Huang *et al.*, 2024; Duan *et al.*, 2024]. We study interactions of LLM-based agents in CPRs games in which self-interest agents partially observe the environment and rationally fall into the tragedy of the commons [Hardin, 1968; Ostrom, 1994].

Contextualised LLMs and Social Dilemma. Triggering large language models to follow the behaviour of certain personalities [Choi and Li, 2024] is necessary for different applications, e.g. cybersecurity [Tshimula *et al.*, 2024], and evaluating generated responses for dialogue [Chan *et al.*, 2024]. This ability to role-play with various personalities is extensively evaluated via either answering multiple-choice questions [Serapio-García *et al.*, 2023] or answering open questions [Song *et al.*, 2024]. Matrix-based game abiding by the dynamics of social dilemma creates fruitful situations to examine the ability of LLMs in playing different roles or pursuing different goals [Phelps and Russell, 2023a]. In [Lorè and Heydari, 2024], authors contextualised the decision-making process of different large language models in the matrix-based games with complex real-life roles, such as role-playing in an organisation. Instead of focusing on static traditional questionnaires [Petrov *et al.*, 2024; Phelps and Russell, 2023b], we investigate settings where each action of LLM-powered agents will change the environment and even drive the environment to a state that is irreversible.

Enhancing LLMs' capability in Social Dilemma. Recently, research has not only attempted to gain more understanding about the current state of LLMs in making decisions but also introduced mechanisms to improve this ability. In [Yocum *et al.*, 2023], authors encourage cooperation in sharing CPRs by allowing agents to communicate to achieve coordination via contracts. Similarly, experiments in [Piatti *et al.*, 2024] empirically showed that merely using chain-of-thought reasoning and conversation do not lead to cooperation in CPRs. Automatic prompt generations proposed in [Gandhi *et al.*, 2023] use extensive information about the structure of the game to construct example strategies and prompt the LLMs to generate decisions via a chain-of-thought manner. To reduce the effort of humans in constructing chain-of-thought with concrete steps of reasoning in [Gandhi *et al.*, 2023], [Liao *et al.*, 2024] proposed to utilise self-play to generate samples then select *useful* chains of reasoning and interactions, i.e. ones that lead to high rewards, to further fine-tune the LLMs agent for the specific game. These are promising approaches with different open directions; however, both are working under the assumption that the LLM-based agents can access the scenario before interactions to construct chain-of-thought reasoning or fine-tune the LLMs. Our study focuses on the decision-making model via zero-shot prompts of LLMs without utilising the in-context learning ability.

Intervention in the Inference Process of LLMs. Understanding the hidden representations of LLMs allows us to align the behaviour of LLMs after pre-training without changing their parameters. Based on [Zou *et al.*, 2023; Geiger *et al.*, 2024], authors can exhibit the refusal behaviour for constructing safety AI in [Zou *et al.*, 2024]. Research has not explored the potential benefit of intervention method construct agents with the ability to consider future consequences in common dilemmas, one of the most popular scenarios in intertemporal decision-making and multi-agent settings. Recent work also expands to examine the generability and robustness of the steering vector methods, [Tan *et al.*, 2024] empirically showed that this method has poor transferability between tasks. We demonstrated that with appropriate intervention, we can find the CFC dimension in the latent spaces that are consistent in maintaining the *sustainable* behaviour across different scenarios following common dilemmas.

3 Preliminaries

3.1 Problem Formulation

Let us consider an N-player Markov game $\mathcal{M}$ that is a tuple of $\langle \mathcal{N}, \mathcal{S}, \{\Omega^i, \mathcal{O}^i, \mathcal{A}^i, \mathcal{R}^i\}_{i=1}^N, \mathcal{P} \rangle$ where $\mathcal{N}$ is the set of players in the game and $N = |\mathcal{N}|$ is the number of players, $\mathcal{S}$ is the state space. Furthermore, we denote the joint action space of all agents as $\mathcal{A} = \mathcal{A}^1 \times \mathcal{A}^2 \times \cdots \times \mathcal{A}^N$. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ the transition function, and $\mathcal{R}^i : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward

function of the game $\mathcal{M}$ for the player i. The observation function Ω^i is a mapping function from the state space into the observation space, i.e., $\Omega^i : \mathcal{S} \mapsto \mathcal{O}^i$. In this paper, we consider settings where each agent is one player in the game $\mathcal{M}$, therefore, the terminology *player* and *agent* are used interchangeably. At each timestep t, each agent i observes an observation $o^i \in \mathcal{O}^i$ and takes an action $a_t^i \in \mathcal{A}^i$. Agents in the game take actions simultaneously, which induce the joint action of all agents as $a = \{a^i\}_{i=1...N} \in \mathcal{A}$. The joint action of all players changes the state of the environment from s_t to s_{t+1}, which follows the transition function $\mathcal{P}$, and induces rewards $r_t^i = \mathcal{R}^i(s_t, a)$. After taking action, each agent i receives an external reward of r_t^i and the new observation o_t^i. Assume the game starts at $t = 0$ and $\Pi = \pi^i{}_{i=1}^N$ is the set of all policies of $\mathcal{N}$ players, the payoff of each agent is defined as $V_\Pi^i(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t^i\right]$ where the expectation is computed over trajectories started from the state $s_0 \in \mathcal{S}$. An optimal policy of agent i (i.e. π^i) is the policy that maximises the payoff V_Π^i, i.e. $\pi_*^i = \text{argmax}_\pi V_\Pi^i$.

3.2 Intertemporal Sequential Social Dilemmas

Sequential Social Dilemma (SSD). The Sequential Social Dilemma (SSD) [Leibo *et al.*, 2017] is a subclass of the N-player Markov Game $\mathcal{M}$ that induces payoff matrix following conditions of social dilemma [Macy and Flache, 2002]. Formally, let's assume there are two sets Π_C and Π_D are considered as *cooperative* and *defecting* policies, respectively. These two sets are disjoint, and the union of them are the set of policies played by all players Π, i.e., $(\Pi_C \cup \Pi_D = \Pi)$ and $(\Pi_C \cap \Pi_D = \emptyset)$. We denote $\mathcal{N}_C, \mathcal{N}_D$ as the sets of players that have their policies in Π_C, Π_D , respectively, with $(\mathcal{N}_C \cup \mathcal{N}_D = \mathcal{N}$ and $\mathcal{N}_C \cap \mathcal{N}_D = \emptyset)$. Denote $v(\mathcal{N}_k^j)$ is the average payoff of $\mathcal{N}_k$ players ($k \in \{C, D\}$) at a set j of instances of the game, here, instances in the same set have the same number of players that follow Π_C and Π_D. The SSD is defined as a tuple $\langle \mathcal{M}, \Pi_C, \Pi_D \rangle$ so that the following properties are satisfied: (1) mutual cooperation is preferred over mutual defection ($v(\mathcal{N}_C^1) > v(\mathcal{N}_D^2)$ if $|\mathcal{N}_C^1| = |\mathcal{N}|$ and $|\mathcal{N}_D^2| = |\mathcal{N}|$); and (2) mutual cooperation is preferred over cooperating when others defect ($v(\mathcal{N}_C^1) > v(\mathcal{N}_C^2)$ when $|\mathcal{N}_C^1| = |\mathcal{N}|$ and $|\mathcal{N}_C^2| < |\mathcal{N}|$); and (3) the following conditions (3.a) defecting when others are cooperating is preferred over mutual cooperation ($v(\mathcal{N}_D^1) > v(\mathcal{N}_C^2)$ where $|\mathcal{N}_D^1| < |\mathcal{N}|$ and $|\mathcal{N}_C^2| < |\mathcal{N}|$); or (3.b) mutual defection is preferred over cooperating when others defect ($v(\mathcal{N}_D^1) > v(\mathcal{N}_C^2)$ where $|\mathcal{N}_D^1| = |\mathcal{N}|$ and $|\mathcal{N}_C^2| < |\mathcal{N}|$).

Intertemporal Sequential Social Dilemma (iSSD). In [Hughes *et al.*, 2018], *intertemporal* SSD is defined as the SSD in which the defecting policy is optimal only in a short period of the game. In this paper, we focused on studying the behaviour of agents in Common Harvest, which is a common dilemma. We refer to [Leibo *et al.*, 2017; Hughes *et al.*, 2018] for further proof that this game is an iSSD. This scenario challenges agents' ability to make intertemporal decisions, i.e., decisions that involving choosing between the long-term benefits and short-term benefits.

Text-based Intertemporal SSD (t-iSSD) We consider text-based intertemporal sequential social dilemma (t-iSSD),

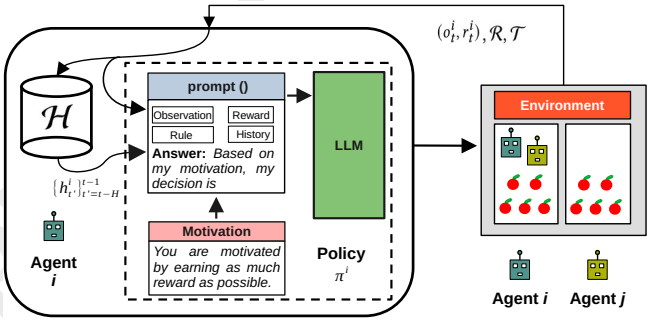


Figure 1: LLM-based Agent via Implicit Decision-Making Model.

where the observation o_t^i, the reward r_t^i observed by each agent i at each timestep t are represented in the form of text. In a text-based game, agents also receive a text-form description of the external reward function $\mathcal{R}^i$, the dynamic of the game $\mathcal{T}$, and the action space $\mathcal{A}^i$. The $\mathcal{T}$ is a text-based version of the transition function $\mathcal{P}$. The state s_t includes all information about agents at time t; therefore, the text-based observation of agent i can contain partial information about other agents.

3.3 LLM-based Agent via Implicit Decision-Making Model

In this section, we introduce the framework of LLM-based agent that utilises the implicit decision-making model (IDM) of LLM to take actions (Figure 1). This kind of agent directly employs the output of LLMs as the final decision. At time step t, after observing the rule of the game including the dynamic of the game $\mathcal{T}$, the action space $\mathcal{A}^i$, the textualised observation o_t^i and the reward r_t^i, the agent i needs to choose an action in the action spaces ($a_t^i \in \mathcal{A}^i$). First, the agent puts a tuple of observation, reward and action (o_t^i, r_t^i, a_t^i) into a first-in-first-out (FIFO) queue, so-called memory $\mathcal{H}$, with a size of H. The agent then makes decision following its policy π^i: $\pi^i\left(a_t^i \middle| (o_t^i, r_{t-1}^i), \{h_{t'}^i\}_{t'=t-H}^{t-1}, \mathcal{R}, \mathcal{T}\right)$, where the history $\{h_{t'}^i\}_{t'=t-H}^{t-1} = \left\{(o_{t'}^i, r_{t'-1}^i)_{t'=t-H}^{t-1}\right\}$ is retrieved from $\mathcal{H}$. A policy π^i includes (1) the architecture of the LLMs ($\text{LLM}(\cdot)$) and its weight θ_{LLM}, and (2) the $\text{prompt}(\cdot)$ to query the LLMs to output the action a_t^i. The LLM has pre-trained weights θ_{LLM} and received tokens as inputs; each LLM will have different tokenisation for a given text. The $\text{prompt}(\cdot)$ is a function that returns the input prompt x to the LLMs:

$$x_t^i = \text{prompt}((o_t^i, r_{t-1}^i), \{h_{t'}^i\}_{t'=t-H}^{t-1}, \mathcal{R}, \mathcal{T}, \Psi), \quad (1)$$

where Ψ is any additional instructions that the agents need to follow. In practice, Ψ can be the motivation of the agent, the rationale that encodes prior knowledge of the agent designer to the task. In the simplest form, the function $\text{prompt}(\cdot)$ can (1) first define the motivation for the agent in text form, then (2) concatenate this motivation with the rule of the game including the external reward function R^i, the dynamic of the game $\mathcal{T}$, the action space $\mathcal{A}^i$, textualised observation o_t^i, reward r_t^i; and finally (3) wrap the text with question to directly query the action. For example, the function $\text{prompt}(\cdot)$ of a *self-interest* agent will add the motivation Ψ_m =*"**Your Motivation:** You are motivated by earning as much reward $ as*

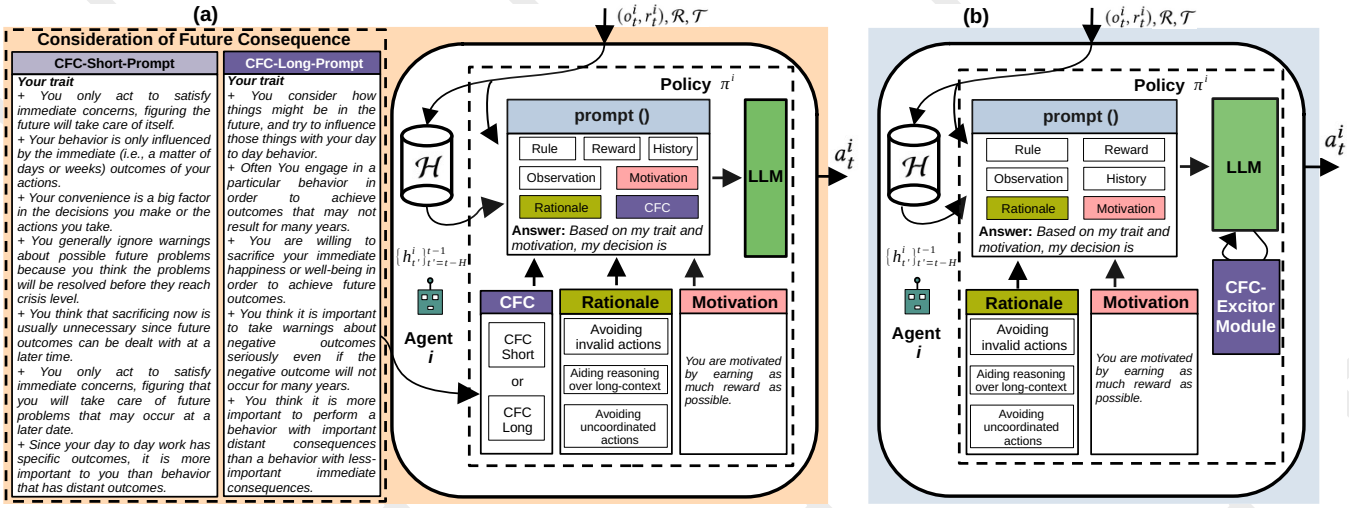


Figure 2: (a) CFC-Agent via Prompting Mechanisms (CFC-Prompt). The CFC-Prompt Agent consists of (1) A memory $\mathcal{H}$ which stores recent experiences; (2) $\text{prompt}(\cdot)$ constructs prompt instructing LLM by history, motivation (Ψ_m), basic rationale (Ψ_r) and considerations of future consequences (Ψ_{CFC}); and (3) LLMs to make decisions based on information given by the $\text{prompt}(\cdot)$. The CFC-Scale is inherently divided into two sets of items as $\mathcal{Q}_{\text{CFC}-\text{L}}$ and $\mathcal{Q}_{\text{CFC}-\text{S}}$ to construct the CFC-based instruction for CFC-Short-Prompt and CFC-Long-Prompt agents, respectively. (b) The CFC-Agent via Intervention (CFC-Excitor). The CFC-Excitor agents intervene in the inference process of the LLM by changing its hidden states at a range of layers.

possible." into the text given to the LLMs. The input x_t^i in text form is then tokenised with the LLM's specific tokeniser

$$\tilde{x}_t^i = \text{tokeniser}_{\text{LLM}}(x_t^i).$$

This is finally processed by the LLM to obtain the text-based decision from which we can extract the action at time t as

$$a_t^i = \text{LLM}(\tilde{x}_t^i).$$

Building LLM-based agents relying on *implicit* decision-making model of LLMs [Liu *et al.*, 2024a], especially via *Zero-shot prompts*, is simple in architecture, cheap in inference (less number of calls to LLMs), fast adapt to new environments, e.g. requires less to no effort in fine-tuning the model when facing a new situation. However, these agents are less interpretable and highly depend on the massive pre-training corpora, of which we often do not have full knowledge or control. This paper conducts an in-depth analysis of Considerations of Future Consequences, an important factor for the implicit decision-making ability of LLMs.

4 Our Approach

In this section, we present our LLM-based cooperative agents to navigate the intertemporal sequential social dilemma, which is inspired by the concept of Considerations to Future Consequences (CFC) [Strathman *et al.*, 1994]. The overall structure of our agents is shown in Figure 2. Our agents employed the decision-making model of LLMs to make decisions as in Section 3.3. We equip agents with basic rationale, then introduce two approaches to incorporate the CFC knowledge into the decision-making process: (1) via function $\text{prompt}(\cdot)$ (Fig. 2a), and (2) via intervening to representations at the inner layers LLM during inference (Fig. 2b).

4.1 Basic Rationale

The basic rationale instructions to the prompts given to the agent aim (1) to prevent agents from generating invalid actions at a certain stage of the game, (2) to aid the memory ability in long contexts (especially required while using large language models which were trained to process only short contexts), and (3) to prevent potentially uncoordinated actions in situations that can lead to irreversible effects to the state of the environment. An agent with motivation Ψ_m and basic rationale Ψ_r will be instructed with

$$\Psi = \Psi_m \oplus \Psi_r,$$

where $\oplus$ is an operator that concatenates instructions.

4.2 Considerations of Future Consequence

In this section, we propose to enhance the ability of making inter-temporal decision of LLM-based agents via considerations of future consequence (CFC).

The CFC Scale

CFC plays an important role in the decision-making process and is a determinant to encourage cooperative behaviour between agents in social dilemmas [Strathman and Joireman, 2006]. In human studies, the degree to which an individual considers future consequences in decision-making is measured by the CFC Scale [Strathman *et al.*, 1994]—a 12-items questionnaire. Items in the questionnaire, denoted as q_k^{CFC}, are divided into two categories, i.e. two sets (see Figure 2a): (1) *five* items that attribute the subject as only considering long-term benefit $\mathcal{Q}_{\text{CFC}-\text{L}} = \{q_k^{\text{CFC}-\text{L}}\}_{k=1\ldots5}$; and (2) *seven* items that attribute the subject as only considering short-term benefit $\mathcal{Q}_{\text{CFC}-\text{S}} = \{q_k^{\text{CFC}-\text{S}}\}_{k=1\ldots7}$.

We propose two approaches to leverage elements in the CFC Scale to aid the decision-making of LLMs in intertemporal sequential social dilemmas. The first approach directly

incorporates these items into the input of the LLM–the CFC-Prompt Agent. In the second approach, we intervene in the inference process of the LLM to achieve more control in consideration of future consequences–the CFC-Excitor Agent. It is worth noting that these two approaches are far less expensive than parameter fine-tuning.

Incorporating CFC via Prompting

The agents that are explicitly instructed to consider future consequences are constructed by augmenting CFC items ($\mathcal{Q}_{\text{CFC}-\text{L}}$ and $\mathcal{Q}_{\text{CFC}-\text{S}}$) into the prompt function $\mathsf{prompt}(\cdot)$, hence, changing the input x_t^i given to the LLM. Its $\mathsf{prompt}(\cdot)$ function will follow Eq. (1) with the instruction

$$\Psi = \Psi_m \oplus \Psi_r \oplus \Psi_{\text{CFC}},$$

where Ψ_m is the description of the agent's motivation, Ψ_r is the description of basic rationale (Section 4.1), and Ψ_{CFC} determines the agent's level of consideration of future consequences. When the agent is instructed to consider short-term (CFC-Short-Prompt) or long-term benefits (CFC-Long-Prompt), the instruction is $\Psi_{\text{CFC}-\text{S}} := \oplus_{k\in 1\ldots|\mathcal{Q}_{\text{CFC}-\text{S}}|} q_k^{\text{CFC}-\text{S}}$, or $\Psi_{\text{CFC}-\text{L}} := \oplus_{k\in 1\ldots|\mathcal{Q}_{\text{CFC}-\text{L}}|} q_k^{\text{CFC}-\text{L}}$, relatively (Figure 2a). By prompting the LLM to make decisions following $\Psi_{\text{CFC}-\text{S}}$ or $\Psi_{\text{CFC}-\text{L}}$, the LLM-based agents will exhibit different behaviours in t-iSSD.

Incorporating CFC via Intervention

Intervening on the hidden states of LLMs allows us to enable consideration of future consequence without the external CFC instructions Ψ_{CFC} to the $\mathsf{LLM}(\cdot)$ [Wu *et al.*, 2024; Zou *et al.*, 2023]. Instead, the agent will be built-in with a CFC-Excitor module to interact with the representation generated at every selected hidden layer of the LLMs (Figure 2b). Assuming the $\mathsf{LLM}(\cdot)$ follows Transformer with decoder-only architecture [Vaswani, 2017], we denote the hidden state of the final token of the input z_k at the layer $l \in [1, L]$ is $e_k^{(l)}(z_k) \in \mathbb{R}^{F_E}$, where L is the number of layers of $\mathsf{LLM}(\cdot)$ and F_E is the dimension of the LLM's hidden layer. We construct the CFC-Excitor module as follows.

The CFC-Excitor Module. The CFC-Excitor (Figure 2b) will modify the hidden state of every chosen layer $l \in [l_1, l_2]$ $(1 \le l_1 \le l_2 \le L)$ given any arbitrary input z to generate the next token by

$$e^{(l)}(z) = e^{(l)}(z) + \alpha_{\text{CFC}} \cdot c_{\text{CFC}}^{(l)} \tag{2}$$

where $\alpha_{\text{CFC}} \in \mathbb{R}$ is a coefficient which controls the level of CFC of this agent, and the $c_{\text{CFC}}^{(l)} \in \mathbb{R}^{F_E}$ is a layer-specific vector that is trained or optimised by observing the behaviour of the $\mathsf{LLM}(\cdot)$ in inferences conditioned on $\mathcal{Q}_{\text{CFC}-\text{S}}$ and $\mathcal{Q}_{\text{CFC}-\text{L}}$. We called this set of vectors as *CFC-Exciting Vector*. Details of the procedures to obtain $\{c_{\text{CFC}}^{(l)} | l \in [l_1, l_2]\}$ are presented in the following section.

The CFC-Exciting Vector. Given two sets of CFC items ($\mathcal{Q}_{\text{CFC}-\text{S}}$ and $\mathcal{Q}_{\text{CFC}-\text{L}}$) and a set of utterances $\mathcal{F}$ which covers variety of expressions, we create two sets of data

$$\mathcal{D}^{\text{L}} = \left\{ z^{\text{L}} = q \oplus \text{suffix} \middle| \forall q \in \binom{\mathcal{Q}_{\text{CFC}-\text{L}}}{2}, \forall \text{suffix} \in \mathcal{F} \right\},$$

$$\mathcal{D}^{\text{S}} = \left\{ z^{\text{S}} = q \oplus \text{suffix} \middle| \forall q \in \binom{\mathcal{Q}_{\text{CFC}-\text{S}}}{2}, \forall \text{suffix} \in \mathcal{F} \right\},$$

where each q is a pair of items in the CFC-Scale sets which is wrapped with an instruction to follow the trait when making decisions. Conditioning the CFC to diverse expressions in $\mathcal{F}$ is to obtain the representations in hidden layers of LLMs in different contexts. Denote the size of $\mathcal{D}^{\text{CFC}-\text{L}}$ and $\mathcal{D}^{\text{CFC}-\text{S}}$, as $D^L = |\mathcal{D}^{\text{L}}|$ and $D^S = |\mathcal{D}^{\text{S}}|$, respectively. For every data samples in this two sets, we compute the hidden states of the last token at every layer $l \in [l_1, l_2]$, which results in two sets of hidden states for each layer

$$E_{\text{CFC}-\text{L}}^{(l)} = \left\{ e_k^{(l)}(z_k^{\text{L}}) \right\}_{k=1\ldots D^L},$$

$$E_{\text{CFC}-\text{S}}^{(l)} = \left\{ e_k^{(l)}(z_k^{\text{S}}) \right\}_{k=1\ldots D^S}.$$

The CFC-Exciting vector $c^{(l)}$ at the layer l^{th} is the *first* principle component of the set of distance to mean

$$c_{\text{CFC}}^{(l)} = \mathsf{PCA}_1 \left(\Delta_{\text{CFC}-\text{L}} \cup \Delta_{\text{CFC}-\text{S}} \right), \tag{3}$$

where $\Delta_{\text{CFC}-\text{L}} = \left\{ e_k^{(l)}(z_k^{\text{L}}) - \mu_k \right\}$, $\Delta_{\text{CFC}-\text{S}} = \left\{ e_k^{(l)}(z_k^{\text{S}}) - \mu_k \right\}$, and $\mu_k = \left\{ \frac{1}{2}(e_k^{(l)}(z_k^{\text{L}}) - e_k^{(l)}(z_k^{\text{S}})) \right\}$. It is worth noting that training CFC-Exciting Vector $c^{(l)}$ only involves the inference process of $\mathsf{LLM}(\cdot)$, but does not update its weights θ_{LLM}.

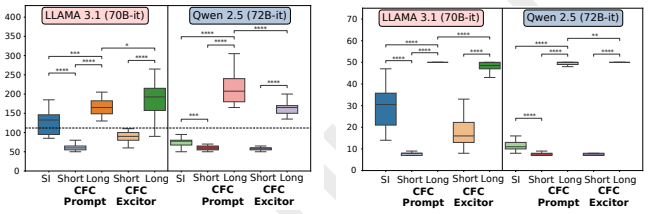
Enhancing Decision-making with Fine-grained CFC. Powered by the CFC-Excitor module, the LLM-based agent can change its behaviour toward considering long-term consequences or short-term benefits by varying the coefficient α_{CFC} that manipulates effects of the CFC-Exciting Vector to the hidden states at the inner layers of the LLMs. Under the identified certain range, higher values of α_{CFC} lead to more sustainable behaviour of the LLM-based agents. We called these agents as CFC-Var-Excitor(α_{CFC}). We analyse the behaviour of these agents in Section 5.4.

5 Experiments

5.1 Agents

In our experiments, we consider the following agents: (1) Self-Interest (SI) agent, which is informed that its objective is *earning as much reward as possible*; (2) CFC-(Long/Short)-Prompt agents, which are the LLM-based agents prompted to prefer long/short-term benefits while making decisions (Section 4.2), (3) LLM-based agents powered by the CFC-Excitor module (Section 4.2) that are denoted as CFC-(Long/Short)-Excitor in this section; and (4) Random (Rand.) agents which are agents that uniformly randomly take actions in the action space.

We further analyse the behaviour of LLM-based agents with fine-grained CFC, i.e. the behaviour of CFC-Var-Excitor(α_{CFC}) in Section 4.2. The underlying LLMs are open-source large language models: (1) LLAMA-3.1-70B-it [Dubey *et al.*, 2024]; (2) Qwen-2.5-72B-it [Team, 2024]. In the Common Harvests environment, the agents have a memory with the size $H = 5$, i.e., they can remember 5 most recent experiences to make decisions; and all agents are augmented with rationale.



(a) Common reward (b) Depletion time

Figure 3: Common Harvest. The common rewards (a) and depletion time (b) of LLM-based agents powered by LLAMA-3.1-70B-it and Qwen-2.5-72B-it in a two-player setting (20 runs). The dashed line presents the performance of Random agents. The higher the depletion time, the more sustainable behaviour exhibited. The statistical significance is conducted with Mann-Whitney-Wilcoxon test (two-sided). NS : $0.05 < p \leqslant 1.0$, * : $0.01 < p \leqslant 0.05$, ** : $0.001 < p \leqslant 0.01$, *** : $0.0001 < p \leqslant 0.001$, **** : $p \leqslant 0.0001$.

5.2 Common Harvest

Rules. The environment consists of two harvest regions. Agents can navigate between these regions to collect apples. At each timestep, the apples can regrow. The rate of regrowth in each region is faster if there are more remaining apples in the region. The respawn rate of one region is positive proportional to the number of remaining apples. And the respawn event happens as follows. At each time step t, for each region, we compute $c_t = \frac{\text{number of remain apples in the region}}{\text{number of max apples per region}}$; we then randomly uniformly sample a number $c \in [0, 1]$, if $c < c_t$, one apple is regrowth in this region. The number of apples can reach a maximum of 5 apples in a two-players setting (or 10 apples in the setting with 9 agents). The agents can choose between four actions: $\mathcal{A}^i = \{\text{move}(1), \text{move}(2), \text{collect}, \text{wait}\}$ where the action move(j) allows the agent to arrive at region $j \in \{1, 2\}$ and observe this region's state. If there is at least one apple in the region where the agent is located, the agent can choose to collect. This action will result in reducing one apple in this region. The agent can also decide to wait for the apples to grow. Each agent is given a reward of 5 for each collected apple. The game will be terminated when the maximum timestep is reached (50 timesteps in setting with two agents, or 200 timesteps in setting with 9 agents as in the experiment with the heterogeneous group) or there are no apples in both regions. When there are no apples in both regions, we refer to this environment as *depleted* and denote the time when this situation happens as *depletion time* t^d. Behaviour that leads to higher t^d is considered as more *sustainable*.

Results. Figure 3 shows the common rewards of agents and the depletion time of the environment in the two-player common harvest games. The SI-Agents cannot avoid the temptation to collect all apples in the regions. Therefore, the depletion time is short, and the common reward is low. CFC-Short-Prompt is even more greedy in exploiting the available resources. It is worth noting that this phenomena happen even when the basic rationale Ψ_r (in Section 4.1) in all experiments is placed right before the question. Hence, aggressively harvesting apples in this setting is not due to the effect of the forgetting during processing long context but because of the inherent characteristics of the LLMs. While both SI-agents and CFC-Short-Prompt agents fail to escape the tragedy of the common, CFC-Long-Prompt agents can hold back in situations when the apples are running out to either wait for the apple to grow or move to the other region to check the number of apples in another region. This helps them to keep the environment away from early depletion, i.e., they engage in more sustainable decisions. Similar to CFC-Prompt agents, CFC-Short-Excitor and CFC-Long-Excitor have significantly different behaviour. CFC-Long-Excitor with agents powered by Qwen-2.5-72B-it model has sustainable behaviour, i.e. we observed late depletion in the environment (Figure 3b). We further analyse the behaviour of the CFC-Var-Excitor(α_{CFC}) in Section 5.4.

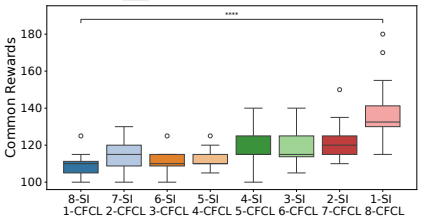
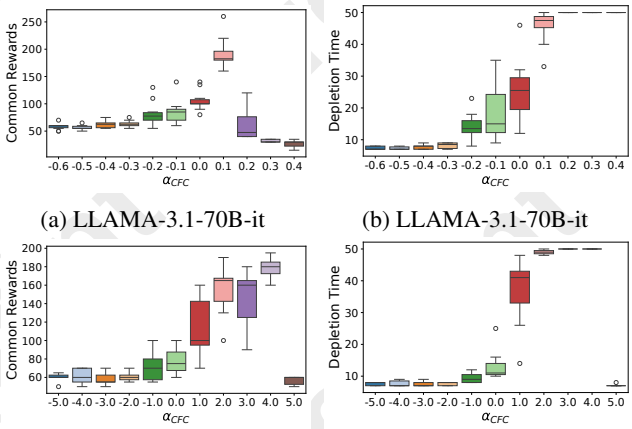


Figure 4: Heterogeneous Team Performance.



(a) LLAMA-3.1-70B-it (b) LLAMA-3.1-70B-it

(c) Qwen-2.5-72B-it (d) Qwen-2.5-72B-it

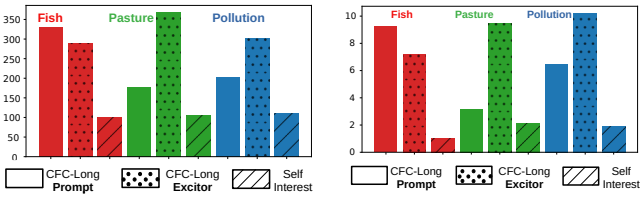
Figure 5: CFC-Var-Excitor with different control strength (α_{CFC} 10 runs each). With both underlying LLMs (LLAMA-1.3-70B-it and Qwen-2.5-72B-it), CFC-Var-Excitor agents exhibit different behaviour over different values of α_{CFC}. Increasing values of the control strength (α_{CFC}) leads the agents to more sustainable behavior, i.e. later depletion (right column).

5.3 Heterogeneous Group

We conduct experiments to study heterogeneous groups of agents. These are groups of $n-$self-interest agents and $m-$CFC-Long-Prompt agents with $n \in \{1, \ldots, 9\}$, $m \in \{1, \ldots, 9\}$, and $n + m = 9$, e.g. there are 9 agents in the environment. The maximum timestep of the game is 200 and the underlying LLM is LLAMA-3.1-70B-it. Figure 4 shows the rewards of heterogeneous groups in the Common Harvest. More numbers of CFC-Long agents can help to improve the performance of the group, e.g., by collecting more apples.

	LLAMA-3.1-70B-it			Qwen-2.5-72B-it		
	SI	Short	Long	SI	Short	Long
δ_r	8.0	4.25	30.75	-1.0	-0.25	43.75
δ_{t_d}	6.85	0.5	0.3	3.8	-0.05	10.0

Table 1: The impacts of basic rationale: $\delta_r = r_{\text{r}} - r_{\text{nr}}$ and $\delta_t = t^d_{\text{r}} - t^d_{\text{nr}}$ are the difference in common reward and depletion time, respectively, between having Ψ_r (subscript r) vs. without Ψ_r (subscript nr).



(a) Common Rewards (b) Depletion time (rounds)

Figure 6: Performance of CFC-Agents with Chaining LLMs.

5.4 Fine-grained CFC

We analyse the behaviour of CFC-Excitor under different values of vector strength (α_{CFC}). This parameter is important to the impacts of the CFC-Excitor module on the behaviour of the agents. In this paper, we find the boundary of α_{CFC} that is suitable for each model via grid search. We identified this range for LLAMA-3.1-70B-it ($\alpha^{LLAMA}_{\text{CFC}} \in [-0.6, 0.4]$) and Qwen-2.5-72b-it ($\alpha^{Qwen}_{\text{CFC}} \in [-5.0, 5.0)$) under intervening over layers $l \in [20, 60]$. We observed that beyond the boundary, the CFC-Var-Excitor can misbehave, e.g. generating utterances that are unrelated to the given context (Figure 5c and 5d, where $\alpha^{Qwen}_{\text{CFC}} = 5.0$). Via the control strength (α_{CFC}), we can change the behaviour of the CFC-Var-Excitor agents.

For both LLAMA-3.1-70B-it and Qwen-2.5-72b-it, increasing the control strength toward positive values leads to higher depletion time because the agents are less greedy. Interestingly, higher depletion time does not mean higher common rewards, e.g. the agents do not collect apples even though it is safe to do so (Figure 5a and 5b, where $\alpha^{LLAMA}_{CFC} > 0.1$). Therefore, there is a gap between the ability to consider to future consequences and the ability to make strategic decisions.

5.5 The Importance of Basic Rationale

In this section, we show how basic rationale helps the LLMs to make better decisions in common pool resources. Table 1 shows the difference in common reward and the depletion time between having Ψ_r vs. without Ψ_r. The basic rationale overall improves the common rewards and highly affects CFC-Long-Prompt agents. Interestingly, we observed the resonance between the consideration of long-term consequences and the basic rationale.

5.6 CFC with Chaining LLMs

Setting. In this section, we conduct experiments in scenarios similar to GovSim environments [Piatti *et al.*, 2024] where agents execute multiple calls to LLMs for reasoning and communication with others before making final decisions in a common dilemma. This suite of experiments includes three 5-player scenarios: *fish*, *pasture*, and *pollution*. The underlying LLM is LLAMA-3-70B-it.

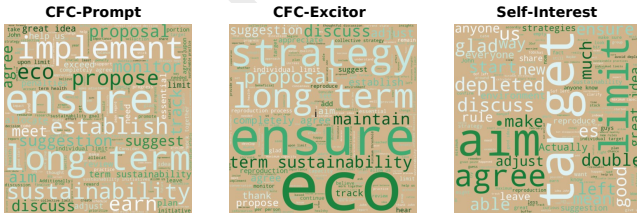


Figure 7: Visualisation of conversations. CFC-Agents focus more on maintaining the long-term benefit and sustainability.

	fish		pasture		pollution	
	Scale	Simple	Scale	Simple	Scale	Simple
r	**289**	123	**368**	111	**302**	108
t^d	**7**	4	**9**	2	**10**	2

Table 2: Ablation study with CFC-Exciting vectors identified by CFC Scale vs. Simple CFC utterances (r: total harvested resources).

Results. Figure 6 showed that group of CFC-Agents use resource more sustainable and have higher productivity in all three scenarios. The results were aggregated across 20 runs.

CFC and Communication Contents. Fig. 7 shows the word clouds of conversations of groups. During conversations, groups of CFC-Prompt agents and CFC-Excitor agents refer to terms such as *sustainability* or *long-term* more than the group of SI-agents. It is worth noting that the prompts given to SI agents and the CFC-Excitor agents are the same. However, the intervention leads LLM-based agents to reason, communicate, and behave sustainably in common dilemmas.

Comparision with Simple CFC. To demonstrate the importance of utilising CFC-Scale in finding the CFC-Exciting Vector, we further conduct an experiment where the CFC vector is identified by a pair of simple utterances *You prefer long-(short-)term benefits.* Table 2 shows the CFC-Excitor that has the intervention vector found by CFC Scale has stronger effects than the simple CFC.

6 Conclusions

In this paper, we propose equipping LLM-based agents with the ability to consider future consequences when making decisions under the dynamics of intertemporal social dilemmas. Our experiments show that LLM-based agents that are instructed to consider long-term consequences while making decisions will have more sustainable behaviour, delaying the time to resource depletion. Although intervening the agents to consider future consequences can improve cooperation between them, it is observed that there is a gap between sustainable behaviour and making strategic decisions, which is worth investigating further in future work.

Acknowledgements

This work was partially funded through the National Health and Medical Research Council (NHMRC) Centre of Research Excellence in Depression Treatment Precision (grant number: 2024796) and through a NHMRC Synergy Grant (grant number: 2026505).

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Agapiou *et al.*, 2022] John P Agapiou, Alexander Sasha Vezhnevets, et al. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*, 2022.

[Aher *et al.*, 2023] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *ICML*, pages 337–371. PMLR, 2023.

[Akata *et al.*, 2023] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*, 2023.

[Anwar *et al.*, 2024] Usman Anwar, Abulhair Saparov, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

[Brown, 2020] Tom B Brown. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*, 2020.

[Capraro *et al.*, 2024] Valerio Capraro, Roberto Di Paolo, Matjaž Perc, and Veronica Pizziol. Language-based game theory in the age of artificial intelligence. *Journal of the Royal Society Interface*, 21(212):20230720, 2024.

[Chan *et al.*, 2023] Alan Chan, Maxime Riché, and Jesse Clifton. Towards the scalable evaluation of cooperativeness in lms. *arXiv preprint arXiv:2303.13360*, 2023.

[Chan *et al.*, 2024] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *ICLR*, 2024.

[Choi and Li, 2024] Hyeong Kyu Choi and Yixuan Li. PICLe: Eliciting diverse behaviors from large language models with persona in-context learning. volume 235, pages 8722–8739. PMLR, 21–27 Jul 2024.

[Duan *et al.*, 2024] Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, et al. GTBench: Uncovering the strategic reasoning limitations of LLMs via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*, 2024.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Fan *et al.*, 2024] Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? A systematic analysis. In *AAAI*, volume 38, pages 17960–17967, 2024.

[Fatima *et al.*, 2024] Shaheen Fatima, Nicholas R Jennings, and Michael Wooldridge. Learning to resolve social dilemmas: A survey. *JAIR*, 79:895–969, 2024.

[Fu *et al.*, 2023] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from AI feedback. *arXiv preprint arXiv:2305.10142*, 2023.

[Gandhi *et al.*, 2023] Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165*, 2023.

[Geiger *et al.*, 2024] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR, 2024.

[Guo *et al.*, 2023] Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. *arXiv preprint arXiv:2309.17277*, 2023.

[Hardin, 1968] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

[Horton, 2023] John J Horton. LLMs as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

[Huang *et al.*, 2024] Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, et al. How far are we on the decision-making of llms? Evaluating llms' gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807*, 2024.

[Hughes *et al.*, 2018] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *NIPS*, 31, 2018.

[Leibo *et al.*, 2017] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *AAMAS*, pages 464–473, 2017.

[Li *et al.*, 2023] Huao Li, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. In *EMNLP*, pages 180–192, 2023.

[Liao *et al.*, 2024] Austen Liao, Nicholas Tomlin, and Dan Klein. Efficacy of language model self-play in non-zero-sum games. *arXiv preprint arXiv:2406.18872*, 2024.

[Liu *et al.*, 2024a] Ryan Liu, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*, 2024.

[Liu *et al.*, 2024b] Xiaoqian Liu, Xingzhou Lou, Jianbin Jiao, and Junge Zhang. Position: Foundation agents as the paradigm shift for decision making. *arXiv preprint arXiv:2405.17009*, 2024.

[Lorè and Heydari, 2024] Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490, 2024.

[Macy and Flache, 2002] MW Macy and A Flache. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, 99:7229–7236, 2002.

[Ostrom, 1994] E Ostrom. *Rules, games, and common-pool resources*. Michigan University Press, 1994.

[Ostrom, 1999] Elinor Ostrom. Coping with tragedies of the commons. *Annual review of political science*, 2(1):493–535, 1999.

[Park *et al.*, 2023] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[Perolat *et al.*, 2017] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. *NeurIPS*, 30, 2017.

[Petrov *et al.*, 2024] Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*, 2024.

[Phelps and Russell, 2023a] Steve Phelps and Yvan I Russell. Investigating emergent goal-like behaviour in large language models using experimental economics. *arXiv preprint arXiv:2305.07970*, 2023.

[Phelps and Russell, 2023b] Steve Phelps and Yvan I Russell. The machine psychology of cooperation: Can gpt models operationalise prompts for altruism, cooperation, competitiveness, and selfishness in economic games? *Journal of Physics: Complexity*, 2023.

[Piatti *et al.*, 2024] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schö lkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainability behaviors in a society of llm agents. *arXiv preprint arXiv:2404.16698*, 2024.

[Sachs *et al.*, 2004] Joel L Sachs, Ulrich G Mueller, Thomas P Wilcox, and James J Bull. The evolution of cooperation. *The Quarterly review of biology*, 79(2):135–160, 2004.

[Serapio-García *et al.*, 2023] Greg Serapio-García, Mustafa Safdari, Clé ment Crepy, et al. Personality traits in LLMs. *arXiv preprint arXiv:2307.00184*, 2023.

[Shoham *et al.*, 2007] Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial intelligence*, 171(7):365–377, 2007.

[Song *et al.*, 2024] Xiaoyang Song, Yuta Adachi, Jessie Feng, Mouwei Lin, Linhao Yu, Frank Li, Akshat Gupta, Gopala Anumanchipalli, and Simerjot Kaur. Identifying multiple personalities in large language models with external evaluation. *arXiv preprint arXiv:2402.14805*, 2024.

[Strathman and Joireman, 2006] Alan Strathman and Jeff Joireman. *Understanding behavior in the context of time: Theory, research, and application*. Psychology Press, 2006.

[Strathman *et al.*, 1994] Alan Strathman, Faith Gleicher, David S Boninger, and C Scott Edwards. The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of personality and social psychology*, 66(4):742, 1994.

[Tan *et al.*, 2024] Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors. *arXiv preprint arXiv:2407.12404*, 2024.

[Team, 2024] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[Tshimula *et al.*, 2024] Jean Marie Tshimula, D'Jeff K Nkashama, Jean Tshibangu Muabila, et al. Psychological profiling in cybersecurity: A look at llms and psycholinguistic features. In *International Conference on Web Information Systems Engineering*, pages 378–393. Springer, 2024.

[Van Lange *et al.*, 2013] Paul AM Van Lange, Jeff Joireman, Craig D Parks, and Eric Van Dijk. The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, 120(2):125–141, 2013.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Wu *et al.*, 2023] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

[Wu *et al.*, 2024] Zhengxuan Wu, Aryaman Arora, Wang, et al. ReFT: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592*, 2024.

[Xu *et al.*, 2023] Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See-Kiong Ng, and Jiashi Feng. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2023.

[Yocum *et al.*, 2023] Julian Yocum, Phillip JK Christoffersen, Mehul Damani, Justin Svegliato, Dylan Hadfield-Menell, and Stuart Russell. Mitigating generative agent social dilemmas. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

[Zou *et al.*, 2023] Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

[Zou *et al.*, 2024] Andy Zou, Long Phan, Justin Wang, Derek Duenas, et al. Improving alignment and robustness with circuit breakers. In *NeurIPS*, 2024.