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

Extended Abstract

Dung Nguyen
Applied Artificial Intelligence Institute (A^2I^2), Deakin University
Geelong, Australia
dung.nguyen@deakin.edu.au

Hung Le
Applied Artificial Intelligence Institute (A^2I^2), Deakin University
Geelong, Australia
thai.le@deakin.edu.au

Kien Do
Applied Artificial Intelligence Institute (A^2I^2), Deakin University
Geelong, Australia
k.do@deakin.edu.au

Sunil Gupta
Applied Artificial Intelligence Institute (A^2I^2), Deakin University
Geelong, Australia
sunil.gupta@deakin.edu.au

Svetha Venkatesh
Applied Artificial Intelligence Institute (A^2I^2), Deakin University
Geelong, Australia
svetha.venkatesh@deakin.edu.au

Truyen Tran
Applied Artificial Intelligence Institute (A^2I^2), Deakin University
Geelong, Australia
truyen.tran@deakin.edu.au

ABSTRACT

Agents built on LLMs have shown versatile capabilities but face difficulties in being cooperative in social dilemma situations. When making decisions under the strain of selecting between long-term consequences and short-term benefits in commonly shared resources, LLM-based agents are vulnerable to *the tragedy of the commons*, i.e. individuals' greed exploitation leads to early depletion. We propose LLM agents that consider future consequences to aid them in navigating intertemporal social dilemmas. We introduce two approaches—prompting and intervention—to equip the agent with the ability to consider future consequences when making a decision, which results in a new kind of agent—*CFC-Agent*. Furthermore, we enable the CFC-Agent to act toward different levels of consideration for future consequences. Our experiments in different settings show that agents that consider future consequences exhibit sustainable behaviour and achieve high common rewards for the population.

KEYWORDS

Social Dilemmas, Cooperation, Multi-agent Interaction, LLMs

ACM Reference Format:
Dung Nguyen, Hung Le, Kien Do, Sunil Gupta, Svetha Venkatesh, and Truyen Tran. 2025. Navigating Social Dilemmas with LLM-based Agents via Consideration of Future Consequences: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Social dilemmas such as common pool resources [5, 9] require cooperation among individuals. We, humans, can effectively resolve dilemmas under different conditions [4, 11]; however, it still is a challenge for artificial agents to do so. Recent AI research in this area typically has long been focused on the paradigm where the sheer amount of training is required for training cooperative behaviour in specific situations, e.g. reinforcement learning agents [2, 6, 7, 10]. Parallel to this development is the rise of large language models (LLMs) [1, 3], allowing us to build coordinated agents in a zero-shot manner. This new paradigm of building artificial agents [8] potentially helps to overcome training issues in achieving reasonable social behaviours, which 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 a low level of cooperation in different settings, including sharing common pool resources. In this paper, we aim to construct a framework for LLM-based multi-agents in which they achieve sustainable use of shared resources 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 Consequence (CFC), defined as *the extent to which individuals consider the potential future outcomes of their current behaviour* [12]. CFC is a personality trait that has been shown to be an important factor in social dilemmas [13, 14]. An instrument to gauge this trait is the CFC Scale, consisting of a list of 12 statements that describe the individuals' considerations of potential consequences [12]. This list is divided into two categories: (1) short-term related items, e.g. *I only act to satisfy immediate concerns, figuring the future will take care of itself*; and (2) long-term related 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 in decision-making in sequential social dilemmas.

Although CFC was extensively studied in social science research, it has not been studied to aid LLMs in making decisions. We present two approaches for integrating CFC into the decision-making process of LLMs. Our first method leverages a cost-effective strategy to induce the desired behaviour through prompting mechanisms, referred to as CFC-Prompt. The second approach involves intervening in the hidden states of LLMs during inference to guide their decisions towards anticipating future outcomes, known as CFC-Excitor. Both methods require only a single call to the LLM for each decision.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

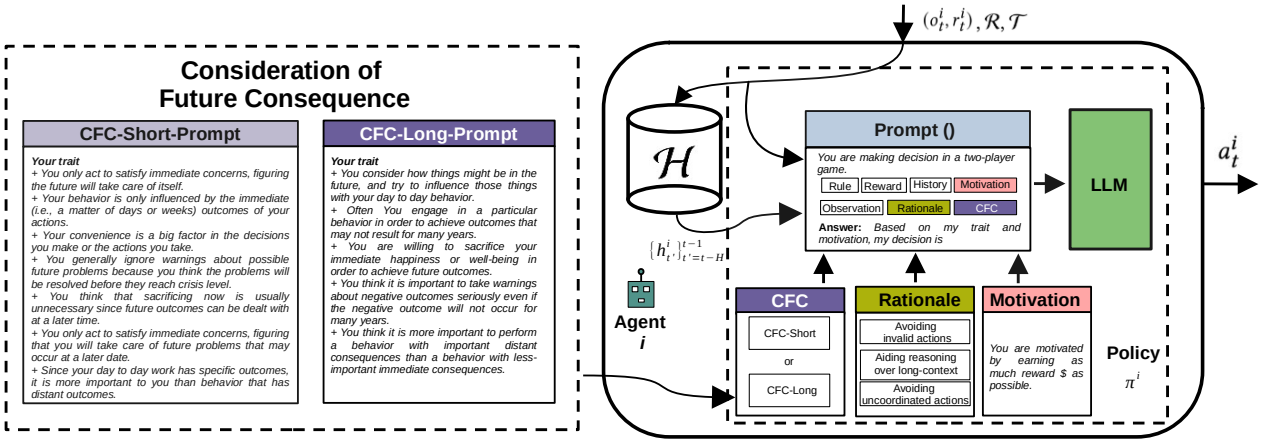

Figure 1: CFC-Agent via Prompting Mechanisms (CFC-Prompt).

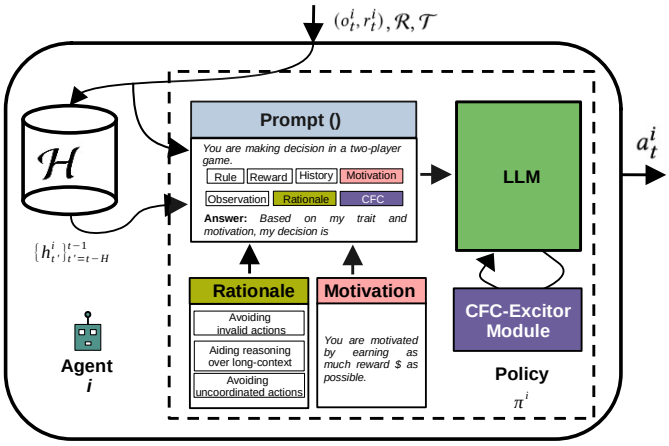

Figure 2: The CFC-Agent via Intervention (CFC-Excitor).

Additionally, we design and examine LLM-based agents capable of considering future consequences at varying levels, utilising a coefficient that regulates the CFC during the intervention.

2 APPROACHES

Our LLM-based cooperative agent is inspired by the concept of Considerations to Future Consequences (CFC) [12]. This is done via: (1) function prompt$(\cdot)$; or (2) via intervening to the features at the inner layers LLM while doing inference.

Incorporating CFC via Prompting. CFC plays an important role in the decision-making process and is a determinant to encourage cooperative behaviour between agents in social dilemmas [13]. In human studies, the degree to which an individual considers future consequences in decision making is measured by the CFC Scale [12]—a 12-items questionnaire. Items in the questionnaire are divided into two categories, i.e. two sets: (1) *five* items that attribute the subject as only considering long-term benefit; and (2) *seven* items that attribute the subject as only considering short-term benefit. The overall structure of our agents is shown in Figure 1.

Incorporating CFC via Intervention. Intervening on the hidden states of LLMs [15, 16] allows us to enable the agents to consider the future consequences of their actions without the external CFC instructions in the LLM$(\cdot)$. In our approach, 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 2). The CFC-Exciting Vector $c^{(l)}$ is identified based on pairs of items in the CFC-Scale sets. It is worth noting that training CFC-Exciting Vector only involves the inference process of LLM$(\cdot)$, but does not need to update its weights θ_{LLM}. 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 the effects of the CFC-Exciting Vector to the hidden states at the inner layers of the LLMs.

3 EXPERIMENTS

Our experiments in Common Harvest, which is a game following the dynamic of the common dilemma, 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. Experiments on heterogeneous populations suggest that having more agents that consider long-term consequences will increase the common reward. Furthermore, our advancements allow us to construct methods to achieve different levels of CFC, and we empirically found that the intervention mechanism can help the LLM-based agents exhibit CFC-related behaviour in a fine-grained manner.

4 CONCLUSIONS

In this paper, we propose to equip LLM-based agents with the ability to consider to future consequence in making decisions under the dynamics of intertemporal social dilemmas. We introduce two approaches to incorporating CFC into the decision-making process of LLMs: (1) prompting mechanisms (CFC-Prompt); and (2) the intervention mechanism (CFC-Excitor).

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñ ez Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, et al. 2022. Melting Pot 2.0. *arXiv preprint arXiv:2211.13746* (2022).

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

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

[5] Garrett Hardin. 1968. The Tragedy of the Commons. *Science* 162, 3859 (1968), 1243–1248. https://doi.org/10.1126/science.162.3859.1243

[6] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in Neural Information Processing Systems* 31 (2018).

[7] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and Multi -Agent Systems*. 464–473.

[8] Xiaoqian Liu, Xingzhou Lou, Jianbin Jiao, and Junge Zhang. [n.d.]. Position: Foundation Agents as the Paradigm Shift for Decision Making. In *Forty-first International Conference on Machine Learning*.

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

[10] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in Neural Information Processing Systems* 30 (2017).

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

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

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

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

[15] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024. ReFT: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592* (2024).

[16] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405* (2023).