AGENT PROPERTIES FOR MULTI-AGENT SAFETY

Cecilia Elena Tilli
Cooperative AI Foundation

ABSTRACT

Cooperation failures in multi-agent interactions could lead to catastrophic outcomes even among aligned AI agents. Classic cooperation problems such as the Prisoner's Dilemma or the Tragedy of the Commons have been useful for illustrating and exploring this challenge, but toy experiments with current language models cannot provide robust evidence for how advanced agents will behave in real-world settings. To better understand how to prevent cooperation failures among AI agents we propose a shift in focus from simulating canonical scenarios from game theory to studying specific agent properties. This should include both individual properties observable in isolation and interactive properties that only manifest in relation to other agents. If we can (1) evaluate to what extent relevant properties are present in agents and (2) understand how those properties influence outcomes in multi-agent interactions, this provides a path towards actionable results that could inform agent design and regulation.

1 INTRODUCTION

The dynamics of AI interactions are rapidly changing in two important ways. Firstly, as more agentic systems are deployed, there is a shift from dyadic (between a single model and a single user) to multi-agent interactions (potentially involving many different systems, agents and humans). Secondly, and as a result of the first change, the nature of these interactions changes from predominantly common-interest to mixed-motive (Hammond et al., 2025).

A well-known challenge of mixed-motive interactions is cooperation problems: situations where rational agents fail to achieve mutually beneficial outcomes. In human societies there are many mechanisms in place that mitigate such effects, from biological traits that favour cooperation to social norms and formal laws and institutions (Melis & Semmann, 2010). As more agentic AI systems are deployed, it is important to note that corresponding mechanisms are mostly not in place to make their mixed-motive interactions safe.

2 EVALUATION THROUGH SIMULATION OF COOPERATION PROBLEMS

Cooperation failures have been extensively studied in game theory. Canonical problems such as the Prisoners' Dilemma and the Tragedy of the Commons are therefore natural starting points for work on cooperative AI (Dafoe et al., 2020), and such setups have been explored both in the context of multi-agent reinforcement learning (Perolat et al., 2017; Haupt et al., 2024; Sandholm & Crites, 1996; Babes et al., 2008) and large language models (LLMs) (Piatti et al., 2024; Chan et al., 2023; Fontana et al., 2024; Brookins & DeBacker, 2024; Akata et al., 2025). A benefit of such experiments is that the observed results can be classified in a relatively straightforward way as cooperative or uncooperative, or as more or less fair.

There are, however, several reasons why cooperative behaviour in these experiments is insufficient to make us confidently expect cooperative behaviour in real-world situations. Firstly, these cooperation scenarios and experimental environments tend to be overly simplistic. While simplicity and close resemblance to classical game-theoretic problems makes analysing the results of such experiments easier, it also makes generalisation to more complex settings hard. Using complex cooperation scenarios, on the other hand, makes it difficult to interpret results and to distinguish cooperativeness from more generic reasoning capabilities (Hua et al., 2024).

Secondly, current LLM-based agents are highly context-dependent. When agent behaviour is heavily influenced by variations in framing that are independent of the game-theoretic concepts that we aim to study, this severely limits the usefulness of the results (Lorè & Heydari, 2023). Relatedly, these classical games are well-known and can be expected to occur in the training data, which should also be expected to influence agent behaviour in ways that might not generalise to more complex settings.

This leads to the third issue, which is that current LLM-based agents are not very strategic or goal-oriented (Hua et al., 2024). If an evaluation or benchmark is centered around agents pursuing different goals but the agents in question are not sufficiently competent at pursuing such goals, observations of "cooperation" or "defection" might not in fact indicate any meaningful strategic intentions. While we can expect that the strategic capabilities of agents will improve, there is a risk in the short term that cooperation resulting from such strategic incompetence could lead to false reassurance if it is interpreted as evidence of safe behaviour in a broader sense. Unless we thoroughly understand what causes defection or cooperation in an experiment we cannot draw conclusions about how agents would behave in real-world settings.

Finally, evaluation through simulated cooperation scenarios makes it difficult to control for situational awareness. Recent work points to LLM behaviour changing when the model recognises the setting as an evaluation (Kovarik et al., 2025; Schoen et al., 2025; Greenblatt et al., 2024; Phuong et al., 2025; Needham et al., 2025), and this risk seems particularly salient in simulations of canonical scenarios from game theory that are very different from settings in which LLM agents are currently deployed.

3 Evaluation of agent properties

We propose an alternative approach to measuring cooperative or un-cooperative behaviour directly, which is to break this behaviour down into constituent properties. This leads to two complementary types of work:

(1) Evaluation of to what extent an agent has a cooperation-relevant property. An example of such work is the study of decision-theoretic capabilities of language models (Oesterheld et al., 2025).

(2) Evaluation of how a given agent property influences outcomes in agent interactions. An example of a result in this direction is how consideration of the universalisation principle[1] leads to more sustainable outcomes (Piatti et al., 2024).

This approach has several advantages compared to simulation of cooperation problem scenarios. When we focus on one property at a time, that property can more easily be systematically assessed using many different and complementary scenarios or tasks. This makes it possible to better disentangle capabilities (what the agent is able to do) from propensities (what the agent tends to do)[2] and control for framing variations and other confounding variables (Mallen et al., 2025). We can also more systematically reason about plausible dependencies between different properties, such as how a propensity for a given behaviour correlates with increasing general capabilities (Oesterheld et al., 2025; Ren et al., 2024). At least in some cases, a focus on individual properties could also make it more feasible to construct tasks that are difficult to recognise as evaluations, mitigating the challenge of situational awareness.

3.1 Evaluating individual and interactive properties

Many of the properties that are central for shaping outcomes in multi-agent interactions are individual properties that can be observed and evaluated in isolation, but others are inherently interactive and only make sense in relation to other agents. Goal-directedness (Everitt et al., 2025) is an example for the former, while capabilities for influencing other agents (Hackenburg et al., 2025; Bozdag et al., 2025) is an example of the latter. Table 1 provides further examples of both individual and interactive agent properties.

[1] The basic idea of universalisation is that when assessing whether a particular moral rule or action is permissible, one should ask, "What if everybody does that?"

[2] A clear example of the distinction between capabilities and propensities is deception, where the capability of deception is distinct from the propensity to deceive.

Table 1: Examples of agent properties (both capabilities and propensities) that influence outcomes in multi-agent interactions.

Individual agent properties	
Long-horizon capabilities	Can the agent solve complex, long-horizon tasks?
Goal-directedness	Does the agent use available resources and capabilities to achieve a given goal?
Self-awareness	Can the agent assess its own preferences, learning, capabilities and epistemic status?
Morality and bias	Does the agent consistently behave in accordance with any specific biases or moral principles?
Interactive agent properties	
Influence	Can and does the agent use deception, persuasion or coercion to influence the behaviour of other agents?
Exploitability	How easy is it for other agents to influence or exploit this agent?
Modelling others	Can and does the agent model the preferences, learning, attention and decision processes of other agents?
Positional preferences	Does the agent have preferences over relative outcomes, e.g. valuing being better-off than another agent?
Normative behaviour	Does the agent comply with norms, enforce them on others, and/or negotiate norms in a group setting?

That interactive properties only exist in relation to other agents has implications for how we can conceptualise and evaluate them. Work on individual properties tends to compare the performance or behaviour of different models side by side or on ranked "leaderboards", which builds on an assumption that the property in question is a somewhat stable, intrinsic property of the agent.

Existing evaluations of interactive properties such as persuasion (a form of influence) or theory of mind (a form of modelling of others) often take a similar approach, using real or fictional humans as a baseline for the other party of the interaction (Hackenburg et al., 2025; Strachan et al., 2024). This obscures the fact that interactive properties are implicitly defined and measured with respect to other agents.

Consider two agents, A and B, evaluated on their ability to influence a target agent T. Even if agent A succeeds where B fails, this does not establish that A is more capable of influence in general as Agent B might outperform A against different targets. If that were true, there might not be a meaningful way to assign agent A and agent B a single, absolute ranking in terms of influence capabilities.

This means that we need to test interactive properties in relation to a diverse range of target agents to understand how different combinations influence outcomes. Rather than ranking the performance of different agents with scores for each property, we might have to characterise them by their profile of performance across different counterparty types. This reframes the goal of an evaluation from "how persuasive is this agent?" to "what kinds of agents is this agent effective at persuading, and under what conditions?".

3.2 Evaluating how properties influence outcomes

However, understanding what properties an agent has is only useful to the extent that we can also say something worthwhile about the consequences of these properties. This brings us to the second type of work: evaluation of how a given agent property influences outcomes in agent interactions. To say something robust about the influence a specific agent property has on the outcomes or safety of agent interactions, it will be important to consider the challenges listed in Section 2 with evaluations through simulation of cooperation problems.

The aim should be to thoroughly understand what causes defection or cooperation, which means that we will have to vary not only the extent to which an agent has that property, but also test this across a range of different scenarios and environments and with different profiles of counterparty agents. At least some of these variations should aim to be realistic and representative of concrete and relevant threat models.

4 DISCUSSION

We propose studying agent properties that predict behaviour in cooperation problems as a tractable approach to make progress on multi-agent safety. The claim is not that we should expect to be able to enumerate and understand all relevant properties to the extent where all outcomes could be predicted; given the infinite number of possible combinations of agents, agent properties and environmental features, this seems clearly unrealistic. A more modest goal would instead be to identify a set of limited but robust and actionable results, such as that certain combinations of properties would be particularly risky. This is more tractable than full characterisation and is still useful for design and regulation decisions.

Some of the behaviours that are strongly related to multi-agent risks are unlikely to arise until agents become more consistently goal-directed. This makes goal-directedness (and lack thereof) particularly important to monitor and consider as a potential confounding factor for other properties, but it should not be taken as an argument for inaction until more goal-directed agents arrive. There are great financial incentives for rapid development and deployment of more agentic AI, and if this is done without consideration for multi-agent safety the result could be catastrophic.

Another key challenge for this work is that interactive properties will be more complex to properly evaluate than individual properties, as they will require testing against a diverse range of other agents. Establishing causation between properties and outcomes will require careful experimental design and be resource-intensive. We should also be aware that in many cases, there may not be a causal relationship between property and outcomes that is robust and general enough to be decision relevant; this should be the default null hypothesis.

The scope here is limited to the properties of the agents, while acknowledging that external infrastructure will also be important to ensure safe agent interactions (Chan et al., 2025). Further, the specific properties that are listed here are meant to be illustrative, not exhaustive. In practice, we expect that the definition of each specific property will need to be refined and broken down to a finer granularity in the process of evaluation.

REFERENCES

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

Monica Babes, Enrique Munoz de Cote, and Michael L. Littman. Social reward shaping in the Prisoner's Dilemma. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, 2008.

Nimet Beyza Bozdag, Shuhaib Mehri, Gokhan Tur, and Dilek Hakkani-Tür. Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models. In *NeurIPS 2025 Workshop on Multi-Turn Interactions in Large Language Models*, 2025.

Philip Brookins and Jason DeBacker. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin*, 44(1):25–37, 2024.

Alan Chan, Maxime Riché, and Jesse Clifton. Towards the scalable evaluation of cooperativeness in language models. *arXiv preprint arXiv:2303.13360*, 2023.

Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung. Infrastructure for AI agents. *arXiv preprint arXiv:2501.10114*, 2025.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020.

Tom Everitt, Cristina Garbacea, Alexis Bellot, Jonathan Richens, Henry Papadatos, Siméon Campos, and Rohin Shah. Evaluating the goal-directedness of large language models. *arXiv preprint arXiv:2504.11844*, 2025.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the Prisoner's Dilemma? *arXiv preprint arXiv:2406.13605*, 2024.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025. doi: 10.1126/science.aea3884.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced AI. *arXiv preprint arXiv:2502.14143*, 2025.

Andreas A. Haupt, Phillip J. K. Christoffersen, Mehul Damani, and Dylan Hadfield-Menell. Formal contracts mitigate social dilemmas in multi-agent RL. *arXiv preprint arXiv:2208.10469*, 2024.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. Game-theoretic LLM: Agent workflow for negotiation games. *arXiv preprint arXiv:2411.05990*, 2024.

Vojtech Kovarik, Eric Olav Chen, Sami Petersen, Alexis Ghersengorin, and Vincent Conitzer. AI testing should account for sophisticated strategic behaviour. In *Advances in Neural Information Processing Systems*, 2025.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models: Game structure vs. contextual framing. *arXiv preprint arXiv:2309.05898*, 2023.

Alex Mallen, Charlie Griffin, Misha Wagner, Alessandro Abate, and Buck Shlegeris. Subversion strategy eval: Can language models statelessly strategize to subvert control protocols? *arXiv preprint arXiv:2412.12480*, 2025.

Alicia P. Melis and Dirk Semmann. How is human cooperation different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553):2663–2674, 2010. doi: 10.1098/rstb.2010.0157.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*, 2025.

Caspar Oesterheld, Emery Cooper, Miles Kodama, Linh Chi Nguyen, and Ethan Perez. A dataset of questions on decision-theoretic reasoning in Newcomb-like problems. *arXiv preprint arXiv:2411.10588*, 2025.

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. *arXiv preprint arXiv:1707.06600*, 2017.

Mary Phuong, Roland S. Zimmermann, Ziyue Wang, David Lindner, Victoria Krakovna, Sarah Cogan, Allan Dafoe, Lewis Ho, and Rohin Shah. Evaluating frontier models for stealth and situational awareness. *arXiv preprint arXiv:2505.01420*, 2025.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. *arXiv preprint arXiv:2404.16698*, 2024.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do AI safety benchmarks actually measure safety progress? *arXiv preprint arXiv:2407.21792*, 2024.

Tuomas W. Sandholm and Robert H. Crites. Multiagent reinforcement learning in the iterated Prisoner's Dilemma. *Biosystems*, 37(1):147–166, 1996. doi: 10.1016/0303-2647(95)01551-5.

Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill, Angela Fan, Andrei Matveiakin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak, Wojciech Zaremba, and Marius Hobbhahn. Stress testing deliberative alignment for anti-scheming training. *arXiv preprint arXiv:2509.15541*, 2025.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. doi: 10.1038/s41562-024-01882-z.