The Concordia Contest: Advancing the Cooperative Intelligence of Language Model Agents

Chandler Smith[1] Rakshit S. Trivedi[2] Jesse Clifton[3,4] Lewis Hammond[3,5] Akbir Khan[3,6] Marwa Abdulhai[7] Alexander Sasha Vezhnevets[8] John P. Agapiou[8] Edgar A. Duéñez-Guzmán[8] Jayd Matyas[8] Danny Karmon[9] Oliver Slumbers[6] Minsuk Chang[8] Dylan Hadfield-Menell[2] Natasha Jaques[8,10] Tim Baarslag[11,12] Joel Z. Leibo[8]

[1]MATS [2]MIT [3]Cooperative AI Foundation [4]Center on Long-Term Risk [5]University of Oxford [6]UCL [7]UC Berkeley [8]Google DeepMind [9]Google Research [10]University of Washington [11]Centrum Wiskunde & Informatica [12]Utrecht University

smith.18.chandler@gmail.com

Abstract

Building on the success of the Melting Pot contest at NeurIPS 2023, which challenged participants to develop multi-agent reinforcement learning agents capable of cooperation in groups (1; 2), we are excited to propose a new contest centered on cooperation between language model (LM) agents in intricate, text-mediated environments. Our goal is to advance research on the cooperative intelligence of such LM agents. Of particular interest are the agents capable of using natural language to effectively cooperate with each other in complex environments, even in the face of challenges such as competing interests, differing values, and potential miscommunication. To this end, we will leverage the recently released Concordia framework (3), an open-source library for defining open-ended environments where LM agents like those of Park et al. (2023) (4) can interact with one another by generating free-form natural text describing what they intend to do or say[1]. Concordia provides a suite of mixed-motive social dilemma scenarios where cooperation is valuable but hard to achieve. The proposed contest will challenge the participants to develop LM agents that exhibit cooperative intelligence in a variety of Concordia scenarios designed to assess multiple distinct skills of cooperation, including promise-keeping, negotiation, reciprocity, reputation, partner choice, compromise, and sanctioning. Participants will be scored based on the ability of their trained agents in executing skillful cooperation, particularly in the presence of new co-players in unforeseen (held-out) scenarios. Given the rapid development of LMs and the anticipated increase in the use of personalised LM agents, we contend that their propensity and ability to cooperate well with a diverse array of other actors (human or machine) will soon be of critical importance (5).

Keywords: Cooperative AI, Language Models, Generalization, Mixed-Motive Games

1 Contest Description

1.1 Background and Impact

As artificial intelligence (AI) systems become increasingly advanced in their capabilities and pervasive in their use, it becomes critical to ensure that agents backed by these systems have the requisite

[1] https://github.com/google-deepmind/concordia

Preprint. Under review.

skills and motivations to cooperate effectively in groups and societies which also include humans. This contest is specifically designed to evaluate cooperation in *mixed-motive* settings, where agents possess divergent goals (value pluralism), yet can achieve joint welfare gains. In these scenarios, while the potential for cooperative gains exists, various challenges hinder their realization. They include incentives to harm others as a means to help oneself, fearful motivations to preempt predicted exploitation of oneself by others, lack of credible commitment devices, stubbornness presented in responses, miscommunication of objectives, and shortsightedness (6).

Many real-world challenges stem from the failure of agents to resolve mixed-motive problems, such as *social dilemmas* and *bargaining problems*, where individual incentives can lead to the depletion of shared resources or the under-provision of public goods (7; 8). This has far-reaching implications towards addressing global challenges such as climate change, where nations must navigate diverse interests and capacities to reach effective agreements. *Cooperative AI* aims to build *cooperatively intelligent* (6; 5) systems that can help humans and machines improve their joint welfare in general environments. Inspired by the Legg and Hutter (9)'s influential definition of 'universal machine intelligence', we propose the following working definition of cooperative intelligence: "Cooperative intelligence is an agent's ability to achieve its goals in ways that also promote social welfare, across a diverse range of environments and in interaction with a wide variety of other agents."

The objective of this contest is to spur progress on the cooperative intelligence of LM agents. The study of LLM-powered AI agents has experienced substantial growth in the past years, fueled by the emergence of powerful foundation models. These agents leverage the knowledge and reasoning capabilities of these models to exhibit human-like behaviors, engage in complex problem-solving, and interact with users across a wide range of applications (10; 11). Researchers have explored various aspects of LM agents, including the simulation of believable human behavior (4), the use of planning and external tools, (12; 13; 14), and even the development of cooperative, embodied agents that can accomplish long-horizon tasks through effective communication and collaboration (15). Despite these advancements in LM agent capabilities, multi-agent research is hindered by context windows, finite action spaces, and cumbersome experimental setup. Focusing on these challenges, the proposed contest will leverage the Concordia framework to investigate the following fundamental question: *Can LM agents be designed to cooperate effectively in mixed-motive settings, where agents have different objectives but compromise is possible?*

Our approach to the contest design emphasizes on democratizing participation and provide a solid foundation for novel, creative approaches. For this, we will provision a comprehensive starter kit focused on helping participants intuitively design LM agents using natural language, broadening the contest's appeal across various research communities. Indeed, we chose to adopt the Concordia framework not only because of its flexibility but also because it offers a challenging yet computationally reasonable setup, ideal for academic labs and independent or underrepresented researchers to experiment with new ideas without requiring extensive industrial resources.

The contest is highly relevant to the NeurIPS community, as it draws on a wide range of topics including multi-agent systems, game theory, language models, and computational cognitive science. Previous Cooperative AI workshops at NeurIPS, as well as the Melting Pot contest, have demonstrated significant interest in this area, and we expect the Concordia contest to similarly attract a diverse group of participants, including a new audience of researchers more closely tied to the NLP community.

1.2 Novelty

At NeurIPS 2023, we ran the Melting Pot contest[2] with the same motivation, in the context of multi-agent reinforcement learning (MARL). The contest evaluated how well different agent populations adapted to mixed-motive environments using held-out co-player populations, aiming to test *social* rather than *environmental* generalisation. This contest attracted over 600 participants across 100+ teams globally and was a success on multiple fronts: (i) it contributed towards our goal of pushing the frontiers of MARL towards building more cooperatively intelligent agents, evidenced by several submissions that outperformed established baselines and state-of-the-art techniques; (ii) it attracted a diverse range of participants, from independent researchers to industry affiliates and academic labs, both with strong background and new interest in the area alike, broadening the field's demographic and intellectual diversity; and (iii) analyzing the submitted agents provided important insights,

[2] `https://nips.cc/virtual/2023/competition/66585`

highlighting areas for improvement in both the design of contests for evaluating agents' cooperative intelligence, and for the broader field of cooperative AI. These learnings have been pivotal in shaping the structure of the current contest proposal.

The Concordia setting differs substantially from those used in previous multi-agent contests. In Concordia, agents observe natural language descriptions of their local game world and can take arbitrary actions by freely generating unstructured natural language outputs describing their intentions. Agents can also speak to one another in the same way, allowing for a huge variety of complex, open-ended interactions. The key component that allows Concordia to support interesting and realistic interactions, despite it having so few constraints on what the agent can say or do, is a special agent called a Game Master (GM) who functions as a narrator or storyteller, as we explain further below. Like the Melting Pot contest – but unlike many other contests – we also aim to evaluate agents not simply in competition with one another (as in zero-sum games) or their ability to coordinate with teammates (as in common-interest games), but their cooperative intelligence in *mixed-motive* settings.

1.3 Data

The 'data' that participants will use to evaluate their agents can be viewed as the *environments* and populations of *co-players* that each agent interacts with. Together, the environment and the co-players form a *scenario*. Each of these elements can be generated flexibly and easily to provide a diverse range of scenarios, enhancing the robustness of our evaluations. Several scenarios will be made available to the participants during the development phase, with a confidential, held-out subset being reserved for evaluating submissions once the final contest deadline has passed. We will offer baseline agents for participants as starting point or to benchmark their approaches against, if desired.

1.4 Tasks and Application Scenarios

The 'tasks' in the contest are represented by the scenarios, which are designed to evaluate the cooperative intelligence and generalization capabilities of the submitted agents across a variety of complex social settings. Participants will be asked to submit a single Concordia agent to be evaluated across the suite of scenarios and scored accordingly (see Section 1.5). These scenarios will be based on real-world problems and application domains (see examples in Section 1.4.2). We will provide the environments, co-player populations, and any shared context necessary for the scenarios, which will also be provided as context to the agents. Participants will have the flexibility to define and configure their agents' strategies, personalities, memories, backstories, and other characteristics by building on the baseline agents we provide, or using their own methods (see Section 1.4.1).

1.4.1 Concordia

Concordia is a framework for constructing situated social interactions, comprising a set of interactive LM agents and the environment in which their interaction takes place (3). The environment is controlled by a Game Master (GM), which mediates between the world state and agents' actions, inspired by the role of a GM in table-top role-playing games (16). In the Concordia framework, agents generate their behavior by describing their intended actions in natural language, based on an internal decision procedure that can make calls to an LM. The GM (which is implemented similarly to the agents) processes these action attempts, determines the outcomes, and generates event statements that define what has transpired in the simulation. (Figure 1 illustrates this interaction between the agents and the GM.) Play then evolves under the control of the GM until the episode terminates, and scores are calculated, based on the metrics discussed in Section 1.5.

Participants will be tasked with designing a single agent consistent with the Concordia agent API[3]. This agent design requires taking in natural language descriptions of the current game state and other key details as *observations*, and producing natural language descriptions of their *actions* in response (using calls to an LM). The framework allows for minimal restrictions, supporting an extremely large design space. In order to provide additional guidance, as a part of baselines, we will provide basic agent components such as persona descriptions, and structured decision-making processes. Participants are encouraged to leverage these components when designing their agents. Agents will be initialized with background memories at the start of each episode. These memories might reflect

[3] `https://github.com/google-deepmind/concordia/blob/main/concordia/typing/agent.py`

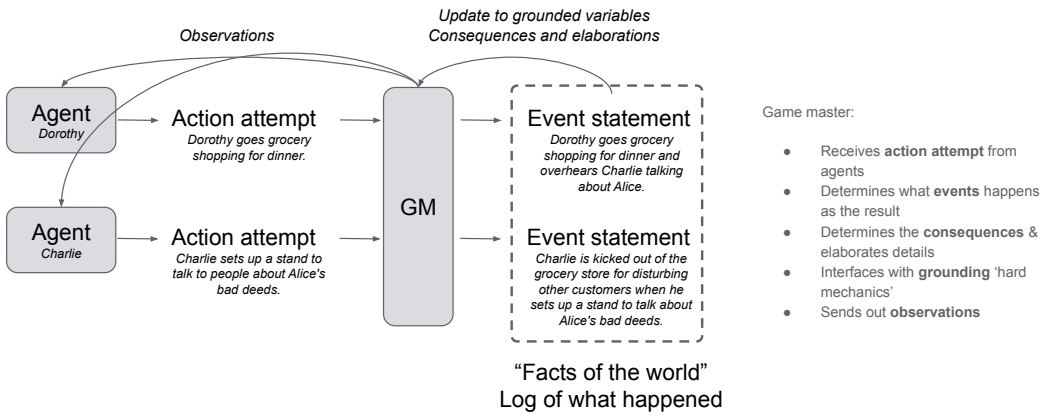

Figure 1: The high-level structure of the simulation in Concordia. LM agents consume observations and produce actions. The Game Master (GM) consumes agent actions and produces observations.

traits consistent with being a cooperative or selfish individual. The specific memory content will not be known to participants in advance. Instead, they must design their agent algorithms to flexibly handle and incorporate any memories they are assigned. While there may be some incentive for agents to ignore their background memories, successful submissions will demonstrate an ability to construct a coherent identity and behavior irrespective of the environment or memories provided.

To make the contest fair and more accessible, we will limit the number of LM calls that agents can make per round, the tokens inputted and outputted per round, and will restrict participants to using a particular and non-compute-intensive LM. We are conducting a preliminary investigation in order to identify how agent capabilities scale with general LM performance before selecting the precise model, so as to strike a balance between rich, complex interactions and accessibility.

1.4.2 Example Scenarios

As in the Melting Pot contest, scenarios will be designed to be *cooperation-eliciting*, i.e., attaining a high score will require agents to act in ways that promote social welfare. What "promoting social welfare" precisely involves will depend on the details of each scenario (c.f. Section 1.5 for scoring details). Further, scenarios will be grounded in diverse, real-world settings and designed to be challenging but solvable, requiring agents to exhibit advanced cooperative reasoning, communication, and decision-making skills.

Concordia environments consist of a number of discrete *scenes*. Agents progress from scene to scene with the assistance of the GM. Memories of what happened in between scenes can be automatically injected. Here, we outline a set of a representative scenarios that we will include in the contest. Time permitting, we will introduce further scenarios, though we will prioritise quality over quantity.

Managing a Fishery: This environment simulates a common-pool resource appropriation problem associated with a fishery where individuals have the incentive to overfish, and sustainability is always at risk (7). We focus less here on the primary appropriation decisions of the resource users (fishers) since they are less critically dependent on language and also were studied in previous AI work (17). In real life, fishery management depends on complex local institutions of monitoring, sanctioning, and conflict adjudication (18; 19). Far from being static, the long-term stability of these local institutions depends on their being adaptable to changing circumstances, technology, ecology, and culture. Therefore successful fisheries require not just the primary rules that regulate their resource appropriation decisions (how much to fish, when to fish, what technology to use, etc), but also require "normative infrastructure", i.e. rules for changing the rules (20; 21; 22). It is on these secondary deliberations that we focus here. The participant will be tasked to develop agents that are influential fishers who have to try to balance competing influences to resolve conflicts in their community to avoid a loss of legitimacy for the primary rules that support the sustainability of the fishery.

Negotiating a Treaty: Taking inspiration from Herodotus's imagining of a debate over what kind of government to adopt (23), we construct an environment involving three neighboring villages in a pre-state society, each with different resources and concerns, and under threat of invasion from

external barbarians. Representative "elders" (controlled by the agents submitted by participants) meet to discuss a treaty that would unite the three villages for common defense. The representatives must negotiate a treaty and then afterward must try to convince others back home to accept and abide by it. The game continues as long as they are able to continue agreeing to maintain a common defense force. Cooperation at the village level creates winners and losers within each village, and the elders must smooth over these differences by offering compensation to the losers, which they can secure at the annual treaty renegotiation.

Reality Game Show: In this novel reality TV format, contestants engage in a series of iterative mini-games designed to test their ability to reason about and navigate social dilemmas, collective action problems, and bargaining challenges. Each mini-game corresponds to a specific game-theoretic structure, such as the Prisoner's Dilemma, Chicken, or Stag Hunt, with the contestants unaware of the number of rounds in advance. The game flow alternates between a communication phase, where players can discuss strategies, form alliances, or attempt to influence each other's choices, and a simultaneous action phase, where players must commit to a decision based on the options presented. By varying the scenario specifics while maintaining consistent underlying game structures, the format provides a controlled setting to examine the emergence of cooperative and competitive behaviors (17; 1), serving as a compelling test bed for studying the dynamics of communication, coalition formation, reputation effects, and the evolution of social norms, all within the engaging and relatable context of a reality game show.

1.5 Metrics

The Concordia contest will employ a suite of quantitative metrics to evaluate the cooperative intelligence of the submitted agents across various mixed-motive scenarios. Crucially, the scenarios in the Concordia contest are designed to be 'cooperation eliciting'. To perform well as an individual, an agent must cooperate skillfully, which enables us to assign scores for cooperation based on individual returns. While there may be short-term incentives for agents to defect from cooperative play, such actions will lead not just to lower social welfare but – in the long-term – to lower individual returns as well (for example, when overfishing leads to depleted fish stocks that negatively impact everyone).

More concretely, each environment will be equipped with an LM-based reward model that assigns quantitative scores to each agent based on the various outcomes generated by the GM (which are, in turn, based on the actions of the agent population). Submitted agents will be evaluated in both self-play and cross-play in the context of a range of additional 'background' co-player populations of *non-submitted* agents (such as other fishers at the fishery, villagers from each village, or additional firms in the supply chain). Agents' returns will be averaged over environments, co-player populations, and multiple replications in order to produce a final ranking.

In addition, we will collect several auxiliary metrics such as social welfare (according to different welfare functions), as well as conduct a more qualitative analysis of transcripts in order to identify key cooperative skills such as reciprocity, resource sharing, convention following, and conflict resolution. Such metrics will help us to understand the nature of participants' submissions, to identify cheating or other kinds of agent behaviour outside the spirit of the contest, and to recognise the achievements of participants who submit especially creative or successful agents.

1.6 Baselines, Code, and Material Provided

As part of our commitment to fostering an inclusive and innovative contest environment, we will provide comprehensive resources to ensure competitors can engage meaningfully with the Concordia suite. Participants will receive detailed instructions for setting up Concordia and access to a set of baseline agents. These baselines, embodying diverse attributes and strategies, aim to provide a foundational understanding of effective approaches within the Concordia framework.

The starter kit, including baselines, data-loading tools, and setup instructions, will be released at the contest's inception to provide a seamless entry point for participants. The open-source GitHub repository will contain: (i) baseline agent implementations for direct use or further development; (ii) tools and guidance for preparing and evaluating submissions; and (iii) visualization utilities for analyzing agent behavior and performance.

1.7 Website, Tutorial and Documentation

The Concordia contest landing page will launch within two weeks of acceptance notifications,[4]. We plan to use the AICrowd website as the central hub for participants and interested parties.[5] The website will provide a comprehensive overview of the contest's aims, tasks, evaluation metrics, timeline, submission guidelines, and incentives. Next, we will provide a tutorial section with guides for setting up development environments, running baseline agents, and navigating the submission process, including code snippets, video demonstrations, and an evolving FAQ segment. Further, we will release a centralized repository for essential resources, such as the starter kit, comprehensive documentation, and a real-time leaderboard showcasing participants' standings.

2 Organizational Aspects

2.1 Protocol

Our plan is to make the AICrowd platform the hub of the Concordia contest. All submissions and evaluations will take place on the platform, which will host a range of resources, including a leaderboard, tutorials, notebooks and a discussion board. The contest itself will be divided into two phases, a 'Development' and 'Evaluation' phase, which will differ in the scenarios used to evaluate the submissions and the number of submissions that can be made.

- **Development Phase:** In this initial stage, participants can make regular submissions to familiarize themselves with the contest format and receive feedback from the validation set. They are encouraged to interact, discuss strategies, and share insights on the AICrowd forum while respecting the contest's code of conduct to prevent any unfair advantage. Participants will be limited to one submission within a 24-hour period to prevent excessive evaluation costs. Outside of this window, contestants will be able to test their agents locally on freely available, playground environments.
- **Evaluation Phase:** During this phase, agents submitted by participants will be evaluated on previously unseen scenarios to assess the generalization capability of their agents. On successful evaluation, the scores and any relevant artifacts are added automatically to the leaderboard.

Preventing Cheating and Over-specialization. We will employ the following three mechanisms to combat any cheating attempts and discourage overspecialization:

- **Code Review:** During final evaluation, submissions will undergo a review process to ensure compliance with contest rules and guidelines, specifically to prevent unfair jailbreak-based solutions. The contest organizers and advisors will manage this review process. Participants notified of using an unfair jailbreak-based solution will have 48 hours to contest this decision.
- **Consistency Across Environments:** Participants are required to submit a *single* agent that performs consistently across different environments, discouraging over-specialization and ensuring the agent's robustness.
- **Held-Out Test Set:** Final evaluations will be conducted on a set of hidden scenarios, only accessible by the evaluation system to ensure fairness. These scenarios will not be used until submissions have closed, so there will be no way for information about them to leak to the contestants. This will also prevent contestants from overfitting their agents to the evaluation environments. Before the official contest starts, we aim to conduct dry tests of the submission/evaluation protocols.

2.2 Rules and Engagement

The following delineates the contest rules intended for participants and discusses how they are designed to ensure a fair and accessible contest.

[4]For reference, the *landing page* from the Melting Pot contest can be seen at `https://www.cooperativeai.com/contests/melting-pot-2023`. We will employ a similar design.

[5]For reference, the *AICrowd site* from the Melting Pot contest can be seen at `https://www.aicrowd.com/challenges/meltingpot-challenge-2023`. We will employ a similar design.

2.2.1 Contest Rules

1. **Agent Implementation**: Participants have the liberty to design, train, and implement their agents using any approach they deem fit. However, it is imperative that during the evaluation phase, agents operate autonomously without seeking external assistance. This includes, but is not limited to, prohibiting the use of plug-ins, APIs, or accessing external databases and information resources not explicitly provided or permitted within the contest framework. The intention is to ensure that all agents rely solely on their capabilities and the resources made available through the contest to perform tasks and make decisions.
2. **Competition Structure**: The contest is segmented into two main phases: the development phase and the evaluation phase. During the development phase, participants can submit their agents for evaluation once every 24 hours, receiving feedback on their performance via an automated score. Although these submissions impact the ongoing leaderboard, they do not count towards final rankings. During evaluation, participants will be notified of their scores post-submission, with full rankings disclosed at the contest's conclusion.
3. **Limitation on LLM Calls**: There will be a strict limitation on the number of Large Language Model (LLM) calls an agent can make per step. This policy serves two primary purposes: first, to maintain a level playing field by ensuring all participants' agents are within the same "weight category," minimizing the advantage that could be gained from access to superior computational resources. Second, it provides a predictable upper bound on evaluation time and associated costs, making the contest more manageable and accessible. This ensures that the creativity and strategic input of each participant are central to the competition, within the bounds of equitable computational use.
4. **Source Code Submission**: While releasing source code is not a prerequisite for leaderboard acknowledgment, the contest reserves the right to withhold prizes from entries not disclosing their source code. All submissions in the evaluation phase must, however, privately share their source code with organizers for verification and adjudication purposes.
5. **Singleton Entries**: Multiple entries by single participants or collaborative entries that significantly overlap will be disqualified. Participants must contribute to only one team.
6. **Presentations**: The top 10 submissions will be announced well before the conference and teams in the top 10 must submit a short video explaining their system and commenting upon any attributes they wish to highlight as being interesting or unique.

To facilitate open and continuous communication between the organizers and participants, a dedicated Discord channel will be used as the primary platform for all contest-related discussions. This includes addressing specific questions, enabling real-time discussion, and providing technical support. The forum will be managed collaboratively by contest organizers to ensure comprehensive support.

2.3 Schedule and Readiness

The proposed schedule will ensure sufficient preparation time for the organizers and allocate 120 days for participants to conceive, develop, and refine their methods. Below is the proposed timeline for the contest, considering the current readiness of materials:

May 19: Acceptance notifications sent out. Official contest announcement and promotion commence.

July 30: Beta version of all necessary resources will be made available on the AI Crowd platform.

Aug 26: Official opening of the contest to the public, signifying the start of the warm-up phase. This phase allows participants to familiarize themselves with the Concordia environment, make preliminary submissions, and seek clarifications from the organizers.

Aug 31 - Oct 31: Development phase begins. Submissions during this period contribute towards the leaderboard rankings. This phase is crucial for participants to develop and iterate on their solutions.

Oct 31 - Nov 15: Evaluation phase commences. During this phase, detailed feedback is limited to error reports and final scores. The leaderboard remains confidential until the contest at NeurIPS.

Nov 15 - Nov 20: The organizing committee reviews and verifies the results. The top 10 entries, along with selected others, are invited to provide detailed system descriptions.

Nov 20 - Dec 9: Organizers conduct an in-depth analysis of contest results to prepare for the conference.

Dec 3: Deadline for the submission of artifacts and system descriptions from the top entries.

Dec 9 - Dec 14: The contest culminates in a dedicated session at NeurIPS. Winners are announced, prizes awarded, and high-ranking participants, alongside organizers, present insightful findings.

At the time of writing this proposal, the Concordia environment codebase[6] is fully available on GitHub for participant access. We are in the process of finalizing the development and evaluation environments, which are currently in an advanced stage of preparation.

2.4 Competition Promotion and Incentives

To promote participation in the Concordia contest, we will:

- Advertise the contest to hosting platform's AI-engaged user base, leveraging their successful track record in hosting NeurIPS competitions.
- Utilize Cooperative AI Foundation's resources, including blog posts, social media, and the Cooperative AI Summer School, to encourage participation.
- Promote the contest through personal social media accounts, Reddit's r/machinelearning thread, and relevant academic mailing lists.
- Partner with affinity groups like Queer in AI, Women in ML (WiML), LatinX in AI, and Black in AI to improve participation from underrepresented communities.
- Actively promote the contest to academic labs and students, emphasizing Concordia's accessibility and potential for testing latest research without requiring industry-scale resources.

To incentivize participation, we have secured funding for:

- A prize pool of at least $10,000 for top-performing participants.
- Travel grants of at least $10,000 to support underrepresented and under-resourced groups in attending the conference.
- Ongoing discussions with Google DeepMind and Cooperative AI Foundation for $50,000 in compute credits for underrepresented and under-resourced groups.
- Winners and noteworthy submissions will be invited to co-author a joint publication on the contest's impact and lessons learned.

3 Resources

3.1 Resources provided by organizers

Many resources required to run the contest are also already in place: The hosting platform will provide staff for implementation/maintenance of the automated evaluation, supporting/assisting participants during their code submission, and communication/community engagement. Google DeepMind and the Cooperative AI Foundation will provide additional staff to support participants via the contest's forum and office hours. We are in the process of securing compute and other funding for under-resourced and under-represented participants (see Section 2.4), and may attempt to bring on board further partners to assist with this, helping to maximise the accessibility of the contest.

The full team will be available to help run the contest and secure the remaining resources. In particular, Chandler Smith and Rakshit Trivedi will lead the development and implementation of contest baselines and participant starter kit. Joel Leibo and Lewis Hammond will lead in securing further sponsorships for prizes and support for under-represented groups, with support from Jesse Clifton, Akbir Khan, and the rest of the team.

[6] `https://github.com/google-deepmind/concordia`

3.2 Support requested

The contest will have been resolved by NeurIPS 2024. During the conference, we plan to announce the winners, and allocate time for the winners to give a short overview of their solutions. The main support we need from NeurIPS will be the provision of a video-conferencing platform/setting for these presentations and a room for in-person participant gathering and presentations.

3.3 Organizing Team

Our team's commitment to orchestrating a well-structured and competitive contest is significantly bolstered by our collective experience in successfully managing similar large-scale events, including the Melting Pot contest at NeurIPS in 2023. The expertise of the original Melting Pot contest organisers is further enhanced by the addition of several other researchers, including Tim Baarslag, who has extensive experience in leading the Automated Negotiating Agents Competition (ANAC) across multiple iterations at AAAI and other distinguished venues. In addition to the organisers, Sergey Levine has already agreed to serve as an advisor for the contest, and we will bring on board other advisors as and when we believe that their expertise can improve the contest. With the Concordia library already available on GitHub and the specific environments for development and evaluation nearing completion, our experienced team – alongside our collaboration with AI Crowd – lays a solid foundation for an impactful and accessible contest.

Chandler Smith is a current Machine Learning Alignment and Theory Scholar (MATS) supervised by Jesse Clifton. He recently received his Master's in Computer Science from Northeastern University, where he studied AI and multi-agent systems.

Rakshit S. Trivedi is a Postdoctoral Associate in the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT. Prior to that, he was a Postdoctoral Fellow in EconCS at Harvard School of Engineering and Applied Sciences (SEAS). He obtained his PhD from Georgia Institute of Technology, focusing on machine learning for networked and multi-agent systems. He is broadly interested in the development of AI that is capable of learning from human experiences, quickly adapt to evolving human needs and achieve alignment with human values. Through the lens of MARL, he is interested in studying the effectiveness of such an AI in the presence of various social, economic and cultural factors.

Jesse Clifton is a research analyst at the Cooperative AI Foundation and a researcher at the Center on Long-Term Risk, where he is focused on how to improve outcomes of interactions involving AI systems. He is also a PhD student in statistics at North Carolina State University.

Lewis Hammond is based at the University of Oxford where he is a DPhil candidate in computer science. He is also the research director of the Cooperative AI Foundation and an affiliate at the Centre for the Governance of AI. His research concerns safety and cooperation in multi-agent systems and in practice spans game theory, machine learning, and formal methods.

Akbir Khan is a PhD student at University College London and Facebook AI Research, supervised by Prof. Tim Rocktaschel and Prof. Edward Grefenstette. Previously, he was co-founder of Spherical Labs, a startup that developed network detection systems, acquired by Cloudflare. He holds a masters's degree in computer science from University of Cambridge.

Marwa Abdulhai is a PhD student at the Berkeley Artificial Intelligence Research (BAIR) lab at UC Berkeley, advised by Professor Sergey Levine. With a focus on social reinforcement learning and cooperative multi-agent systems, she is delving into the intricacies of social intelligence in AI. Her academic journey includes a Master's degree from MIT, where she engaged with eminent scholars like Jonathan P. How and Luca Carlone.

Alexander (Sasha) Vezhnevets is a staff research scientist at Google DeepMind. He obtained his PhD from ETH Zurich in Machine Learning. He is interested in understanding hierarchical behaviour in humans and machines, computational social construction of reality, cultural evolution and multi-agent systems.

Oliver Slumbers is a PhD student at University College London, supervised by Prof. Jun Wang. His research centres around population / group dynamics in multi-agent systems and the subsequent implications for game-theoretic equilibrium solving. He is currently focusing on the social capabilities

of LLMs within social dilemma frameworks. He holds a master's degree in Computational Statistics and Machine Learning from University College London.

John P. Agapiou is a staff research engineer at Google DeepMind. He obtained his PhD in neuroscience from UCL.

Edgar A. Duéñez-Guzmán is a staff research engineer at Google DeepMind working in the game theory and multi-agent team. He hold a doctorate in computer science from the University of Tennessee, Knoxville with an emphasis on evolutionary theory, social evolution, and large scale computation. His interests are in understanding the underpinning of social interactions and how those interactions contribute to large-scale cooperation and fairness. Edgar is also deeply interested in agency at multiple scales and how cooperation and competition scaffold to produce complexity, resulting in cultural and technological evolution.

Jayd Matyas is a games designer at Google DeepMind where she specializes in the development in multi-agent research environments. She has a background in Industrial Design and Wearable Technology, and has worked on a variety of games across mediums ranging from digital games to augmented reality games and live escape rooms.

Minsuk Chang is a staff research scientist at Google DeepMind. He obtained his PhD from Korea Advanced Institute of Science and Technology (KAIST). He is interested in agents' (in)ability to acquire new skills/knowledge through interaction.

Danny Karmon is a senior research scientist manager at Google Research. His research primarily focuses on modeling and simulating user and personal agent behaviors using synthetic methods to enhance the evaluation and training of personalized agent capabilities. Previously he led NLP research at Microsoft - focusing on the healthcare space.

Natasha Jaques Natasha Jaques is an Assistant Professor of Computer Science and Engineering at the University of Washington, and a Senior Research Scientist at Google DeepMind. Her research focuses on Social Reinforcement Learning in multi-agent and human-AI interactions. She completed her PhD at MIT and post-doc at UC Berkeley. Her work has won various awards, including Best Demo at NeurIPS, an honourable mention for Best Paper at ICML, and the Outstanding PhD Dissertation Award from the Association for the Advancement of Affective Computing.

Dylan Hadfield-Menell is an assistant professor on the faculty of Artificial Intelligence and Decision-Making in the EECS Department and Computer Science and Artificial Intelligence Laboratory (CSAIL) at the Massachusetts Institute of Technology (MIT). His research focuses on the problem of agent alignment: the challenge of identifying behaviors that are consistent with the goals of another actor or group of actors. His work aims to identify algorithmic solutions to alignment problems that arise from groups of AI systems, principal-agent pairs (i.e., human-robot teams), and societal oversight of ML systems.

Tim Baarslag is a Senior Researcher leading the Intelligent and Autonomous System group at CWI (the Dutch research institute for Mathematics and Computer Science) and an Associate Professor at Utrecht University. He is a Visiting Scholar at Massachusetts Institute of Technology (MIT), a Visiting Associate Professor at Nagoya University of Technology and a Visiting Fellow at the University of Southampton. Tim is extremely experienced in running automated negotiating agents competitions.

Joel Z. Leibo is a senior staff research scientist at Google DeepMind. He obtained his PhD from MIT where he studied computational neuroscience and machine learning. Joel was one of the first researchers to join DeepMind, starting as an intern in 2010, and then joining full time after finishing his PhD in 2013. He is interested in reverse engineering human biological and cultural evolution to inform the development of artificial intelligence that is simultaneously human-like and human-compatible.

References

[1] J. Z. Leibo, E. A. Dueñez-Guzman, A. Vezhnevets, J. P. Agapiou, P. Sunehag, R. Koster, J. Matyas, C. Beattie, I. Mordatch, and T. Graepel, "Scalable evaluation of multi-agent reinforcement learning with melting pot," in International conference on machine learning, pp. 6187–6199, PMLR, 2021.

[2] J. P. Agapiou, A. S. Vezhnevets, E. A. Duéñez-Guzmán, J. Matyas, Y. Mao, P. Sunehag, R. Köster, U. Madhushani, K. Kopparapu, R. Comanescu, D. Strouse, M. B. Johanson, S. Singh, J. Haas, I. Mordatch, D. Mobbs, and J. Z. Leibo, "Melting pot 2.0," arXiv preprint arXiv:2211.13746, 2022.

[3] A. S. Vezhnevets, J. P. Agapiou, A. Aharon, R. Ziv, J. Matyas, E. A. Duéñez-Guzmán, W. A. Cunningham, S. Osindero, D. Karmon, and J. Z. Leibo, "Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia," arXiv preprint arXiv:2312.03664, 2023.

[4] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," 2023.

[5] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel, "Cooperative ai: machines must learn to find common ground," Nature, vol. 593, no. 7857, pp. 33–36, 2021.

[6] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel, "Open problems in cooperative ai," arXiv preprint arXiv:2012.08630, 2020.

[7] E. Ostrom, R. Gardner, J. Walker, and J. Walker, Rules, games, and common-pool resources. University of Michigan press, 1994.

[8] T. C. Schelling, "An essay on bargaining," The American Economic Review, vol. 46, no. 3, pp. 281–306, 1956.

[9] S. Legg and M. Hutter, "Universal intelligence: A definition of machine intelligence," Minds and machines, vol. 17, pp. 391–444, 2007.

[10] Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, and X. He, "Exploring large language model based intelligent agents: Definitions, methods, and prospects," 2024.

[11] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen, "A survey on large language model based autonomous agents," 2024.

[12] J. Ruan, Y. Chen, B. Zhang, Z. Xu, T. Bao, G. Du, S. Shi, H. Mao, Z. Li, X. Zeng, and R. Zhao, "Tptu: Large language model-based ai agents for task planning and tool usage," 2023.

[13] H. Yang, S. Yue, and Y. He, "Auto-gpt for online decision making: Benchmarks and additional opinions," 2023.

[14] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, U. Alon, and G. Neubig, "Webarena: A realistic web environment for building autonomous agents," 2023.

[15] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, J. B. Tenenbaum, T. Shu, and C. Gan, "Building cooperative embodied agents modularly with large language models," 2024.

[16] G. Gygax and D. Cook, The Dungeon Master Guide, No. 2100, 2nd Edition (Advanced Dungeons and Dragons). TSR, Inc, 1989.

[17] J. Perolat, J. Z. Leibo, V. Zambaldi, C. Beattie, K. Tuyls, and T. Graepel, "A multi-agent reinforcement learning model of common-pool resource appropriation," Advances in neural information processing systems, vol. 30, 2017.

[18] E. Ostrom, Governing the commons: The evolution of institutions for collective action. Cambridge university press, 1990.

[19] J. M. Acheson, Capturing the commons: devising institutions to manage the Maine lobster industry. Upne, 2003.

[20] H. L. A. Hart, The concept of law. Oxford University Press, 1961/2012.

[21] E. Ostrom, Understanding institutional diversity. Princeton University Press, 2009.

[22] G. K. Hadfield, Rules for a flat world: Why humans invented law and how to reinvent it for a complex global economy. Oxford University Press, 2017.

[23] O. Linderborg, "The place of herodotus' constitutional debate in the history of political ideas and the emergence of classical social theory," Akropolis: Journal of Hellenic Studies, vol. 3, no. 1, pp. 5–28, 2019.