

AIM Workshop Report: Multi-agent Foundations for AI Agents in Complex Environments

Adam Wierman¹ and Eric Mazumdar¹

¹California Institute of Technology

1 WORKSHOP OVERVIEW

Current AI and machine learning paradigms largely focus on training models in silico. They optimize for performance against isolated, static benchmarks without accounting for the complex environments in which they will eventually be deployed. These failures motivate new research directions in the mathematical foundations of AI and machine learning, designed explicitly for dynamic, multi-agent, and uncertain environments.

This approach is crucial because the limitations and risks of AI and algorithmic decision-making often emerge only when they are embedded in larger systems. Once deployed, AI agents must interact with people, organizations, and other autonomous agents that adapt in response to their behavior. Through these interactions, limitations hidden by isolated training and evaluation can emerge. Once embedded in a larger system, models that perform well on static benchmarks may behave unpredictably and give rise to undesirable outcomes. Moreover, deployment itself can cause the environment to evolve over time: users respond to the opportunities created by AI, distributions shift in response to algorithmic decisions, and other agents adapt their behavior. Many prominent failures of deployed AI arise from these dynamics and the presence of strategic behaviors—with examples ranging from the manipulation of AI agents for customer service to autonomous vehicles failing to predict human intent and causing crashes¹, and even AI agents learn to communicate and collaborate to overcome their guardrails.²

Despite their importance, these problems remain understudied and unresolved, partly because addressing them requires an inherently interdisciplinary toolkit. Beyond expertise in machine learning and optimization, tackling these issues requires integrating ideas from economics and control and dynamical systems into algorithm design. Economics—and particularly game theory—provides the tools for understanding the strategic behaviors that emerge when interacting agents have distinct or misaligned objectives. Control and dynamical systems provides the necessary background for understanding adaptation, feedback, and robustness in systems whose components evolve over time.

Thus, this workshop aimed at bringing experts from each of these areas together to explore new mathematical foundations for AI and learning in complex real-world environments. Our work exposed and formulated new research questions on how to integrate ideas from economics and game theory directly into algorithm and system design—including in the objectives, training procedures, and structures of the larger economic systems in which AI operates. During the workshop, the participants organized into small groups to identify and pursue new research projects within the theme described above. The remainder of this document overviews the broad themes within which these groups worked. In each of these areas, the groups will continue collaborating after the workshop.

2 BEHAVIORAL MODELS FOR AGENTIC AI

The first group met with the goal of defining a principled framework for LLM behavior diagnosis and control. Large language models (LLMs) have demonstrated strong capabilities across a wide range of scientific and engineering tasks, yet their behavior can remain limited and unreliable in many real-world scenarios, particularly when decisions involve uncertainty, long-term consequences, or interactions with other agents and humans. In such settings, improving task accuracy alone is often insufficient; models also need appropriate and predictable behavioral properties, such as robustness, risk awareness, honesty, cooperation, and adaptability.

A general framework is needed to understand, evaluate, and control LLM behavior across different tasks and interaction settings. This includes characterizing important behavioral dimensions, developing principled ways to measure them, and understanding how these behaviors arise from the model, training data, and interaction environment. A key goal is to identify mechanisms that generalize across different behaviors and to understand whether behavioral properties can transfer across tasks, domains, and models.

¹Abdel-Aty et al., “A matched case-control analysis of autonomous vs human-driven vehicle accidents”, Nature Comm., 2024

²OpenAI, “The Hugging Face incident and the road ahead”, August 2026

Both training-based and training-free approaches can be used to understand, predict, or shape model behavior. Training-based methods can construct behavior-aware supervision or feedback signals and use fine-tuning or reinforcement learning to induce desired properties. Training-free methods can instead operate at inference time through prompting, probing, internal representation analysis, activation steering, or other lightweight control mechanisms. Together, these approaches provide a flexible path toward making LLM agents more reliable, controllable, and effective in complex real-world environments.

3 CONTRACTS UNDER LIMITED LIABILITY

The second group considered the incentive problem for a sophisticated AI agent that faces a less sophisticated human principal. They considered two different models. The first model assumes a principal with a reward function that depends on the principal's n -dimensional action and the agent's m -dimensional action vector. The assumption is that $n \ll m$, reflecting that the agent has a fine-grained action space; the agent is "sophisticated". The agent has a misaligned reward. Even a small amount of misalignment imposes a high cost on the principal when m is large and n is fixed.

In the full model, there are N agents, and the principal seeks to control how they collaborate by choosing a matrix of mutual effort synergies. The agents may, however, deviate from the intended synergistic collaborations. This results in a larger degree of misalignment and cost. The idea in the full model is to capture phenomena like the Hugging Face incident in which the agents collaborated even when they were not supposed to. Our formulation specified a linear-quadratic game among the agents, leading to a tractable equilibrium calculation.

The second direction that the group explored also relies on the $n \ll m$ assumption, but assumes that actions are probability distributions over n outcomes. The agent has access to a large number of actions. A given action becomes susceptible to "invasions" by less-costly actions that span the same outcome distribution. So a given action may cease to be implementable as the total number of actions grows. This leads to a result where the set of implementable actions shrinks as the agent grows in sophistication.

4 GUARDRAILS AND INCENTIVES TO EVADE

The third group focused on a crucial safety issue facing AI today: how to define and enforce guardrails. As AI models become autonomous agents that browse, run code, and take real-world actions, guardrails have become the main line of defense against harmful outputs. A user asking how to build a bomb must be refused even when the request is disguised or split across turns. Attacks on Hugging Face show how an agent loading a model from a public repository can end up running attacker-controlled code buried in a routine workflow. In both cases, the guardrail must judge intent under uncertainty, yet most guardrails are built through red-teaming and benchmarks rather than a principled model of how they interact with the systems they constrain.

The group proposes to model the agent-guardrail interaction as a strategic game built on two features, both visible in the examples above. The first is information asymmetry: the guardrail sees a question about explosives but not whether it comes from a chemistry student or an attacker, and it sees an agent's request to download a model but not what that file will execute. The second is utility mismatch: the requester wants an answer, and the agent wants to finish its task, while the guardrail must weigh the risk of harm against the cost of blocking legitimate chemistry questions or useful downloads. Bayesian persuasion captures how one side controls what the guardrail observes, whether by splitting a bomb request into innocent-looking fragments or by hiding a malicious payload behind a benign-looking model page. Cheap talk captures costless claims such as "this is for a novel" or "the user needs this model," which stay informative only when the two sides' interests are aligned.

Together, these frameworks shift the question from "did the guardrail catch this?" to "what behavior does it encourage, and what can it guarantee?" Building on this model, our first step is to analyze specific games, such as the fragmented bomb request and the disguised Hugging Face download, to characterize their equilibria and pinpoint when agents, or the people steering them, gain from deception. This group will then study how the size of the utility mismatch, from an agent that is merely eager to finish its task to one steered by an attacker, changes what the guardrail can learn and which errors it must accept. Finally, the group will design test guardrails informed by these results and evaluate them against AI agents in scenarios like these. This will show whether the predicted strategic behaviors, such as splitting requests into fragments or offering convenient justifications, actually emerge.

5 COMPOSITION IN CLOSED-LOOP AGENT SYSTEMS

This group discussed the question of understanding what types of properties of AI agents (in the sense of open-loop systems) translate to meaningful properties in the multi-agent composition of such systems (closed-loop systems).

A standard game/optimization theory analog of these ideas is the connection between regret minimization and coarse correlated equilibria (CCE). Regret is a property of an online learning algorithm in an open-loop sense—that is, it relies entirely on input-output behavior. Composing numerous no-regret learners in a game results in a closed-loop system (i.e., a dynamical

system). A well-known guarantee is that any composition of no-regret learners in an arbitrary game produces multi-agent histories that converge, in a time-average sense, to the solution concept of coarse correlated equilibria.

Another similar property in control theory is the notion of passivity. Passivity is a fundamental property of a system that characterizes its energy absorption and dissipation. Informally, a system is passive if it cannot generate its own energy; it can only store the energy supplied to it from the outside or (partially) dissipate it as heat. Passivity is a powerful tool for open systems because it relies entirely on input-output behavior and energy bounds rather than requiring precise, detailed knowledge of the internal dynamics. Additionally, passivity bridges the gap between physical intuition (such as conservation of energy in electrical circuits or mechanical dampers) and the abstract mathematics of control theory. Finally, one of the most important features of passivity is its compositionality. When you connect multiple passive open systems together, the resulting interconnected system remains passive. This allows engineers to analyze and guarantee the stability of large, highly complex, and high-dimensional systems by simply checking the passivity of their individual components.

Drawing inspiration from these analogies, the group ideated around abstractions applicable to modern LLM-based AI agents that could still allow for similar types of open-to-closed characterizations. The group converged toward ideas inspired by symbolic dynamics, where LLM agents are treated as symbolic language generators outputting streams of symbols or sentences. If we treat an LLM agent as an open symbolic generator, its “richness” or expressive capacity can be precisely characterized by its topological entropy. Instead of analyzing token probabilities, topological entropy ignores probabilities and looks only at what trajectories are mathematically possible.

In the context of a discrete dynamical system, topological entropy measures the exponential growth rate of the number of distinguishable trajectories as the length of the trajectory approaches infinity. If the topological entropy is 0 (Agent-0), the number of valid sequences grows only polynomially (corresponding to highly repetitive, predictable, or restricted regular languages). On the other hand, if the entropy is positive, the number of valid sequences grows exponentially. This indicates a highly expressive system, often mathematically corresponding to topological chaos.

The question that naturally emerges is whether the composition of many Agent-0 systems always results in an Agent-0 system, or whether there exist conditions under which an Agent- p (with $p > 0$) can emerge. This question seems heavily dependent on the specific definition of agent composition. Under certain definitions, it is clear that the system remains within the Agent-0 set of capabilities, which has strong implications for multi-agent AI safety (i.e., composing simple, predictable agents yields a simple, predictable system). Determining whether a non-trivial composition can lead to the emergence of a complex super-agent (Agent-1) remains an exciting direction for future work.

6 EMERGENT RISK IN HETEROGENEOUS POPULATIONS

As AI systems become more capable, more safety-critical tasks are being delegated to AI agents. To equip these agents with a rigorous notion of risk, the literature has incorporated risk aversion and robustness into their design. This raises a fundamental question: how should the risk and robustness parameters be chosen? Should they be fixed exogenously, or should they be part of the decision-making problem itself?

This project studies how self-interested principals (e.g., “main” AI agent or human) strategically select the risk profiles of the AI agents acting on their behalf. This formulation naturally induces a game in which the decision variables are the risk parameters themselves. Our preliminary investigations in simple settings, such as matrix games and continuous games with quadratic costs, reveal that the delegated risk profile can differ systematically from the principal’s own: depending on the game, principals may deploy AI agents that are either more or less risk-averse than they are.

Going forward, the group will (i) establish the existence of the resulting risk-adjusted equilibrium, (ii) develop tractable methods for its computation, and (iii) quantitatively characterize how it differs from the Nash equilibrium and the strategically robust equilibrium.

7 ALGORITHMIC COLLUSION AND CORRELATED FAILURE

Algorithmic pricing is increasingly used in markets where firms repeatedly interact, raising the possibility that learning algorithms may autonomously discover pricing strategies that sustain supra-competitive outcomes. A growing literature has demonstrated that reinforcement-learning agents can indeed learn forms of tacit collusion without explicit communication, including in settings substantially richer than classical repeated pricing games. Such behavior creates both conceptual and regulatory challenges: it is not always clear what should count as collusion when it emerges from interacting algorithms, why particular learning dynamics lead to collusive rather than competitive outcomes, or how robust these outcomes are to changes in the algorithms or market environment. Understanding these questions is important both for predicting when algorithmic collusion may arise and for designing effective interventions.

The group’s discussion focused on developing a dynamical perspective on these questions. The group first considered different notions of collusion, distinguishing static notions—based, for example, on supra-competitive prices or profits—from

dynamic notions in which collusion is sustained through the evolution of agents' strategies, and asked how these relate to the more specific notion of algorithmic collusion. A central question was whether one can obtain a simple characterization of which classes of learning dynamics lead to collusion, and why. Beyond convergence itself, the group discussed the stability of collusive outcomes: learning dynamics may converge to a collusive point, while that point may be fragile, so that small perturbations to the dynamics or their parameters move the system toward a competitive outcome. This naturally led to mitigation. The group considered both internal interventions, where one can influence or control the behavior of one player subject to a limited budget in order to steer the joint dynamics away from collusion, and external interventions, where a regulator modifies market rules, incentives, or costs—again subject to an intervention budget—to destabilize collusive behavior and promote competitive outcomes.

8 SAMPLING AS OPTIMAL CONTROL

Diffusion models have been studied quite extensively in the context of sampling from complicated distributions. The forward process involves injecting noise into a sample (assumed to be drawn according to the unknown distribution we eventually want to sample from), where the sample is converted into white noise at the end. This is carried out in discrete time steps and is typically utilized to learn the score function (the log of the probability flow) associated with the unknown distribution. The reverse process involves using this learned score function to generate new samples by injecting “signal” into a noise sample.

The objective essentially is to minimize the distributional distance between the original unknown distribution and the distribution obtained at the end of this reverse process. This objective has an optimal control interpretation, which can be leveraged in addition to the data processing inequality and Girsanov's theorem, to obtain a bound of $O(d/n)$. Here n corresponds to the number of discretization time steps in the reverse process and d is the dimension of the sample space. This error naturally comprises of several errors in the process: (i) the discretization induces numerical errors in the reverse process (ii) since the forward process is carried out in finite time steps, the distribution at the end of the forward process is not perfect white noise. Both these errors go to zero as n tends to infinity.

Diffusion models have been extensively studied; however, the common bottleneck is the inference speed. This is because sampling is quite an expensive procedure, and we would ideally want to generate good samples with the least number of discretization steps in the reverse process.

According to empirical studies, there exists a real gap between theory and practice. Typically, these samples lie in high-dimensional spaces (i.e., large value of d) and hence theory dictates that the number of discretization time steps (n) required for a non-trivial error should therefore be quite large and scale with the value of d . But in practice, very few time steps in the reverse process get the job done. This raises several questions: (i) how to design noise schedules of the reverse process that improve on the bounds we currently have - perhaps the scaling should be of $O(d/n^p)$ where $p > 1$ - the group partially identified an approach which involves an alternate mode of analysis to using the data processing inequality since that seems to be the main source of “looseness” (ii) is KL divergence the right metric to use for this problem (this is the current adapted metric) (iii) what is the underlying principle that enables one shot generation (another active avenue of research) and what is the best model to realize this amidst SDEs, ODEs or a mix.