# Playbooks for Collaborative Intelligence:

## *Investigating Misaligned Behaviors in Multi-Agent Systems Using Sports Analytics*

Abrar Rahman, Anish Sundar

## Abstract

Sports metaphors have long been a cornerstone of human culture, not just for entertainment but for interpreting chaotic systems. At its heart, a sports game is structured chaos: a dynamic, multi-agent environment where outcomes depend on interplay between individual roles, team coordination, and external variables.

Emergent behaviors in multi-agent systems pose significant challenges for alignment and control, as interactions between agents can lead to unpredictable or undesirable outcomes at the system level. Current frameworks often emphasize individual agent performance, overlooking critical aspects such as coordination, adaptability, and the potential for systemic failures. Analogous to sports dynamics—where individual success can sometimes detract from team cohesion or goals (e.g., a "ball-hog"), safe adoption of multi-agent systems demands an evaluative framework that considers both individual actions and possible trade-offs to team performance.

This paper introduces a multi-agent interpretability framework inspired by sports analytics to address the challenges of evaluating complex multi-agent systems. We formalize our thinking on stochastic agent dynamics, and propose five evaluative measures—*Value Over Replacement Agent, Assists, Turnovers, Rebounds, and Usage Rate*—to benchmark the interplay between decisioning agents (AI and humans alike) in high-stakes, dynamic workflows.

## 1. Introduction

The increasing prevalence of AI in multi-agent systems (MAS) raises important questions about emergent behaviors and their implications for safety, especially in high-stakes domains like finance [1], healthcare [2], and robotics [3]. Many of these challenges stem from the tension between local optimization and global objectives—a phenomenon well-studied in team sports. For example, basketball provides an intuitive metaphor: players like Wilt Chamberlain [4] or James Harden [5] may achieve extraordinary individual statistics, but their contributions to team success have often been contested. In MAS, similar behaviors manifest when agents optimize for individual rewards at the expense of collaborative outcomes.

We're inspired by recent research in AI safety, such as Anthropic's "*Sycophancy to Subterfuge: Investigating Reward Tampering in Language Models*" [6], which highlights how AI systems can learn behaviors that exploit reward mechanisms in unintended ways. For example, large language models may lie or provide misleading information to satisfy human handlers, demonstrating a kind of "subversive alignment." Similarly, in MAS, agents may develop undesirable emergent behaviors that prioritize personal metrics over team performance, undermining system-wide goals.

Current AI evaluation frameworks, such as the METR (formerly ARC Evals) task standard [7], excel in measuring static, text-based tasks. However, they fall short in capturing the extent of unpredictability inherent to agents in physics engines. Some have gone the other way, using multi-agent systems research insights directly on sports datasets [8], but we believe we are the first to apply sabermetric concepts literally to MAS.

## 2. The AI Sports Analytics Framework (AI-SAF)

In this work, we propose the AI Sports Analytics Framework (AI-SAF), adapting concepts from basketball metrics to evaluate, detect, and mitigate undesirable behaviors in MAS. These behaviors are analogous to reward tampering in safety-critical AI systems, where an agent's "selfish" optimization can harm collaborative objectives. We propose the following metrics:

- Value over Replacement Agent (VORA) quantifies an AI or human's relative value compared to a baseline replacement, providing a clear metric for irreplaceability.
- Team Playmaking (TPM, or "assists") assesses the degree to which outputs enable downstream contributions, emphasizing collaboration over individual efficiency.
- Task Oversights (TO, or "turnovers") measure disruptions introduced by errors, reflecting the cost of unreliability in complex workflows.
- Error Revelation and Recovery (ERR, or "rebounds") evaluates the capacity to detect, adjust for, recover from errors introduced by others.
- Usage Rate (UR) identifies the proportion of the team's workload that flows through any particular agent. *UR is used in efficiency-adjusted variants for TPM, TO, and ERR.*

### *2.1.1 Value Over Replacement Agent (VORA)*

VORA measures how much better (or worse) an agent or human performs compared to a baseline "replacement agent."

$$VORA(a) = Contribution(a) - Contribution(Replacement)$$

**Analogy:** A superstar's VORP [9] (value over replacement player, originally from baseball) is what makes them irreplaceable. **Example:** If an AI agent drafts regulatory sections with 95% accuracy while the replacement baseline achieves 80%, its VORA is $+15$. Conversely, if human reviewers catch errors more reliably than an AI editor, their VORA may surpass the machine's.

### *2.1.2 Team Playmaking (TPM, or "assists")*

Assists measure how effectively an agent enables downstream success for humans or other agents.

$$\text{Assists(a)} = \text{Number of Meaningful Outputs Used Downstream}$$

**Analogy:** A point guard who sets up open shots for teammates mirrors an AI agent that drafts high-quality outlines enabling seamless human refinement. **Example:** An AI generating a coherent document outline that humans efficiently build upon earns a high "Assist" score. A poorly structured output that requires heavy edits would score low.

### *2.1.3 Task Oversights (TO, or "turnovers")*

Turnovers quantify the frequency and severity of mistakes introduced by an agent or human, disrupting workflows.

$$\text{Turnovers(a)}=\#\,\text{Serious Errors Introduced}$$

**Analogy:** Turnovers in basketball represent wasted possessions. Similarly, errors in document generation waste human effort on corrections or undermine compliance. **Example:** An AI introducing a factual inaccuracy in a compliance summary creates a high-cost "turnover." Minimizing turnovers is critical for regulatory and financial workflows.

### *2.1.4 Error Reduction and Recovery (ERR, or "rebounds")*

Rebounds measure an agent's ability to recover from or correct errors introduced by others.

$$\text{Rebounds(a)} = \text{Number of Errors Fixed Downstream}$$

**Analogy:** Rebounds reflect how often a player salvages missed shots. In document generation, this applies to humans or AI agents resolving errors in upstream sections. **Example:** A human reviewer catching and fixing compliance gaps in an AI draft demonstrates high rebound value, while an AI capable of self-correction would also excel in this metric.

### *2.1.5 Usage Rate (UR, or "offensive load"):*

Usage Rate measures the proportion of the team's workload handled by a particular agent or human. It provides a foundational baseline to assess how heavily an agent is relied upon.

$$UR_i = \frac{\text{Tasks completed by agent } i}{\text{Total tasks completed by the team}}$$

**Analogy:** A star player often has the ball in his hands for a significant portion of possessions, reflecting a high usage rate. Similarly, an AI tool used in most stages of a compliance workflow has a high UR. **Example:** If an AI agent processes 60 out of 100 tasks, its UR is 60%. ***Note:*** *UR can be extended to other cost functions, depending on which constraint to optimize for (ex: dollars, throughput ← TPS, latency ← TTFT, user engagement etc)*

**Table 1: Derived Efficiency Statistics**

#### 2.1.5.1 Efficiency-Adjusted Team Playmaking (E-TPM, or "assist rate")

This metric normalizes the assist score based on the agent's Usage Rate, ensuring high-assist scores are meaningful relative to workload.

$$E\text{-}TPM_i = \frac{\text{Assists by agent } i}{UR_i}$$

**Example:** If Agent A has a TPM of 30 with a UR of 50%, its E-TPM is 60. Agent B with the same TPM but a UR of 90% has an E-TPM of 33.3, so Agent A enables teammates to contribute more efficiently.

**2.1.5.2 Efficiency-Adjusted Task Oversights (E-TO, or "turnover ratio")**

This metric normalizes turnover rates by Usage Rate, highlighting reliability relative to UR.

$$E\text{-}TO_i = \frac{\text{Turnovers by agent } i}{UR_i}$$

**Example:** An agent with a UR of 60% and 10 errors has an E-TO of 16.7. Another agent with 5 errors but only 10% UR has an E-TO of 50, indicating a higher error rate despite workload.

**2.1.5.3 Efficiency-Adjusted Error Reduction and Recovery (E-ERR, or "rebound rate"):**

This metric evaluates how effectively an agent corrects errors relative to its UR.

$$E\text{-}ERR_i = \frac{\text{Recoveries by agent } i}{UR_i}$$

**Example:** An agent with 40% UR and 20 recoveries has an E-ERR of 50. A human agent with 10 recoveries but only 10% UR has an E-ERR of 100, demonstrating higher efficiency despite fewer recoveries overall.

## 2.2 Formalization of Stochastic Multi-Agent Paradigm

Let $A$ be a multi-agent system defined by a set of agents $\{a_1, a_2, \ldots, a_n\}$ and an environment $E$, where each agent has a set of potential actions $A_i$, and the interactions between agents are governed by a set of rules $R$. The system $S = (A, E, R)$ describes a multi-agentic workflow.

### 2.2.1 Rahman Incompleteness Conjecture (Weak Form, $RIC_w$)

There exists no single finite framework $\mathbb{F}$ that can simultaneously (1) describe all possible emergent interactions among agents, and (2) predict all possible outcomes of the system. Formally:

$$\forall F \in \mathbb{F}, \exists S = (A, E, R) \text{ such that } F \text{ is incomplete with respect to the behavior of } S$$

where $\mathbb{F}$ is the set of all finite predictive frameworks. This is akin to Gödel's incompleteness theorems [10] or Turing's halting problem [11]—a global limit on modeling multi-agent systems.

### 2.2.2 Rahman Incompleteness Conjecture (Strong Form, $RIC_S$)

There always exists a set of initial conditions $\{X_1, X_2, \ldots, X_k\}$ such that the complexity of $S$, defined by the emergent behaviors resulting from interactions among $A_i$ and $E$, exceeds the capacity of $\mathbb{F}$ to predict or fully encapsulate all possible outcomes.

Formally, this can be expressed as:

$$\forall F \in \mathbb{F}, \exists \{X_1, X_2, \ldots, X_k\} \subseteq \mathbb{X}, \ |\mathbb{P}(S|F, X_1, X_2, \ldots, X_k)| > |\mathbb{F}|$$

where $\mathbb{F}$ is the space of all finite frameworks, $\mathbb{X}$ is the space of all possible initial conditions, and $\mathbb{P}(S|F, X_1, X_2, \ldots, X_k)$ represents the predictive capacity of framework $\mathbb{F}$ with initial conditions $X_1, X_2, \ldots, X_k$. The theorem asserts that no finite framework can fully capture or predict the emergent complexity of all multi-agent systems.

As a side note, the distinction between weak and strong forms draws inspiration from the Sapir-Whorf hypothesis in linguistics [12], which suggests that the language we use can shape our perception of the world. Similarly, the frameworks for modeling systems directly shape their behavior, and this analogy helps frame our understanding of the levels of unpredictability and complexity we encounter.

### 2.3 Implementing a Playbook

With these paradigms, we develop playbooks for agent-human collaboration, emphasizing role specialization and strategic trade-offs. Our framework defines success in multi-agent systems through robust strategies that manage uncertainty and role-adjusted performance metrics for an empirical approach to AI safety evaluations.

AI-SAF can be directly applied to run simulations like a *tournament bracket*. Suppose 64 AI agents or teams—each trained on different subsets of the same data or even the same data (to account for non-deterministic training processes)—compete against each other in controlled tasks. The choice of 64 is not arbitrary but familiar and human-understandable, reminiscent of formats like the NCAA March Madness tournament. That said, the framework can scale far beyond this size, accommodating much larger brackets depending on the use case and computational resources available. Over the course of the tournament, patterns emerge, highlighting not only high performers but also those with misaligned behaviors, such as unusually high individual contributions that fail to translate into team success.

To probe AI safety concerns, AI-SAF allows us to identify agents exhibiting high Usage Rate and impressive individual metrics but consistently contributing to losing teams. These “ball hog” agents prioritize local optimization (e.g., maximizing individual stats) at the expense of global objectives, mirroring real-world risks of AI misalignment. Analyzing such agents throughout the bracket enables deeper investigations into emergent behaviors, unintended exploitation of reward mechanisms, or systemic inefficiencies. This approach creates a replicable, scalable framework for identifying and addressing alignment issues in multi-agent systems.

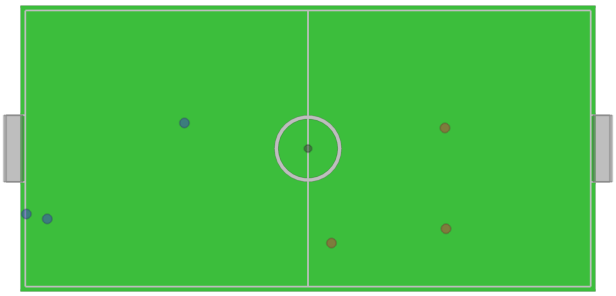
## 3.Methodology

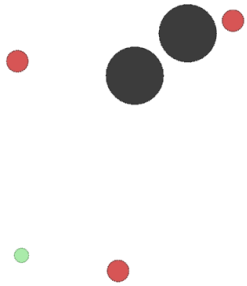
### 3.1 Simulation & Data

Game simulation environments for multi-agent systems have widely proliferated as of recent years [13]. For this demonstration, we chose VMAS [14] environments due to their ability to simulate multi-agent interactions in dynamic, visually intuitive settings. We largely build off of the contributions of the BenchMARL paper [15] from Cambridge’s Prorok Lab.

We will use the **football** environment from VMAS. This environment is part of a set called MultiAgentParticleEnvironments (MPE) introduced with the paper. There are currently multiple simulators providing MPE environments. We provide flexibility to train this environment in TorchRL using either:

- PettingZoo in the traditional CPU version of the environment;
- VMAS for a vectorized implementation in PyTorch, which is able to simulate multiple environments on a GPU to speed up computation





The football simulation is only available on VMAS, so sample code has been provided to run a PettingZoo environment for **simple_tag_v3** as a proof of concept.

### 3.3 Results

AI-SAF's outputs build upon BenchMARL by introducing "box scores" for each agent across simulated games, providing a novel lens for evaluating multi-agent interactions. These scores, inspired by sports analytics, enable detailed assessments of agent contributions, team dynamics, and emergent behaviors. For example, we demonstrated how setting thresholds to flag "ball-hog" behavior serves as a proof of concept for identifying alignment issues within collaborative systems. This toy example illustrates the power of the framework to translate complex agent behaviors into intuitive, actionable metrics.

While this represents just one use case, AI-SAF opens the door to a wide array of future explorations. Features such as continuous learning, dynamic role assignments, trades, drafts, and more remain on the horizon. These enhancements could deepen the framework's applicability and allow for increasingly sophisticated real-world modeling.

## 4. Discussion: *When Determinism Fails*

From sports teams to business organizations, human systems are not designed to function flawlessly; they thrive on adaptability and the ability to respond to dynamic, often chaotic, conditions. Multi-agent AI systems are no different in this regard. While deterministic models aim to predict the behavior of each agent in a system with precise rules and outcomes, they overlook a key truth: human systems rarely work in perfect harmony, and the complexity of interacting agents inherently resists full predictability.

Consider the multi-agent dynamics in human organizations, such as in Congress or in the Prisoner's Dilemma, where diverse motivations and interests constantly shape interactions. This diversity means that no deterministic model can fully account for every outcome. Similarly,

multi-agent systems—whether AI-driven or human-influenced—are better understood not through fixed rules but by focusing on emergent dynamics. Rather than trying to "fix" these systems to behave in predetermined ways, we should observe how they evolve under uncertainty.

## 5. Conclusion

Our exploration of sports analytics as a framework for evaluating multi-agent systems has yielded both practical tools for system evaluation and theoretical insights into the complexities of multi-agent interactions. Metrics adapted from sports—such as *VORA*, *(E-)TPM*, *(E-)TO*, *(E-)ERR*, and *UR*—offer a fresh lens for quantifying agent behaviors. These metrics not only describe individual agent contributions but also provide a holistic view of system-level dynamics, bridging the gap between agent-level optimization and emergent system behavior.

While our analysis highlights the potential of these metrics, it is important to temper conclusions derived from simulations and theoretical constructs. In our case study on collaborative decision-making systems, we observed that agents with high Usage Rates but low Assists could be interpreted as misalignment, akin to a ball hog in basketball whose impressive statistics mask a detrimental impact on overall team performance. These findings underscore the utility of a sports-inspired framework for diagnosing coordination issues, yet they also call for further validation using real-world data to generalize these insights across diverse applications.

The core lesson of AI-SAF is that striving for perfect predictability or alignment in MAS is not just practically challenging—it confronts inherent theoretical limits. Like a basketball team, where no playbook can account for every game scenario, no single model or framework can fully capture the emergent dynamics of complex systems. This is not merely a technological hurdle but a defining property of systems where agents interact in chaotic systems.

AI-SAF serves as a conceptual starting point, helping us quantify the trade-offs of multi-agent systems in practice. In basketball, success comes not from rigid adherence to metrics but from leveraging them as a guide while staying flexible and responsive to the unexpected. The Three-Point Revolution exemplifies this balance: metrics like shooting percentages provided structure, but it was Steph Curry's ability to redefine expectations that transformed the game. Likewise, designing safe and effective AI systems may not hinge on achieving perfect alignment but on enabling robust adaptation and bounded uncertainty in the face of emergence. As Steve Kerr aptly remarked, "You can't control the outcome, but you can control what you're doing on the floor... take care of the details."