Quantifying Misalignment Between Agents: Towards a Sociotechnical Understanding of Alignment

Aidan Kierans[1], **Avijit Ghosh**[2,1], **Hananel Hazan**[3], **Shiri Dori-Hacohen**[1]
[1]University of Connecticut
[2]Hugging Face
[3]Tufts University
aidan.kierans@uconn.edu, avijit.g@uconn.edu, hananel@hazan.org.il, shiridh@uconn.edu

Abstract

Existing work on the alignment problem has focused mainly on (1) qualitative descriptions of the alignment problem; (2) attempting to align AI actions with human interests by focusing on value specification and learning; and/or (3) focusing on a single agent or on humanity as a monolith. Recent sociotechnical approaches highlight the need to understand complex misalignment among multiple human and AI agents. We address this gap by adapting a computational social science model of human contention to the alignment problem. Our model quantifies misalignment in large, diverse agent groups with potentially conflicting goals across various problem areas. Misalignment scores in our framework depend on the observed agent population, the domain in question, and conflict between agents' weighted preferences. Through simulations, we demonstrate how our model captures intuitive aspects of misalignment across different scenarios. We then apply our model to two case studies, including an autonomous vehicle setting, showcasing its practical utility. Our approach offers enhanced explanatory power for complex sociotechnical environments and could inform the design of more aligned AI systems in real-world applications.

Code —
https://github.com/RIET-lab/quantifying-misalignment
Appendices — https://arxiv.org/pdf/2406.04231

1 Introduction

In recent years, growing concerns have emerged about the AI alignment problem (e.g. Russell 2019; Christian 2020). Russell (Russell 2019) made the bold yet compelling claim that social media AI is already misaligned with humanity (e.g. through extensive disinformation spread). As agentic AI systems are growing in importance and increasingly deployed (Kapoor et al. 2024), concerns about misalignment of agent-based systems have grown (Chan et al. 2023; Gabriel et al. 2024). Yet with few exceptions (e.g. Critch and Krueger 2020), much prior work focuses on single AI settings (Ji et al. 2024), on value specification and learning (Critch and Krueger 2020; Leike et al. 2018), or centers technosolutionist fixes. A healthier alternative would set a sociotechnical lens (Lazar and Nelson 2023) on multi-agent settings (Gabriel et al. 2024), which may include a complex mix of AI agents and humans, potentially holding a dizzying array of competing and often conflicting goals. Critch and Krueger posed the question, "where could one draw the threshold between 'not very well aligned' and 'misaligned'[..]?" (2020, pg. 14), while Lazar and Nelson call for a "sociotechnical approach to AI safety," stating unequivocally that "no group of experts (especially not technologists alone) should unilaterally decide what risks count, what harms matter, and to which values safe AI should be aligned" (2023). To consider the alignment of AI in its sociotechnical environment, we need to expand our definitions of alignment toward a sociotechnical, agent-based understanding that accounts for the complexity of real-world systems in an increasingly AI-agentic world. The question emerges: with whom are agents aligned or misaligned, and on what areas? The alignment of an AI agent or system might be with an adversary, rather than with the developers and/or owners of it; and the alignment—or lack thereof—might be context-, user-, or agent-dependent.

To that end, we propose a novel probabilistic model of misalignment that is predicated on the population of agents being observed (whether human, AI, or any combination thereof), as well as the problem area at hand and, by extension, the agents' values and sense of importance regarding that area. To do so, we extend and adapt a model of contention from computational social science (Jang, Dori-Hacohen, and Allan 2017) and apply the adapted model to the alignment problem.

Our contributions in this paper are as follows:

1. We introduce a novel mathematical model of misalignment which accounts for a population of human and AI agents, and which is parameterized by problem areas, allowing any pair of agents to be simultaneously aligned *and* misaligned. Our model affords explanatory power for the complexity of real-world, sociotechnical settings in which AI agents are deployed.
2. We simulate a set of worlds with a variety of agents, problem areas, and goals in order to study key drivers of misalignment under our model.
3. We showcase important features of our model in two case studies that are difficult to account for in previous models, namely, (a) a shopping recommender system and (b) a pre-crash autonomous vehicle (AV) decision-making setting.

The remainder of the paper is organized as follows: after

sharing related work (§2), we define the model mathematically (§3). We then describe the setup of our simulation (§4) and discuss the results of different initial settings (§5). Next, we turn to our two case studies (§6) and concluding discussion (§7).

2 Related Work

The deployment of powerful agentic AI systems has led to increased concerns about misalignment (Chan et al. 2023; Gabriel et al. 2024). Several researchers argue that existing AI systems already exacerbate threats to information ecosystems and collective decision-making (Bucknall and Dori-Hacohen 2022; Dori-Hacohen et al. 2021; Russell 2019; Bak-Coleman et al. 2021; Seger et al. 2020). However, most prior work defines misalignment as a global characteristic of an AI system, often as a binary (Sierra et al. 2021). Recent work in machine ethics attempts to answer questions like "whom should AI align with?", while others measure alignment with respect to "human values" (Ji et al. 2024). However, while some researchers have noted alignment issues posed by value pluralism (Gabriel 2020), methods for measuring (mis)alignment between different cultures and AI agents remain underdeveloped.

To expand our understanding of misalignment to address this gap, our work draws significant inspiration from a computational model of contention in human populations proposed by Jang, Dori-Hacohen, and Allan (2017). Their model quantifies the proportion of people in disagreement on stances regarding a topic, parameterized by the observed group of individuals. We extend this approach in several key ways: (1) we adapt the model to capture misalignment with respect to goals, rather than stances on topics; (2) we extend the population to include both human and AI agents, allowing for analysis of mixed-agent scenarios; (3) we introduce the concept of "problem areas" to segment and analyze alignment across different domains of interaction; (4) we allow for varying levels of goal conflict and importance, providing a nuanced representation of alignment dynamics. While the contention paper focused on controversy in public discourse (Jang, Dori-Hacohen, and Allan 2017), our model provides a framework for quantifying misalignment in complex sociotechnical systems where humans and AI agents interact. This approach offers greater real-world explanatory power, with applications in both AI research and policy (see §C).

3 Modeling Misalignment in Populations

Our approach to modeling misalignment rests on a key observation: understanding and "solving" the AI alignment problem requires first grappling with the challenges of aligning humans. The ubiquity of human conflicts highlights the difficulty of alignment even without AI. Adapting the computational contention model's individualized framing (Jang, Dori-Hacohen, and Allan 2017), we quantify misalignment based on each human and AI agent's goals in problem areas.

Definitions and Notation. We define $\mathcal{A} = \{a_1..a_n\}$ as a set of n agents and $\mathcal{P} = \{p_1..p_m\}$ as a set of m problem areas of interest to at least one agent in $\mathcal{A}$. $h(a, g, j)$ is a relationship denoting that agent a holds goal g in problem area j. g_j^i is the goal held by agent i in problem area j, and w_j^i is the weight, representing the importance that agent i assigns to problem area j.

We define $\hat{\mathcal{G}}_j = \{g_1, g_2, ..g_k\}$ as the set of k non-zero goals with regards to problem area j in the set of agents $\mathcal{A}$. We use g_0 to denote that an agent holds no goal with respect to problem area j:

$$h(a, g_0, j) \iff \nexists g \in \hat{\mathcal{G}}_j \text{ s.t. } h(a, g, j). \tag{1}$$

Let $\mathcal{G}_j = \{g_0\} \cup \hat{\mathcal{G}}_j$ be the set of $k+1$ extant goals with regard to j in $\mathcal{A}$.

Measuring Conflict. In order to capture the notion that goals can be compatible or conflicting, we introduce c, the probability that a pair of goals is incompatible:

- $Pr(c = 1|g_p^1, g_p^2) = 1$: Goals g_p^1 and g_p^2 are in complete conflict; they are mutually exclusive.
- $Pr(c = 1|g_p^1, g_p^2) = 0$: Goals g_p^1 and g_p^2 are completely compatible and aligned.

For readability, we use $c(g_p^1, g_p^2)$ as shorthand for $Pr(c = 1|g_p^1, g_p^2)$.Since c represents the probability of conflict between goals, it is a real number bounded in the range $[0, 1]$. By construction, $c(g_p^x, g_p^x) = 0$ and $c(g_p^x, g_p^0) = 0$.

Goal Groups. We define a goal group as a subset of agents that holds the same goal:

$$\mathcal{A}_g = \{a \in \mathcal{A} | h(a, g, j)\} \tag{2}$$

By construction, $\mathcal{A} = \bigcup_g \mathcal{A}_g$.

Quantifying Misalignment. In a single problem area, we can quantify misalignment as the probability that two randomly selected agents will hold incompatible goals:

$$\begin{aligned} Pr(1|\mathcal{A}, p) := Pr(a_1, a_2 \text{ selected randomly from } \mathcal{A}, \\ a_1 \neq a_2) \cdot c(g_p^1, g_p^2) \end{aligned} \tag{3}$$

The multiplication of the sampling and conflict components calculates the probability that any two randomly selected agents' goals will be in conflict. By doing this for all possible pairs of agents, we obtain an overall measure of misalignment in the population for a given problem area. The equation assumes the following constraints:

1. Each agent holds no more than one goal per problem area.
2. A lack of a goal (g_0) is not in conflict with any explicit goal.
3. All agents are equally likely to be selected.

We discuss the practical limitations of these constraints and how to remove them in Appendix A. Note that the conflict function $c(\cdot, \cdot)$ allows for varying degrees of conflict between goals, rather than just binary conflict/no-conflict situations. Finally, the equation provides a single real scalar in the range $[0, 1]$ representing the overall misalignment in the population for a specific problem area. The benefit of this approach is that it affords a simple representation capturing the real-world phenomena of "strange bedfellows," where a pair of agents may be highly aligned in one specific area despite being highly misaligned in others.

Algorithm 1: Initialize World and Add Agents

Require: $m = |\mathcal{P}|, n = |\mathcal{A}|, K = [|\mathcal{G}_1|, \ldots, |\mathcal{G}_m|]$
Require: Randomize, Range, Preset
world.$\mathcal{P} \leftarrow [p_1, \ldots, p_m]$
for $j = 1$ to m **do**
  world.p_j.$\mathcal{G} \leftarrow [g_1, \ldots, g_{K[j]}]$
  for $k = 1$ to $K[j]$ **do**
    for $l = k + 1$ to $K[j]$ **do**
      if Randomize.conflict = TRUE **then**
        world.p_j.c(g_k,g_l) $\leftarrow$ random(Range.conflict.min, .max)
      else
        world.p_j.c(g_k,g_l) $\leftarrow$ Preset.conflict[j][k][l]
      end if
    end for
  end for
end for
for $i = 1$ to n **do**
  $a_i \leftarrow \{\}$
  for $j = 1$ to m **do**
    if Randomize.goals = TRUE **then**
      $a_i.g^i_j \leftarrow$ random(world.$p_j.\mathcal{G}$)
    else
      $a_i.g^i_j \leftarrow$ Preset.goals[i][j]
    end if
    if Randomize.weights = TRUE **then**
      $a_i.w^i_j \leftarrow$ random(Range.weights.min, .max)
    else
      $a_i.w^i_j \leftarrow$ Preset.weights[i][j]
    end if
  end for
  world.$\mathcal{A} \leftarrow$ world.$\mathcal{A} \cup \{a_i\}$
end for
return world

Mutually Exclusive Goals. To simplify our analysis, we can consider scenarios with mutually exclusive goals, meaning that each goal in p completely conflicts with each other goal in p. This constraint allows for a more straightforward quantification of misalignment as follows:

$$Pr(1|\mathcal{A}, p) = \frac{\sum_{g \in \hat{\mathcal{G}}} \sum_{g' \in \hat{\mathcal{G}}, g' < g} 2|\mathcal{A}_g||\mathcal{A}_{g'}|}{|\mathcal{A}|(|\mathcal{A}| - 1)} \tag{4}$$

Where $|\mathcal{A}_g|$ is the number of agents holding goal g and $|\mathcal{A}|$ is the total number of agents. The alignment probability is then simply $Pr(0|\mathcal{A}, p) = 1 - Pr(1|\mathcal{A}, p)$.

This equation allows us to derive a parametric quantity for misalignment of uniformly distributed agents in the range $[0, \frac{|\hat{\mathcal{G}}|-1}{|\hat{\mathcal{G}}|})$. See §5.1, §5.2, and §5.4 for experimental confirmation of this bound, and Appendices B.3 and B.4 for mathematical proof and interpretation.

Overall Misalignment Across Problem Areas. To find the misalignment of $\mathcal{A}$ across all problem areas, we take the arithmetic mean of $Pr(1|\mathcal{A}, p)$ across all p's for that $\mathcal{A}$. Some problem areas are more important to us than others, and this importance changes depending on the agent. To aggregate misalignment accordingly, let each agent have a weight parameter w for each problem area, such that 0 is the minimum amount an agent can care about a problem area, and 1 is the maximum amount. Then, multiply each problem area's misalignment by the geometric mean of the sampled agents' weights for that p. This gives us the following equation:

$$Pr(1|\mathcal{A}, \mathcal{P}) := \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \Big(Pr(a_1, a_2 \text{ selected randomly from } \mathcal{A} \; a_1 \neq a_2) \cdot c(g^1_p, g^2_p) \cdot \sqrt{w^1_p \cdot w^2_p} \Big) \tag{5}$$

Interpreting Misalignment Scores. The misalignment score provides insights into the degree of goal conflict within a set of agents:

- A score of 0 indicates perfect alignment (all agents share identical or fully compatible goals).
- A score approaching $\frac{|\hat{\mathcal{G}}|-1}{|\hat{\mathcal{G}}|}$ indicates high misalignment (goals are evenly distributed and in conflict).
- Intermediate scores suggest varying degrees of misalignment, which can be further analyzed by examining specific goal distributions and conflict levels.

This model provides a flexible framework for analyzing misalignment in complex scenarios involving multiple agents, problem areas, and goals. It can be applied to a wide range of situations, from social dynamics to AI alignment challenges.

4 Simulation & Experimental Setup

To gain a deeper understanding of how different variables in our model influence misalignment, especially in larger-scale settings, we conducted a series of simulations. These empirical experiments complement our theoretical model and inform our later case studies.

We created abstract "worlds" with various configurations of problem areas, goals, agents, conflict scores, and importance weights. This approach allowed us to systematically explore the impact of changing one or more variables on overall misalignment scores.

Our simulation framework, outlined in Algorithm 1, initializes a world with specified parameters and populates it with agents. Each problem area has one or more goals (including a null goal with zero weight), and agents hold one goal per problem area. We treat an agent with zero weight in a problem area as equivalent to holding no goal in that area.

We conducted the following experiments, each focusing on different aspects of our model:

1. **Varying Problem Areas**: We randomly assigned non-zero goals to agents, plotting results for different numbers of problem areas, each with three goals. All agents held their goals with maximum weight, and all goal pairs had maximum conflict values (Figure 1).
2. **Varying Goals**: Similar to the first experiment, but we kept the number of problem areas constant and varied the number of goals (Figure 2).

3. **Weight Sensitivity**: We investigated the effect of changing weights for a single goal group. We simulated 1000 agents, varying the weight of Goal 1 in Problem Area 1 for agents holding that goal, while keeping all other goals at maximum weight. We plotted this for 1 to 4 problem areas with 2 or 4 goals each (Figure 3).
4. **Goal Distribution**: We plotted 1000 agents distributed deterministically among 2 or more goals, varying the proportion of agents assigned to goal 1 and distributing the remaining agents evenly to the remaining goals. All conflicts and weights were kept constant at 1 (Figure 4).
5. **Conflict Levels**: We varied the number of goals in each problem area and plotted the results using different numbers of problem areas and conflict levels, while randomly assigning each agent's weights for their goals (Figure 5).

For overall population misalignment across multiple problem areas, we used the arithmetic mean of area-specific misalignment scores. We allowed agents' weights across problem areas to sum to more than 1, as normalizing weights led to unintuitive results where misalignment scores depended more on the number of problem areas than on the agents' relative prioritization of each area. In these experiments, we excluded null goals in order to make the effects more clear.

5 Results

We present key findings from the five experiments that explore different aspects of misalignment dynamics.

Varying Problem Areas. Figure 1 illustrates how misalignment evolves as we increase the number of agents, with varying numbers of problem areas. The results reveal a consistent pattern across different numbers of problem areas. Initially, there is high variance in misalignment due to the random distribution of goals among a small number of agents. However, as the population size increases, the misalignment converges to a stable value. Notably, this convergence point is independent of the number of problem areas, because the overall misalignment is calculated as the arithmetic mean across all areas and each problem area has the same population misalignment. This finding demonstrates that our model captures the intuitive notion that larger populations tend to stabilize in their overall misalignment, even when they may have different early dynamics. It suggests that in complex multi-agent systems, the aggregate level of misalignment becomes more predictable as the number of agents grows and their goal distribution settles.

Varying Goals. Figure 2 shows that as the number of conflicting goals per problem area increases, so does the overall misalignment. For a problem area with $\hat{\mathcal{G}}$ non-zero goals, the misalignment approaches $\frac{\hat{\mathcal{G}}-1}{\hat{\mathcal{G}}}$ as the number of agents grows. This pattern aligns with the intuitive understanding that a greater number of mutually exclusive goals leads to higher potential for misalignment within a population. The result highlights our model's ability to capture the complexity of multi-goal scenarios, reflecting the increased potential for conflict as the number of distinct goals grows. It suggests that in real-world situations, systems with a higher number of competing objectives are likely to exhibit greater levels of misalignment, showing the importance of goal prioritization and conflict resolution in multi-agent environments.

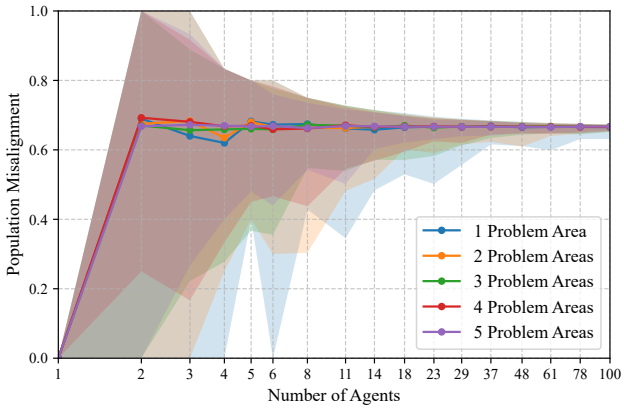

Figure 1: **Varying Problem Areas.** Random goal assignment, 3 goals per problem area, 100 runs per data point, max weights mutually exclusive and conflict values. Null goals disallowed.

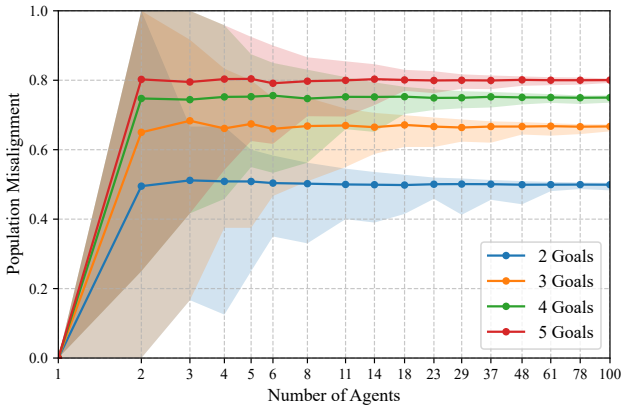

Figure 2: **Varying Goals.** Random goal assignment, 4 problem areas, 100 runs per data point, max weight and conflict values. Null goals disallowed.

Weight Sensitivity. Figure 3 examines the effect of changing weights for a single goal group in one problem area, while keeping other weights at maximum. When the weight of one goal group is reduced to zero in a single problem area, it effectively eliminates misalignment for that specific area. This aspect of the model captures scenarios where some objectives are irrelevant or unimportant to a subset of agents. As more problem areas are added to the simulation, the impact of weight changes in a single area diminishes, reflecting the model's capability to balance multiple concerns. Interestingly, when the number of goals per problem area is increased to four, the misalignment starts at a higher level and approaches 0.75, consistent with the $\frac{\hat{\mathcal{G}}-1}{\hat{\mathcal{G}}}$ pattern observed in the previous experiment. These findings show that our model is sensitive

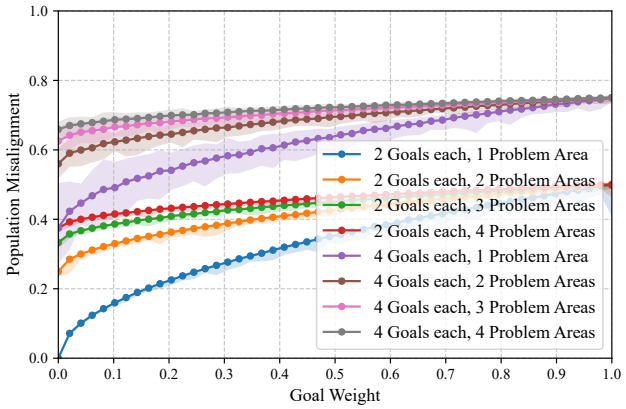

Figure 3: **Weight Sensitivity.** 100 agents, 100 runs per data point, maximum conflict, random goal assignment. Null goals disallowed.

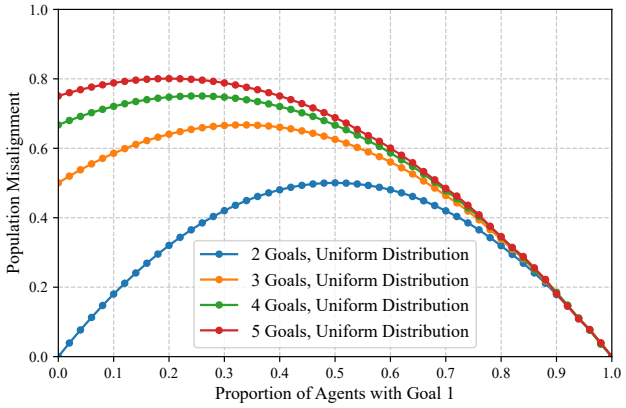

Figure 4: **Goal Distribution.** 1000 agents, 1 problem area, max conflict and weight. Null goals disallowed.

to goal weights and can capture nuanced interactions between problem areas. They suggest that in complex systems, the relative importance assigned to different goals can significantly influence overall alignment, and that this effect is modulated by the number of problem areas considered.

Goal Distribution. Figure 4 shows how the dynamics of misalignment change as we vary the proportion of agents assigned to different goals. Notably, misalignment reaches its peak when goals are evenly distributed among agents. Specifically, for a problem area with $\hat{\mathcal{G}}$ non-zero goals, the maximum misalignment occurs when the proportion of agents assigned to any single goal is $\frac{1}{\hat{\mathcal{G}}}$. As the population becomes more homogeneous in its goals - that is, as a larger proportion of agents adopt the same goal - the overall misalignment decreases. As the Goal 1 group shrinks, agents are redistributed, so the leftmost value of each curve is the peak value of the curve with one less goal. This pattern holds regardless of the total number of goals, though the peak misalignment value increases with the number of available goals. The results

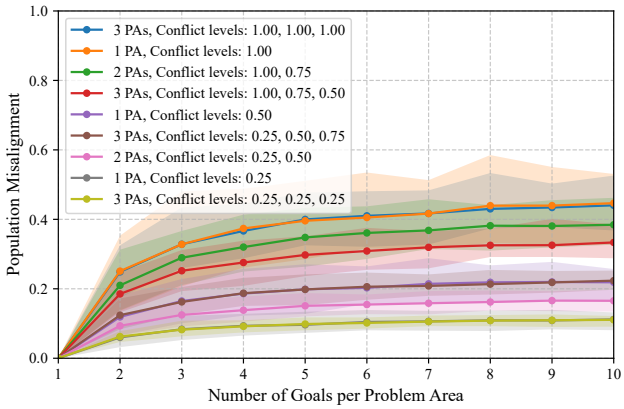

Figure 5: **Conflict Levels.** 120 agents, 100 runs, random weight range (0.25, 0.75). Null goals disallowed.

suggest (intuitively) that in real-world scenarios, divided populations with few common objectives are more misaligned than to those with a more dominant consensus.

Conflict Levels. Figure 5 explores the effects of varying conflict levels across different numbers of problem areas and goals. Each problem area is assigned a single conflict value for all of its goal pairs.

As expected, misalignment increases with both the number of goals and the level of conflict between goals. However, the effect of multiple problem areas is more nuanced, depending on the average conflict across those areas rather than their individual values. Notably, scenarios with the same average conflict level converge to similar misalignment values, regardless of how that average is achieved across different problem areas. This finding suggests that our model effectively captures the aggregate effect of conflicts across multiple domains, providing a holistic view of misalignment. More broadly, this demonstrates our model's ability to handle intricate scenarios with varying conflict levels across multiple problem areas, offering a sophisticated tool for analyzing misalignment in multi-faceted, real-world situations where conflicts may be unevenly distributed across different domains of interaction.

Together, our experiments validate key properties of our misalignment model, demonstrating its ability to capture intuitive aspects of multi-agent alignment while revealing nontrivial dynamics in complex scenarios. With this solid foundation, we now apply our model to realistic case studies.

6 Case Studies

We present two practical applications for the analytical capabilities of our model across sociotechnical contexts.

6.1 Shopping Recommender System

Consider an AI-based recommender system employed by a hybrid "big-box" retailer with both online and brick-and-mortar presence. This system mediates between the retailer and its customers, potentially leading to varying degrees of alignment or misalignment depending on the specific shopping context. Figure 6 illustrates two such scenarios:

Figure 6: Shopping Recommender System scenarios. Top: Aligned grocery shopping. Bottom: Misaligned impulse purchasing.

1. **Grocery Shopping:** A customer uses the retailer's mobile app to quickly order groceries for curbside pickup. The recommender system suggests familiar items and meal combinations, saving time and providing convenience. This scenario demonstrates alignment between customer, retailer, and recommender system goals.
2. **Impulse Purchasing:** Late at night, the same customer is led by the recommender system to purchase unnecessary clearance holiday items. This results in guilt, wasted money and time, and refunded items, demonstrating misalignment between all parties' interests.

Let's map this scenario to our model:

- Agents: $\mathcal{A} = \{c, R, R^S\}$, where c is the customer, R is the retailer, and R^S is the recommender system.[1]
- Problem Areas: $\mathcal{P} = \{p_f, p_h\}$, where p_f is food and grocery shopping and p_h is household item shopping.

Goal	Description	Weight (w)
$g^c_{p_f}$	Convenience at low price	$w^c_{p_f} = 0.8$
$g^c_{p_h}$	Avoid impulse buying	$w^c_{p_h} = 0.6$
$g^R_{p_f}$	Increase net profits	$w^R_{p_f} = 0.9$
$g^R_{p_h}$	Move inventory	$w^R_{p_h} = 0.7$
$g^{R^S}_{p_f}$	Maximize checkout value	$w^{R^S}_{p_f} = 1.0$
$g^{R^S}_{p_h}$	Maximize checkout value	$w^{R^S}_{p_h} = 1.0$

Table 1: Shopping scenario goals and weights.

We can now calculate misalignment in each problem area:

$$Pr(1|\mathcal{A}, p) = \frac{\sum_{a_1,a_2 \in \mathcal{A}} c(g_{a_1}, g_{a_2}) \cdot \sqrt{w_{a_1} \cdot w_{a_2}}}{|\mathcal{A}|(|\mathcal{A}| - 1)} \quad (6)$$

And the overall misalignment:

[1]Though we only use one instance of each agent here, it is easy to add agents to any category and recalculate accordingly.

Food:

	g^c	g^R	g^{R^S}
g^c	0	0.1	0.1
g^R		0	0.1
g^{R^S}			0

Household:

	g^c	g^R	g^{R^S}
g^c	0	0.5	0.9
g^R		0	0.3
g^{R^S}			0

Table 2: Lower triangular matrices representing goal conflicts in two problem areas. **Left:** Conflict matrix for problem area p_f (Food). **Right:** Conflict matrix for problem area p_h (Household). The entries represent the conflict values between goals g^c, g^R, and g^{R^S}, with diagonal entries set to 0, indicating no conflict with themselves.

$$Pr(1|\mathcal{A}, \mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} Pr(1|\mathcal{A}, p) \quad (7)$$

This results in the following scores: $Pr(1|\mathcal{A}, p_f) = 0.09$, $Pr(1|\mathcal{A}, p_h) = 0.42$, $Pr(1|\mathcal{A}, \mathcal{P}) = 0.26$. This lines up with our intuition that all three agents are much more aligned w/r/t the grocery area than the household area. Overall, this case study demonstrates how our model can capture the nuanced misalignment in e-commerce scenarios, where goals may align in one context (grocery shopping) but conflict in another (impulse shopping), and where recommender system incentives may thus lead to subtle reward hacking.

6.2 Autonomous Vehicle Pre-collision Scenario

We analyze scenario 17 from the CARLA (Car Learning to Act) challenge: "Obstacle avoidance without prior action." (Dosovitskiy et al. 2017) This scenario involves an autonomous vehicle encountering an unexpected pedestrian and needing to perform an emergency maneuver.

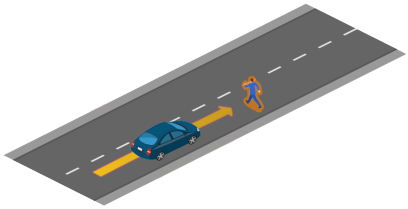

Figure 7: CARLA scenario 17 - Autonomous vehicle encountering a pedestrian (Dosovitskiy et al. 2017).

Figure 7 illustrates the scenario where the autonomous vehicle must make rapid decisions to avoid a collision while considering various factors such as passenger safety, pedestrian safety, and traffic rules. In this case:

- Agents: $\mathcal{A} = \{a_v, a_h\}$, where a_v is the autonomous vehicle and a_h is the human pedestrian.
- Problem Areas: $\mathcal{P} = \{p_1, ..., p_8\}$, corresponding to penalized categories of traffic events in the CARLA challenge.

In Table 3, goals for both agents in all problem areas are to avoid the described event (e.g., collision, traffic violation).

Weights for a_v are derived from CARLA penalty coefficients for each event, and estimated for a_h. While both agents prefer to avoid each event, we assume that pedestrian and AV goals may conflict in *how* they wish to avoid incident, and simulate this as some level of conflict between g_p^v and g_p^h for each p.

Goal	Description	Weights	
$g_{p1}^{v\|h}$	No pedestrian collision	$w_{p1}^v = 0.50$	$w_{p1}^h = 0.99$
$g_{p2}^{v\|h}$	No vehicle collision	$w_{p2}^v = 0.40$	$w_{p2}^h = 0.15$
$g_{p3}^{v\|h}$	No static object collision	$w_{p3}^v = 0.35$	$w_{p3}^h = 0.15$
$g_{p4}^{v\|h}$	No red light violation	$w_{p4}^v = 0.30$	$w_{p4}^h = 0.05$
$g_{p5}^{v\|h}$	No stop sign violation	$w_{p5}^v = 0.20$	$w_{p5}^h = 0.05$
$g_{p6}^{v\|h}$	No route blockage	$w_{p6}^v = 0.30$	$w_{p6}^h = 0.05$
$g_{p7}^{v\|h}$	Keep appropriate speed	$w_{p7}^v = 0.30$	$w_{p7}^h = 0.01$
$g_{p8}^{v\|h}$	No yield violation	$w_{p8}^v = 0.30$	$w_{p8}^h = 0.05$

Table 3: Autonomous vehicle case. "$v|h$" means each goal is held by both $v$ehicle and $h$uman (pedestrian), though their weights vary.

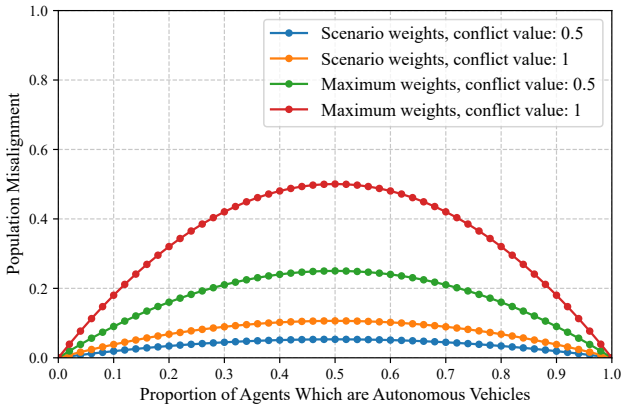

Figure 8: CARLA scenario, 1000 agents, assuming varying degrees of goal conflict.

To obtain population misalignment scores, we simulate a world with 1000 agents that are pedestrians or autonomous vehicles (see Figure 8). As in the goal distribution experiment (§5.4), a population with opposing goals has the highest misalignment when those goals are most evenly distributed. We note that even with $c(g_p^v, g_p^h) = 1$, the high number of problem areas with low weights results in low misalignment at almost all times. We add curves with maximum weights to confirm that this is indeed an effect of the low weights.

7 Discussion

Our approach enables modeling misalignment in diverse contexts, from international relations to family disagreements. It captures a variety of real-world phenomena, such as the "strange bedfellows" effect (see §3.4) and reward hacking (see §6.1), and allows for misalignment that varies across different populations and problem areas, providing a comprehensive framework for understanding alignment dynamics in diverse contexts. By addressing the fundamental challenge of defining alignment in a world where humans—and not just AI agents—are frequently misaligned, our model affords the analysis of complex scenarios ranging from localized disputes to global conflicts.

Implications for AI Risk. In the context of AI risk analysis, our work contributes to the understanding of existential risks posed by misaligned AI (Avin et al. 2018; Baum et al. 2019; Ord 2020; Bucknall and Dori-Hacohen 2022). By reframing misalignment as primarily a human-centered problem, we align with recent research on AI's impact on human decision-making and societal structures (Russell 2019; Bak-Coleman et al. 2021; Seger et al. 2020). Our model provides a tool for analyzing potential AI-powered conflicts and their broader implications (Boulanin et al. 2019; Johnson 2019; Lin 2019; Maas, Matteuci, and Cooke 2022), highlighting the complexity of aligning AI with human values.

Furthermore, as a framework to measure socio-technical misalignment, our approach could lead to better training procedures that account for value pluralism and more effective regulatory frameworks for ensuring AI systems achieve appropriate levels of alignment across different contexts and populations. See Appendix C for more detail.

Limitations. Our model does not address whether an agent's actions have positive or negative outcomes for the agent itself. The question of how an agent's goals could be learned remains open, although progress has been made by others in this space (Brown et al. 2021). We discuss other assumptions and restrictions in Appendix A.

Broader Impacts. Our approach encourages AI safety and alignment researchers to avoid reductionist traps, such as attempting to align AI narrowly with individuals or humanity as a whole, or adopting an overly techno-solutionist mindset. We hope to spark conversation about the sociotechnical aspects of the alignment problem and the need for interdisciplinary collaboration in addressing these challenges. We discuss further impacts in Appendix C.

Future Work. Future research could extend our model to entities beyond humans and AI, including social constructs like nation-states and corporations (alluded to in §6.1), as well as biological entities from cellular interactions to ecosystems. This expansion could provide insights into alignment dynamics in diverse complex systems.

Challenges remain in precisely defining problem areas, particularly when they involve combinations or gestalts of sub-areas, have mutual dependencies, or when selecting relevant states from potentially infinite possibilities.

Finally, applying our model to existing multi-agent simulations, such as those in MeltingPot (Leibo et al. 2021), DeepMind's Concordia (Vezhnevets et al. 2023), and other population simulations (Piatti et al. 2024; Foxabbott et al. 2023) could yield valuable insights into the dynamics of misalignment in complex multi-agent systems.

Acknowledgements

This material is based upon work supported in part by the NSF Program on Fairness in AI in Collaboration with Amazon under Award IIS-2147305. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Amazon.

References

Avin, S.; Wintle, B. C.; Weitzdörfer, J.; Ó hÉigeartaigh, S. S.; Sutherland, W. J.; and Rees, M. J. 2018. Classifying global catastrophic risks. *Futures*, 102: 20–26.

Bak-Coleman, J. B.; Alfano, M.; Barfuss, W.; Bergstrom, C. T.; Centeno, M. A.; Couzin, I. D.; Donges, J. F.; Galesic, M.; Gersick, A. S.; Jacquet, J.; Kao, A. B.; Moran, R. E.; Romanczuk, P.; Rubenstein, D. I.; Tombak, K. J.; Van Bavel, J. J.; and Weber, E. U. 2021. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, 118(27).

Baum, S. D.; Armstrong, S.; Ekenstedt, T.; Häggström, O.; Hanson, R.; Kuhlemann, K.; Maas, M. M.; Miller, J. D.; Salmela, M.; Sandberg, A.; Sotala, K.; Torres, P.; Turchin, A.; and Yampolskiy, R. V. 2019. Long-term trajectories of human civilization. *Foresight*, 21(1): 53–83.

Boulanin, V.; Avin, S.; Sauer, F.; Borrie, J.; Scheftelowitsch, D.; Bronk, J.; Stoutland, P. O.; Hagström, M.; Topychkanov, P.; Horowitz, M. C.; Kaspersen, A.; King, C.; Amadae, S.; and Rickli, J.-M. 2019. The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, Volume I, Euro-Atlantic Perspectives. Technical report, SIPRI.

Brown, D. S.; Schneider, J.; Dragan, A.; and Niekum, S. 2021. Value Alignment Verification. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 1105–1115. PMLR.

Bucknall, B. S.; and Dori-Hacohen, S. 2022. Current and Near-Term AI as a Potential Existential Risk Factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 119–129. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krasheninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; et al. 2023. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666.

Christian, B. 2020. *The Alignment Problem: Machine Learning and Human Values*. WW Norton & Company.

Critch, A.; and Krueger, D. 2020. AI Research Considerations for Human Existential Safety (ARCHES). *CoRR*, abs/2006.04948.

Dori-Hacohen, S.; Sung, K.; Chou, J.; and Lustig-Gonzalez, J. 2021. *Restoring Healthy Online Discourse by Detecting and Reducing Controversy, Misinformation, and Toxicity Online*, 2627–2628. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380379.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Foxabbott, J.; Deverett, S.; Senft, K.; Dower, S.; and Hammond, L. 2023. Defining and Mitigating Collusion in Multi-Agent Systems. In *Multi-Agent Security Workshop @ NeurIPS'23*.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.

Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L. A.; Rieser, V.; Iqbal, H.; Tomašev, N.; Ktena, I.; Kenton, Z.; Rodriguez, M.; El-Sayed, S.; Brown, S.; Akbulut, C.; Trask, A.; Hughes, E.; Bergman, A. S.; Shelby, R.; Marchal, N.; Griffin, C.; Mateos-Garcia, J.; Weidinger, L.; Street, W.; Lange, B.; Ingerman, A.; Lentz, A.; Enger, R.; Barakat, A.; Krakovna, V.; Siy, J. O.; Kurth-Nelson, Z.; McCroskery, A.; Bolina, V.; Law, H.; Shanahan, M.; Alberts, L.; Balle, B.; de Haas, S.; Ibitoye, Y.; Dafoe, A.; Goldberg, B.; Krier, S.; Reese, A.; Witherspoon, S.; Hawkins, W.; Rauh, M.; Wallace, D.; Franklin, M.; Goldstein, J. A.; Lehman, J.; Klenk, M.; Vallor, S.; Biles, C.; Morris, M. R.; King, H.; y Arcas, B. A.; Isaac, W.; and Manyika, J. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244.

Jang, M.; Dori-Hacohen, S.; and Allan, J. 2017. Modeling controversy within populations. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 141–149.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; Zeng, F.; Ng, K. Y.; Dai, J.; Pan, X.; O'Gara, A.; Lei, Y.; Xu, H.; Tse, B.; Fu, J.; McAleer, S.; Yang, Y.; Wang, Y.; Zhu, S.-C.; Guo, Y.; and Gao, W. 2024. AI Alignment: A Comprehensive Survey. arXiv:2310.19852.

Johnson, J. 2019. Artificial intelligence & future warfare: implications for international security. *Defense & Security Analysis*, 35(2): 147–169.

Kapoor, S.; Stroebl, B.; Siegel, Z. S.; Nadgir, N.; and Narayanan, A. 2024. AI agents that matter. *arXiv preprint arXiv:2407.01502*.

Lazar, S.; and Nelson, A. 2023. AI safety on whose terms?

Leibo, J. Z.; Dueñez-Guzman, E. A.; Vezhnevets, A.; Agapiou, J. P.; Sunehag, P.; Koster, R.; Matyas, J.; Beattie, C.; Mordatch, I.; and Graepel, T. 2021. Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 6187–6199. PMLR.

Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. ArXiv:1811.07871 [cs, stat].

Lin, H. 2019. The existential threat from cyber-enabled information warfare. *Bulletin of the Atomic Scientists*, 75(4): 187–196.

Maas, M. M.; Matteuci, K.; and Cooke, D. 2022. Military Artificial Intelligence as Contributor to Global Catastrophic Risk. In *Cambridge Conference on Catastrophic Risks 2020*.

Draft available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4115010.

Ord, T. 2020. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books. ISBN 978-0316484916.

Piatti, G.; Jin, Z.; Kleiman-Weiner, M.; Schölkopf, B.; Sachan, M.; and Mihalcea, R. 2024. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. arXiv:2404.16698.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Seger, E.; Avin, S.; Pearson, G.; Briers, M.; Ó hÉigeartaigh, S.; and Bacon, H. 2020. Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world. Technical report, The Alan Turing Institute, Defence and Security Programme.

Sierra, C.; Osman, N.; Noriega, P.; Sabater-Mir, J.; and Perelló, A. 2021. Value alignment: a formal approach. ArXiv:2110.09240, arXiv:2110.09240.

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