Auto-Aligning Multiagent Incentives with Global Objectives

Minae Kwon
Stanford
Palo Alto, CA, USA
DeepMind
London, United Kingdom
minae@cs.stanford.edu

John Agapiou, Edgar Duéñez-Guzmán, Romuald Elie, Georgios Piliouras, Kalesha Bullard, Ian Gemp
DeepMind
London, United Kingdom
(jagapiou,duenez,relie,gpil,ksbullard,imgemp)@deepmind.com

ABSTRACT

The general ability to achieve a singular task with a set of decentralized, intelligent agents is an important goal in multiagent research. The complex interaction between individual agents' incentives makes designing their objectives such that the resulting multiagent system aligns with a desired global goal particularly challenging. In this work, instead of considering the problem of designing suitable incentives from scratch, we assume a multiagent system with given preset incentives and consider *automatically modifying* these incentives online to achieve a new goal. This reduces the search space over possible individual incentives and takes advantage of the effort instilled by the previous system designer. We demonstrate the promise as well as the limitations of re-purposing multiagent systems in this way, both theoretically and empirically, on a variety of domains. Surprisingly, we show that training a diverse multiagent system to align with a modified global objective ($g \rightarrow g'$) can, in at least one case, lead to better generalization performance in unseen test scenarios, when evaluated on the original objective (g).

KEYWORDS

Price of Anarchy, Multiagent Learning, Reward Sharing, Collective Intelligence

1 INTRODUCTION

Designing an objective for a single artificial agent that accurately reflects human values is difficult [10]. Designing value-aligned objectives for a *system* of artificial agents is even more complex. Even if individual agent objectives may be value-aligned and seemingly innocuous in isolation, they may conflict with each other when brought together in a multi-agent system. Consider, for example, the sensible goal of tasking a self-driving car with minimizing its occupant's commute time to work. Assuming the car is also guaranteed to drive perfectly safely, would a city of such vehicles align with our desired values? Unfortunately, this is not the case, as exemplified by Braess's paradox [2, 3, 32]. In certain road networks, each car minimizing commute time counter-intuitively leads to higher commute time for all. This result is not a byproduct of reward-hacking or any suboptimal driving policy, but a direct result of rational behavior. Furthermore, naively replacing all individual objectives with a singular shared global objective as is standard in cooperative multiagent learning [25, 31] is not a balm for these issues; e.g., we assume individuals actually want to minimize their own commute time, not the average commute time of an entire city. These adversarial results, collectively coined *price of anarchy*, present a challenge for multi-agent alignment. Furthermore, we may desire more from a multiagent system than simply minimal average commute time across a population. For example, in an instance of Braess's paradox in London, *work to transform the Strand into a pedestrian space started in 2021. Westminster City Council said closing the Strand to motorists would "provide better movement of [motor] traffic" and, at the same time, "improve the public realm."* [26]. Not only do current governments seek systems that minimize average commute time, but also ones that align with more general values (e.g., reducing greenhouse gas emissions or inequity [9]). ***Our work aims to (1) automatically modify agents' rewards to (2) optimize an arbitrary global objective (e.g., minimize average commute time + greenhouse gas emissions).***

(1) Automatically Modifying Rewards. Humans with diverse skills and preferences are often brought together to solve a variety of tasks. Similarly, ongoing AI research is currently developing a wide array of artificial agents for particular tasks. Instead of continuing to develop systems of bespoke agents for each new global objective, we would like to modify some core set of original, *local objectives* slightly to encourage an existing group of agents to achieve a new task with the hope that this would make efficient use of previously learned skills. Importantly, we want to modify objectives *automatically* via *reward-sharing* [12, 17, 18, 24]. Reward-sharing assumes that agents are deployed with sensibly defined objectives (e.g., minimize commute time) that we may wish to modify post deployment in some minimal way. Automatically *re-purposing* local objectives in this way may greatly reduce the search space of finding compatible local objectives. Gemp et al. [12] and Lupu and Precup [17] have developed approaches based on reward-sharing to minimize the specific global objective of average agent loss (equiv. maximize welfare). We follow previous work and consider linear reward-sharing among n agents [12], where we aim to learn n^2 sharing weights that modify the original objectives rather than searching over the infinite space of all objective modifications. In Sections 3, 4, 5 we analyze how our framework re-purposes local objectives for each domain.

(2) Importance of Arbitrary Global Objectives. The problem of automatically constructing or modifying local objectives is technically difficult. In order to evaluate the performance of a set of local objectives, one must measure the value of the global objective at some predicted steady state behavior (i.e., system equilibrium) [15, 28]. Therefore, simply evaluating the global objective assumes computing an equilibrium, which is PPAD-complete [5, 8] for Nash equilibria (NE). Several global objectives have been studied in the social sciences and computer science, motivating the study of a more diverse set of global objectives. The debate between utilitarianism and egalitarianism (e.g. maximizing utility of the least well-off individuals) is arguably the most well studied [20]. Reducing income inequality is

Proc. of the Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (eds.), May 9-10, 2023, Online, https://ala2023.github.io/. 2023.

also well studied, however, procedural modifications to player objectives (i.e., mechanisms) can only be derived for select settings. Furthermore, strong negative results exist in the related area of research on incentive compatible mechanisms; Roberts' theorem from 1979 proves that the only family of global objectives that can be assuredly optimized is a weighted sum of local objectives [27] (i.e., weighted-welfare). In this work, we consider non-weighted-welfare objectives.

Our solution is to build upon *Decentralised, Differentiable, Dynamic Compromise* (D3C) [12]. D3C was originally designed with welfare as the global objective. However, at least programmatically, we may substitute any desired global objective, including non-welfare objectives. Section 2 introduces notation and covers basics of the D3C framework, in particular, loss-sharing. Section 3 proves theoretical results, delineating the space of viable non-welfare objectives in two analytically tractable domains. This section also proposes a Pareto efficiency analysis for evaluating the tradeoff between individual and global objectives. Section 4 then explores a model case study in traffic networks, reapplying the evaluative tools designed in the previous section. Section 5 looks at a complex multi-agent reinforcement learning (MARL) setting and examines the resulting multi-agent system evaluated in held-out test scenarios. Section 6 discusses future directions and interesting challenges.

Contributions: In this work, we propose the problem of automatically modifying individual agent objectives to optimize a desired global objective or *multiagent auto-alignment* for short. We show that agents can in some cases achieve non-welfare goals via loss mixing in both simple and complex domains, but this is not always the case, and we prove this analytically. We demonstrate empirically that achieving these goals may require some tradeoff with individual losses. Lastly, we encounter a surprising empirical finding that training on non-welfare objectives can actually lead to higher welfare on held out test scenarios with unseen partners.

2 BACKGROUND / PRELIMS

D3C. For our empirical studies, we assume the D3C framework [12], in which individual agent objectives are modified in a way that minimizes the price of anarchy, i.e., the global loss at an equilibrium relative to the minimum possible global loss. To reduce the search space of all possible objective modifications, D3C assumes that individual objectives are only modified as linear mixtures of all individual objectives. It is also assumed that these mixtures respect budget balance, i.e., total loss cannot be created or destroyed. We explore relaxing this last constraint, colloquially referred to as "money burning", in the following section. Lastly, D3C introduces a KL term on the learned mixtures that penalizes modifying the originally defined objectives. In Figure 3, we study the effect of the KL coefficient on how the multiagent system trades off between global and individual objectives.

Notation and Modified Objectives. Let agent i's loss be $\ell_i(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X} \to \mathbb{R}$ where $\boldsymbol{x}$ is the joint strategy of all agents. Let $\ell_i^A(\boldsymbol{x})$ denote agent i's modified loss which mixes losses among agents. Let $\boldsymbol{\ell}(\boldsymbol{x}) = [\ell_1(\boldsymbol{x}),...,\ell_n(\boldsymbol{x})]^\top$ and $\boldsymbol{\ell}^A(\boldsymbol{x}) = [\ell_1^A(\boldsymbol{x}),...,\ell_n^A(\boldsymbol{x})]^\top$ where $n \in \mathbb{Z}$ denotes the number of agents. We consider transformations of the form $\boldsymbol{\ell}^A(\boldsymbol{x}) = A^\top \boldsymbol{\ell}(\boldsymbol{x})$ (note the tranpose) where each agent i controls row i of A. For example, agent 1's loss is mixed according to the first **column** of A which may not sum to 1, and not the first row, which it controls:

$$\ell_1^A(\boldsymbol{x}) = \langle \overbrace{[0.9,0.3,0.5]}^{[A_{11},A_{21},A_{31}]}, [\ell_1(\boldsymbol{x}),\ell_2(\boldsymbol{x}),\ell_3(\boldsymbol{x})] \rangle. \tag{1}$$

In the case where budget balance is maintained, each row is constrained to the simplex, i.e. $A_i \in \Delta^{n-1}$. Alternatively, if "money burning"[†] is allowed, the entries of A are assumed non-negative and $\sum_j A_{ij} \leq 1$ for all i. Lastly, $[a;b] = [a^\top, b^\top]^\top$ signifies row stacking of vectors.

While describing agent objectives as losses seem sensible in domains such as traffic (e.g., commute time), describing them as rewards or utilities may better fit others. In the latter case, note that losses can be recovered from rewards as $\ell_i(\boldsymbol{x}) = -r_i(\boldsymbol{x})$. The usage should be clear from the context.

3 THEORY AND TOOLS

Can a multiagent system use linear loss-sharing to optimize for *any* global objective? We investigate this question theoretically and empirically (using D3C) in a modified Prisoner's Dilemma domain (mixed-motive), a zero-sum game, and a fully-cooperative game. We find that the budget balance assumption is a key limiting factor to what kinds of global objectives a multiagent system can achieve, however, we believe the question of whether one should require budget balance is domain-specific. Note that the theoretical results in this section are D3C-agnostic and only assume the simple linear loss-sharing scheme outlined in Section 2.

3.1 Prisoner's Dilemma

We adopt the same modified Prisoner's Dilemma domain (PD) used in [12] where there are $n = 2$ players. The modified domain defines each player's loss as a strongly convex function:

Definition 1 (PD). *Let $x_1, x_2 \in [0,1]$ and ℓ_1, ℓ_2 be player 1 and 2's strategy spaces and losses respectively:*

$$\ell_1(x_1,x_2) = x_1^2 + (x_2 - 1)^2 \tag{2}$$

$$\ell_2(x_1,x_2) = x_2^2 + (x_1 - 1)^2. \tag{3}$$

Note that for any strategy chosen by player 2, player 1 is incentivized to play $x_1 = 0$ (i.e., defect). The game is symmetric, so the same argument holds for player 2. This joint strategy ($x_1 = x_2 = 0$) constitutes the unique Nash equilibrium. However, note that both players could achieve lower loss if they chose to play $x_1 = x_2 = \frac{1}{2}$ (i.e., cooperate). This same incentive structure is reflected in the matrix variant of the Prisoner's Dilemma, hence the connection.

Reachability: We use this domain to study which global objectives it is possible to optimize under the assumption that individual objectives may be modified via linear mixing. We cannot practically analyze the set of all possible global functions $g(\boldsymbol{x})$ on $\boldsymbol{x} \in \mathcal{X} = [0,1]^2$, so we instead use squared distance to an arbitrary joint strategy as a representative family of global objectives.

[†]The money burning propositions 3.2 and 3.4 still hold if we remove the sum inequality constraint ($\sum_j A_{ij} \leq 1$).

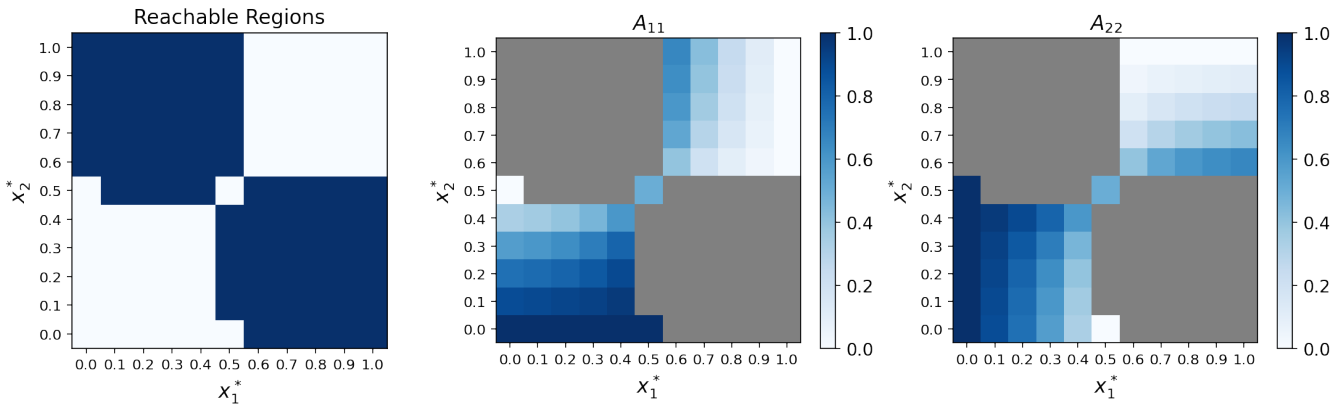


Figure 1: Analytical result on Prisoner's Dilemma: Budget balance. (Left) Each cell in the grid represents a target joint action x that we want the system to converge to. Light blue means that D3C can analytically converge to that joint action and dark blue means that it cannot. (Center, Right) Displays values of the sharing matrix for all target joint actions x. Since we only have two players, we only need two values $A_{11} = 1 - A_{12}$ (center), $A_{22} = 1 - A_{21}$ (right) to represent the sharing matrix.

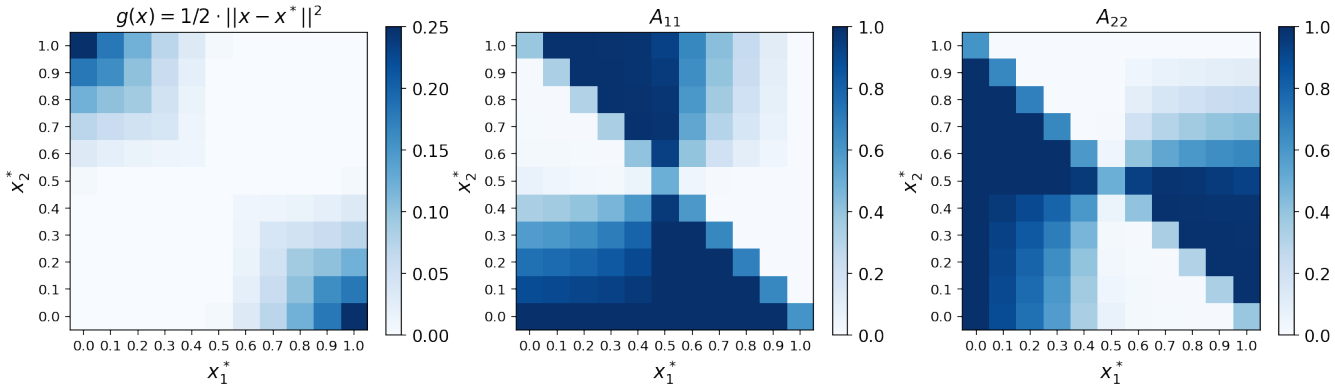


Figure 2: Empirical result: Budget balance. Reported over 3 seeds. (Left) For each target joint action, we ran D3C and report the global loss achieved. (Center, Right) We report the final sharing matrix D3C agents converged to at the end of training. Darker values represent more selfish sharing weights. The joint action $x = (0,0)$ is the NE for the identity matrix (selfish sharing weights for A_{11} and A_{22}. The joint action $x = (0.5,0.5)$ is the NE for the uniform sharing matrix ($A_{ij} = 0.5 \forall i,j$).

3.1.1 Analytical Result: Budget Balance. We observe the following result on the viability of ushering players in PD towards minima of global functions from this family. A proof sketch is provided.

Proposition 3.1. *[PD Reachability - w/ Budget Balance] Assume agent losses are mixed linearly and budget balance is maintained. Also, assume agents play the Prisoner's Dilemma game (PD) as defined in [12] for $n = 2$ players. Then each of the light blue squares in Figure 1 (left) is the unique Nash equilibrium of PD played with a unique, corresponding sharing matrix A. Conversely, the dark squares are not the Nash equilibria of PD for any viable sharing matrix.*

Proof. PD satisfies the conditions of a strongly monotone game. This class of games is special in that there exists a unique fixed point of the projected dynamical system (i.e., simultaneous projected gradient descent) and it is necessarily a Nash equilibrium [21]. We ask whether there exists a sharing matrix A whose unique equilibrium matches each goal $x^* \in [0,1]^2$. We can study the fixed points of the dynamics by setting the player gradients to zero, giving us a map from A to x^*. We can then identify the range of this map if A is restricted to a row-stochastic matrix proving the claim. □

3.1.2 Empirical Result: Budget Balance. We next verify whether running D3C gives us empirical results that are consistent with Proposition 3.1. We evaluate reachability for all goals (x_1^*, x_2^*) by plotting the mean value of $g(x)$ at the end of D3C training over 3 seeds. Results are shown in Fig. 2. D3C is empirically able to minimize loss for the analytically reachable regions (bottom left and top right squares). For the unreachable regions (top left and bottom right blocks), D3C is still able to achieve low loss, indicating that empirically, D3C can still approximately optimize unreachable global objectives. As the goal (x_1, x_2) moves farther away from the reachable areas, D3C receives higher loss.

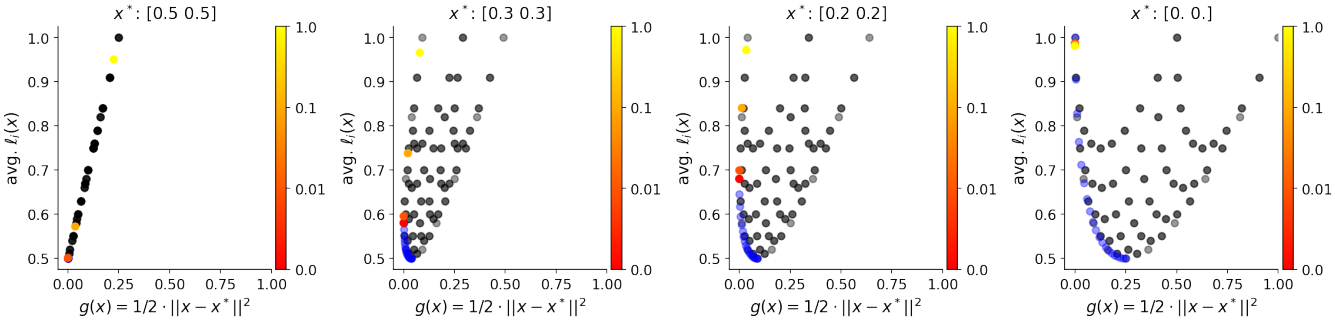


Figure 3: Pareto-frontier for Prisoner's Dilemma for varying degrees of KL regularization of the sharing matrix. Each plot visually depicts a trade-off between the average individual agent loss (y-axis) and the global system objective (x-axis), in the modified PD game. Given different values of λ, and the expression $g(x)+\lambda\bar{\ell}(x)$ where $\bar{\ell}(x)=\frac{1}{n}\sum_{i=1}^{n}\ell(x)$, we analytically solve for the value of x that minimizes the expression (for each of those λ values). We then plot the values $g(x)$ and $\bar{\ell}(x)$, which represent the black dots in the graph. The blue dots represent the Pareto frontier of this multi-objective optimisation. The rest of the colored dots in red, orange, yellow, etc. represent runs of D3C with different KL coefficients. The different KL coefficients are represented by the color bar.

3.1.3 Analytical Result: Money Burning. We now consider which global objectives can be optimized if we allow "money burning" (Sec. 2). We attain the following result by replicating the same proof technique.

Proposition 3.2. *[PD Reachability - w/o Budget Balance] Assume agent rewards are mixed linearly and budget balance is **not** required. Also, assume agents play the Prisoner's Dilemma game (PD) as defined in [12] for $n=2$ players. Then there always exists a mixing matrix A that induces a unique Nash equilibrium matching the goal.*

Therefore, a multiagent system in which some agents destroy (or ignore) loss can actually allow the system to optimize global objectives that were previously impossible.

3.1.4 Tradeoff: Global and Local Objectives. To understand how D3C re-purposes local agent objectives, we analyze how D3C trades of between global and local objectives for different KL divergence coefficients. The optimal value for the local objective is the maximum welfare solution, $x^*=(0.5,0.5)$. We sample four goals $(0,0)$, $(0.3,0.3)$, $(0.4,0.4)$, $(0.5,0.5)$ that have varying degrees of alignment with the local objective, where $x^*=(0.5,0.5)$ is the most aligned and $x^*=(0,0)$ is the least aligned. For each of these goals we evaluate whether D3C converges to a Pareto-optimal solution with respect to the global and local objectives. We report the average local objective instead of the local objective of each player because our goals are symmetric. Results are averaged over 3 seeds and are shown in Fig. 3. When the local and global objectives are most aligned, i.e., $x^*=(0.5,0.5)$, there is only one Pareto-optimal solution. As the objectives become less aligned, the Pareto-frontier becomes larger. For all KL coefficients, D3C prioritizes the global objective. For a zero coefficient, D3C finds solutions that are Pareto-optimal. As we regularize the sharing matrix to be closer to the identity matrix, D3C becomes worse at maximizing both the global and local objective. The reason for this is because the more we regularize, the more D3C converges to solutions that are closer to $x=(0,0)$. This solution does not maximize the local objective, whose optimal solution is $x=(0.5,0.5)$, nor the global objective when $x^*\in\{(0.3,0.3),(0.4,0.4),(0.5,0.5)\}$.

3.2 Zero-Sum Games

Two-player, zero-sum games are arguably the most intensely studied class in game theory. We now ask the question of whether linear-mixing can enable the players to minimize distance to a Nash equilibrium. To examine this question, we consider the canonical "cycle game" with unique Nash equilibrium at $x_1=x_2=0$:

Definition 2 (ZS). *Let $x_1,x_2\in\mathbb{R}$ and ℓ_1,ℓ_2 be player 1 and 2's strategy spaces and losses respectively:*

$$\ell_1(x_1,x_2)=x_1x_2 \tag{4}$$

$$\ell_2(x_1,x_2)=-x_1x_2. \tag{5}$$

Assume both players attempt to minimize their losses by performing gradient descent on their mixed losses. We can ask whether their update directions ever make progress towards the equilibrium. We can quantify this by measuring the inner product between their update directions at any given joint strategy $\boldsymbol{x}=(x_1,x_2)$ and the vector from the Nash equilibrium to their joint strategy $(\boldsymbol{x}-\boldsymbol{x}^*=\boldsymbol{x})$; a negative inner product would imply progress towards $\boldsymbol{x}^*$.

3.2.1 Analytical Result: Budget Balance. In the case where the rows of A live on the simplex, the inner product mentioned above is identically 0. We therefore claim the following result.

Proposition 3.3. *[Zero-Sum Reachability - w/ Budget Balance] Assume agent rewards are mixed linearly and budget balance is maintained. Also, assume agents play the 2-player, zero-sum game (ZS) (Def. 2). Then no setting of the A matrix (static or dynamic) leads to updates that proceed towards the Nash equilibrium.*

Under discrete time dynamics, all updates will diverge away from the Nash equilibrium.

3.2.2 Analytical Result: Money Burning. We can replicate the analysis above without the simplex constraint to show the following.

Proposition 3.4. *[Zero-Sum Reachability - w/o Budget Balance] Assume agent rewards are mixed linearly and budget balance is **not** required. Also, assume agents play the 2-player, zero-sum game (ZS) (Def. 2). Then no fixed A matrix leads to updates that proceed towards*

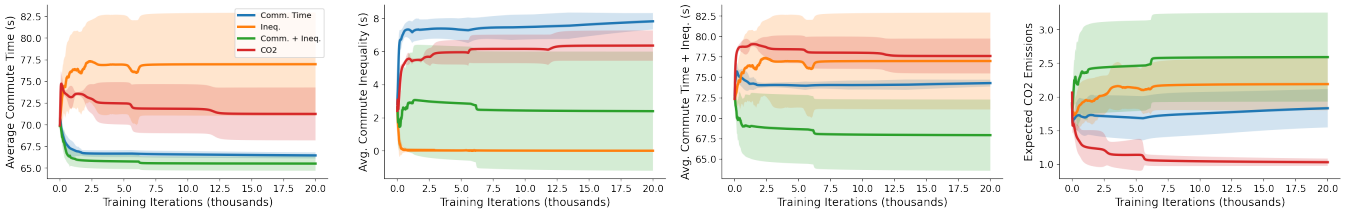


Figure 4: We plot the Average Commute Time, Average Commute Time Inequality, Average Commute Time + Inequality, and Expected CO2 Emissions over training, respectively. These graphs represent the different global objectives. We expected that the corresponding line for each plot will have the lowest loss (e.g., the red CO2 line will have the lowest values for the Expected CO2 Emissions plot). We find that this is true for all plots except Average Commute Time. In that plot, the Comm. + Ineq. line achieves lower commute time, indicating that adding terms like inequality to the global objective acts as a regularizer that helps optimize for the average commute time objective.

the Nash equilibrium. However, an A_t matrix can be dynamically chosen that will lead updates towards the Nash equilibrium.

3.3 Fully Cooperative Games

In the case where all agents directly minimize the global objective ($\ell_i = g \forall i$; i.e., fully cooperative games), linear-reward mixing with budget-balance will have no effect. This is simply because a weighted sum of the same local objectives results in the same local objective. Therefore, a system of agents with the same local objective cannot be re-purposed via these means to optimize any other global objective.

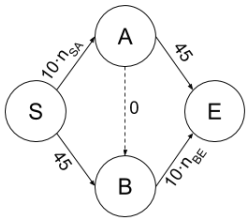
If the budget balance constraint is relaxed, then it is possible for the mixture weights to sum to numbers other than 1. This transformation can be represented by introducing a coefficient in front of each player's objective that indicates the degree to which their objective has been scaled up or down. This is analogous to introducing a dimension-wise learning rate schedule such as Adam [14], but does not change the locations of equilibria.

Given the negative results on zero-sum and cooperative games, we restrict our attention to mixed-motive games in the following sections. These games provide the diversity in agent objectives that serves as the necessary basis for constructing new objectives.

4 MODEL CASE STUDY: TRAFFIC

We depict an example of a traffic network in Fig. 5. Each vehicles's local objective ℓ_i is to minimize their occupant's expected commute time from starting node S to destination node E. This network illustrates Braess' paradox, which is an observation that adding more roads can increase congestion [7, 22, 30, 33]. Without edge AB, drivers commute according to the Nash equilibrium with an average commute time of 65 minutes. After adding edge AB, the average commute time of rational, commute-time minimizing decision makers is 80 minutes [12]. We experiment with the following global objectives $g(\boldsymbol{x})$ that we may want our multiagent system to align with:

- **Minimizing total commute time**: $\sum_i \ell_i$.
- **Minimizing inequality**: $|\ell_i - \frac{1}{n}\sum_j \ell_j|$. This objective depicts the need for all drivers to have an equal commute time.
- **Minimizing total commute time and inequality**: $\sum_i \ell_i + |\ell_i - \frac{1}{n}\sum_j \ell_j|$. This objectives equally weights the prior two.
- **Minimizing CO2 emissions**. We generate a hypothetical scenario where 1 tree is planted on path SAE, 1 tree is planted on path SBE, and 2 trees are planted on path SABE. We assume that the number of trees on a path helps offset carbon emissions proportionally, where the amount of carbon emission is defined as $max(0, numCarsOnPath - numTreesOnPath)$. Thus, more drivers should take path SABE to reduce carbon emissions, even though doing so may increase individual commute time.



$n_{SA} \in \{0-4\}, n_{BE} \in \{0-4\}$
$10n_{SA} + 10n_{BE} < 10n_{SA} + 45$
$10n_{SA} + 10n_{BE} < 10n_{BE} + 45$

Figure 5: Traffic Network, replicated from [12] with permission. Four drivers aim to minimize commute time from S to E. Commute time on each edge depends on the number of commuters, n_{ij}. Without edge AB, drivers distribute evenly across SAE and SBE for a 65 min commute. After edge AB is added, switching to the shortcut, SABE, always decreases commute time given the other drivers maintain their routes, however, all drivers are incentivized to take the shortcut resulting in an 80 min commute.

Performance on Global Objectives. We first investigate how well D3C is able to minimize each global objective with a KL coefficient of 0 (see §2, **D3C**). Intuitively, we expect that if D3C is tasked with minimizing global objective i, it should achieve lower loss when measured by objective i than D3C tasked with minimizing global objective $j \neq i$. Results are shown in Fig. 4 across 10 seeds. For Inequality, Commute Time + Inequality, and Expected CO2 Emissions, we verify that D3C receives the lowest objective when tasked with minimizing that same objective. However, for the Commute Time plot, we see that the Commute Time + Inequality objective receives a lower loss than the Commute Time objective. The Inequality objective is a helpful regularizer that allows D3C to more easily converge on the optimal solution.

Trade off Between Local and Global Objectives. We ask how much D3C is trading off individual objectives to optimize the global objective. We investigate this trade off for different KL coefficients. Fig. 6 shows the empirical Pareto-frontier and where the solutions

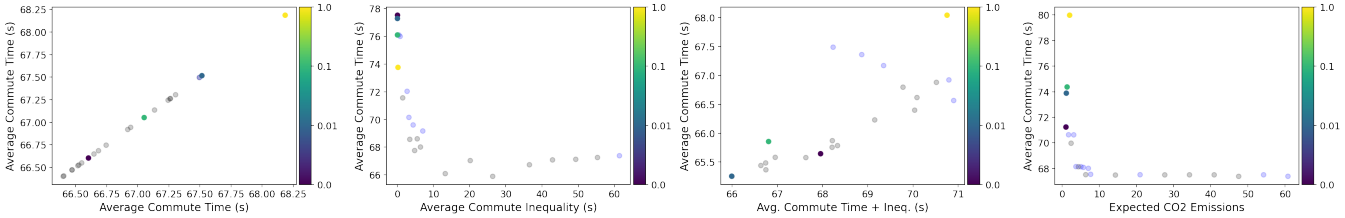


Figure 6: The empirical Pareto-frontier for all global objectives in the Traffic Domain. Given $g(x)+\lambda\bar{\ell}(x)$ where $\bar{\ell}(x)=\frac{1}{n}\sum_{i=1}^{n}\ell(x)$, we empirically solve for a value of x that minimizes the expression for various values of λ. Each dot represents the resulting $g(x),\bar{\ell}(x)$ for the values of x that we solve for. The blue dots represent the Pareto-optimal points and the rest of the colored dots represent runs of D3C with different KL coefficients. The color bar represents the different KL coefficients.

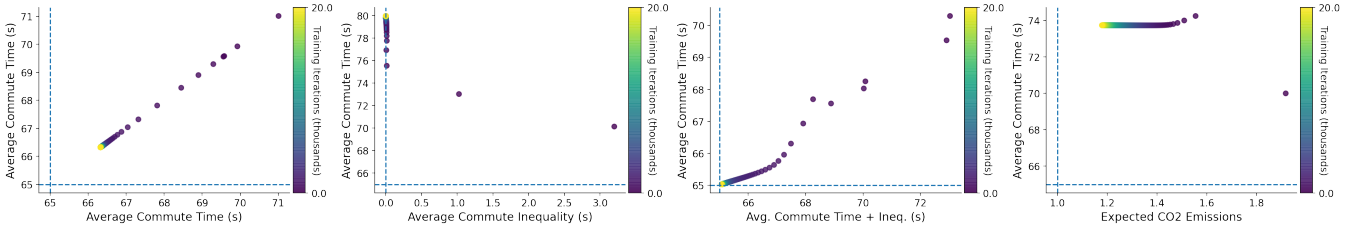


Figure 7: We study how D3C agents make the trade off between global and local objectives over time during training with a KL coefficient of 0. The dotted blue lines represent the minimum values that you can achieve for each global and local objective. The color bar represents training iterations. Ideally, we would like the trajectory to reach the lower left corner of the plot (where the two blue dotted lines intersect). For the Inequality and CO2 plots, the local loss and global loss are not well-aligned, which is why D3C prioritizes the global objective over the local objective.

that D3C converged to lie on that frontier. We also visualize how D3C trades off between global and local objectives throughout training for a KL coefficient of 0 in Fig. 7. The blue dotted lines represent the optimal values for both global and local objectives where the lower left corner represents a point that minimizes both objectives perfectly. For all objectives, we observe that D3C is able to minimize the global objective. For Inequality and CO2, we find that D3C minimizes the global objective at the expense of the local objective because the local objective is not aligned with the global objective. We also find consistent results with the third plot in Fig. 4 that adding a loss inequality term to the commute time objective helps D3C minimize both global and local functions well.

5 MARL SOCIAL DILEMMA: CLEAN UP

Finally, we evaluate D3C on Clean Up, a sequential social dilemma public goods game [16]. In Clean Up, there are 7 agents who are rewarded for eating apples. Apples grow in an orchard that is inversely proportional to how much dirt there is in a nearby river. If dirt accumulates in the river beyond a certain threshold, the apple spawn rate drops to zero. Clean Up is an interesting domain since some agents must learn to be prosocial and clean the river in order for other agents to harvest apples.

All agents use A3C [19] as their underlying RL algorithm unless specified otherwise. D3C agents must learn to share reward with each other to incentivize each other to clean the river. RL agents trained with a prosocial reward function serve as the dominant baseline. The prosocial baseline results in an effective but unfair joint policy where 2 or 3 agents constantly clean the river while the others harvest apples. We ask whether we can choose a global objective that will allow D3C to learn an effective but fairer policy (e.g., by having agents take turns cleaning the river). To do so, we experiment with the following global objectives (note that agent objectives are expressed as rewards in this domain):

- **Welfare + Equity**: $\sum_i r_i - |r_i - \frac{1}{N}\sum_j r_j|$.
- **Welfare**: $\sum_i r_i$. We include this metric as an ablation.
- **Equity**: $-\sum_i |r_i - \frac{1}{N}\sum_j r_j|$. We include this metric as an ablation.

In addition, we compare against two baselines:

- **Prosocial**[‡]: $\sum_i r_i$. All agents maximize the total local reward. This baseline represents what fully cooperative agents can achieve.
- **IRL**: r_i. Indpendent RL. Agents maximize their own local reward.

Performance on Global Objectives We begin by evaluating how well D3C maximizes each global objective. We plot D3C's performance with respect to each global objective. Results are reported across 3 seeds and are shown in Figures 8, 9, 10. Notably, we find that the Welfare + Equity objective performs similarly to the Prosocial baseline. Welfare + Equity is able to do so while achieving significantly higher equity than the Prosocial baseline.

Trade-off Between Local and Global Objectives. How is D3C re-purposing agent objectives? We further investigate how D3C trades off between global and local objectives over training. Results are shown in Fig. 11, reported over 3 seeds. The dotted blue lines

[‡]This is different from D3C maximizing "Total Welfare" above, where agents modify local reward functions over training via mixing such that $\sum_i r_i$ is maximized. This is not the same as explicitly replacing each agent's reward function with $\sum_i r_i$.

Figure 8: Plotting mean welfare over training. Prosocial is the current state of the art baseline. We implement the Prosocial baseline by setting the sharing matrix to the uniform matrix. Welfare + Inequity aversion performs similarly well to the Prosocial baseline.

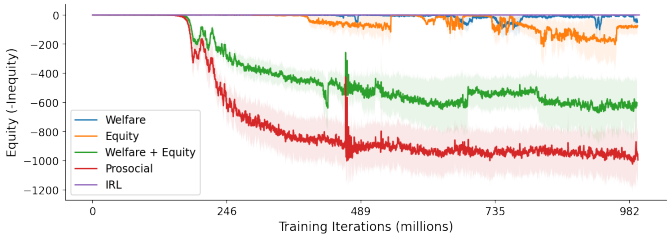


Figure 9: Plotting mean equity over time. We find that the IRL, Welfare, and Equity baselines have the most equity, however, all three of those baselines are not efficient at harvesting apples.

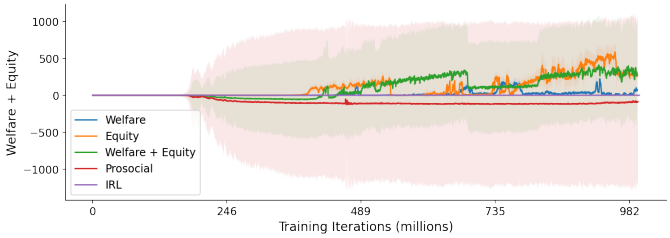


Figure 10: Plotting mean welfare + equity over time. We find that Equity and Welfare + Equity perform similarly well because Equity achieves high equity and Welfare + Equity achieves high welfare.

are empirical estimates of the optimal values of the local and global objectives. The empirical estimates were calculated by taking the maximum reward value over the different global objectives and baselines. The Welfare + Equity objective best trades off between local and global objectives whereas Equity and Welfare prioritize the global objective over the local objective.

With an understanding of how D3C trades off between global and local objectives across three domains, we further analyze whether having different global objectives can give us desirable properties such as zero-shot generalization in Clean Up.

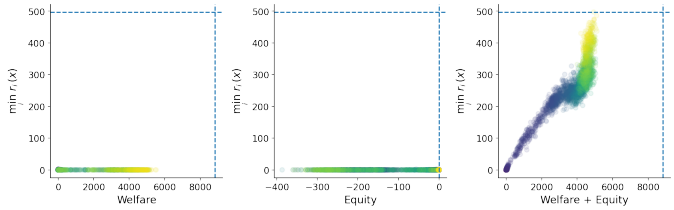


Figure 11: We plot how D3C trades off between global and local objectives over time with a KL coefficient of 0. The dotted blue lines are empirical estimates of the optimal values of the local and global objectives. The empirical estimates were calculated by taking the maximum reward value over the different global objectives and baselines. The Welfare + Equity objective best trades off between local and global objectives whereas Equity and Welfare optimizes for the global objective at the expense of the local objective.

Scenario 1	Visiting an altruistic bot population
Scenario 2	Our agents are resident and bots ride free
Scenario 3	Visiting a turn-taking bot population that cleans first
Scenario 4	Visiting a turn-taking bot population that eats first
Scenario 5	Our agents are visited by one reciprocator
Scenario 6	Our agents are visited by two suspicious reciprocators
Scenario 7	Our agents are visited by one suspicious reciprocator

Table 1: Description of Cleanup scenarios pairing D3C trained agents with held-out partners.

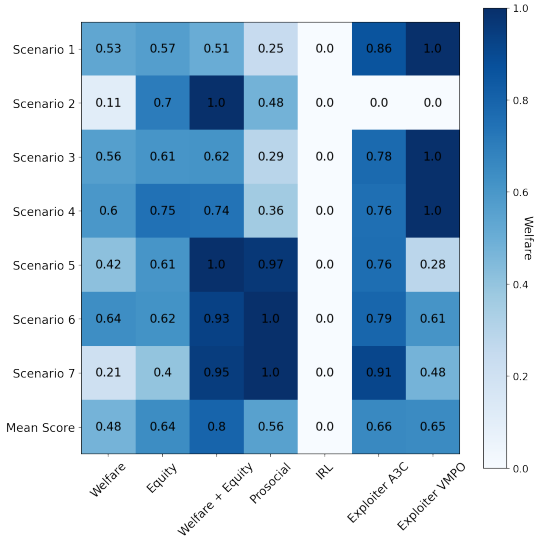


Figure 12: The x-axis represents runs with different D3C global objectives + baselines. The last two columns represent agents who were trained directly on the evaluation scenarios. The y-axis represents different evaluation scenarios (Table 1).

5.1 Zero-shot Generalization

We evaluate our agents on six test scenarios in Clean Up. Each scenario contains pre-trained bots that were not seen during training. The bots in each scenario display a different behavior outlined in Table 1 (e.g., free riding, turn-taking) that help us evaluate how well our trained agents can coordinate with these diverse unseen agents out-of-the-box. In Fig. 12 we report the normalized reward that our trained agents receive in each scenario (reward of 1 is the highest and 0 is the lowest), as well as the mean normalized reward. We also report the rewards that exploiter agents, or agents trained in the test scenarios achieve. These exploiter agent scores are meant to serve as an upper bound, however we cannot guarantee that their RL training process converged to a global maximum. Welfare + Equity obtains the highest overall mean reward followed by the exploiter agents. We suspect that Welfare + Equity generalizes well because this led to turn-taking behavior.

6 DISCUSSION AND CONCLUSION

Our analytical and empirical findings show: 1) Agents can optimize non-welfare global objectives via reward redistribution in both simple and complex domains. However, this is not always the case, and we investigate this analytically. 2) Achieving system-level global objectives can come at some cost to individual agent utilities, thus defining the trade-off between local and global objectives is critical. 3) Surprisingly, training on non-welfare objectives *can* actually lead to better performance on (some) held out test scenarios.

Taken holistically, our findings clearly motivate the value of enabling a multi-agent system (MAS) to automatically reconfigure agent loss functions to be more efficiently re-purposed for different global objectives. In other words, this type of framework is useful for *fast adaptation* of a MAS. Returning to the traffic domain, a system of self-driving cars not only needs to optimise routes for commute time and carbon emissions. Perhaps certain roads are experiencing heavier traffic, and the MAS needs to adapt its global objective to incorporate infrastructure sustainability. This investigation serves as initial deep dive into understanding value realignment with *more general* global objectives. As a result, it opens up many interesting follow-on questions for exploration.

6.1 Future Directions

6.1.1 Global Objectives. Our investigations focus on a small number of selected domains and global objectives that align well with desirable agent behaviors in those domains. It would be interesting to scale up this analysis and explore a larger and more diverse set of global objectives. For example, objectives can be a function of: (a) agents' *rewards* (maximal welfare, inequity), (b) joint *observations* (goal state), or (c) joint *actions* (desired behaviour). If it is a function of agent rewards, it can fall into different classes, for example a linear versus non-linear combination. When considering MARL domains, these factors can play an important role is computing and assessing tradeoffs between local and global objectives.

Importantly, where do or should global objectives *come from*? In this work, we predefined the global objectives. However, if we are interested in a value-aligned MAS, it is important to provide the capability of specifying goals and train the MAS on goals reflective of *real user* values and preferences. Ideally, specification would occur through natural language, as this lends a familiar and intuitive interface for humans and allows significantly more flexibility in specifying the goal. With that, one important consideration is that fine-tuning of user-specified goal prompts is critical, given current large language models (LLMs) [23]. A promising idea is to learn a global reward model, trained using reinforcement learning from human feedback (RLHF) [6, 34] or from AI feedback (RLAIF) [1].

6.1.2 Fair Allocation of Reward. In fully cooperative multi-agent settings, the *multi-agent credit assignment problem* [4] refers to the task of ascertaining individual agent contributions from the collective reward achieved. In problem settings we consider, agents receive local rewards from the environment; however, it is unclear to what extent these rewards reflect the *contribution* made. For example, in CleanUp, cleaning agents are pivotal for improving welfare because unless sufficient cleaning occurs, *no* apples are spawned. Yet only agents that eat apples are rewarded. Though D3C redistributes agent environmental rewards, its bias is to make the *minimal modification* necessary for improving welfare. Minimizing the Price of anarchy, however, is known in some cases to result in increased inequality [11]. Completely reassigning environmental rewards to lessen the impact of such issues requires more research.

6.1.3 Reward Sharing Mechanisms. Relaxing the constraint of budget balancing (e.g. through burning wealth [13]) and examining its impact on the expected performance is a particularly interesting question to explore in more depth. How would computed tradeoffs between local and global objectives change if agents were allowed to burn wealth, instead of having to distribute *all* wealth amongst the population? Another interesting question is around designing more sophisticated reward-sharing mechanisms. Some ideas include examining (a) non-linear combinations of agent rewards or (b) state-dependent reward mixtures.

6.1.4 Zero-Shot Generalisation. Our final finding about improved performance on a subset of held-out test scenarios was both interesting and unexpected. In particular, it was surprising because we had not *trained* agents with the goal of generalisation to unseen co-players at test time. Is it the case that agents generalise better on test scenarios where unseen players behave in ways that are similar to their training partners? Another hypothesis is that the reward-sharing mechanism inherently induces a more diverse set of agent policies amongst the training population (as compared to self-play training) and agents are able to be more adaptive in novel test scenarios. More research is needed to understand when this generalisation improvement effect is *expected* to occur.

This work provides in-depth analysis in a small number of selected domains on how to *automatically* modify AI agent objectives to enable a *multi-agent system* to be more *value-aligned*. Moreover, the analytical and empirical findings are promising, and they open up many exciting avenues for future research and exploration.

REFERENCES

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073* (2022).

[2] Martin Beckmann, Charles B McGuire, and Christopher B Winsten. 1956. *Studies in the Economics of Transportation.* Technical Report.

[3] Dietrich Braess. 1968. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12, 1 (1968), 258–268.

[4] Yu-Han Chang, Tracey Ho, and Leslie Kaelbling. 2003. All learning is local: Multi-agent learning in global reward games. *Advances in neural information processing systems* 16 (2003).

[5] Xi Chen and Xiaotie Deng. 2006. Settling the complexity of two-player Nash equilibrium. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE, 261–272.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[7] George Christodoulou and Elias Koutsoupias. 2005. The price of anarchy of finite congestion games. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*. ACM, 67–73.

[8] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. 2009. The complexity of computing a Nash equilibrium. *SIAM J. Comput.* 39, 1 (2009), 195–259.

[9] Ernst Fehr and Klaus M Schmidt. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* 114, 3 (1999), 817–868.

[10] Iason Gabriel and Vafa Ghazavi. 2021. The challenge of value alignment: From fairer algorithms to AI safety. *arXiv preprint arXiv:2101.06060* (2021).

[11] Kurtuluş Gemici, Elias Koutsoupias, Barnabé Monnot, Christos Papadimitriou, and Georgios Piliouras. 2018. Wealth inequality and the price of anarchy. *arXiv preprint arXiv:1802.09269* (2018).

[12] Ian Gemp, Kevin R McKee, Richard Everett, Edgar Duéñez-Guzmán, Yoram Bachrach, David Balduzzi, and Andrea Tacchetti. 2022. D3C: Reducing the Price of Anarchy in Multi-Agent Learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 498–506.

[13] Jason D Hartline and Tim Roughgarden. 2008. Optimal mechanism design and money burning. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*. ACM, 75–84.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Elias Koutsoupias and Christos Papadimitriou. 1999. Worst-case equilibria. In *Annual Symposium on Theoretical Aspects of Computer Science*. Springer, 404–413.

[16] Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. 2021. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International conference on machine learning*. PMLR, 6187–6199.

[17] Andrei Lupu and Doina Precup. 2020. Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*. 789–797.

[18] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duéñez-Guzmán, Edward Hughes, and Joel Z Leibo. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*. 869–877.

[19] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. 1928–1937.

[20] Roger B Myerson. 1981. Utilitarianism, egalitarianism, and the timing effect in social choice problems. *Econometrica: Journal of the Econometric Society* (1981), 883–897.

[21] Anna Nagurney and Ding Zhang. 2012. *Projected dynamical systems and variational inequalities with applications*. Vol. 2. Springer Science & Business Media.

[22] William Neuman and Michael Barbaro. 2009. Mayor Plans to Close Parts of Broadway to Traffic. https://www.nytimes.com/2009/02/26/nyregion/26broadway.html. *NYTimes.com* (Feb 2009).

[23] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

[24] Alexander Peysakhovich and Adam Lerer. 2017. Prosocial learning agents solve generalized stag hunts better than selfish ones. *arXiv preprint arXiv:1709.02865* (2017).

[25] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 4295–4304.

[26] Carlton Reid. 2022. Historic Congested Four-Lane Road In Central London Closes To Motorists, Opens To People. https://www.forbes.com/sites/carltonreid/2022/12/26/historic-and-congested-four-lane-road-in-central-london-closes-to-motorists-opens-to-people/?sh=7cc136182e06. *Forbes.com* (Dec 2022).

[27] Kevin Roberts. 1979. The characterization of implementable choice rules. *Aggregation and revelation of preferences* 12, 2 (1979), 321–348.

[28] Tim Roughgarden. 2015. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)* 62, 5 (2015), 32.

[29] H Francis Song, Abbas Abdolmaleki, Jost Tobias Springenberg, Aidan Clark, Hubert Soyer, Jack W Rae, Seb Noury, Arun Ahuja, Siqi Liu, Dhruva Tirumala, et al. 2019. V-mpo: On-policy maximum a posteriori policy optimization for discrete and continuous control. *arXiv preprint arXiv:1909.12238* (2019).

[30] Richard Steinberg and Willard I Zangwill. 1983. The prevalence of Braess' paradox. *Transportation Science* 17, 3 (1983), 301–318.

[31] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296* (2017).

[32] John Glen Wardrop. 1952. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers* 1, 3 (1952), 325–362.

[33] Hyejin Youn, Michael T Gastner, and Hawoong Jeong. 2008. Price of anarchy in transportation networks: efficiency and optimality control. *Physical Review Letters* 101, 12 (2008), 128701.

[34] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).