

The Battling Influencers Game: Nash Equilibria Structure of a Potential Game and Implications to Value Alignment

Young Wu, Yancheng Zhu, Jin-Yi Cai, Xiaojin Zhu

Department of Computer Sciences, University of Wisconsin–Madison

Abstract

When multiple influencers attempt to compete for a receiver’s attention, their influencing strategies must account for the presence of one another. We introduce the Battling Influencers Game (BIG), a multi-player simultaneous-move general-sum game, to provide a game-theoretic characterization of this social phenomenon. We prove that BIG is a potential game, that it has either one or an infinite number of pure Nash equilibria (NEs), and these pure NEs can be found by convex optimization. Interestingly, we also prove that at any pure NE, all (except at most one) influencers must exaggerate their actions to the maximum extent. In other words, it is rational for the influencers to be non-truthful and extreme because they anticipate other influencers to cancel out part of their influence. We discuss the implications of BIG to value alignment.

1. Introduction

Life is full of agents who want to influence others: Food truck vendors entice us with BBQ samples; Social media influencers review selective pickleball brands to persuade us; Editors publish op-eds to sway public opinions. When multiple influencers with conflicting interests battle for our attention, intuitively they would be strategic and adjust their actions to account for the presence of one another in order to be effective.

This paper presents a game theoretic definition of the *Battling Influencers Game (BIG)*. We model the influencers as players in a multi-player simultaneous-move general-sum game. Our main technical result is that BIG is a potential game with special pure Nash Equilibria structures. Consequently, we can predict how rational influencers would adjust their strategies in the battle: **exaggeration is inevitable**. This prediction may shed new computational light on the genesis of misinformation.

As a use case, BIG can be applied to the AI value alignment problem. The receivers of the influence were traditionally

people, but can extend to AI value alignment algorithms. However, unlike in standard machine learning, our focus is not on the value alignment algorithm itself. Instead, BIG predicts how battling alignment-data providers could be motivated to intentionally produce training data that do not truthfully reflect their values. While out of scope for the current paper, our insight can help design future value alignment algorithms to remove such incentives.

2. Related Work

Our work provides a game theoretic model of the numerical example and informal theorem in section 5 of (Park et al., 2024), in particular, we also assume strategic data providers (which we call influencers) to large language models (which we call receivers), and our model leads to results consistent with their example where the data providers untruthfully report their opinions. In addition, we prove the existence of pure strategy Nash equilibria of this class of games, and we show the property that almost all influencers maximally exaggerate their preferences in every equilibrium. Our work is also closely related to (Hao & Duan, 2024), which models the interaction between multiple influencers by a dynamic Bayesian game. They described the phenomenon of strategic extreme exaggeration, which is also discussed in (Sun et al., 2024), (Soumalias et al., 2024), (Conitzer et al., 2024), and (Roughgarden & Schrijvers, 2017) for various applications, but they do not explicitly compute the equilibria of the original game or quantify the amount of exaggeration. In comparison, we use a static game with known influencer types and we are able to provide better characterizations of the set of equilibria of the game.

Value alignment aims to make language models produce outputs that are more aligned with human values. Existing training frameworks tailored for this purpose, such as (Ouyang et al., 2022) and (Rafailov et al., 2024), collect preference data from humans and train a large language model to follow users’ intent. However, research in this direction mostly focuses on the algorithmic aspects of value alignment and does not emphasize the heterogeneity of human values. There has been work that studies how to make LLMs align with diverse preferences of different demo-

graphic groups (Bakker et al., 2022; Chen et al., 2024), but they do not have a rigorous game theoretic foundation that characterizes strategic behaviors of preference data providers. There is also recent work by (Munos et al., 2023), (Swamy et al., 2024), and (Rosset et al., 2024) on learning a pairwise or general preference model, where they used the term Nash learning or Direct Nash Optimization. The players in their zero-sum games are not strategic data providers, and they use the Nash equilibrium solution concept mainly as an optimization technique to solve their minimax problem.

3. Problem Definition

The Battling Influencer Game (BIG) is an n -player simultaneous-move game. The players are the n influencers. The players have a common continuous action space $\mathcal{X} \subset \mathbb{R}^d$ which is compact and convex. Let $\mathbf{x}_i \in \mathcal{X}$ be the action of the i th player for $i \in [n] := \{1, 2, \dots, n\}$. For example, \mathbf{x}_i could be the embedding vector of the text corpus that influencer i produces. A joint action, or pure strategy profile, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ denotes the simultaneous action choice of all influencers. As in standard literature, we also write $\mathbf{x} = (\mathbf{x}_i, \mathbf{x}_{-i})$ when we want to emphasize player i .

For the narrative, we posit a receiver whom the influencers want to influence. The receiver is not a strategic agent and not a player of the game. Like an impressionable person, the receiver aggregates to various degrees the inputs it receives from all influencers. In this paper we consider affine receivers of the form

$$\hat{\mathbf{x}} := w_0 \mathbf{x}_0 + \sum_{i=1}^n w_i \mathbf{x}_i, \quad (1)$$

where $w_i, i \in [n]$ is a real-valued (not necessarily normalized) weight that signifies how much influence influencer i has on the receiver. For example, a company which can afford to buy more ads has a larger w_i compared to a company with a smaller budget. $\mathbf{x}_0 \in \mathcal{X}$ is a bias term that, together with $w_0 \in \mathbb{R}$, denotes a fixed, constant ‘‘background’’ influence that is beyond the control of the n influencers. The receiver (1) is common knowledge to all players.

The n influencers each has a target $\mathbf{t}_i \in \mathbb{R}^d$ (i.e. the target is not restricted to \mathcal{X}). For example, \mathbf{t}_i could be the embedding vector of the published party manifesto of political party i . The goal of influencer i is to drive the receiver’s $\hat{\mathbf{x}}$ close to \mathbf{t}_i . This goal is reflected in the loss (negative utility) function of influencer i . For concreteness, here we consider squared 2-norm as the loss:

$$\ell_i(\mathbf{x}) := \|\hat{\mathbf{x}} - \mathbf{t}_i\|_2^2 = \|w_0 \mathbf{x}_0 + \sum_{i=1}^n w_i \mathbf{x}_i - \mathbf{t}_i\|_2^2. \quad (2)$$

(See Section 5.1 for an alternative using cosine similarity.) As rational agents, the influencers want to selfishly minimize

their own $\ell_i(\mathbf{x})$. The fact that the receiver’s $\hat{\mathbf{x}}$ is defined by the joint action \mathbf{x} couples the influencers together in a general-sum game. It is due to the possible differences in $\mathbf{t}_1, \dots, \mathbf{t}_n$ that the influencers battle one another.

The above narrative can be abstracted into the following formal definition of BIG, where the receiver becomes implicit:

Definition 1 (Battling Influencer Game (BIG)). The Battling Influencer Game is an n -player general-sum game $G = (n, \mathcal{X}, \{\ell_i\}_{i=1}^n)$. where n is the number of influencers, $\mathcal{X} \subset \mathbb{R}^d$ is a compact and convex action space, and $\ell_i : \mathcal{X}^n \mapsto \mathbb{R}$ taking the form of equation (2) is player i ’s loss function. The parameters $\mathbf{x}_0, \{w_i\}_{i=0}^n, \{\mathbf{t}_i\}_{i=1}^n$ of the loss functions are common knowledge to all players.

In the rest of the paper, we are interested in finding the pure strategy Nash equilibria (NEs) of the game G . We then characterize properties of these NEs, interpreting them in the context of influencers.

4. Pure NEs of BIG and Their Properties

Definition 2. A pure strategy Nash equilibrium of the game $G = (n, \mathcal{X}, \{\ell_i\}_{i=1}^n)$ is a strategy profile $\mathbf{x} \in \mathcal{X}^n$ satisfying,

$$\ell_i(\mathbf{x}_i, \mathbf{x}_{-i}) \leq \ell_i(\mathbf{y}, \mathbf{x}_{-i}), \forall \mathbf{y} \in \mathcal{X}, i \in [n]. \quad (3)$$

A mixed strategy Nash equilibrium is a pure strategy Nash equilibrium of the mixed extension $G' = (n, \Delta\mathcal{X}, \{\ell_i\}_{i=1}^n)$ where the set of actions for each player is a distribution (called a mixed strategy) over the original action set \mathcal{X} , that is, a mixed strategy profile $s_{1:n}$ with $s_i \in \Delta\mathcal{X}$ satisfying,

$$\mathbb{E}[\ell_i(s_i, s_{-i})] \leq \mathbb{E}[\ell_i(s', s_{-i})], \forall s' \in \Delta\mathcal{X}, i \in [n]. \quad (4)$$

In general, for finite games, for example, when \mathcal{X} is finite, there exists at least one mixed strategy Nash equilibrium, but computing the Nash equilibrium is PPAD-complete (Polynomial Parity Arguments on Directed graphs). When \mathcal{X} is not finite, which is usually the case for our BIG problem, a mixed strategy Nash equilibrium is not guaranteed to exist.

Potential games are games with a special structure and allow strong results on pure Nash equilibria. We will show BIG is a potential game. The difficulty is in finding the potential function. The following theorem provides a constructive proof.

Theorem 1 (Potential Game). *The Battling Influencers Game G is a potential game with the potential function*

$$\phi(\mathbf{x}) = \phi(\mathbf{x}_1, \dots, \mathbf{x}_n) := \left\| \sum_{i=0}^n w_i \mathbf{x}_i \right\|_2^2 - 2 \sum_{i=1}^n w_i \mathbf{t}_i^\top \mathbf{x}_i. \quad (5)$$

Proof. We need to show if any player i deviates from action \mathbf{x}_i to any action $\mathbf{y} \in \mathcal{X}$, we have $\ell_i(\mathbf{x}_i, \mathbf{x}_{-i}) - \ell_i(\mathbf{y}, \mathbf{x}_{-i}) = \phi(\mathbf{x}_i, \mathbf{x}_{-i}) - \phi(\mathbf{y}, \mathbf{x}_{-i})$. To this end, define an auxiliary variable \mathbf{z} that does not depend on \mathbf{x}_i or \mathbf{y} :

$$\mathbf{z} := w_0 \mathbf{x}_0 + \sum_{j \neq i} w_j \mathbf{x}_j.$$

Then $\ell_i(\mathbf{x}_i, \mathbf{x}_{-i}) = \|w_i \mathbf{x}_i + (\mathbf{z} - \mathbf{t}_i)\|^2 = \|w_i \mathbf{x}_i\|^2 + 2w_i \mathbf{x}_i^\top (\mathbf{z} - \mathbf{t}_i) + \|\mathbf{z} - \mathbf{t}_i\|^2$, and

$$\begin{aligned} \ell_i(\mathbf{x}_i, \mathbf{x}_{-i}) - \ell_i(\mathbf{y}, \mathbf{x}_{-i}) &= \|w_i \mathbf{x}_i\|^2 + 2w_i \mathbf{x}_i^\top (\mathbf{z} - \mathbf{t}_i) - \|w_i \mathbf{y}\|^2 - 2w_i \mathbf{y}^\top (\mathbf{z} - \mathbf{t}_i). \end{aligned}$$

On the other hand,

$$\begin{aligned} \phi(\mathbf{x}_i, \mathbf{x}_{-i}) &= \|w_i \mathbf{x}_i + \mathbf{z}\|^2 - 2w_i \mathbf{t}_i^\top \mathbf{x}_i - 2 \sum_{j \neq i} w_j \mathbf{t}_j^\top \mathbf{x}_j \\ &= \|w_i \mathbf{x}_i\|^2 + 2w_i (\mathbf{z} - \mathbf{t}_i)^\top \mathbf{x}_i + \|\mathbf{z}\|^2 - 2 \sum_{j \neq i} w_j \mathbf{t}_j^\top \mathbf{x}_j. \end{aligned}$$

The last two terms do not depend on \mathbf{x}_i . Hence

$$\begin{aligned} \phi(\mathbf{x}_i, \mathbf{x}_{-i}) - \phi(\mathbf{y}, \mathbf{x}_{-i}) &= \|w_i \mathbf{x}_i\|^2 + 2w_i (\mathbf{z} - \mathbf{t}_i)^\top \mathbf{x}_i - \|w_i \mathbf{y}\|^2 - 2w_i (\mathbf{z} - \mathbf{t}_i)^\top \mathbf{y} \\ &= \ell_i(\mathbf{x}_i, \mathbf{x}_{-i}) - \ell_i(\mathbf{y}, \mathbf{x}_{-i}). \end{aligned}$$

□

We next show that $\mathbf{x} \in \mathcal{X}^n$ is a pure NE of G if and only if \mathbf{x} is a minimum of ϕ restricted to the domain \mathcal{X}^n .

Proposition 2 (Pure NEs \iff minima). *The set of pure Nash equilibria in G is*

$$\text{pNE}(G) = \underset{\mathbf{x} \in \mathcal{X}^n}{\text{argmin}} \phi(\mathbf{x}). \quad (6)$$

Proof. Our potential function ϕ is the sum of two convex functions in \mathbf{x} and hence convex. Since the domain \mathcal{X}^n is convex and ϕ is smooth and convex, by Theorem 1 of (Neyman, 1997) and its corollary, the set of pure Nash equilibria coincides with the minima of the potential function on \mathcal{X}^n . □

We remark that by definition \mathcal{X} is compact and convex, thus \mathcal{X}^n is bounded and closed. The potential function $\phi(\mathbf{x})$ may not have a global minimum on the extended domain \mathbb{R}^{nd} (it could diverge to $-\infty$ there), but on \mathcal{X}^n it will have at least one minimum (perhaps on the boundary). In fact, we have the following guarantee.

Corollary 3 (Cardinality of $\text{pNE}(G)$). *G has either one pure NE or infinite pure NEs.*

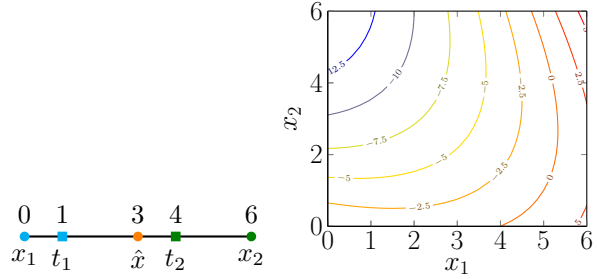


Figure 1. (left) Both players maximally exaggerate their actions with pure NE $(x_1 = 0, x_2 = 6)$. (right) the potential function ϕ

Proof. Since the set of minima of the convex potential function on a compact domain \mathcal{X}^n is non-empty, there is at least one pure Nash equilibrium. Since the set of minima of the convex potential function on a convex domain \mathcal{X}^n is convex (corollary in (Neyman, 1997)), any linear combination of two distinct pure Nash equilibria is another pure Nash equilibrium. □

We now provide a few illustrative examples of BIG.

Example 1 (Two influencers with 1D actions). There are $n = 2$ influencers whose individual action space is $\mathcal{X} = [a, b]$ (we use $a = 0, b = 6$ in Figures 1 and 2). The receiver takes the average: $\hat{x} = \frac{x_1 + x_2}{2}$. If the targets satisfy $t_1 < \frac{a+b}{2}$ and $t_2 > \frac{a+b}{2}$ as in Figure 1, then there is a unique pure NE: $\text{pNE}(G) = \{(x_1 = a, x_2 = b)\}$. Note $(x_1 = t_1, x_2 = t_2)$ is in general not a NE: if $x_2 = t_2$, the first influencer will want to take a more extreme action $x_1 < t_1$ to drive the receiver $\hat{x} = \frac{x_1 + x_2}{2}$ closer to its target t_1 , and vice versa for the second influencer, *ad infinitum* until they hit the boundary. In other words, **both influencers must exaggerate their actions to the maximum extent possible**. In fact, this is the well-known best response dynamics in potential games which we discuss later. As a result, both influencers will end up playing at the opposite boundary of \mathcal{X} . No one is entirely happy: the receiver ends up in the middle $\hat{x} = \frac{a+b}{2}$ so no influencer achieves their target. Still, this is the best each influencer can do under the presence of other influencers.

A more nuanced situation happens if t_1, t_2 are on the same side of $\frac{a+b}{2}$, for example the left side in Figure 2. There is still a unique pure NE; at the NE both influencers still need to misrepresent their target. However, the influencer whose target is closer to the center point (in this example, t_2) can claim victory: The other influencer simply runs out of more left-leaning actions and has to stop at the left boundary $x_1 = a$. The winning influencer best-responds with $x_2 = 2t_2 - a$, so that the receiver will end up at its target $\hat{x} = t_2$. Thus $\text{pNE}(G) = \{(x_1 = a, x_2 = 2t_2 - a)\}$ in this example. Interestingly, if in addition $t_2 > (a + \frac{a+b}{2})/2$ then influencer 2 indeed has a left-leaning target but has

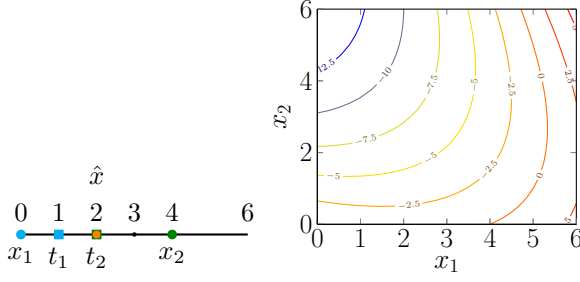


Figure 2. Pure NE ($x_1 = 0, x_2 = 4$). Influencer 2 not at boundary.

to misrepresent itself as right-leaning ($x_2 > \frac{a+b}{2}$) to the receiver. We will generalize this example in Theorem 4, where we show at most one influencer can be interior.

Example 2 (2D actions). Same as Example 1 but let $d = 2$ and $\mathcal{X} = [-a, a] \times [-a, a]$. The game may now have an infinite number of pure NEs. For example, let the targets be $\mathbf{t}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{t}_2 = -\mathbf{t}_1$ as in Figure 3 (left). Then $\text{pNE}(G) = \left\{ \left(\begin{bmatrix} -a \\ -z \end{bmatrix}, \begin{bmatrix} a \\ z \end{bmatrix} \right) : z \in [-a, a] \right\}$. These infinite many pure NEs are indicated by the blue and green line segments, paired through the origin. All of them are exaggerations from both influencers in terms of the x -axis. All of them result in the receiver arriving at the origin.

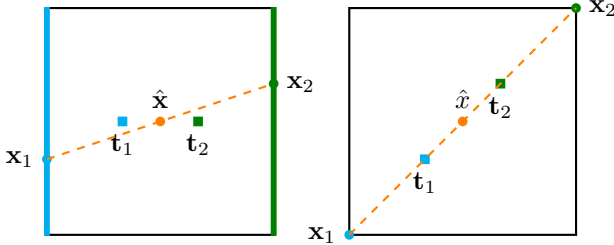


Figure 3. Examples of infinite (left) and unique (right) pure Nash equilibria in $d = 2$

In contrast, if the targets are arranged as in Figure 3 (right), there will only be a unique pure NE: $\text{pNE}(G) = \left\{ \left(\begin{bmatrix} -a \\ -a \end{bmatrix}, \begin{bmatrix} a \\ a \end{bmatrix} \right) \right\}$.

As seen from these examples, in BIG the pure NEs often involve all influencers misrepresenting their true target to the receiver (i.e. $\mathbf{x}_i \neq \mathbf{t}_i$). Furthermore, such misrepresentation often takes the form of *extreme exaggeration*, in the sense that an influencer's rational action \mathbf{x}_i at any pure NE is often pushed to the boundary of the action space \mathcal{X} so they cannot exaggerate the action further. Our next theorem precisely quantifies this phenomenon. We prove that, provided that the influencers' targets $\mathbf{t}_1 \dots \mathbf{t}_n$ are all distinct, **at any pure**

NE extreme exaggeration is necessary for all but at most one influencer. It is possible that all influencers must perform extreme exaggeration. Furthermore, if there exists one influencer (say influencer i^*) who does not, then it is the winner in that the receiver will end up at its target \mathbf{t}_{i^*} ; Still, this winning influencer in general also need to misrepresent its target $\mathbf{x}_{i^*} \neq \mathbf{t}_{i^*}$ to the receiver, it is just that \mathbf{x}_{i^*} is in the interior of \mathcal{X} and not extreme.

Theorem 4 (All-But-At-Most-One Extreme Exaggeration). *If $\{\mathbf{t}_i\}_{i=1}^n$ are all distinct, then every pure NE $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ satisfies the property that,*

$$|\{i \in [n] : \mathbf{x}_i \in \text{int } \mathcal{X}\}| \leq 1. \quad (7)$$

Furthermore, if $\mathbf{x}_{i^*} \in \text{int } \mathcal{X}$ for some $i^* \in [n]$, then,

$$\hat{\mathbf{x}} = \mathbf{t}_{i^*}. \quad (8)$$

Proof. For any $i \in [n]$,

$$\nabla_{\mathbf{x}_i} \phi = 2w_i \sum_{k=0}^n w_k \mathbf{x}_k - 2w_i \mathbf{t}_i. \quad (9)$$

Suppose $\exists i, j \in [n], i \neq j : \mathbf{x}_i, \mathbf{x}_j \in \text{int } \mathcal{X}$. Then

$$\nabla_{\mathbf{x}_i} \phi = \mathbf{0} = \nabla_{\mathbf{x}_j} \phi \quad (10)$$

$$\Rightarrow \mathbf{t}_i = \sum_{k=0}^n w_k \mathbf{x}_k = \mathbf{t}_j, \quad (11)$$

a contradiction. Equation (8) follows from $\nabla_{\mathbf{x}_{i^*}} \phi = \mathbf{0}$ and the definition of the receiver (1). \square

Finally, we remark on algorithms to find pure NEs for BIG. Due to Proposition 2, any convex optimization algorithm that minimizes the convex function ϕ over the convex set \mathcal{X}^n with strong guarantees can be utilized to find a pure NE (Boyd & Vandenberghe, 2004). Meanwhile, in the game theory community the best-response dynamics is a traditional algorithm for finding a pure NE in potential games: (Roughgarden, 2010)

Definition 3 (Best Response Dynamics). Starting from an arbitrary $\mathbf{x}^{(0)} \in \mathcal{X}^n$, the best response sequence $\{\mathbf{x}^{(t)}\}_{t=1}^{\infty}$ converges to a pure Nash equilibrium of G , where,

$$\mathbf{x}_i^{(t)} = \begin{cases} \operatorname{argmin}_{\mathbf{x}_i} \phi(\mathbf{x}_i, \mathbf{x}_{-i}^{(t-1)}) & \text{if } i = t \pmod n \\ \mathbf{x}_i^{(t-1)} & \text{otherwise.} \end{cases} \quad (12)$$

Of course, best response dynamics correspond to coordinate descent on ϕ . One interesting observation that is relevant for BIG is that, under best response dynamics, no influencer needs to know other influencers' targets. This removes the requirement that $\mathbf{t}_1, \dots, \mathbf{t}_n$ must be common knowledge.

Concretely, the influencers may carry out the best response dynamics as a learning dynamics in a distributed fashion over time, with no two influencers simultaneously updating their actions. When influencer i updates its own action \mathbf{x}_i , it observes other player's most recent actions \mathbf{x}_{-i} but does not need to know their targets \mathbf{t}_{-i} . This is because minimizing the potential function ϕ along the \mathbf{x}_i direction is equivalent to minimizing its own loss function ℓ_i :

$$\operatorname{argmin}_{\mathbf{x}_i} \phi(\mathbf{x}_i, \mathbf{x}_{-i}) = \operatorname{argmin}_{\mathbf{x}_i} \ell_i(\mathbf{x}_i, \mathbf{x}_{-i}). \quad (13)$$

Therefore, the best response dynamics may offer a computational account on how influencers in the real world iteratively adjust their actions based on actions of other influencers without knowing the other influencers' true intentions, and still reaching an equilibrium.

5. Extensions of BIG

5.1. An Alternative Player Loss Function

Up to now influencer i 's loss function (2) is based on the Euclidean distance between its target point $\mathbf{t}_i \in \mathcal{X}$ and the receiver $\hat{\mathbf{x}}$. In some applications, the following *negative inner product loss* can be more appropriate:

$$\ell_i(\mathbf{x}) := -\mathbf{t}_i^\top \hat{\mathbf{x}}. \quad (14)$$

That is, influencer i has a small loss if the receiver $\hat{\mathbf{x}}$ has a large projection onto the direction of target direction \mathbf{t}_i .

BIG with this negative inner product loss (14) has an even stronger guarantee: the game has a *Weakly Dominant Strategy Equilibrium*.

Definition 4 (Weakly Dominant Strategy Equilibrium (wDSE)). An action profile $\mathbf{x} = (\mathbf{x}_1 \dots \mathbf{x}_n) \in \mathcal{X}^n$ is a wDSE if for every player i ,

$$\ell_i(\mathbf{x}_i, \mathbf{x}_{-i}) \leq \ell_i(\mathbf{y}, \mathbf{x}_{-i}), \quad \forall \mathbf{y} \in \mathcal{X}, \mathbf{x}_{-i} \in \mathcal{X}^{n-1}. \quad (15)$$

Remark 1. The term weakly dominant strategy equilibrium is used in Chapter 4.5 of (Tadelis, 2013), and it is also called dominant strategy equilibrium in Chapter 10.3 of (Osborne, 1994), and dominant strategy solution in Chapter 1.3.1 of (Roughgarden, 2010). As noted in (Osborne, 1994), an action in a wDSE is not required to weakly dominate all other actions for a player since the player could have multiple actions that are equivalent, all of which dominate the remaining actions. wDSE is also a weaker solution concept compared to (strictly) dominant strategy equilibrium, which requires strict inequality everywhere,

$$\ell_i(\mathbf{x}_i, \mathbf{x}_{-i}) < \ell_i(\mathbf{y}, \mathbf{x}_{-i}), \quad \forall \mathbf{y} \in \mathcal{X}, \mathbf{x}_{-i} \in \mathcal{X}^{n-1}. \quad (16)$$

Theorem 5 (Existence of wDSE). *Under loss (14), game G has a wDSE \mathbf{x}_i^* satisfying,*

$$\mathbf{x}_i^* \in \operatorname{argmax}_{\mathbf{x}_i \in \mathcal{X}} w_i \mathbf{t}_i^\top \mathbf{x}_i, \quad \forall i \in [n]. \quad (17)$$

Proof. Given an arbitrary joint action from other players \mathbf{x}_{-i} , player i 's best response is

$$BR(\mathbf{x}_{-i}) := \operatorname{argmin}_{\mathbf{x}_i \in \mathcal{X}} \ell_i(\mathbf{x}_i, \mathbf{x}_{-i}) \quad (18)$$

$$= \operatorname{argmin}_{\mathbf{x}_i \in \mathcal{X}} -\mathbf{t}_i^\top \hat{\mathbf{x}} \quad (19)$$

$$= \operatorname{argmin}_{\mathbf{x}_i \in \mathcal{X}} -\mathbf{t}_i^\top \left(w_0 \mathbf{x}_0 + \sum_{j=1}^n w_j \mathbf{x}_j \right) \quad (20)$$

$$= \operatorname{argmin}_{\mathbf{x}_i \in \mathcal{X}} -w_i \mathbf{t}_i^\top \mathbf{x}_i + \text{const} = \mathbf{x}_i^*. \quad (21)$$

The best response is independent of \mathbf{x}_{-i} . \square

One significant benefit of this wDSE is that each influencer can compute their own \mathbf{x}_i^* without the knowledge of other influencers' weights w_j and target \mathbf{t}_j , $\forall j \neq i$. This removes the requirement that $\mathbf{x}_0, \{w_{0:n}\}, \{\mathbf{t}_{1:n}\}$ are common knowledge from Definition 1. The wDSE action \mathbf{x}_i^* in (17) is determined in part by the sign of w_i . When $w_i < 0$ this is akin to reverse psychology: knowing that the receiver will flip its action direction in (1), influencer i should intentionally go against its own target. We remark that \mathbf{x}_i^* may not be exactly along the direction of $w_i \mathbf{t}_i$ since it depends on the domain \mathcal{X} . Nonetheless, computing \mathbf{x}_i^* is a convex optimization problem—maximizing a linear function over a convex set—and thus efficient.

Example 3 (Unique wDSE). In this example, $d = 2$ and \mathcal{X} is given by the diamond shape in Figure 4 (left), with $\mathbf{t}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{t}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The unique wDSE is the action profile containing the extreme points in \mathcal{X} in the direction of \mathbf{t}_1 and \mathbf{t}_2 .

Example 4 (Infinite wDSE). In this example, $d = 2$ and $\mathcal{X} = [-a, a] \times [-a, a]$, with $\mathbf{t}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{t}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. All actions for player 1 along the blue line are equivalent and lead to the same loss regardless of player 2's action, and all other actions in \mathcal{X} except for these actions along the blue line are strictly dominated. Similarly, all actions for player 2 along the green line are equivalent and strictly dominate all other actions. As a result, given Definition 4, any pair of actions, one on the blue line for player 1 and one on the green line for player 2, is a wDSE. There are infinite number of them.

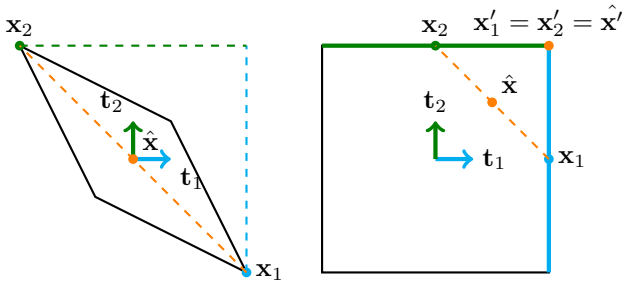


Figure 4. Examples of unique wDSE (left) and infinite number of wDSEs (right) in $d = 2$

5.2. Finite Action Space

So far, we have assumed that the influencers' action space \mathcal{X} is a compact and convex (hence infinite unless singleton) subset of \mathbb{R}^d . In some applications, the influencers are restricted to picking their actions from a finite \mathcal{X} instead. For example, \mathcal{X} may be the collection of news articles published by all professional news agencies within the past 24 hours, and each influencer may select a handful of such news articles to place on a social media user (the receiver)'s timeline. This motivates the extension to finite action space:

Definition 5 (BIG with finite action space). Battling Influencer Game with finite action space is an n -player general-sum game $F = (n, \{\mathcal{D}^{(k_i)}\}_{i=1}^n, \{\ell_i\}_{i=1}^n)$, where the action space of player i , $\mathcal{D}^{(k_i)}$ is the set of all subsets containing k_i elements (optionally allow repeats) from a finite $\mathcal{X} \subset \mathbb{R}^d$. The loss function of player i is given by $\ell_i(\mathbf{x}) = \|\hat{\mathbf{x}} - \mathbf{t}_i\|_2^2$ with

$$\hat{\mathbf{x}} = w_0 \mathbf{x}_0 + \sum_{i=1}^n w_i \sum_{j=1}^{k_i} \mathbf{x}_i^{(j)}, \quad (22)$$

where $\mathbf{x}_i^{(j)}$ is the j th element in player i 's chosen subset. The parameters $\mathbf{x}_0, \{w_i\}_{i=0}^n, \{\mathbf{t}_i\}_{i=1}^n$ and $\{k_i\}_{i=0}^n$ are common knowledge to all players.

In the original BIG G (Definition 1) where \mathcal{X} is convex, allowing the players to choose multiple items would not affect the results since choosing multiple items is equivalent to choosing the mean of these items, which is still in \mathcal{X} . In the new game, the average of items in \mathcal{X} may not be in \mathcal{X} . However, the new game can be interpreted as each influencer picks one "meta item" \mathbf{x}'_i instead of k_i items,

$$\text{with } w'_i = w_i k_i, \mathbf{x}'_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \mathbf{x}_i^{(j)}.$$

Proposition 6. F is also a potential game, with a potential

$$\text{function } \phi \left(\left(\left\{ \left\{ \mathbf{x}_i^{(j)} \right\}_{j=1}^{k_i} \right\}_{i=1}^n \right) \right) := \left\| w_0 \mathbf{x}_0 + \sum_{i=1}^n w'_i \mathbf{x}'_i \right\|_2^2 - 2 \sum_{i=1}^n w'_i \mathbf{t}_i^\top \mathbf{x}'_i. \quad (23)$$

Therefore, the new game F also has at least one pure NE. Since \mathcal{X} is finite and the number of players is finite, F is a finite game. A pure NE can be found through best response dynamics (Roughgarden, 2010). Unlike the continuous case, each iteration of best response dynamics can be costly to compute for large values of k_i since it involves solving a variant of the subset sum problem.

Proposition 6 shows that the new game F must have at least one pure NE. We now show a non-trivial example where F can have an exponential number of pure NEs even when there are only $n = 2$ players. This is true even if \mathcal{X} contains distinct elements, and the influencers cannot repeat chosen elements.

Example 5 (Many pure NEs in F). Consider an instance of game F with $n = 2, d = 1, t_1 = -\frac{1}{4}, t_2 = \frac{1}{4}, k_1 = k_2 = k$, the receiver takes simple average $\hat{x} = \frac{1}{2k} \left(\sum_{j=1}^k x_1^{(j)} + \sum_{j=1}^k x_2^{(j)} \right)$. Let $\mathcal{X} = \mathcal{X}_- \cup \mathcal{X}_+$ where

$$\mathcal{X}_- := \{-2k \cdot 2^i\}_{i=0}^{|\mathcal{X}|/2-1}, \mathcal{X}_+ := \{2k \cdot 2^i\}_{i=0}^{|\mathcal{X}|/2-1}. \quad (24)$$

Assume $|\mathcal{X}| \geq 2k$. Consider any player 2 action \mathbf{x}_2 which is a subset of \mathcal{X}_+ of size k . Note $\frac{1}{2k} \sum_{j=1}^k x_2^{(j)}$ is a positive integer with \mathbf{x}_2 indexing its binary representation. For example, if $k = 2$ and $\mathbf{x}_2 = \{2k \cdot 2^0, 2k \cdot 2^1\}$ then this integer is 3. What is player 1's best response to player 2? Given player 1's target $t_1 = -\frac{1}{4}$, player 1 should take action \mathbf{x}_1 which selects the corresponding negative items in \mathcal{X}_- , for example $\mathbf{x}_1 = \{-2k \cdot 2^0, -2k \cdot 2^1\}$. The joint action $\mathbf{x}_1, \mathbf{x}_2$ brings the receiver to $\hat{x} = 0$. This is the best that player 1 can do: any other \mathbf{x}'_1 (i.e. a subset of size k of \mathcal{X}) is not a best response because it changes \hat{x} to a different integer, which is farther away from t_1 compared to $\hat{x} = 0$. Conversely, \mathbf{x}_2 is also the best response to that \mathbf{x}_1 . Therefore, $(\mathbf{x}_1, \mathbf{x}_2)$ forms a pure NE. Now, player 2 could have started with $\binom{|\mathcal{X}|/2}{k}$ different \mathbf{x}_2 subsets, each corresponds to a different pure NE. Therefore, if we allow k to grow with $|\mathcal{X}|$ such as $k = |\mathcal{X}|/4$, this game has an exponential number of pure NEs.

6. Implications of BIG to Value Alignment

6.1. Heterogeneous Value Alignment as a Game

We now show empirically that a stylized version of value alignment can be modeled by BIG. We are interested in the

setting where multiple people with heterogeneous values provide feedback data (Santurkar et al., 2023; Bakker et al., 2022; Chen et al., 2024), and that they are aware of the presence of one another. Our focus is not on the value alignment algorithm itself, which is fixed and only plays the role of the receiver in BIG. Instead, we focus on how these people may become strategic in providing their feedback. Specifically, **our analysis on BIG implies that rational people will exaggerate their own value stance in anticipation of being “canceled out” by one another.** This may help explain one source of misinformation in social discourse.

Concretely, we connect value alignment and BIG as follows. Let the n people be the n players in BIG. We simplify the feedback process by assuming only a single prompt or context which we denote c (to avoid notation conflict with actions \mathbf{x} in BIG). We adopt the standard ideal point model (Coombs, 1950; Jamieson & Nowak, 2011; Singla et al., 2016; Xu & Davenport, 2020): For this context c , an ideal point $\theta \in \mathbb{R}^d$, and a response $\mathbf{y} \in \mathbb{R}^d$ (e.g. a document embedding vector), we define the reward model

$$r_\theta(\mathbf{y}) := -\|\mathbf{y} - \theta\|^2 \quad (25)$$

to measure how good the response \mathbf{y} is for the prompt c according to the ideal point θ . The closer \mathbf{y} is to θ , the higher the reward. Given a pair of responses \mathbf{y}, \mathbf{y}' , a player with ideal point θ draws a stochastic binary pairwise judgment label $z \in \{-1, 1\}$ according to the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952):

$$P(z | \mathbf{y}, \mathbf{y}', \theta) = \frac{1}{1 + \exp(-z(r_\theta(\mathbf{y}) - r_\theta(\mathbf{y}'))))} \quad (26)$$

where $z = 1$ means $\mathbf{y} \succeq \mathbf{y}'$ and $z = -1$ means $\mathbf{y} \prec \mathbf{y}'$ to this player. Each player i will label N_i tuples of the form $(\mathbf{y}, \mathbf{y}', z)$ where we assume \mathbf{y}, \mathbf{y}' are i.i.d. from some response distribution P_Y . Finally, the union of tuples from all players are given to the value alignment algorithm as training data.

Crucially, in BIG the players can misreport their values. Each player $i \in [n]$ has a *true ideal point* $\mathbf{t}_i \in \mathbb{R}^d$ which represents their true value. These are their targets in BIG, because they hope the value alignment algorithm ultimately arrives at \mathbf{t}_i as well. The players have heterogeneous values if the $\mathbf{t}_1 \dots \mathbf{t}_n$ are distinct. If the players were truthful, they would each use $\theta = \mathbf{t}_i$ in (26) when labeling their tuples. However, BIG allows each player i to choose a *fake ideal point* $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$. Player i instead uses $\theta = \mathbf{x}_i$ in (26) to label its N_i tuples. Note the action space \mathcal{X} is the space of ideal points. An action, namely player i choosing ideal point \mathbf{x}_i , is eventually reflected in the N_i tuples (in particular the z 's) from that player for value alignment.

The value alignment algorithm plays the role of the receiver in BIG. But unlike the affine function (1), we adopt the

standard maximum likelihood estimate for a global ideal point model parameter $\hat{\mathbf{x}}$, trained from the union of tuples annotated by all players $(\mathbf{y}, \mathbf{y}', z)_{1:N}$ where $N = \sum_{i=1}^n N_i$:

$$\hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x} \in \mathbb{R}^d} \sum_{j=1}^N \log P(z_j | \mathbf{y}_j, \mathbf{y}'_j, \theta = \mathbf{x}). \quad (27)$$

We note that the training data, being a mixture of BTL, is outside the model family in (27); (27) itself seems highly nonlinear and depends implicitly on the response distribution P_Y . Nonetheless, our empirical results indicate that the MLE $\hat{\mathbf{x}}$ is approximately a linear combination of individual player fake ideal points $\mathbf{x}_1 \dots \mathbf{x}_n$. This allows us to make predictions on player strategic behaviors based on earlier analysis on BIG. In particular, we predict that given the opportunity the players will evolve their fake ideal points $\{\mathbf{x}_i\}$ similar to best-response dynamics; that they do this to make the value alignment algorithm’s global ideal point (27) closer to their true ideal points $\{\mathbf{t}_i\}$; and that they will end up in a Nash equilibrium where their fake ideal points $\{\mathbf{x}_i\}$ are exaggerations of their true values $\{\mathbf{t}_i\}$. We next present an experiment to support these predictions.

6.2. A Value Alignment Experiment

Let there be $n = 2$ players with ideal point action space $\mathcal{X} = [-1, 1]$. Their true ideal points are $t_1 = -0.1, t_2 = 0.3$ respectively. We draw $N_1 = 10000$ i.i.d. pairs from the response distribution for player 1 to label: $(\mathbf{y}, \mathbf{y}') \sim P_Y = \text{uniform}[-10, 10]^2$, and another $N_2 = 10000$ i.i.d. pairs for player 2.

In iteration 0, both players start truthfully. Player 1 starts at its true ideal point $x_1 = t_1$ to annotate its preferences on the N_1 pairs. That is, for each $(\mathbf{y}, \mathbf{y}')$ pair player 1 draws a Bernoulli ± 1 -valued label z according to (26) with $\theta = x_1 = t_1$. We show player 1’s annotated dataset $(\mathbf{y}, \mathbf{y}', z)_{N_1}$ in Figure 5(left). For visualization purpose, we zoom in to the center region $[-3, 3]^2$ and also randomly thinned out the data so that the noisy nature of z is easier to see. Similarly, player 2 annotates its N_2 pairs according to (26) with $\theta = x_2 = t_2$ (Figure 5 right).

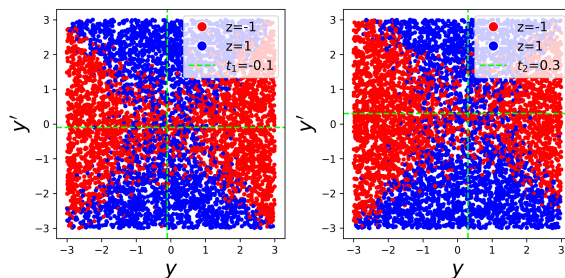


Figure 5. Pairwise preference labels z when both players are truthful. Left: player 1 with $x_1 = t_1$, right: player 2 with $x_2 = t_2$.

These $N_1 + N_2$ preference tuples are given to the value alignment algorithm (the receiver). The receiver numerically solves for the MLE by (27). We show the log likelihood surface in Figure 6(right). The MLE is at $\hat{x} = 0.103$, shown in the same figure(left). We observe that the receiver, despite maximizing the log likelihood, can be well-approximated by an affine receiver

$$\hat{x} = \frac{1}{2}x_1 + \frac{1}{2}x_2. \quad (28)$$

This concludes iteration 0.

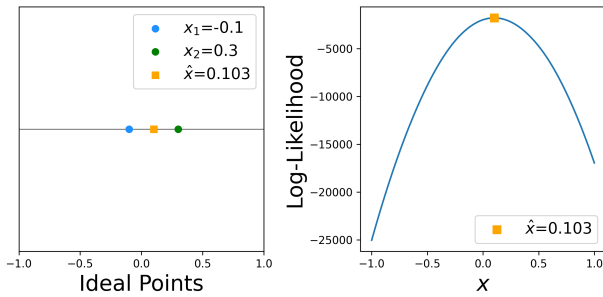


Figure 6. Receiver's MLE \hat{x} under truthful players.

In iteration 1, imagine player 1 observes all training data and the value alignment algorithm output from iteration 0 (i.e. a global ideal point at $\hat{x} = 0.103$). It realizes that the output is far from its true ideal point $t_1 = -0.1$. For the sake of exposition, we now allow player 1 to re-label its N_1 tuples using a different (fake) ideal point x_1 . Player 2's data remain fixed. Then we will run value alignment algorithm again. If player 1 can simulate the value alignment algorithm, it can perform a binary search in x_1 to best-respond to player 2, with the goal to move value alignment algorithm output to t_1 . We show this binary search in Figure 7. After 6 binary search steps player 1 finds that $x_1 = -0.536$ is good: together with player 2's data this indeed moves value alignment algorithm output to $\hat{x} = -0.101$, very close to player 1 target $t_1 = -0.1$. This is one iteration of empirical best-response by player 1. Again, the empirical based-response $x_1 = -0.536$ is close to the theoretical best-response under the affine receiver (28), which is $x_1 = 2t_1 - x_2 = -0.5$.

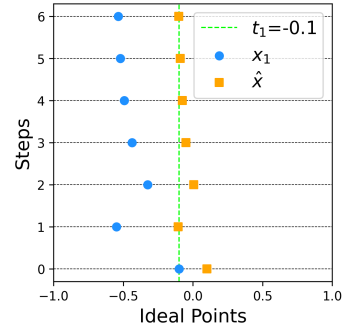


Figure 7. Player 1 conducts a binary search to find the empirical best response x_1 .

In subsequent iterations, we allow alternating players to perform such empirical best-response. Figure 8 shows the dynamics. In iteration 2, player 2 is able to somewhat drag value alignment output \hat{x} back towards its target. However, player 2 is limited by the action space: even with the maximum fake ideal point $x_2 = 1$ the output only moves back to $\hat{x} = 0.209$, not enough to reach its target $t_2 = 0.3$. In iteration 3, player 1 fights back with the minimum fake ideal point $x_1 = -1$, dragging value alignment output to $\hat{x} = 0.024$, closer but not reaching its target $t_1 = -0.1$. This is when the dynamics converges to a Nash equilibrium, where neither player can make further improvements. Observe at the NE ($x_1 = -1, x_2 = 1$) the player's fake ideal points are much exaggerated versions of their true ideal points $t_1 = -0.1, t_2 = 0.3$, respectively.

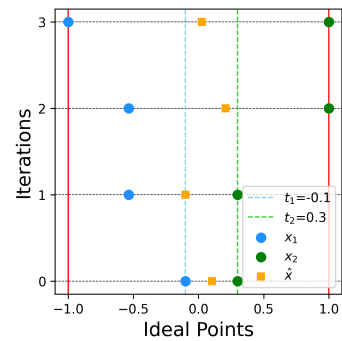


Figure 8. Empirical best-response dynamics converges to an exaggerating Nash equilibrium in iteration 3.

In Figure 9 we provide a visualization of the final preference labels z generated by the players after reaching the Nash equilibrium ($x_1 = -1, x_2 = 1$). This figure is to be contrasted with Figure 5. Now both players are untruthful and produce preference labels according to their fake, exaggerated ideal points. This shift illustrates the effect of BIG.

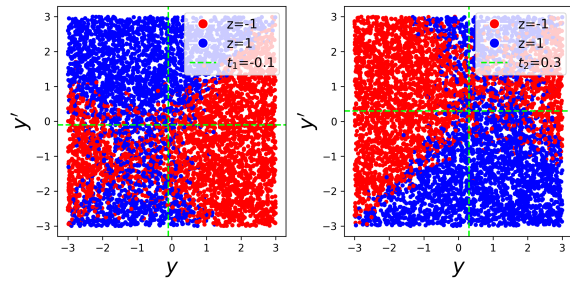


Figure 9. Preference labels z at the Nash equilibrium showing untruthfulness. Left: player 1 exaggerates with $x_1 = -1$; Right: player 2 exaggerates with $x_2 = 1$.

7. Conclusion and Future Work

We proved that a version of the battling influencers game is a potential game, and characterized its pure Nash equilibrium structures. As a use case, our game applies to standard value alignment via learning from preference feedback. Consequently, we rationalized a strategic behavior (exaggeration) in alignment data providers. Future work will focus on mechanism design to remove incentives for such strategic behaviors.

Acknowledgment This project was supported in part by NSF grants 1836978, 2023239, 2202457, 2331669, ARO MURI W911NF2110317, and AF CoE FA9550-18-1-0166.

References

- Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.
- Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.
- Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Chen, D., Chen, Y., Rege, A., and Vinayak, R. V. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*, 2024.
- Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.
- Coombs, C. H. Psychological scaling without a unit of measurement. *Psychological review*, 57(3), 1950.
- Hao, S. and Duan, L. Online learning from strategic human feedback in llm fine-tuning. *arXiv preprint arXiv:2412.16834*, 2024.
- Jamieson, K. G. and Nowak, R. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011.
- Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Michi, A., et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.
- Neyman, A. Correlated equilibrium and potential games. *International Journal of Game Theory*, 26(2):223–227, 1997.
- Osborne, M. J. *A course in game theory*. MIT Press, 1994.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Park, C., Liu, M., Kong, D., Zhang, K., and Ozdaglar, A. E. Rlhf from heterogeneous feedback via personalization and preference aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024.
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- Rosset, C., Cheng, C.-A., Mitra, A., Santacrose, M., Awadallah, A., and Xie, T. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.
- Roughgarden, T. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.
- Roughgarden, T. and Schrijvers, O. Online prediction with selfish experts. *Advances in Neural Information Processing Systems*, 30, 2017.
- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

-
- Singla, A., Tschitschek, S., and Krause, A. Actively learning hemimetrics with applications to eliciting user preferences. In *International Conference on Machine Learning*, pp. 412–420. PMLR, 2016.
- Soumalias, E., Curry, M. J., and Seuken, S. Truthful aggregation of llms with an application to online advertising. *arXiv preprint arXiv:2405.05905*, 2024.
- Sun, H., Chen, Y., Wang, S., Chen, W., and Deng, X. Mechanism design for llm fine-tuning with multiple reward models. *arXiv preprint arXiv:2405.16276*, 2024.
- Swamy, G., Dann, C., Kidambi, R., Wu, Z. S., and Agarwal, A. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.
- Tadelis, S. *Game Theory: An Introduction*. Princeton University Press, 2013.
- Xu, A. and Davenport, M. Simultaneous preference and metric learning from paired comparisons. *Advances in Neural Information Processing Systems*, 33:454–465, 2020.