Conditions for Altruistic Perversity in Two-Strategy Population Games

Colton Hill, Philip N. Brown, and Keith Paarporn

***Abstract*— Self-interested behavior from individuals can collectively lead to poor societal outcomes. These outcomes can seemingly be improved through the actions of altruistic agents, which benefit other agents in the system. However, it is known in specific contexts that altruistic agents can actually induce worse outcomes compared to a fully selfish population — a phenomenon we term altruistic perversity. This paper provides a holistic investigation into the necessary conditions that give rise to altruistic perversity. In particular, we study the class of two-strategy population games where one sub-population is altruistic and the other is selfish. We find that a population game can admit altruistic perversity only if the associated social welfare function is convex and the altruistic population is sufficiently large. Our results are a first step in establishing a connection between properties of nominal agent interactions and the potential impacts from altruistic behaviors.**

I. INTRODUCTION

In systems with a large number of interacting individuals, such as infrastructure and transportation networks, the payoff experienced by agents depends on the actions of other agents in the system. When all agents select strategies to maximize their own payoff (commonly referred to as *selfish*), it is well-known that the resulting system welfare can be suboptimal [1], [2]. Whether by nature or by design, agents may also exhibit behaviors that benefit other agents in the system. These *altruistic* agents are present in several domains of study, ranging from evolutionary biology [3]–[5] (e.g. the social amoeba D. discoideum [6]) and pandemic mitigation [7], [8] to the design of socio-technical systems [9]–[13] (e.g. the use of autonomous vehicles). Experimental research in economics has observed altruistic behavior [14], and the effects of fully-adopted altruism has been studied in a wide variety of games [15].

Game theory offers principled approaches that have been extensively utilized to study the inefficiencies that arise from the actions of selfish agents relative to a system's optimal operation [16]. From this perspective, a pertinent question to investigate is: in general, how does the presence of altruistic agents impact the social welfare of the system? Indeed, social welfare is guaranteed to improve from altruistic behaviors in certain contexts – for example, in non-atomic congestion games where all agents (at least partially) consider their impact on overall welfare [15].

However, the benefits from altruism do not generally apply in other settings. Counter-intuitively, it has been shown that significant negative effects can arise in systems with mixed altruistic and selfish populations. That is, the effect of altruistic behavior can be *perverse* in games with heterogeneous populations [17], [18]. The potential harm caused by altruism can be quantified with the *perversity index*, which measures the ratio between the social welfare in the presence of altruistic agents, with that of the social welfare that would arise if all agents behaved selfishly [19].

Much of the work done regarding perversity in games focuses specifically on the class of congestion games, where subsidies and tolls [20], choosing routes in consideration of the impact on aggregate road congestion [10], and uncertainty [18] effectively measure how altruism impacts the quality of social welfare. In series-parallel networks with arbitrary cost functions, it is known that the worst-case perversity arises when exactly half of the population is altruistic, and that the perversity increases as a function of the steepness of the cost functions [12]. However, altruism (even in only a fraction of the population) is guaranteed to improve social welfare in congestion games with serially-linearly-independent networks and affine cost functions, provided all agents have access to all roads [18]. Significant contributions have been made towards characterizing altruism and conditions for perversity; however, the results often come with assumptions that restrict generality.

The primary motivation of this paper is to study the emergence of altruistic perversity in general contexts that go beyond the well-studied congestion games literature. Specifically, we use a more general context (population games) as we seek to identify conditions on the type of agent interactions that admit welfare degradation (or improvement) in the presence of altruistic agents. This paper represents a first step in this direction, as we consider the impact of altruistic behavior for the entire class of 2×2 population games. This class of games encompasses a wide variety of nominal agent interactions, from Prisoner's Dilemma, Coordination, to Anti-Coordination games.

Our main result (Theorem 3.1) asserts that altruistic perversity can only occur if the function expressing social welfare is convex with respect to the population state. Interestingly, perversity can occur only for a sufficiently large altruistic population. Consequently, even all-altruistic populations have the potential to exhibit perversity. Conversely, games with a concave welfare function cannot exhibit altruistic perversity – the behavior of altruists can only improve societal outcomes in these cases. We provide a detailed illustration of these phenomena in a case study of population games based on the *Prisoner's Dilemma*.

*This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-23-1-0171, the National Science Foundation under award number ECCS-2013779, and the Committee for Research and Creative Works at UCCS.

The authors are with the University of Colorado at Colorado Springs, CO 80918, USA. {chill13, pbrown2, kpaarpor}@uccs.edu

II. Model

A. Symmetric Two-Strategy Population Game with Heterogeneous Types Presented in Normal Form

We consider a heterogeneous population consisting of a unit mass of agents, where each agent is either *altruistic* or *selfish*. Altruistic agents make up mass p_{a}, and selfish agents agents comprise mass p_{s}, so that $p_{\mathrm{a}} + p_{\mathrm{s}} = 1$. In symmetric two-strategy games, a 2×2 matrix can be used to represent the payoff of any outcome from the perspective of a row player. Agents can either *cooperate* by choosing the first row strategy, or *defect* by choosing the second row strategy, and the resulting payoff depends on whether other agents cooperate or defect (the first and second column, respectively). Thus, we write $\mathcal{S} \coloneqq \{C, D\}$ to denote the cooperate and defect strategies available to all agents, where the payoffs are denoted by the matrix:

$$A = \begin{matrix} \\ C \\ D \end{matrix} \begin{matrix} \begin{matrix} C & D \end{matrix} \\ \begin{bmatrix} R & S \\ T & P \end{bmatrix} \end{matrix}, \tag{1}$$

and we may assume without loss of generality that $R, S, T, P \in \mathbb{R}_{\geq 0}$. For $\tau \in \{\mathrm{a}, \mathrm{s}\}$, we write $\mathcal{X}_\tau \coloneqq \{\boldsymbol{x}_\tau \in \mathbb{R}^2_{\geq 0} : \sum_{i \in \mathcal{S}} x_{i,\tau} = p_\tau\}$ to denote the set of *population states* for altruistic and selfish agents. Thus $\mathcal{X} \coloneqq \mathcal{X}_{\mathrm{a}} \times \mathcal{X}_{\mathrm{s}}$ is the set of all population states, and the tuple $\boldsymbol{x} = (\boldsymbol{x}_{\mathrm{a}}, \boldsymbol{x}_{\mathrm{s}}) \in \mathcal{X}$ is a *population state* for altruistic and selfish agents.

All agents can either cooperate or defect, so the payoff for selecting a strategy depends on how many agents of both types choose the same strategy. Given population state $\boldsymbol{x}$, the *utilization level* is a column vector where each entry is the sum of altruistic and selfish agents selecting the corresponding strategy in $\boldsymbol{x}$. We denote the utilization level by $\boldsymbol{u}(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^2$, where $u_i(x_i) = x_{i,\mathrm{a}} + x_{i,\mathrm{s}}$ for each $i \in \mathcal{S}$. When the context is clear, we write $\boldsymbol{u}$ to denote $\boldsymbol{u}(\boldsymbol{x})$. Since there are two strategies, we may represent $\boldsymbol{u}$ by $u \in \mathbb{R}$, where $\boldsymbol{u} = \begin{bmatrix} u & 1-u \end{bmatrix}^\top$. Here, u is the fraction of agents cooperating, and $1 - u$ is the fraction of agents who defect.

We consider the set of altruistic and selfish populations and their strategies as established and identify a game with the payoffs experienced by agents for their decisions. The *payoff function* is a continuous mapping that associates the utilization level for a population state with a payoff vector:

$$f(\boldsymbol{u}(\boldsymbol{x})) : \mathcal{X} \to \mathbb{R}^2. \tag{2}$$

Since the payoffs agents receive is based on the matrix defined by (1), we write $f(\boldsymbol{u}) \coloneqq A\boldsymbol{u}$. We then measure the *total social welfare*, given a population state $\boldsymbol{x}$, by

$$\begin{aligned} W(\boldsymbol{u}) &\coloneqq \boldsymbol{u}^\top A \boldsymbol{u} \\ &= (R+P-(S+T))u^2 + (S+T-2P)u + P, \end{aligned} \tag{3}$$

where $\boldsymbol{u}^\top$ is the transpose of the utilization level for $\boldsymbol{x}$. The payoffs experienced by agents is determined by their type. Selfish agents are concerned only with maximizing their own payoff, so they aim to select the strategy that maximizes the actual payoff for a given utilization level $\boldsymbol{u}$:

$$\begin{aligned} f_{\mathrm{s}}(\boldsymbol{u}) &\coloneqq A\boldsymbol{u} \\ &= \begin{bmatrix} Ru + S(1-u) \\ Tu + P(1-u) \end{bmatrix} \\ &= \begin{bmatrix} f_{C,\mathrm{s}}(u) \\ f_{D,\mathrm{s}}(u) \end{bmatrix}. \end{aligned} \tag{4}$$

In contrast, altruistic agents are concerned with increasing social welfare. Since each agent is infinitesimal, and there are only two strategies to choose from, they select the strategy that is in the direction of increased social welfare. The gradient of the social welfare function (3), projected onto the unit simplex, represents the desired payoff for altruists:

$$\begin{aligned} f_{\mathrm{a}}(\boldsymbol{u}) &\coloneqq \nabla_{\boldsymbol{u}} W(\boldsymbol{u}) \\ &= \begin{bmatrix} (2R-(S+T))u + (S+T-2P)(1-u) \\ (S+T-2R)u + (2P-(S+T))(1-u) \end{bmatrix} \\ &= \begin{bmatrix} f_{C,\mathrm{a}}(u) \\ f_{D,\mathrm{a}}(u) \end{bmatrix}, \end{aligned} \tag{5}$$

where $f_{C,\mathrm{a}}(u) = -f_{D,\mathrm{a}}(u)$. An instance of a population game with selfish and altruistic types is fully specified by the tuple $G = \big(\mathcal{S}, f_{\tau \in \{\mathrm{a},\mathrm{s}\}}, p_{\mathrm{a}}\big)$.

A standard solution concept for population games is the Nash equilibrium, which describes a state in which no agent can benefit from unilaterally changing their strategy.

Definition 1: A *Nash equilibrium* is a population state $\boldsymbol{x} \in \mathcal{X}$ such that for each type $\tau \in \{\mathrm{a}, \mathrm{s}\}$:

$$x_{i,\tau} > 0 \implies f_{i,\tau}(\boldsymbol{u}(\boldsymbol{x})) \geq f_{i',\tau}(\boldsymbol{u}(\boldsymbol{x})) \;\; \forall i, i' \in \mathcal{S}, \tag{6}$$

where a population state corresponding to a Nash equilibrium is denoted $\boldsymbol{x}^{\mathrm{ne}} = (\boldsymbol{x}^{\mathrm{ne}}_{\mathrm{a}}, \boldsymbol{x}^{\mathrm{ne}}_{\mathrm{s}})$. For each $\tau \in \{\mathrm{a}, \mathrm{s}\}$, we may represent $\boldsymbol{x}_\tau$ by $x_\tau \in \mathbb{R}$, since $\boldsymbol{x}_\tau = \begin{bmatrix} x_\tau & p_\tau - x_\tau \end{bmatrix}^\top$, where x_τ is the fraction of agents cooperating, and $p_\tau - x_\tau$ is the fraction of agents defecting. The utilization level that corresponds to a Nash equilibrium, $\boldsymbol{u}(\boldsymbol{x}^{\mathrm{ne}})$, is often denoted $\boldsymbol{u}^{\mathrm{ne}}$ (or simply u^{ne}). Since only two strategies are available, all Nash equilibria must satisfy one of the following conditions:

$$\begin{aligned} x^{\mathrm{ne}}_\tau = 0 &\iff f_{C,\tau}(u^{\mathrm{ne}}) < f_{D,\tau}(u^{\mathrm{ne}}), \\ x^{\mathrm{ne}}_\tau \in (0, p_\tau) &\iff f_{C,\tau}(u^{\mathrm{ne}}) = f_{D,\tau}(u^{\mathrm{ne}}), \\ x^{\mathrm{ne}}_\tau = p_\tau &\iff f_{C,\tau}(u^{\mathrm{ne}}) > f_{D,\tau}(u^{\mathrm{ne}}). \end{aligned} \tag{7}$$

The linearity of the payoff functions implies that there is only one Nash equilibrium $u^*_\tau \in [0, 1]$ for each $\tau \in \{\mathrm{a}, \mathrm{s}\}$ such that $f_{C,\tau}(u^*_\tau) = f_{D,\tau}(u^*_\tau)$:

$$u^*_{\mathrm{s}} \coloneqq \frac{P - S}{R + P - (S + T)}, \tag{8}$$

and

$$u^*_{\mathrm{a}} \coloneqq \frac{2P - (S + T)}{2(R + P - (S + T))}. \tag{9}$$

The case that $f_{C,\tau}(u) = f_{D,\tau}(u)$ for all $u \in [0, 1]$ is trivial, since it implies $W(\boldsymbol{u})$ is constant. For a game $G = \big(\mathcal{S}, f_{\tau \in \{\mathrm{a},\mathrm{s}\}}, p_{\mathrm{a}}\big)$, we write the set of population states that result in a Nash equilibrium for all agents as

$\mathcal{X}^{\mathrm{ne}}(G) \subseteq \mathcal{X}$. The set of Nash equilibria for an all-selfish version of G is denoted $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}}(G) \coloneqq \mathcal{X}^{\mathrm{ne}}\big(\mathcal{S}, f_{\tau\in\{\mathrm{a},\mathrm{s}\}}, 0\big)$, and the corresponding set for an all-altruistic version of G is denoted $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}}(G) \coloneqq \mathcal{X}^{\mathrm{ne}}\big(\mathcal{S}, f_{\tau\in\{\mathrm{a},\mathrm{s}\}}, 1\big)$. We often write $\mathcal{X}^{\mathrm{ne}}(G)$, $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}}(G)$, and $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}}(G)$ as $\mathcal{X}^{\mathrm{ne}}$, $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}}$, and $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}}$ (respectively) when the dependence on G is clear.

B. Performance Metric: Perversity Index

In this paper, we study the *perversity index* [17] to understand the effects of heterogeneous altruism. The perversity index captures the potential negative impact the presence of altruism has in a population game, relative to its all-selfish counterpart (i.e. $p_{\mathrm{a}} = 0$). The perversity index is defined as the worst-case ratio of the social welfare of a heterogeneous Nash equilibrium with that of the social welfare that arises from an all-selfish Nash equilibrium:

$$\mathrm{PI}(G) \coloneqq \frac{\min_{\boldsymbol{x}\in\mathcal{X}^{\mathrm{ne}}(G)} W(\boldsymbol{u}(\boldsymbol{x}))}{\max_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}}(G)} W(\boldsymbol{u}(\boldsymbol{x}))}. \tag{10}$$

$\mathrm{PI}(G) < 1$ indicates the presence of altruists can hurt social welfare at equilibrium – here, we say that the game exhibits altruistic perversity. Likewise, $\mathrm{PI}(G) > 1$ indicates the presence of altruists improves social welfare at equilibrium.

III. Classifying Games with Perversity in Symmetric Two-Strategy Population Games

One might expect that introducing altruistic agents in a population game would lead to Nash equilibria with improved social welfare. We show that this need not be the case. Indeed, we seek to identify necessary conditions on the underlying population game, specifically the payoff functions and welfare, that admit worsened social welfare in the presence of altruists compared to an all-selfish population. That is, we seek to classify games G that admit altruistic perversity, $\mathrm{PI}(G) < 1$. Our main result is given below.

Theorem 3.1: Let G be a heterogeneous symmetric two-strategy population game. If the presence of altruistic agents in G admits altruistic perversity, i.e. $\mathrm{PI}(G) < 1$, then the welfare function defined by (3) is convex.

An example of the altruistic perversity characterized by this result is presented in Section IV. The proof (completed via the contrapositive result) of Theorem 3.1 is presented in section V. The contrapositive states that if the welfare function is concave, then the perversity index is greater than 1. This implies the presence of altruists cannot degrade equilibrium welfare in games with concave welfare functions. Since altruists choose actions in the direction of the welfare gradient, they act as a local gradient ascent on W, so a sufficient amount of altruists will lead to welfare maximization.

Conversely, when W is convex, social welfare is maximized at an extreme point where all agents play the same action. In this case, altruists can still increase welfare, and are at equilibrium when W is maximized. However, altruists now have the potential to induce deteriorated welfare because the local minimizer of W coincides with a Nash equilibrium for altruists. In particular, if u_{a}^* is feasible and exceeds how many selfish agents cooperate, and the altruistic population exceeds u_{a}^*, the worst-case welfare is achieved since u_{a}^* coincides with the global minimum of W. Counter-intuitively, this means that perverse outcomes do not emerge unless there is a sufficiently large population of altruists.

In the next section, we concretely illustrate the altruistic perversity that emerges when the underlying population game is a Prisoner's Dilemma.

IV. Case Study: Prisoner's Dilemma

Here, we present the *Prisoner's Dilemma* population game as an example of the perversity that can arise as described in Theorem 3.1. Suppose the entries in the payoff matrix defined by (1) satisfy $S < P < R < T$, then the symmetric two-strategy population game becomes a Prisoner's Dilemma, which we denote $\mathrm{PD}(p_{\mathrm{a}})$. In the all-selfish version of $\mathrm{PD}(p_{\mathrm{a}})$ (i.e. $p_{\mathrm{a}} = 0$), the only Nash equilibrium is when all agents defect, i.e. $u_{\mathrm{s}}^{\mathrm{ne}} = 0$. Thus, all agents get the punishment payoff, and the social welfare is $W(0) = P$. So, the perversity index defined by (10) becomes

$$\mathrm{PI}(\mathrm{PD}(p_{\mathrm{a}})) = \min_{\boldsymbol{x}\in\mathcal{X}^{\mathrm{ne}}(G)} \frac{W(\boldsymbol{u}^{\mathrm{ne}})}{P}. \tag{11}$$

The result below fully characterizes (11) as a function of the altruistic mass, p_{a}.

Proposition 4.1: Let $\mathrm{PD}(p_{\mathrm{a}})$ be a heterogeneous symmetric two-strategy Prisoner's Dilemma population game, where p_{a} is the mass of altruistic agents. Define $\delta \coloneqq R + P - (S+T)$ and $\beta \coloneqq S + T - 2P$. If $W(u)$ is convex, then the perversity index is given by:

$$\mathrm{PI}(\mathrm{PD}(p_{\mathrm{a}})) = \begin{cases} 1, & \text{if } p_{\mathrm{a}} < u_{\mathrm{a}}^* \\ 1 - \frac{\beta^2}{4P\delta}, & \text{if } p_{\mathrm{a}} \geq u_{\mathrm{a}}^* \end{cases}, \tag{12}$$

and if $W(u)$ is concave, then the perversity index is given by:

$$\mathrm{PI}(\mathrm{PD}(p_{\mathrm{a}})) = \begin{cases} \frac{\delta p_{\mathrm{a}}^2 + \beta p_{\mathrm{a}}}{P} + 1, & \text{if } p_{\mathrm{a}} < u_{\mathrm{a}}^* \\ 1 - \frac{\beta^2}{4P\delta}, & \text{if } p_{\mathrm{a}} \geq u_{\mathrm{a}}^* \end{cases}. \tag{13}$$

The proof is presented in the Appendix, but a short discussion is presented here to describe equilibria and payoffs to agents in this type of game. Here, selfish agents defect regardless of whether altruists cooperate or defect, emphasizing the intuition behind Theorem 3.1 that the improvement (or degradation) of social welfare is dependent on the choices altruists make. Since selfish agents always defect, and since $P < R$, the payoff function defined by (5) that altruists use is accurately informing them of where the locally maximum welfare is as expected. However, if welfare is convex, a flaw arises because altruists are indifferent about cooperating or defecting at the global minimum for welfare, because the payoff at this point is 0 regardless of the strategy altruists select. Fig. 1a depicts the payoffs to altruists and Fig. 1b depicts the altruistic perversity that arises. The case that welfare is concave does not suffer from this issue, as the point at which altruists are indifferent is actually the global maximum for welfare, and it is the only point at which each altruist is content with their decision. The perversity index in this case is depicted in Fig. 1c.

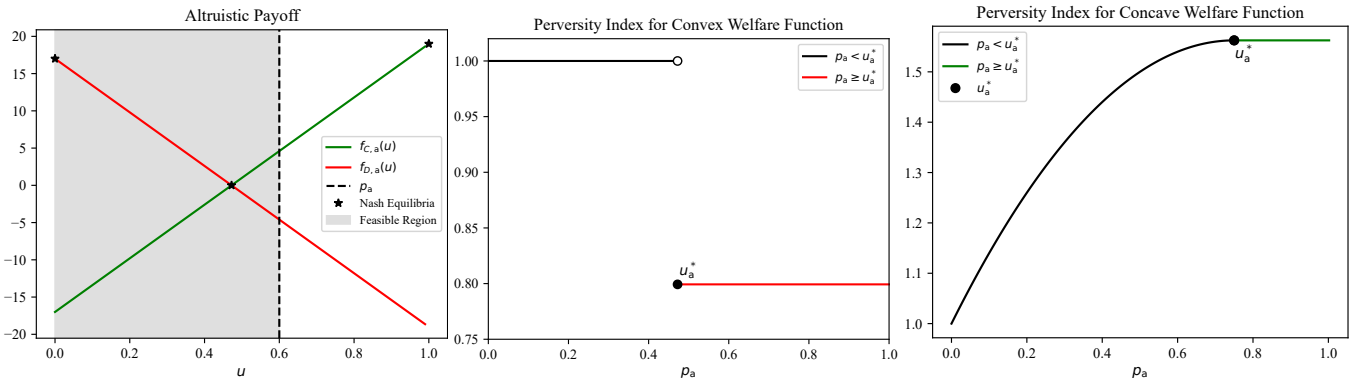

Fig. 1: Fig. 1a characterizes the payoff functions and possible Nash equilibria for altruists in an example game where the welfare function is convex: $R = 21$, $S = 1$, $T = 22$, $P = 20$. The stars represent the Nash equilibria available to altruists when $p_a = 1$, and the shaded area contains feasible sub-population states for a given altruistic population. Fig. 1b represents the perversity index as a function of the altruistic population, $\mathrm{PI}(p_a)$, for the same example game. Here, the perversity index is a piecewise constant function since, if their population is too small, altruists choose to defect just like selfish agents. If their population exceeds u_a^*, altruists may choose the mixed Nash equilibrium, which results in the worst-case welfare. In this example, altruistic perversity can significantly degrade welfare, resulting in a 20% drop in performance. Fig. 1c represents $\mathrm{PI}(p_a)$ for an example game where the welfare function is concave: $R = 3$, $S = 1$, $T = 6$, $P = 2$. Here, the perversity index is continuous because the behavior of the altruistic payoffs is unlike that of Fig. 1a; altruists cooperate until the population is large enough to choose the mixed Nash equilibrium, resulting in the best-case welfare.

V. Proof of Theorem 3.1

We first provide a brief outline of the proof, then present a lemma and discuss its importance. The proof is accomplished by showing the contrapositive: if the welfare function defined by (3) is strictly concave, then $\mathrm{PI}(G) \geq 1$. The contrapositive is proved with the following cases where u_a^* is defined by (9):

- Case 1: if $u_a^* \leq 0$, then $\mathrm{PI}(G) \geq 1$.
- Case 2: if $u_a^* \geq 1$, then $\mathrm{PI}(G) \geq 1$.
- Case 3: if $u_a^* \in (0,1)$ and $u_a^* \leq p_a$, then $\mathrm{PI}(G) \geq 1$.
- Case 4: if $u_a^* \in (0,1)$ such that $u_a^* > p_a$, and $\mathcal{X}_s^{ne} \subseteq \{0,1\}$, then $\mathrm{PI}(G) \geq 1$.
- Case 5: if $u_a^* \in (0,1)$ such that $u_a^* > p_a$, and $\mathcal{X}_s^{ne} = \{u_s^*\}$, then $\mathrm{PI}(G) \geq 1$.

Our sole lemma characterizes potential selfish and altruistic Nash equilibria under concave social welfare functions.

Lemma 5.1: Let G be a symmetric two-strategy population game. If $W(u)$ is strictly concave, then $|\mathcal{X}_s^{ne}| = 1$, and $|\mathcal{X}_a^{ne}| = 1$.

The proof of Lemma 5.1 appears in the Appendix. The implication is that each population of agents has only one Nash equilibrium that they are trying to reach. Thus, the lemma is useful in showing that, in heterogeneous games, agents of each type still only have one Nash equilibrium. Intuitively, this means that each altruistic and selfish agent is making decisions with limited regard to what others are doing. We proceed with the proof of the main result:

Proof of Theorem 3.1: Let $\boldsymbol{u}_s^{ne}$ and $\boldsymbol{u}^{ne}$ be the utilization level for an all-selfish Nash equilibrium and a heterogeneous Nash equilibrium, respectively. We use $\boldsymbol{x}^{ne} = \left[(x_a^{ne}, x_s^{ne}) \quad (p_a - x_a^{ne}, p_s - x_s^{ne})\right]$, to denote a heterogeneous Nash equilibrium where $x_a^{ne}, x_s^{ne} \in [0,1]$. Hence, the heterogeneous utilization level is

$$\begin{aligned}\boldsymbol{u}^{ne} &= \begin{bmatrix} u^{ne} & 1-u^{ne}\end{bmatrix}^\top \\ &= \begin{bmatrix} x_a^{ne} + x_s^{ne} & 1-(x_a^{ne}+x_s^{ne})\end{bmatrix}^\top.\end{aligned}$$

Since $W(u)$ is strictly concave, it is known that u_a^* is the global maximum, so $W(u_a^*) \geq W(u)$ for all u. The following cases complete the proof; we provide intuition here, and proceed with the proof of each case in the appendix.

Case 1: If $u_a^* \leq 0$, then $\mathrm{PI}(G) \geq 1$.

Case 1 implies that altruists always choose to defect in a game where the social welfare function is decreasing from 0 to 1. Hence, the number of agents cooperating will always be less in the heterogeneous game than in the all-selfish game.

Case 2: If $u_a^* \geq 1$, then $\mathrm{PI}(G) \geq 1$.

Case 2 implies that altruists always cooperate in a game where the social welfare function is increasing from 0 to 1. Thus, the number of agents cooperating will always be higher in the heterogeneous game than in the all-selfish game.

Case 3: If $u_a^* \in (0,1)$ such that $u_a^* \leq p_a$, then $\mathrm{PI}(G) \geq 1$.

Case 3 shows that if the mass of altruistic agents is large enough, they are guaranteed to move social welfare closer to their preferred mixed Nash equilibrium, the maximum social welfare, than selfish agents would do on their own.

Case 4: If $u_a^* \in (0,1)$ such that $u_a^* > p_a$, and $\mathcal{X}_s^{ne} \subseteq \{1,0\}$, then $\mathrm{PI}(G) \geq 1$.

Case 4 implies that even a relatively small population of altruists is able to improve the overall social welfare, regardless of whether the selfish agents cooperate or defect.

Case 5: If $u_a^* \in (0,1)$ such that $u_a^* > p_a$, and $\mathcal{X}_s^{ne} = \{u_s^*\}$, then $\mathrm{PI}(G) \geq 1$.

Case 5 also implies that a relatively small population of altruists is able to improve overall social welfare, in the instance that selfish agents now prefer a mixed Nash equilibrium.

Cases 1-5 show that if W is strictly concave, then $\mathrm{PI}(G) \geq 1$ for any value obtained by $u_a^* \in \mathbb{R}$. Hence the contrapositive is shown: if $\mathrm{PI}(G) < 1$, then W is convex. ■

VI. Conclusions

We have provided general conditions for when the presence of altruistic agents can actually worsen social welfare

in the class of two-strategy population games. These results are an initial step to identifying how the structure of agent interactions in a population may dictate whether altruistic behavior improves or degrades social welfare. Future work warrants the investigation into even more generalized relationships between agents and social welfare. Arbitrary n-strategy population games with m-population types, where each population uniquely weighs how much it maximizes its own payoff versus social welfare, as well as stable outcomes associated with evolutionary dynamics will be studied.

APPENDIX

$W(u)$ being strictly concave has the following implication, which is stated here for convenience:

$$R + P - (S + T) < 0. \tag{14}$$

First, we include the proof of Lemma 5.1.

Proof of Lemma 5.1: Suppose to the contrary that the claim is false. Since the payoff functions for agents of both types is affine, the possible cardinality of $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}}$ and $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}}$ is 1, 3, or ∞. If the cardinality of $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}}$ or $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}}$ is ∞, the implication is that the welfare function is constant, so we may assume $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}} = \{1, u_{\mathrm{s}}^*, 0\}$, or $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}} = \{1, u_{\mathrm{a}}^*, 0\}$. Suppose first that $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}} = \{1, u_{\mathrm{s}}^*, 0\}$. Since 1 is a Nash equilibrium, $f_{D,\mathrm{s}}(1) \leq f_{C,\mathrm{s}}(1)$, i.e. $T \leq R$. Similarly, since 0 is a Nash equilibrium, $f_{C,\mathrm{s}}(0) \leq f_{D,\mathrm{s}}(0)$, i.e. $S \leq P$. Thus $S + T \leq P + T \leq P + R$, i.e. $R + P - (S + T) \geq 0$, contradicting (14). Now suppose $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}} = \{1, u_{\mathrm{a}}^*, 0\}$. Since 1 is a Nash equilibrium, $f_{D,\mathrm{a}}(1) \leq f_{C,\mathrm{a}}(1)$, i.e. $S+T-2R \leq 2R - (S + T)$. Similarly, since 0 is a Nash equilibrium, $f_{C,\mathrm{a}}(0) \leq f_{D,\mathrm{a}}(0)$, i.e. $S + T - 2P \leq 2P - (S + T)$. But then $0 \leq 2(R + P - (S + T))$, contradicting (14). ■

Proof of Cases 1-5 for Theorem 3.1

Proof of Case 1: Since $u_{\mathrm{a}}^* \leq 0$ is the global maximum, we have that $u_{\mathrm{a}}^* \leq u$, and thus $W(u) \leq W(u_{\mathrm{a}}^*)$ for all $u \in [0, 1]$. Then $2P-(S+T) \geq 0$ implies $f_{C,\mathrm{a}}(0) = S + T - 2P \leq 0$. Now, (14) implies $f_{C,\mathrm{a}}(u)$ is decreasing, so $f_{C,\mathrm{a}}(u) \leq 0$ for all u. Since $f_{D,\mathrm{a}}(u) = -f_{C,\mathrm{a}}(u)$, $x_{\mathrm{a}}^{\mathrm{ne}} = 0$ is the only Nash equilibrium for altruists. So, for any $p_{\mathrm{a}} \in [0, 1]$, we have that altruists always defect, so $u^{\mathrm{ne}} \leq u_{\mathrm{s}}^{\mathrm{ne}}$. Thus $W(u_{\mathrm{s}}^{\mathrm{ne}}) \leq W(u^{\mathrm{ne}})$, since $W(u)$ is decreasing for all $u \in [0, 1]$. ■

Proof of Case 2: Since $u_{\mathrm{a}}^* \geq 1$ is the global maximum, we have that $u \leq u_{\mathrm{a}}^*$, and thus $W(u) \leq W(u_{\mathrm{a}}^*)$ for all $u \in [0, 1]$. Also, $2R - (S + T) \geq 0$ implies $f_{C,\mathrm{a}}(1) = 2R - (S + T) \geq 0$. Since $u_{\mathrm{a}}^* > 0$, and by (14), it is the case that $f_{C,\mathrm{a}}(0) = S+T-2P \geq 0$. Hence $f_{C,\mathrm{a}}(u) \geq 0$ for all u. Since $f_{D,\mathrm{a}}(u) = -f_{C,\mathrm{a}}(u)$, $x_{\mathrm{a}}^{\mathrm{ne}} = p_{\mathrm{a}}$ is the only Nash equilibrium for altruists. So, for any $p_{\mathrm{a}} \in [0, 1]$, we have that altruists always cooperate, so that $u_{\mathrm{s}}^{\mathrm{ne}} \leq u^{\mathrm{ne}}$. Hence, $W(u_{\mathrm{s}}^{\mathrm{ne}}) \leq W(u^{\mathrm{ne}})$. ■

Proof of Case 3: Since $u_{\mathrm{a}}^* \in (0, 1)$, $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}} = \{u_{\mathrm{a}}^*\}$ by Lemma 5.1. If $x_{\mathrm{s}}^{\mathrm{ne}} \leq u_{\mathrm{a}}^*$, then we claim $x_{\mathrm{a}}^{\mathrm{ne}} = u_{\mathrm{a}}^* - x_{\mathrm{s}}^{\mathrm{ne}}$. We can see that $x_{\mathrm{a}}^{\mathrm{ne}}$ is feasible since $0 \leq x_{\mathrm{a}}^{\mathrm{ne}} \leq u_{\mathrm{a}}^* \leq p_{\mathrm{a}}$, and

$$\begin{aligned} f_{C,\mathrm{a}}(x_{\mathrm{a}}^{\mathrm{ne}} + x_{\mathrm{s}}^{\mathrm{ne}}) &= f_{C,\mathrm{a}}(u_{\mathrm{a}}^*) \\ &= f_{D,\mathrm{a}}(u_{\mathrm{a}}^*). \end{aligned}$$

Further, this is the only Nash equilibrium by Lemma 5.1, thus $W(u_{\mathrm{s}}^{\mathrm{ne}}) \leq W(u^{\mathrm{ne}}) = W(u_{\mathrm{a}}^*)$. If $x_{\mathrm{s}}^{\mathrm{ne}} > u_{\mathrm{a}}^*$, then we claim $x_{\mathrm{a}}^{\mathrm{ne}} = 0$; to be clear, this implies altruists defect (choose $f_{D,\mathrm{a}}(u)$). It is clear that $x_{\mathrm{a}}^{\mathrm{ne}}$ is feasible, and since $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}} = \{u_{\mathrm{a}}^*\}$ and $u_{\mathrm{a}}^* < x_{\mathrm{s}}^{\mathrm{ne}}$, we have that

$$\begin{aligned} f_{C,\mathrm{a}}(u^{\mathrm{ne}}) &\leq f_{C,\mathrm{a}}(u_{\mathrm{a}}^*) \\ &= f_{D,\mathrm{a}}(u_{\mathrm{a}}^*) \\ &\leq f_{D,\mathrm{a}}(u^{\mathrm{ne}}). \end{aligned}$$

Thus, $x_{\mathrm{a}}^{\mathrm{ne}} = 0$ is the only Nash equilibrium by Lemma 5.1, and it follows that $u^{\mathrm{ne}} \leq u_{\mathrm{s}}^{\mathrm{ne}}$. Hence, we have that $W(u_{\mathrm{s}}^{\mathrm{ne}}) \leq W(u^{\mathrm{ne}})$ since $W(u)$ is decreasing for $u_{\mathrm{a}}^* \leq u$. ■

Proof of Case 4: By Lemma 5.1, $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}} = \{u_{\mathrm{a}}^*\}$ since $u_{\mathrm{a}}^* \in (0, 1)$, and $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}}$ is equal to the set containing only one element of $\{1, 0\}$. If $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}} = \{1\}$, then $u_{\mathrm{s}}^{\mathrm{ne}} = 1$, and $f_{C,\mathrm{s}}(1) \geq f_{D,\mathrm{s}}(1)$, i.e. $R \geq T$. Thus, by (14), it must also be the case that $f_{C,\mathrm{s}}(0) = S \geq P = f_{D,\mathrm{s}}(0)$. Hence, $f_{C,\mathrm{s}}(u) \geq f_{D,\mathrm{s}}(u)$ for all u, and so $x_{\mathrm{s}}^{\mathrm{ne}} = p_{\mathrm{s}}$ is the only heterogeneous Nash equilibrium for selfish agents. Now, $u^{\mathrm{ne}} \leq u_{\mathrm{s}}^{\mathrm{ne}}$, and since $\mathcal{X}_{\mathrm{a}}^{\mathrm{ne}} = \{u_{\mathrm{a}}^*\}$, we have that $u_{\mathrm{a}}^* \leq u^{\mathrm{ne}} \leq u_{\mathrm{s}}^{\mathrm{ne}} = 1$, so that $W(u_{\mathrm{s}}^{\mathrm{ne}}) \leq W(u^{\mathrm{ne}})$. If $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}} = \{0\}$, then $u_{\mathrm{s}}^{\mathrm{ne}} = 0$, so $f_{D,\mathrm{s}}(0) \geq f_{C,\mathrm{s}}(0)$, i.e. $P \geq S$. Thus, by (14), it must also be the case that $f_{D,\mathrm{s}}(1) = T \geq R = f_{C,\mathrm{s}}(1)$. Thus $f_{D,\mathrm{s}}(u) \geq f_{C,\mathrm{s}}(u)$ for all u, and so $x_{\mathrm{s}}^{\mathrm{ne}} = 0$ is the only heterogeneous Nash equilibrium for selfish agents. It is clear that $u_{\mathrm{s}}^{\mathrm{ne}} \leq u^{\mathrm{ne}}$, and since $p_{\mathrm{a}} < u_{\mathrm{a}}^*$, it follows that $0 = u_{\mathrm{s}}^{\mathrm{ne}} \leq u^{\mathrm{ne}} \leq u_{\mathrm{a}}^*$. Thus $W(u_{\mathrm{s}}^{\mathrm{ne}}) \leq W(u^{\mathrm{ne}})$. ■

Proof of Case 5: If $x_{\mathrm{a}}^{\mathrm{ne}} \leq u_{\mathrm{s}}^*$ and $u_{\mathrm{s}}^* - x_{\mathrm{a}}^{\mathrm{ne}} \leq p_{\mathrm{s}}$, then we claim that $x_{\mathrm{s}}^{\mathrm{ne}} = u_{\mathrm{s}}^* - x_{\mathrm{a}}^{\mathrm{ne}}$. It is clear that $x_{\mathrm{s}}^{\mathrm{ne}}$ is feasible since $0 \leq x_{\mathrm{s}}^{\mathrm{ne}} \leq u_{\mathrm{s}}^* \leq p_{\mathrm{s}}$, and

$$\begin{aligned} f_{C,\mathrm{s}}(u^{\mathrm{ne}}) &= f_{C,\mathrm{s}}(u_{\mathrm{s}}^*) \\ &= f_{D,\mathrm{s}}(u_{\mathrm{s}}^*). \end{aligned}$$

Note that this is the only Nash equilibrium by Lemma 5.1. Thus, $W(u^{\mathrm{ne}}) = W(u_{\mathrm{s}}^*) = W(u_{\mathrm{s}}^{\mathrm{ne}})$, i.e. $W(u_{\mathrm{s}}^{\mathrm{ne}}) \leq W(u^{\mathrm{ne}})$ trivially. If $x_{\mathrm{a}}^{\mathrm{ne}} \leq u_{\mathrm{s}}^*$ and $u_{\mathrm{s}}^* - x_{\mathrm{a}}^{\mathrm{ne}} > p_{\mathrm{s}}$, then we claim that $x_{\mathrm{s}}^{\mathrm{ne}} = p_{\mathrm{s}}$. It is clear that $x_{\mathrm{s}}^{\mathrm{ne}}$ is feasible, and

$$\begin{aligned} f_{D,\mathrm{s}}(x_{\mathrm{s}}^{\mathrm{ne}} + x_{\mathrm{a}}^{\mathrm{ne}}) &\leq f_{D,\mathrm{s}}(u_{\mathrm{s}}^* - x_{\mathrm{a}}^{\mathrm{ne}} + x_{\mathrm{a}}^{\mathrm{ne}}) \\ &= f_{D,\mathrm{s}}(u_{\mathrm{s}}^*) \\ &= f_{C,\mathrm{s}}(u_{\mathrm{s}}^*) \\ &\leq f_{C,\mathrm{s}}(p_{\mathrm{s}} + x_{\mathrm{a}}^{\mathrm{ne}}) \\ &= f_{C,\mathrm{s}}(u^{\mathrm{ne}}). \end{aligned}$$

This is also the only Nash equilibrium by Lemma 5.1. Now, $u^{\mathrm{ne}} = x_{\mathrm{s}}^{\mathrm{ne}} + x_{\mathrm{a}}^{\mathrm{ne}} < u_{\mathrm{s}}^* - x_{\mathrm{a}}^{\mathrm{ne}} + x_{\mathrm{a}}^{\mathrm{ne}} = u_{\mathrm{s}}^* = u_{\mathrm{s}}^{\mathrm{ne}}$, i.e. $u^{\mathrm{ne}} < u_{\mathrm{s}}^{\mathrm{ne}}$. Also, $u^{\mathrm{ne}} = p_{\mathrm{s}} + x_{\mathrm{a}}^{\mathrm{ne}} \geq u_{\mathrm{a}}^*$ (otherwise, altruists are not at Nash equilibrium or $u_{\mathrm{a}}^* \geq 1$, both contradictions). Hence $u_{\mathrm{a}}^* \leq u^{\mathrm{ne}} \leq u_{\mathrm{s}}^{\mathrm{ne}}$, so that $W(u_{\mathrm{s}}^{\mathrm{ne}}) \leq W(u^{\mathrm{ne}})$. If $x_{\mathrm{a}}^{\mathrm{ne}} > u_{\mathrm{s}}^*$, then we claim $x_{\mathrm{s}}^{\mathrm{ne}} = 0$ (selfish agents defect and choose $f_{D,\mathrm{s}}(u)$). It is clear that $x_{\mathrm{s}}^{\mathrm{ne}}$ is feasible, and since $\mathcal{X}_{\mathrm{s}}^{\mathrm{ne}} = \{u_{\mathrm{s}}^*\}$ and $u_{\mathrm{s}}^* < x_{\mathrm{a}}^{\mathrm{ne}}$, we have that

$$\begin{aligned} f_{C,\mathrm{s}}(u^{\mathrm{ne}}) &< f_{C,\mathrm{s}}(u_{\mathrm{s}}^*) \\ &= f_{D,\mathrm{s}}(u_{\mathrm{s}}^*) \\ &< f_{D,\mathrm{s}}(u^{\mathrm{ne}}). \end{aligned}$$

Now, $u_{\mathrm{s}}^{\mathrm{ne}} = u_{\mathrm{s}}^* < x_{\mathrm{a}}^{\mathrm{ne}} = u^{\mathrm{ne}}$, and $x_{\mathrm{a}}^{\mathrm{ne}} \leq p_{\mathrm{a}} < u_{\mathrm{a}}^*$, so $u_{\mathrm{s}}^{\mathrm{ne}} < u^{\mathrm{ne}} < u_{\mathrm{a}}^*$. Hence $W(u_{\mathrm{s}}^{\mathrm{ne}}) \leq W(u^{\mathrm{ne}})$. ■

Finally, we include the proof of Proposition 4.1.

Proof of Proposition 4.1: Since $S < P$ and $R < T$, defecting is a dominant strategy for selfish agents, i.e. $f_{D,\mathrm{s}}(u) > f_{C,\mathrm{s}}(u)$ for all $u \in [0,1]$. Thus, in any Nash equilibrium, $x_{\mathrm{s}}^{\mathrm{ne}} = 0$. Next, we identify the values of $x_{\mathrm{a}}^{\mathrm{ne}} \in [0, p_{\mathrm{a}}]$ that result in a Nash equilibrium for altruists. In particular, $x_{\mathrm{a}}^{\mathrm{ne}}$ must satisfy one of the Nash equilibrium conditions in (7). Hence, the social welfare at equilibrium in a Prisoner's Dilemma is characterized by

$$W(u^{\mathrm{ne}}) = \begin{cases} P & \text{if } x_{\mathrm{a}}^{\mathrm{ne}} = 0 \\ P - \frac{\beta^2}{4\delta} & \text{if } x_{\mathrm{a}}^{\mathrm{ne}} = u_{\mathrm{a}}^* \\ \delta p_{\mathrm{a}}^2 + \beta p_{\mathrm{a}} + P & \text{if } x_{\mathrm{a}}^{\mathrm{ne}} = p_{\mathrm{a}} \end{cases}. \tag{15}$$

First, assume $\delta = 0$. Then $f_{C,\mathrm{a}}(u) = S + T - 2P > 0$ (since $R > P$), so $x_{\mathrm{a}}^{\mathrm{ne}} = p_{\mathrm{a}}$ is the only Nash equilibrium:

$$\mathrm{PI}(\mathrm{PD}(p_{\mathrm{a}})) = \frac{\delta p_{\mathrm{a}}^2 + \beta p_{\mathrm{a}}}{P} + 1. \tag{16}$$

Since $\beta > 0$ and $\delta = 0$, $\mathrm{PI}(\mathrm{PD}(p_{\mathrm{a}})) \geq 1$.

We next consider $\delta > 0$. Then $u_{\mathrm{a}}^* < 1$, since $R > P$. Now, if $u_{\mathrm{a}}^* \leq 0$, then $x_{\mathrm{a}}^{\mathrm{ne}} = 0$ and so $\mathrm{PI}(\mathrm{PD}(p_{\mathrm{a}})) = 1$. Hence we can just consider $u_{\mathrm{a}}^* \in (0,1)$. Then, social welfare attains the global minimum value of

$$W(u_{\mathrm{a}}^*) = P - \frac{(2P - (S+T))^2}{4(R + P - (S+T))}, \tag{17}$$

and attains a local maximum value of $W(1) = R$ (since $R > P$). It holds that $f_{D,\mathrm{a}}(0) = 2P - (S+T) > 0$, and recall $f_{C,\mathrm{a}}(u) = -f_{D,\mathrm{a}}(u)$. Thus, $x^{\mathrm{ne}} = (0,0)$ is a Nash equilibrium for any $p_{\mathrm{a}} \leq u_{\mathrm{a}}^*$. Now, because the payoff functions are affine, the only equilibrium $x_{\mathrm{a}}^{\mathrm{ne}} \in (0,1)$ for which $f_{C,\mathrm{a}}(x_{\mathrm{a}}^{\mathrm{ne}}) = f_{C,\mathrm{a}}(x_{\mathrm{a}}^{\mathrm{ne}})$ is u_{a}^*. Therefore, if $u_{\mathrm{a}}^* \leq p_{\mathrm{a}}$, then $\boldsymbol{x}^{\mathrm{ne}} = (u_{\mathrm{a}}^*, 0)$. Further, $x_{\mathrm{a}}^{\mathrm{ne}} = p_{\mathrm{a}}$ (i.e. $p_{\mathrm{a}} \leq u_{\mathrm{a}}^*$) if and only if $f_{C,\mathrm{a}}(p_{\mathrm{a}}) \geq f_{D,\mathrm{a}}(p_{\mathrm{a}})$. Hence, for $\delta > 0$, the set of Nash equilibria for $\mathrm{PD}(p_{\mathrm{a}})$ is summarized by $x_{\mathrm{a}}^{\mathrm{ne}}$ as follows:

$$x_{\mathrm{a}}^{\mathrm{ne}} = \begin{cases} 0, & \text{if } p_{\mathrm{a}} < u_{\mathrm{a}}^* \\ \{0, u_{\mathrm{a}}^*, p_{\mathrm{a}}\}, & \text{if } p_{\mathrm{a}} \geq u_{\mathrm{a}}^* \end{cases}. \tag{18}$$

Thus, when W is strictly convex, the resulting perversity index given by (12) is obtained. To see that $\mathrm{PI}(\mathrm{PD}(p_{\mathrm{a}})) \leq 1$, notice that $\frac{\beta^2}{4P\delta} \geq 0$, since $\delta > 0$ and $\beta^2 \geq 0$.

Finally, we consider when W is strictly concave ($\delta < 0$). It can be shown W attains the global maximum value of

$$W(u_{\mathrm{a}}^*) = P - \frac{(2P - (S+T))^2}{4(R + P - (S+T))}, \tag{19}$$

and local minimum value of $W(0) = P$. Now, we need to identify the values $x_{\mathrm{a}}^{\mathrm{ne}}$ can attain. Since $x_{\mathrm{s}}^{\mathrm{ne}} = 0$, and by Lemma 5.1, we know that

$$x_{\mathrm{a}}^{\mathrm{ne}} = \begin{cases} p_{\mathrm{a}}, & \text{if } p_{\mathrm{a}} < u_{\mathrm{a}}^* \\ u_{\mathrm{a}}^*, & \text{if } p_{\mathrm{a}} \geq u_{\mathrm{a}}^* \end{cases}. \tag{20}$$

Thus, when W is strictly concave, the perversity index given by (13) is obtained. To see that $\mathrm{PI}(\mathrm{PD}(p_{\mathrm{a}})) \geq 1$, notice that since $\delta < 0$ and $\beta^2 \geq 0$, it follows that $\frac{\beta^2}{4P\delta} \leq 0$. ■

REFERENCES

[1] T. Roughgarden, *Selfish Routing and the Price of Anarchy*. MIT press, 2005.

[2] G. Hardin, "The tragedy of the commons," *Science*, vol. 162, no. 3859, pp. 1243–1248, 1968.

[3] W. D. Hamilton, "The Evolution of Altruistic Behavior," *The American Naturalist*, vol. 97, no. 896, pp. 354–356, 1963.

[4] L. Lehmann and L. Keller, "The evolution of cooperation and altruism–a general framework and a classification of models," *Journal of evolutionary biology*, vol. 19, no. 5, pp. 1365–1376, 2006.

[5] B. Kerr, P. Godfrey-Smith, and M. W. Feldman, "What is altruism?," *Trends in ecology & evolution*, vol. 19, no. 3, pp. 135–140, 2004.

[6] J. E. Strassmann, Y. Zhu, and D. C. Queller, "Altruism and social cheating in the social amoeba dictyostelium discoideum," *Nature*, vol. 408, no. 6815, pp. 965–967, 2000.

[7] P. N. Brown, B. Collins, C. Hill, G. Barboza, and L. Hines, "Individual altruism cannot overcome congestion effects in a global pandemic game," in *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–6, 2022.

[8] I. Dahmouni and E. Kanani Kuchesfehani, "Necessity of social distancing in pandemic control: A dynamic game theory approach," *Dynamic Games and Applications*, vol. 12, no. 1, pp. 237–257, 2022.

[9] I. Caragiannis, C. Kaklamanis, P. Kanellopoulos, M. Kyropoulou, and E. Papaioannou, "The impact of altruism on the efficiency of atomic congestion games," in *Trustworthly Global Computing: 5th International Symposium, TGC 2010, Munich, Germany, February 24-26, 2010, Revised Selected Papers 5*, pp. 172–188, Springer, 2010.

[10] R. Li, P. N. Brown, and R. Horowitz, "Employing altruistic vehicles at on-ramps to improve the social traffic conditions," in *2021 American Control Conference (ACC)*, pp. 4547–4552, IEEE, 2021.

[11] E. Bıyık, D. A. Lazar, R. Pedarsani, and D. Sadigh, "Altruistic autonomy: Beating congestion on shared roads," in *Algorithmic Foundations of Robotics XIII: Proceedings of the 13th Workshop on the Algorithmic Foundations of Robotics 13*, pp. 887–904, Springer, 2020.

[12] P. N. Brown, "When altruism is worse than anarchy in nonatomic congestion games," in *2021 American Control Conference (ACC)*, pp. 4503–4508, IEEE, 2021.

[13] C. Hill and P. N. Brown, "The tradeoff between altruism and anarchy in transportation networks," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1442–1447, IEEE, 2023.

[14] E. Fehr and K. M. Schmidt, "Chapter 8 The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories," in *Handbook of the Economics of Giving, Altruism and Reciprocity*, vol. 1, pp. 615–691, Elsevier, 2006.

[15] P.-A. Chen, B. D. Keijzer, D. Kempe, and G. Schäfer, "Altruism and its impact on the price of anarchy," *ACM Transactions on Economics and Computation (TEAC)*, vol. 2, no. 4, pp. 1–45, 2014.

[16] T. Roughgarden, "Intrinsic Robustness of the Price of Anarchy," *Journal of the ACM*, vol. 62, no. 5, pp. 32:1–32:42, 2015-11-02.

[17] P. N. Brown and J. R. Marden, "Can taxes improve congestion on all networks?," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 4, pp. 1643–1653, 2020.

[18] S. Sekar, L. Zheng, L. J. Ratliff, and B. Zhang, "Uncertainty in multicommodity routing networks: When does it help?," *IEEE Transactions on Automatic Control*, vol. 65, no. 11, pp. 4600–4615, 2019.

[19] P. N. Brown and J. R. Marden, "The benefit of perversity in taxation mechanisms for distributed routing," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 6229–6234, 2017-12.

[20] B. L. Ferguson, P. N. Brown, and J. R. Marden, "Carrots or sticks? the effectiveness of subsidies and tolls in congestion games," in *2020 American Control Conference (ACC)*, pp. 1853–1858, IEEE, 2020.