Establishing Nearly Universal Cooperation in Finitely Repeated Games via Limited-Altruism Types

Yuan Deng
Duke University
ericdy@cs.duke.edu

Vincent Conitzer
Duke University
conitzer@cs.duke.edu

Abstract

Is it possible to introduce a small number of agents into an environment, in such a way that an equilibrium results in which almost everyone (including the original agents) cooperates almost all the time? This is a compelling question for those interested in the design of beneficial game-theoretic AI, and it may also provide insights into how to get human societies to function better. We investigate this broad question in the specific context of finitely repeated games, and obtain a mostly positive answer. Our main novel technical tool is the use of *limited altruism (LA) types*, which behave altruistically towards other LA agents but not towards selfish agents. The uncertainty about which type of agent one is facing turns out to be essential in establishing cooperation. We provide characterizations in several families of games of which LA types are effective for our purposes.

1 Introduction

One of the main messages of game theory is that self-interested behavior can result in Pareto-dominated outcomes. There are various standard ways to attempt to address this problem, such as prohibiting undesired behavior or taxing behavior according to the externalities it imposes. Still, these approaches are often difficult to implement effectively in practice. This is especially the case when we consider open and distributed systems without a single entity that can exercise control. The problem is exacerbated when such systems are populated by artificially intelligent agents that cannot necessarily be traced back to, or be stopped by, any entity in the "real" world. Is there still something that we can do to avoid undesirable outcomes?

In many environments, we have the ability to design a small (but only a small) fraction of the agents. These agents may act in an environment in which other entities are developing their own AI agents, or in an environment with human actors. Can we use this very limited influence, i.e., our ability to design a small fraction of the agents, to create a system-wide equilibrium that is significantly more desirable? In particular, can we design natural *preferences* for these agents in such a way that good equilibrium behavior results?

As a concrete example, consider the finitely repeated prisoner's dilemma (one round of which is as in Figure 1). If both agents are selfish types that attempt to maximize only their own payoffs, the only equilibrium is for both players to defect in all rounds. (By backward induction, in the last round defection is a dominant strategy; therefore, actions in the previous rounds have no influence on what happens in the last round. Thus, the argument carries on for previous rounds.)

What if we introduce some agents of an *altruistic* type—for example, agents that act to maximize the *average* utility across agents? Such an agent will, in the last round, always want to cooperate, because (it can be checked that) this improves the average utility by $1/2$ regardless of what the other

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	$1,1$	$-2,3$
	Defect	$3,-2$	$0,0$

Figure 1: Prisoner's dilemma.

agent does. If all other agents are selfish (and therefore will defect in the last round), then it follows that behavior in the last round is fully determined regardless of what happened before it. We can then take the same analysis to the second-to-last round, etc., to conclude that the only equilibrium is for selfish types to always defect and altruistic types to always cooperate. While this may be an improvement, it is only a small one, given that by assumption we can add only a small fraction of agents of a different type.

Can we do better by adding a few agents of another type? Specifically, can we use such agents to get the selfish agents to (sometimes) cooperate? It turns out that we can, by adding what we call agents of a *limited altruism* (LA) type. Such agents will behave altruistically towards other agents of that same type (for example, caring about the average of their utilities when facing such a type), but selfishly towards selfish agents (caring only about their own utility in that case). A key aspect is that agents should not know what type of agent they are playing against.

To make this concrete, suppose both agents have probability 0.8 to be of the LA type (that cares about the average utility when facing another LA type), in the sense of a Bayesian game. (This, of course, is a large fraction of the agents, not a small one, but it will help in conveying the intuition.) If there is only a single round, then by a simple calculation, we can show that a selfish type will defect but an LA type will be willing to cooperate: an LA type gets $0.8 \cdot 1 + 0.2 \cdot (-2) = 0.4$ for cooperating, whereas deviating to defecting would still only give $0.8 \cdot 0.5 + 0.2 \cdot 0 = 0.4$. Let us extend this to a two-period version without discounting. In this case, we can get a selfish type to cooperate in the first round, too. This is because playing defect in the first round would reveal the player to be a selfish type, resulting in the opponent not cooperating in the last round even if it is an LA type (because its altruism is directed only at other LA types). Thus, a selfish type obtains $1.0 \cdot 1 + 0.8 \cdot 3 + 0.2 \cdot 0 = 3.4$ for cooperating in the first round (and then defecting in the second), whereas deviating to defecting immediately would give only $1.0 \cdot 3 + 1.0 \cdot 0 = 3$. It follows that for a T-period version without discounting, we can obtain cooperation, i.e., both players always playing Cooperate no matter which types they are of, for the first $(T - 1)$ periods.

Can we still achieve cooperation when agents only have probability $2/3$ to be of the LA type? If there is only a single round, the LA type has incentive to deviate now: he only gets $2/3 \cdot 1 + 1/3 \cdot (-2) = 0$ by playing Cooperate, whereas deviating gives him $2/3 \cdot 0.5 + 1/3 \cdot 0 = 1/3$. What if there are two periods? Note that *if* the belief that the other player is of the LA type remains unchanged across both periods, then each player will always play Defect, no matter which type he is of, by backward induction. But what if we select strategies so that the belief changes?

Consider a strategy profile such that a player of the LA type never plays Defect first, and once a player has played Defect, the other player believes that that player is selfish and punishes him by playing Defect. In particular, suppose that in the first round, a selfish player plays each action with probability $1/2$. Then, if a player of the LA type observes that the other player played Cooperate in the first round, then he believes that the other player has probability $(2/3)/(2/3 + 1/2 \cdot 1/3) = 0.8$ to be of the LA type. By the analysis from before, this LA player is willing to play Cooperate given this belief. Now let us consider the resulting incentives for the selfish type in the first round. Under our construction, in the first round, the other player has probability $1/6$ to be of the selfish type and play Defect, probability $1/6$ to be of the selfish type and play Cooperate, and probability $2/3$ to be of the LA type and play Cooperate. Therefore, playing Cooperate in the first round (and Defect in the second) gives a selfish type $1/6 \cdot (-2+0) + 1/6 \cdot (1+0) + 2/3 \cdot (1+3) = 5/2$, while playing Defect gives him $1/6 \cdot (0 + 0) + 1/6 \cdot (3 + 0) + 2/3 \cdot (3 + 0) = 5/2$. Therefore, a selfish player is indifferent between playing Cooperate and playing Defect. As a result, the selfish player is indeed willing to mix between the actions in the first round.

Building further on this, if there are three periods, in the first period, a selfish player is willing to play Cooperate with probability 1: he obtains $1 + 5/2 = 7/2$ by playing Cooperate in the first round, compared to only 3 by playing Defect. Finally, it can be verified that the players of the LA type are also best off cooperating, and that everyone is best off defecting once someone has deviated.

Intuitively, our approach is as follows. We make the selfish types indifferent between cooperating and defecting in certain rounds, so that in equilibrium, a selfish type defects *with some probability* in those rounds. As a result, selfish players reveal their type with some probability in each round by defecting; hence, conditional on nobody having defected, the belief that the other player is of the LA type increases over time. This way, we can attain cooperation even if the initial probability of LA types is low.

How far can we push this? Is it possible to attain cooperation with an arbitrarily small fraction of LA types? In any game? We show that under certain conditions, the answer is "yes"—but for this we need LA types whose preferences are somewhat different from the one described above, which we call LA_{avg}. For example, we show that LA_{avg} does not work (in terms of attaining a "yes" answer to the above questions) for all parameterizations of the prisoner's dilemma, but LA_{min}—a type that cares about the *minimum* of the two players' payoffs when facing another such type—does. We show that LA_{min} (as well a variant of the LA_{avg} type) also works in *independent-effect* games, where the effect of one player's chosen action on both players' payoffs is independent of the other player's action. Finally, we show that for general games (satisfying the relevant conditions), no definition of an LA type that only cares about a function of the payoffs obtained can work; but, if we allow the LA type also to care about whether the two players played the same action, then there are several definitions of the LA type that work. In all cases, we also provide general results on precisely what definitions of the LA type can work.

1.1 Two Interpretations of Our Results

There are (at least) two natural ways to interpret our results. One is from the perspective of actually designing agent types—for example, these types may be artificially intelligent software agents that we design precisely for the purpose of introducing them in small numbers into an existing game, in order to steer equilibrium behavior towards a socially optimal outcome. Under this interpretation, we believe our results are very satisfactory, as we exhibit quite natural agent types that do the job.

Another interpretation of our results is as potentially describing, at a very abstract level, a phenomenon that we observe in our human world. Some of our LA types might be considered a reasonable reflection of how some people approach strategic decisions; they want to be kind to others who would be similarly kind to them. For example, most people would feel at least a little bad about their choice to defect when it emerges that the other player cooperated, while they would not feel bad if the other player defected. And we observe that even such limited kindness, even when practiced by relatively few people, can have a strong positive effect on the final outcome under some circumstances. Indeed, experimentally, we do find that people cooperate in finitely repeated games Andreoni and Miller (1993); Cooper et al. (1996); Axelrod et al. (1987) (though other explanations of this could also be given) Under this interpretation, we believe our results are somewhat satisfactory: they can be said to give a reasonable, if quite abstract and limited, account of this phenomenon. However, from this perspective, we find that the types that we show to be universally successful (for example, LA_{min} in prisoner's dilemma games, and $LA_{\text{coordinate}}$, which we will introduce later, in general), while at least somewhat natural, are not quite as natural to us as some other types (such as LA_{avg}) that in fact turn out not to be universally successful.

1.2 Related Works

In the economics and computation literature, a closely related topic is that of *Stackelberg routing* (Roughgarden, 2004). In routing games, it is well known that selfish routing can result in equilibria that are socially suboptimal, i.e., there is a nontrivial price of anarchy (Roughgarden and Tardos, 2002). The Stackelberg routing model allows a central player to directly control the behavior of a fraction of the traffic before anyone else moves. It has been shown that this can significantly reduce the price of anarchy (Roughgarden, 2004). While there are significant conceptual similarities between this work and ours, there are also quite a few significant differences, including the following. We do not assume any distinguished (e.g., Stackelberg leadership) role for the player types that we introduce; we specify a simple utility function for these types and let them play optimally accordingly, rather than specifying their behavior directly; we aim to get arbitrarily close to social optimality rather than to just obtain a constant price of anarchy; and finally, of course, we are in a very different setting, namely finitely repeated games, rather than network routing games (which is what allows us to aim to get arbitrarily close to optimality, by increasing the number of rounds). Also closely related, Chen et al. (2014) analyze the price of anarchy of traffic routing under the assumption that users are partially altruistic. They show that if the average level of altruism in the population is large, then the price of anarchy is small.

Technically more closely related to our work is older work in the literature on finitely repeated prisoner's dilemma games Kreps et al. (1982). They point out that in two-person finitely repeated prisoner's dilemma games, if both players are uncertain about the other player's utility function—specifically allowing that with a small probability, a player is happy to play Cooperate if the other player does so as well—then there exists a sequential equilibrium where both players play Cooperate until the last few stages of the game (though they do not work out the equilibrium in full in their paper, just noting that "the details of this equilibrium are quite complex"). There are several differences between our work and theirs. First, we focus primarily on types that care about the *payoffs* of both players, rather than directly about whether both players play a given action. We find such types more natural; in particular, if we consider the deployment of agents in rich and open-ended settings where it may initially not even be clear which courses of action are most desirable, it seems easier to specify utilities as a function of the payoffs. We succeed at this aim in prisoner's dilemma games as well as in independent-effect games, though in the fully general case, we also use types that care whether the players chose the same action—and we show that in general such a move is in fact necessary, in the sense that here we cannot (always) succeed with types that only care about payoffs. (Kreps et al. (1982) do not study independent-effect games or general games.) Additionally, in their model, a player's utility solely depends on his own type and the actions played, while in our case an LA type has a u^{LA} payoff function only if the other player is also an LA type. This is the key trick that allows us to construct effective types that care only about the payoffs, and such conditional altruism seems quite natural. Moreover, we obtain concise *characterizations* of desired sequential equilibrium in which both players play Cooperate for most rounds, which allows us to identify natural types that are universally successful and extends our results to more general games than prisoner's dilemma.

Our work is also related to work by Maskin and Fudenberg (1986), which already shows that cooperation can be obtained in finitely repeated games using a small fraction of "behavioral" types. A key difference between that work and ours is that that work requires a very direct specification of how the behavioral types act (rather than having those types act strategically with respect to a natural utility function), whereas we set out to find general and natural utility functions for our "limited altruism" types—in the sense of how these types value outcomes when playing with another LA type—that work universally across games, when the LA types pursue this utility strategically. In our opinion, again, when we consider deploying agents to rich, complex environments, it is more useful to be able to just specify a natural utility function rather than having to specify behavior directly. The downside of this approach is that we need significantly more technical machinery to make this work. In addition, we prove desirable convergence rates (in the sense of how many rounds of play are needed for desirable play) as a function of the fraction ε of LA types; our rate is $O(\log 1/\varepsilon)$ whereas the result by Maskin and Fudenberg (1986) gives $O(1/\varepsilon)$.

2 Preliminaries

A finitely repeated game $G(T) = \langle G^1, \cdots, G^T \rangle$ is a T-fold repetition of a base game G, where $G^t = G$ for all $1 \leq t \leq T$. Throughout, we restrict our attention to the case where G is a two-player, symmetric, normal-form game. A symmetric normal-form game is defined by $G = \langle X, u_1^S, u_2^S \rangle$, where X is the set of pure strategies (for both players 1 and 2) and their utility functions are given as $u_i^S : X \times X \rightarrow R$. As usual, we use $-i$ to denote the player other than player i. Here, $u_i^S(x_i, x_{-i})$ denotes player i's utility when player i plays pure strategy $x_i \in X$ and player $-i$ plays pure strategy $x_{-i} \in X$. In a symmetric game, $u_1^S(x_1, x_2) = u_2^S(x_1', x_2')$ if $x_1 = x_2'$ and $x_2 = x_1'$.

2.1 Limited Altruism Type

We now generalize to a Bayesian repeated game where each player has one of two possible types. Specifically, the set of types is $\Theta = \{S, LA\}$, where S is the *selfish* type and LA is the *limited-altruism* type. Given a repeated game as defined above, the payoff that i receives in a round of the game when i is of type S is simply $u_i^S(x_i, x_{-i})$ (as before). However, payoffs are changed for the LA type, but *only if* the other player is also of the LA type; if the other player is of the S type, then the payoffs for the former (LA) player are again simply $u_i^S(x_i, x_{-i})$. On the other hand, if *both* players are of the LA type, then the utility functions change to $u_1^{LA}, u_2^{LA} : X \times X \rightarrow R$ for player 1 and 2, respectively. More specifically, we are interested in LA types whose utilities in this case are defined by a function of the selfish payoffs, where this function can be applied to *any* game G.

Definition 2.1 (Universal LA type)**.** *An LA type LA_ϕ is* universal *if there exists a function f_ϕ such that for every possible u_1^S and u_2^S, and all $x_1, x_2 \in X$, $u_i^{LA}(x_1, x_2) = f_\phi(u_i^S(x_1, x_2), u_{-i}^S(x_1, x_2))$*

One example would be LA_{avg}, for which $f_{\text{avg}}(u_i^S(x_1, x_2), u_{-i}^S(x_1, x_2)) = \frac{u_i^S(x_1,x_2)+u_{-i}^S(x_1,x_2)}{2}$. This type cares about the average of the two players' selfish payoffs if the other player is also of the LA_{avg} type, and just about her own payoff otherwise. Another example is LA_{min}, for which $f_{\text{min}}(u_i^S(x_1, x_2), u_{-i}^S(x_1, x_2)) = \min\{u_i^S(x_1, x_2), u_{-i}^S(x_1, x_2)\}$. This type cares about the minimum of the two players' selfish payoffs if the other player is also of the LA_{min} type, and just about her own payoff otherwise. Note that we will not consider any cases where, for example, an LA_{avg} type plays against an LA_{min} type. All our LA types are always the same, but we are interested in which function ϕ gives us desirable properties. That is, if we get to *design* the LA type and insert a small fraction of agents of that type into a population of selfish types, what function ϕ should we use for the LA type?

We make three assumptions on f_ϕ, all of which are satisfied by both LA_{avg} and LA_{min}. The first can be achieved by normalization.

Assumption 1. $f_\phi(0, 0) = 0$.

The second assumption says that f_ϕ is scale invariant.

Assumption 2. $f_\phi(\alpha \cdot v, \alpha \cdot w) = \alpha \cdot f_\phi(v, w)$ *for* $\alpha > 0$.

This second assumption rules out, for example, an LA type that, when facing another LA type, very much wants the other player to get utility at least 1, but after that becomes completely selfish. Such a type would start acting differently if we (say) doubled all payoffs.

Assumption 3. $f_\phi(v, v) \geq v$.

This third assumption rules out, for example, an LA type that has generally lower payoffs when playing against another LA type than when playing against an S type (and therefore would focus relatively more on the case where it faces an S type).

We assume that the type of each player, θ_i for player i, is private and only known to the player himself, which is identically and independently drawn once, before any rounds are played, according to $\Pr[S] = 1 - \varepsilon^{init}$ and $\Pr[LA] = \varepsilon^{init}$. θ_i remains the same throughout game play. At the end of each round, each player observes both players' actions, and hence knows what the selfish payoffs are, but does *not* observe her actual payoffs. Otherwise, a (say) LA_{avg} type might observe that an outcome was realized with selfish payoffs 1 and 2, and her own actual payoff was 1.5, and therefore the other player must be an LA type as well. Instead, in our model, an LA player is generally unsure about her actual accrued payoffs (which we can think of as being revealed after the last round of the game). An S player, on the other hand, always knows her accrued payoffs since they do not depend on the other player's type.

2.2 Histories, Strategies and Utilities

Denote the history for the first t rounds by h^t so that $h^0 = ()$. Since player i has no direct information about player $-i$'s type, his action for the t-th round depends on θ_i and h^{t-1} only. Thus, (θ_i, h^{t-1}) corresponds to an information set. A strategy σ_i for player i is a function that takes as input an information set, i.e., the player's type and a history of play, and selects a strategy in $\Delta(X)$, where $\Delta(X)$ is the set of distributions over X. At the t-th round, $\sigma_i(x_i \mid \theta_i, h^{t-1})$ defines the probability of player i choosing action $x_i \in X$ given his type θ_i and the history h^{t-1}. A realized action $a_i^t \in X$ is drawn from the distribution $\sigma_i^t(\cdot \mid \theta_i, h^{t-1})$. Let $a^t = (a_1^t, a_2^t)$ be the resulting action profile in the t-th round. We then update the history to $h^t = (h^{t-1}, a^t)$.

Player i's utility is determined by the final game history h^T, his own type θ_i, and his opponent's type θ_{-i}:

$$u_i(h^T, \theta_i, \theta_{-i}) = \sum_{t=1}^{T} u_i^{\mathcal{C}}(a_i^t, a_{-i}^t)$$

where $\mathcal{C} = LA$ if $\theta_1 = \theta_2 = LA$; otherwise, $\mathcal{C} = S$.

2.3 Beliefs and Equilibria

How should the players play? One candidate solution concept is *Bayes-Nash equilibrium*, in which each player plays a best response (in terms of expected utility over all rounds combined) given his own type and his prior belief about the other player's type. However, this fails to ensure that players play optimally off the equilibrium path of play, and thus may involve threats that are not credible. Moreover, even subgame perfection is not sufficient: in general, there will not be cleanly separated subgames in the extensive form (where the current node is common knowledge) because players have uncertainty about each other's type. Hence, we need a more sophisticated equilibrium refinement notion.

For such refinements, we need to specify player i's belief assessment $\mu(\theta_{-i} \mid \theta_i, h^{t-1})$ about player $-i$'s type, given his own type θ_i and the history h^t. In line with the literature Fudenberg and Tirole (1991), we require that there is a single joint distribution over (θ_i, θ_{-i}) given h^{t-1}, satisfying

$$\mu_i(\theta_{-i} \mid \theta_i, h^{t-1}) \cdot \mu(\theta_i \mid h^{t-1}) = \mu(\theta_i, \theta_{-i} \mid h^{t-1})$$

where μ is the marginal probability of θ_i, θ_{-i} given h^{t-1}. This allows us to obtain private beliefs (with the subscript i) from the joint distribution (without the subscript). It also allows us to simplify the notation by setting $\mu_i(\theta_{-i} \mid \cdot, \cdot) = \mu(\theta_{-i} \mid \cdot, \cdot)$. We are now ready to introduce a refined equilibrium notion.

2.4 Perfect Bayesian Equilibrium (PBE)

Definition 2.2. *Fudenberg and Tirole (1991) A pair (μ, σ) of a belief assessment μ and a strategy profile σ is* consistent *if for all histories h^{t-1}:*

- *Bayes' rule is used to update beliefs whenever possible: for each player i and for each $a_i' \in X$, if there exists a θ_i' with $\mu(\theta_i' \mid h^{t-1}) > 0$ and $\sigma_i(a_i' \mid \theta_i', h^{t-1}) > 0$, then*

$$\mu(\theta_i \mid h^{t-1}, a^t) = \frac{\mu(\theta_i \mid h^{t-1}) \cdot \sigma_i(a_i^t \mid \theta_i, h^{t-1})}{\sum_{\hat{\theta}_i \in \Theta} \mu(\hat{\theta}_i \mid h^{t-1}) \cdot \sigma_i(a_i^t \mid \hat{\theta}_i, h^{t-1})}$$

- *The posterior beliefs are independent:*

$$\mu(\theta_i, \theta_{-i} \mid h^t) = \mu(\theta_i \mid h^t) \cdot \mu(\theta_{-i} \mid h^t) \quad \textit{for all } \theta_i, \theta_{-i}, \textit{ and } h^t$$

- *The beliefs about player i at the beginning of the $t+1$-st round only depend on h^{t-1} and player i's action at the t-th round:*

$$\mu(\theta_i \mid h^{t-1}, a^t) = \mu(\theta_i \mid h^{t-1}, \hat{a}^t) \quad \textit{for all } \theta_i, \textit{ and all } a_i^t = \hat{a}_i^t.$$

Definition 2.3. *Fudenberg and Tirole (1991) A pair of a strategy profile σ and a belief assessment μ is a* perfect Bayesian equilibrium (PBE) *if (1) (μ, σ) is consistent; (2) σ is a best response in the subgame starting from h^{t-1} to the belief $\mu(\cdot \mid \theta_i, h^{t-1})$ under the assumption that the other player will follow σ from this point.*

Fudenberg and Tirole (1991) point out that if each player has only two possible types, both types have nonzero prior probability, and types are independent, then the sets of perfect Bayesian equilibria and sequential equilibria Kreps and Wilson (1982) coincide. Therefore, in our setting, the sets of perfect Bayesian equilibria equals to the set of sequential equilibria.

For convenience, let $u_i(\theta_i, h^{t-1})$ be player i's expected utility in the subgame starting from history h^{t-1} if he is of type θ_i and $u_i(a_i^t, \theta_i, h^{t-1})$ is the expected utility if in addition he plays a_i^t at the t-th round. Note that both of these depend on (μ, σ), but this is suppressed in the notation.

2.5 Desired LA Type

Which LA types "work"? Specifically, we are interested in LA types such that even with only a small fraction of such types in the population, we will obtain cooperative play (by all players, not only the LA types) in the vast majority of rounds, provided that there are enough rounds. We now make this formal.

We restrict attention to games with a dominant strategy (for selfish types in a single round of the game), i.e., we assume that there exists a dominant strategy x_D in game G such that for all $x', y \in X$, $u_i^S(x_D, y) \geq u_i^S(x', y)$. We assume $u_i^S(x_D, x_D) = 0$ (without loss of generality, due to normalization). Define the cooperative strategy to be $x_C = \arg\max_{x \in X} u_i^S(x, x)$; we assume that (x_C, x_C) *uniquely* maximizes the social welfare among all entries of the game. Let us define a measurement of the quality of an equilibrium.

Definition 2.4 (k-desired PBE)**.** *Given G, T, u^{LA} (i.e., our choice of ϕ), and ε^{init}, we say that a pair (μ, σ) of a belief assessment and a strategy profile of $G(T)$ is a* k-desired PBE *if it is a PBE and the expected number of rounds in which both players choose x_C is at least k.*

Based on this, we can define a notion of which LA types do the job that we would like them to do. Specifically, we would like to see that an *arbitrarily small* (but positive) fraction of LA types enable an equilibrium in which we have cooperation in *almost all* rounds, provided that the number of rounds is sufficiently large.

Definition 2.5 (Desired Universal LA type)**.** *An universal LA type defined by u^{LA} (i.e., a choice of ϕ) is* desired, *if for any game G and for any $\delta > 0$, there exists $0 < \varepsilon^{init} < \delta$ and a sequence $(k(1), \cdots, k(T), \cdots)$ such that $\lim_{T \to \infty} k(T)/T = 1$ and there exists a $k(T)$-desired PBE for all T.*

3 Warm-up: Prisoner's Dilemma

Most of our ideas are well illustrated by the prisoner's dilemma (see Table 1), on which we focus in this section. We will generalize beyond the prisoner's dilemma in Section 4 and 5.

In the prisoner's dilemma, the dominant strategy x_D is Defect while the cooperative strategy x_C is Cooperate. The general form of u^{LA} is as follows (see Table 1b), where e, y and z depend on the specific LA type. (E.g., $y^{LA_{\text{avg}}} = u_i^{LA_{\text{avg}}}(\text{Cooperate}, \text{Defect}) = (b-c)/2$, and $y^{LA_{\min}} = u_i^{LA_{\min}}(\text{Cooperate}, \text{Defect}) = -c$.)

	Cooperate	Defect
Cooperate	a, a	$-c$, b
Defect	b, $-c$	0,0

(a) u^S: $b > a > 0$ and $c > 0$

	Cooperate	Defect
Cooperate	e, e	y, z
Defect	z, y	0,0

(b) u^{LA}: $e > z$ and $e \geq a$

Table 1: Utility function in the prisoner's dilemma

3.1 Framework

Suppose $\mu(\theta_1 = S \mid h^{t-1}) = \mu(\theta_2 = S \mid h^{t-1}) = 1 - \varepsilon^{init}$ (where ε^{init} is small) for all $1 \leq t \leq T$. Then, in the last round, the dominant strategy for both players is to play Defect, no matter which types they have. By backward induction, both players would play Defect in every round.

Therefore, to construct a k-desired Bayesian equilibrium of $G(T)$ with k close to T, the belief about the probability that the other player is of type LA should be large for some histories. This can be achieved as follows. If a player is of type S, he should play a mixed action that reveals him to be of type S with some probability. For example, consider a strategy profile in which both players choose to play Cooperate for the first $(t-1)$ rounds, no matter which type they have. At the t-th round, a player of type LA keeps playing Cooperate while a player of type S chooses to play either action with probability $1/2$. Then, at the beginning of the $(t+1)$-th round, for a player who witnessed the other player playing Cooperate in the t-th round, his belief that the other player is of type LA becomes $2\varepsilon^{init}/(\varepsilon^{init}+1) > \varepsilon^{init}$. Of course, we can only achieve this in equilibrium if it is in the interest of a player of type S to play a mixed action. That is, the expected utility for playing either action must be the same.

We restrict our attention to symmetric strategy profiles and belief assessments that satisfy the following assumption, which says that LA types will never defect first, and if someone has defected then everyone believes that that player has type S and plays Defect afterwards.

Assumption 4. *We say that the (symmetric) profile* (μ, σ) satisfies Assumption 4 *if for every possible history* h^{t-1}*, Bayes' rule is used to update beliefs whenever possible and*

- *if there does not exist* $a_j^{t'} \neq x_C$ *in* h^{t-1} *for any* $j \in \{1,2\}$*, then* $\sigma_i(x_C \mid LA, h^{t-1}) = 1$ *for both* $i \in \{1,2\}$*;*
- *if there exists* $a_j^{t'} \neq x_C$ *in* h^{t-1} *for some* $j \in \{1,2\}$*: if* $a_i^{t_{min}} \neq x_C$*, then* $\mu(\theta_i = S \mid h^{t-1}) = 1$*, where* $t_{min} = \min\{t' \mid a_j^{t'} \neq x_C \text{ for some } j \in \{1,2\}\}$*; in other words, a player that defected* first *is believed to be selfish with probability* 1*;*
- *if there exists* $a_j^{t'} \neq x_C$ *in* h^{t-1} *for some* $j \in \{1,2\}$*,* $\sigma_i(x_D \mid \cdot, h^{t-1}) = 1$ *for both* $i \in \{1,2\}$.

3.2 Construction

Assumption 4 requires that once a player has played Defect, then both players will keep playing Defect until the end. This leaves only the actions after histories in which no players have played Defect unspecified. Let h_C^t denote the unique such history with t rounds. We can now construct a perfect Bayesian equilibrium by a backward induction on h_C^t. Along the way, this will require assumptions on the specific LA type (i.e., how the values of e, y and z in Table 1b are derived from a, b, and c in Table 1a); these assumptions then in the end combine to give us a sufficient condition for an LA type to be desired.

Since the profile is symmetric, let $\varepsilon_t = \mu(\theta_1 = LA \mid \cdot, h_C^{t-1}) = \mu(\theta_2 = LA \mid \cdot, h_C^{t-1})$.

3.2.1 The Last Round: h_C^{T-1}

Since, with a single round remaining, Defect is a dominant strategy for a player of type S, $\sigma_i(\text{Defect} \mid S, h_C^{T-1}) = 1$. As for a player of type LA, Assumption 4 requires that $\sigma_i(\text{Cooperate} \mid LA, h_C^{T-1}) = 1$. In order for an LA player to in fact be willing to cooperate in this case require Cooperate to be a best response when S types play Defect and LA types play Cooperate:

$$\begin{aligned}&\varepsilon_T \cdot u_i^{LA}(\text{Cooperate}, \text{Cooperate}) + (1-\varepsilon_T) \cdot u_i^S(\text{Cooperate}, \text{Defect})\\ \geq\ &\varepsilon_T \cdot u_i^{LA}(\text{Defect}, \text{Cooperate}) + (1-\varepsilon_T) \cdot u_i^S(\text{Defect}, \text{Defect}).\end{aligned}$$

which is

$$\varepsilon_T \cdot e + (1-\varepsilon_T) \cdot (-c) \geq \varepsilon_T \cdot z + (1-\varepsilon_T) \cdot 0 \Leftrightarrow \varepsilon_T \geq \frac{c}{e+c-z} \tag{1}$$

Given that LA types play Cooperate in the last round, $u_i(S, h_C^{T-1}) = \varepsilon_T \cdot b$.

3.2.2 Constructing the Strategy for S by Induction

By Assumption 4, ε_t is non-decreasing as t increases. Consider t such that $\varepsilon_t \leq \varepsilon_{t+1} < 1$, i.e., a player of type S may play a mixed action in the t-th round. Let $p_{S,C}^t = \sigma_i(\text{Cooperate} \mid S, h_C^{t-1})$. By Assumption 4 and Bayes' rule, we have

$$\varepsilon_{t+1} = \frac{\varepsilon_t}{p_{S,C}^t \cdot (1-\varepsilon_t) + \varepsilon_t} \Leftrightarrow p_{S,C}^t = \frac{\varepsilon_t \cdot (1-\varepsilon_{t+1})}{\varepsilon_{t+1} \cdot (1-\varepsilon_t)}$$

Denote by p_C^t the probability that a player plays Cooperate in the t-th round (conditioning on h_C^{t-1}):

$$p_C^t = \Pr[a_i^t = \text{Cooperate} \mid h_C^{t-1}] = (1-\varepsilon_t) \cdot p_{S,C}^t + \varepsilon_t = \frac{\varepsilon_t}{\varepsilon_{t+1}}$$

Note that $p_C^T = \varepsilon_T$. The following definition specifies which sequences of cooperation probabilities are feasible based on the strategic constraints for players of type S.

Definition 3.1. *We say a sequence of $p_C(T) = (p_C^1, \cdots, p_C^T)$ is* feasible *if the following conditions hold for $t < T$:*

- *If $p_C^t < 1$, then a player of type S is willing to play a mixed action in the t-th round:*

$$u_i(S, h_C^{t-1}) = u_i(\textit{Cooperate}, S, h_C^{t-1}) = u_i(\textit{Defect}, S, h_C^{t-1});$$

- *If $p_C^t = 1$, then a player of type S weakly prefers to play Cooperate in the t-th round:*

$$u_i(S, h_C^{t-1}) = u_i(\textit{Cooperate}, S, h_C^{t-1}) \geq u_i(\textit{Defect}, S, h_C^{t-1}).$$

By Assumption 4 that both players keep playing Defect once a player plays Defect (and given that we have normalized the payoff of both players defecting to 0), we have that for all t:

$$\begin{aligned} & u_i(\text{Defect}, S, h_C^{t-1}) \\ = \; & p_C^t \cdot u_i^S(\text{Defect}, \text{Cooperate}) + (1 - p_C^t) \cdot u_i^S(\text{Defect}, \text{Defect}) + (T - t) \cdot u_i^S(\text{Defect}, \text{Defect}) \\ = \; & p_C^t \cdot u_i^S(\text{Defect}, \text{Cooperate}) = p_C^t \cdot b \end{aligned}$$

Lemma 3.1. *When $p_C^t = 1$ and $p_C^{t+1} < 1$, $p_C(T)$ is $u_i(\textit{Cooperate}, S, h_C^{t-1}) \geq u_i(\textit{Defect}, S, h_C^{t-1})$ if and only if $p_C^{t+1} \geq 1 - \frac{a}{b}$.*

Proof. Note that $p_C^t = 1$ and $p_C^{t+1} < 1$ indicate that $u_i(S, h_C^{t-1}) = u_i(\text{Cooperate}, S, h_C^{t-1}) \geq u_i(\text{Defect}, S, h_C^{t-1})$ and $u_i(S, h_C^t) = u_i(\text{Defect}, S, h_C^t)$. We have

$$\begin{aligned} u_i(\text{Cooperate}, S, h_C^{t-1}) &= u_i^S(\text{Cooperate}, \text{Cooperate}) + u_i(S, h_C^t) \\ &= u_i^S(\text{Cooperate}, \text{Cooperate}) + u_i(\text{Defect}, S, h_C^t) \\ &= a + p_C^{t+1} \cdot b \end{aligned}$$

Thus, in order to ensure $u_i(\text{Cooperate}, S, h_C^{t-1}) \geq u_i(\text{Defect}, S, h_C^{t-1})$, we must have $a + p_C^{t+1} \cdot b \geq p_C^t \cdot b = b$, which is equivalent to, $p_C^{t+1} \geq 1 - \frac{a}{b}$. □

Corollary 3.1. *When $p_C^{t+k} < 1$, and for all $t \leq t' < t+k$, $p_C^{t'} = 1$, $u_i(\textit{Cooperate}, S, h_C^{t-1}) \geq u_i(\textit{Defect}, S, h_C^{t-1})$ if $p_C^{t+k} \geq 1 - \frac{a}{b}$.*

Proof. When a player of type S plays Cooperate in round t, his utility is

$$\begin{aligned} u_i(\text{Cooperate}, S, h_C^{t-1}) &= k \cdot u_i^S(\text{Cooperate}, \text{Cooperate}) + u_i(S, h_C^{t+k-1}) \\ &= k \cdot u_i^S(\text{Cooperate}, \text{Cooperate}) + u_i(\text{Defect}, S, h_C^{t+k-1}) \\ &= k \cdot a + p_C^{t+k} \cdot b \geq a + p_C^{t+k} \cdot b \geq b \\ &= u_i(\text{Defect}, S, h_C^{t-1}) \end{aligned}$$

□

Lemma 3.2. *When $p_C^t < 1$, $p_C^{t+k} < 1$, and for all $t < t' < t+k$, $p_C^{t'} = 1$, $u_i(\textit{Cooperate}, S, h_C^{t-1}) = u_i(\textit{Defect}, S, h_C^{t-1})$ if and only if*

$$p_C^t = \frac{c}{k \cdot a - (1 - p_C^{t+k}) \cdot b + c}. \tag{2}$$

Proof. $p_C^t < 1$ and $p_C^{t+k} < 1$ imply that a player of type S weakly prefers to defect at $(t+k)$, and is indifferent between defecting and cooperating at t (a player of type S must cooperate with positive probability in round t, because otherwise we must have $p_C^{t+k} = 1$). Therefore, we must have $u_i(S, h_C^t) = u_i(\text{Defect}, S, h_C^{t-1}) = u_i(\text{Cooperate}, S, h_C^{t-1})$ and $u_i(S, h_C^{t+k-1}) = u_i(\text{Defect}, S, h_C^{t+k-1})$. By Assumption 4 that players of type LA do not defect first, we have:

$$\begin{aligned} u_i(\text{Cooperate}, S, h_C^{t-1}) &= p_C^t \cdot \Big(k \cdot u_i^S(\text{Cooperate}, \text{Cooperate}) + u_i(S, h_C^{t+k-1})\Big) \\ &\quad + (1 - p_C^t) \cdot \big(u_i^S(\text{Cooperate}, \text{Defect}) + (T - t) \cdot u_i^S(\text{Defect}, \text{Defect})\big) \\ &= p_C^t \cdot \Big(k \cdot u_i^S(\text{Cooperate}, \text{Cooperate}) + u_i(\text{Defect}, S, h_C^{t+k-1})\Big) \\ &\quad + (1 - p_C^t) \cdot \big(u_i^S(\text{Cooperate}, \text{Defect}) + (T - t) \cdot u_i^S(\text{Defect}, \text{Defect})\big) \\ &= p_C^t \cdot (k \cdot a + p_C^{t+k} \cdot b) + (1 - p_C^t) \cdot (-c) \end{aligned}$$

By the fact that $u_i(\text{Cooperate}, S, h_C^{t-1}) = u_i(\text{Defect}, S, h_C^{t-1})$, we have

$$p_C^t \cdot (k \cdot a + p_C^{t+k} \cdot b) + (1 - p_C^t) \cdot (-c) = p_C^t \cdot b \Leftrightarrow p_C^t = \frac{c}{k \cdot a - (1 - p_C^{t+k}) \cdot b + c} \qquad \square$$

Corollary 3.2. *$p_C(T)$ is feasible only if for all $1 \le t \le T$, $p_C^t \ge 1 - \frac{a}{b}$.*

Proof. For the sake of contradiction, assume that there exists a t such that $p_C^t < 1 - \frac{a}{b}$. By Lemma 3.1, p_C^{t-1} must be less than 1. Moreover, by Lemma 3.2 (with $k = 1$), we have $p_C^{t-1} = \frac{c}{a-(1-p_C^t)\cdot b+c} > 1$ producing the sought contradiction. □

Combining Lemma 3.1, 3.2 and Corollary 3.1, we can conclude:

Theorem 3.1. *Under Assumption 4, a sequence $p_C(T)$ is feasible if and only if*

- *When $p_C^t = 1$ and $p_C^{t+1} < 1$, we have $p_C^{t+1} \ge 1 - \frac{a}{b}$;*
- *When $p_C^t < 1$, $p_C^{t+k} < 1$, and for all $t < t' < t + k$, $p_C^{t'} = 1$, we have $p_C^t = \frac{c}{k\cdot a-(1-p_C^{t+k})\cdot b+c}$.*

Now, we are ready to construct a sequence $p_C(T)$ that is feasible. We consider two different kinds of sequence $p_C(T)$, one in Lemma 3.3 and one in Lemma 3.4. In both cases, we also show that the belief that a player is of the LA type can be lower than δ even only $O(\log \frac{1}{\delta})$ rounds before the end. That is, we do not need many rounds in order to accommodate a low initial fraction of LA types.

The first of the two lemmas requires that the effect of an action on payoffs is independent of which action the other player chooses.

Lemma 3.3. *Suppose $c = b - a$. Then, let $p_C(T) = (1, \cdots, 1, p_C^{t^*}, \cdots, p_C^T)$ be such that for all $t^* \le t < T$, $p_C^t = \frac{c}{a-(1-p_C^{t+1})\cdot b+c}$ and $p_C^T \ge 1 - \frac{a}{b}$. Then, $p_C(T)$ is feasible. Moreover, for any $\delta > 0$, $\varepsilon_{T-O(\log \frac{1}{\delta})} \le \delta$.*

Proof. By Theorem 3.1, since $p_C^t = \frac{c}{a-(1-p_C^{t+1})\cdot b+c}$ for all $t^* \le t < T$, all that remains to show is that $p_C^{t^*} \ge 1 - \frac{a}{b}$. By induction, assume $1 - \frac{a}{b} \le p_C^t \le 1$ for $t \ge \bar{t}$ (where the base case $\bar{t} = T$ is satisfied by assumption). For $t = \bar{t} - 1$, we have $p_C^t = \frac{c}{a-(1-p_C^{t+1})\cdot b+c} \ge \frac{c}{a+c} = 1 - \frac{a}{b}$ where the last inequality is due to the assumption that $c = b - a$.

As for the convergence rate, rearranging (2) when $k = 1$, we have $b \cdot p_C^t \cdot p_C^{t+1} = (b - a - c) \cdot p_c^t + c = c$. Therefore, $p_C^t \cdot p_C^{t+1} = \frac{\varepsilon_t}{\varepsilon_{t+2}} \le \frac{c}{b}$. Thus, for any $\delta > 0$, we have $\varepsilon_{T-2\cdot\log_{b/c}\frac{1}{\delta}} = \varepsilon_{T-2\cdot\log_{c/b}\delta} \le \delta$. □

Definition 3.2 ($\gamma^*(G)$)**.** *Given a game G such that $a = u_i^S(x_C, x_C)$, $b = u_i^S(x_D, x_C)$, $-c = u_i^S(x_C, x_D)$ and $0 = u_i^S(x_D, x_D)$. Let $\gamma^*(G)$ be the positive root of the quadratic equation $f(\gamma) = b \cdot \gamma^2 + (a - b + c) \cdot \gamma - c = 0$.*

Proposition 3.1. *Given a game G, $\gamma^*(G) \ge \max(1 - \frac{b}{a}, \frac{c}{a+c})$.*

The proof is in Appendix. When G is clear in the context, we write γ^* for convenience.

Lemma 3.4. *Let $p_C(T) = (1, \cdots, 1, p_C^{t^*}, \cdots, p_C^T)$ be such that for all $t^* \le t < T$, $p_C^t = \frac{c}{a-(1-p_C^{t+1})\cdot b+c}$ and $p_C^T = \gamma^*$. Then, $p_C(T)$ is feasible. Moreover, for any $\delta > 0$, $\varepsilon_{T-O(\log \frac{1}{\delta})} \le \delta$.*

Proof. Note that if $p_C^T = \gamma^*$, then for all $t^* \le t < T$, $p_C^t = \gamma^*$, due to the following proof by induction. We have $f(\gamma) = b \cdot \gamma^2 + (a - b + c) \cdot \gamma - c = 0$, which is equivalent to $\frac{c}{a-(1-\gamma)\cdot b+c} = \gamma$. Hence, given that $p_C^{t+1} = \gamma^*$, we have $p_C^t = \frac{c}{a-(1-\gamma^*)\cdot b+c} = \gamma^*$. Hence, we know that $p_C^{t^*} = \gamma^*$. All that remains to show is that $p_C^{t^*} = \gamma^* \ge 1 - \frac{a}{b}$. By Propostion 3.1, we have $1 - \frac{a}{b} < \gamma^*$.

As for the convergence rate, note that $p_C^t = \frac{\varepsilon_t}{\varepsilon_{t+1}} = \gamma^*$. Therefore, for any $\delta > 0$, we have $\varepsilon_{T-\log_{1/\gamma^*}\frac{1}{\delta}} = \varepsilon_{T-\log_{\gamma^*}\delta} \le \delta$. □

3.2.3 Verifying Incentives for LA by Induction

We have discussed how to construct the sequence of cooperation probabilities in such a way as to ensure that the selfish types do not deviate. However, we also need to ensure that the limited-altruism types do not deviate. This is what the following lemma achieves. It states that, if the sequence is feasible (from the perspective of getting the selfish types not to deviate) and is such that an LA type is willing to cooperate in the last round (assuming no defections have taken place), then an LA type is willing to cooperate in all rounds (assuming no defections have taken place).

Lemma 3.5. *If $p_C(T)$ is feasible and $p_C^T \geq \frac{c}{e+c-z}$, then for all $1 \leq t \leq T$, $u_i(LA, h_C^{t-1}) = u_i(\textit{Cooperate}, LA, h_C^{t-1}) \geq u_i(\textit{Defect}, LA, h_C^{t-1})$.*

Proof. We prove by induction. Suppose $u_i(LA, h_C^{t'-1}) = u_i(\text{Cooperate}, LA, h_C^{t'-1}) \geq u_i(\text{Defect}, LA, h_C^{t'-1})$ for $t' > t$; we will show it is also true for $t' = t$. The base case $t = T$ is true since by the assumption in the lemma, $p_C^T = \varepsilon_T \geq \frac{c}{e+c-z}$, which satisfies (1). As for the t-th round, first note that if a player of type LA plays Defect, his utility is

$$\begin{aligned} u_i(\text{Defect}, LA, h_C^{t-1}) = \;& \varepsilon_t \cdot u_i^{LA}(\text{Defect}, \text{Cooperate}) + (p_C^t - \varepsilon_t) \cdot u_i^S(\text{Defect}, \text{Cooperate}) \\ & + (1 - p_C^t) \cdot u_i^S(\text{Defect}, \text{Defect}) + (T - t) \cdot u_i^S(\text{Defect}, \text{Defect}) \\ = \;& \varepsilon_t \cdot z + (p_C^t - \varepsilon_t) \cdot b \end{aligned}$$

(Note that this makes use of the fact that players will always defect after a defection has taken place; we will return to discussing whether this is indeed optimal for the players in Section 3.2.4.)

All that remains to show is that $u_i(\text{Cooperate}, LA, h_C^{t-1})$ is always at least as large as this expression. We do this by considering two cases.

Case (1): If $p_C^t = 1$, then we have $\varepsilon_t = \varepsilon_{t+1}$ and:

$$\begin{aligned} u_i(\text{Cooperate}, LA, h_C^{t-1}) = \;& \varepsilon_t \cdot u_i^{LA}(\text{Cooperate}, \text{Cooperate}) + (1 - \varepsilon_t) \cdot u_i^S(\text{Cooperate}, \text{Cooperate}) + u_i(LA, h_C^t) \\ \geq \;& a + u_i(LA, h_C^t) \geq a + u_i(\text{Defect}, LA, h_C^t) \\ = \;& a + \varepsilon_{t+1} \cdot z + (p_C^{t+1} - \varepsilon_{t+1}) \cdot b = a + \varepsilon_t \cdot z + (p_C^{t+1} - \varepsilon_t) \cdot b \\ \geq \;& \varepsilon_t \cdot z + (1 - \varepsilon_t) \cdot b = u_i(\text{Defect}, LA, h_C^{t-1}) \end{aligned}$$

where the last inequality is due to the fact that for all t', $p_C^{t'} \geq 1 - \frac{a}{b}$ by Corollary 3.2.

Case (2): If $p_C^t < 1$, let $k > 0$ be such that $p_C^{t+k} < 1$ and $p_C^{t'} = 1$ for all $t < t' < t + k$. (Such a k must exist because $p_C^T < 1$ and $t < T$.) We have $p_C^t \cdot \varepsilon_{t+k} = \varepsilon_t$ and:

$$\begin{aligned} u_i(\text{Cooperate}, LA, h_C^{t-1}) = \;& \varepsilon_t \cdot k \cdot u_i^{LA}(\text{Cooperate}, \text{Cooperate}) + (p_C^t - \varepsilon_t) \cdot k \cdot u_i^S(\text{Cooperate}, \text{Cooperate}) \\ & + p_C^t \cdot u_i(LA, h_C^{t+k-1}) + (1 - p_C^t) \cdot u_i^S(\text{Cooperate}, \text{Defect}) \\ & + (1 - p_C^t) \cdot (T - t) \cdot u_i^S(\text{Defect}, \text{Defect}) \\ \geq \;& p_C^t \cdot (k \cdot a + u_i(LA, h_C^{t+k-1})) + (1 - p_C^t) \cdot (-c) \\ \geq \;& p_C^t \cdot (k \cdot a + u_i(\text{Defect}, LA, h_C^{t+k-1})) + (1 - p_C^t) \cdot (-c) \\ = \;& p_C^t \cdot (k \cdot a + \varepsilon_{t+k} \cdot z + (p_C^{t+k} - \varepsilon_{t+k}) \cdot b) + (1 - p_C^t) \cdot (-c) \\ = \;& p_C^t \cdot k \cdot a + \varepsilon_t \cdot z + (p_C^t \cdot p_C^{t+k} - \varepsilon_t) \cdot b + (1 - p_C^t) \cdot (-c) \end{aligned}$$

By Theorem 3.1, we have

$$p_C^t = \frac{c}{k \cdot a - (1 - p_C^{t+k}) \cdot b + c} \Leftrightarrow p_C^t \cdot k \cdot a + (p_C^t \cdot p_C^{t+k} - p_C^t) \cdot b + (1 - p_C^t) \cdot (-c) = 0$$

Therefore,

$$\begin{aligned} u_i(\text{Cooperate}, LA, h_C^{t-1}) \geq\ & p_C^t \cdot k \cdot a + \varepsilon_t \cdot z + (p_C^t \cdot p_C^{t+k} - \varepsilon_t) \cdot b + (1 - p_C^t) \cdot (-c) \\ =\ & \Big(p_C^t \cdot k \cdot a + (p_C^t \cdot p_C^{t+k} - p_C^t) \cdot b + (1 - p_C^t) \cdot (-c)\Big) + \big(\varepsilon_t \cdot z + (p_C^t - \varepsilon_t) \cdot b\big) \\ =\ & u_i(\text{Defect}, LA, h_C^{t-1}) \end{aligned}$$ □

3.2.4 Verifying other Histories

Finally, we verify that, if a deviation has taken place, players indeed wish to defect forever. Given $h \notin \{h_C^0, \cdots, h_C^{t-1}\}$, by Assumption 4, $\mu(\theta_j = S \mid h) = 1$ for some $j \in \{1, 2\}$. That means that a player that did *not* deviate will believe (with probability 1) that the other player is of type S. Defection is clearly a best response for this player that did not deviate, no matter which type she is of, since she believes that she is facing utility function u^S with probability 1 (and that the other player will defect forever regardless). Similarly, if the deviating player is of type S, it is always a best response for him to keep playing Defect. Therefore, all that remains to check is that if a player of type LA deviated, it is also a best response for that player to keep playing Defect for the remaining of the game. This is less straightforward, because such a player may still believe there is some probability ε that the other player is of type LA as well; and, in fact, it is not unconditionally true. The next lemma precisely identifies when it is true.

Lemma 3.6. *Given a feasible sequence $p_C(T)$, the following two statements are equivalent:*

- *For any possible history $h \notin \{h_C^0, \cdots, h_C^{t-1}\}$ in which i has deviated, $u_i(\text{Defect}, LA, h) \geq u_i(\text{Cooperate}, LA, h)$.*
- *(x_D, x_D) forms a Nash equilibrium in u^{LA} (when it is common knowledge that both players are of type LA) or $p_C^T \leq \frac{c}{c+y}$.*

Proof. Given a history h in which i has deviated, let $\varepsilon = \mu(\theta_{-i} = LA \mid h)$. We need to make sure that for any such ε the following holds (noting that the other player's play in future rounds will be unaffected anyway, so we can just focus on play in a single round):

$$\begin{aligned} &\varepsilon \cdot u_i^{LA}(\text{Defect}, \text{Defect}) + (1 - \varepsilon) \cdot u_i^S(\text{Defect}, \text{Defect}) \\ \geq\ &\varepsilon \cdot u_i^{LA}(\text{Cooperate}, \text{Defect}) + (1 - \varepsilon) \cdot u_i^S(\text{Cooperate}, \text{Defect}) \end{aligned}$$

That is, we need to make sure that $\varepsilon \cdot y + (1 - \varepsilon) \cdot (-c) \leq 0$. If (x_D, x_D) forms a Nash equilibrium in u^{LA}, (i.e., $y \leq 0$), the inequality holds for any ε. On the other hand, if $y > 0$, since $\varepsilon_T \geq \varepsilon_t$ for all $t \leq T$, we have $\varepsilon_T \cdot y + (1 - \varepsilon_T) \cdot (-c) \geq \varepsilon_t \cdot y + (1 - \varepsilon_t) \cdot (-c)$. Note that ε_T is the belief that would result from the LA player deviating in the round right before the last one, so this is one of the values of ε for which we need to ensure the inequality which represents that it is better to continue defecting. By the previous inequality, this also suffices to guarantee it for other values of ε (assuming $y > 0$). But in fact, $\varepsilon_T \cdot y + (1 - \varepsilon_T) \cdot (-c) \leq 0 \Leftrightarrow p_C^T \leq \frac{c}{c+y}$. □

3.2.5 Characterization

We now have the tools to characterize the conditions under which our construction will indeed induce a PBE. Recall that our construction focuses on establishing an equilibrium in which LA types never deviate first and both players will defect forever after deviation (by Assumption 4); therefore, a candidate equilibrium is described by the sequence $p_C(T)$. Specifically, combining Theorem 3.1 and Lemma 3.5, 3.6, we can conclude that

Theorem 3.2. *Under Assumption 4, a sequence $p_C(T)$ induces a PBE if and only if*

- *If* $p_C^t = 1$ *and* $p_C^{t+1} < 1$*, then* $p_C^{t+1} \geq 1 - \frac{a}{b}$*.*
- *If* $p_C^t < 1$*,* $p_C^{t+k} < 1$*, and for all* $t < t' < t+k$*,* $p_C^{t'} = 1$*, then* $p_C^t = \frac{c}{k \cdot e - (1-p_C^{t+k}) \cdot b + c}$*.*
- $p_C^T \geq \frac{c}{e+c-z}$*.*
- (x_D, x_D) *forms a Nash equilibrium or* $p_C^T \leq \frac{c}{c+y}$*.*

3.3 Desired LA Types for Prisoner's Dilemma

Theorem 3.2 allows us to assess which LA types are desired for the class of prisoner's dilemma games. LA_{avg} works for *some* values of a, b, and c in the prisoner's dilemma, but not all, as the next example illustrates.

Example 1. *Consider the prisoner's dilemma game in Table 2, where* $b = 3a + \delta$ *and* $c = a + 3\delta$*, for small* $\delta > 0$*.*

	Cooperate	Defect
Cooperate	a, a	$-a - 3\delta, 3a + \delta$
Defect	$3a + \delta, -a - 3\delta$	0,0

Table 2: An example of a prisoner's dilemma where LA_{avg} fails to induce cooperation.

For LA_{avg}*, by Theorem 3.2, we require* $p_C^T \geq \frac{c}{e+c-z} = \frac{a+3\delta}{a+4\delta}$*. Therefore, as* $\delta \to 0$*, we need* $p_C^T \to 1$*. Let* $k > 0$ *be the minimum integer such that* $\varepsilon_{T-k} < 1$*. By Lemma 3.2, we have* $p_C^{T-k} = \frac{c}{k \cdot a - (1-p_C^T) \cdot b + c}$*, which, as* $\delta \to 0$*, approaches a value no greater than* $\frac{c}{a+c}$*, which is strictly less than* $1 - \frac{a}{b}$ *(because by construction we have* $c < b - a$*). By Corollary 3.2, it follows that for sufficiently small* δ*, there exists no feasible* $p_C(T)$*.*

In contrast, it turns out that LA_{min} is in fact desired for all prisoner's dilemma games. To prove this, we combine Theorem 3.2 with the construction of feasible sequences from Lemma 3.4.

Theorem 3.3. LA_{min} *is desired for prisoner's dilemma games.*

Proof. Consider $p_C(T) = (1, \cdots, 1, p_C^{t^*}, \cdots, p_C^T)$ in which for all $t^* \leq t < T$, $p_C^t = \frac{c}{a-(1-p_C^{t+1}) \cdot b + c}$ and $p_C^T = \gamma^*$. By Theorem 3.2 and Lemma 3.4, all that remains to show is that (1) $p_C^T \geq \frac{c}{e+c-z}$ and (2) $y \leq 0$ or $p_C^T \leq \frac{c}{c+y}$. (2) follows from the fact that for LA_{min}, we have $e = a$ and $y = z = -c < 0$. As for (1), using Proposition 3.1 and (again) the fact that $z = -c$, we have $\frac{c}{e+c-z} = \frac{c}{a+2c} \leq \frac{c}{a+c} \leq \gamma^* = p_C^T$. □

4 Independent-Effect Games

We now consider another class of games that is incomparable to the class of prisoner's dilemma games, namely the class of *independent-effect* games. The idea here is that the effect of one player's actions on the utilities of all players is independent of what any other player plays.

Definition 4.1 (Independent-Effect Games)**.** *A symmetric two-player game* G *is an* independent-effect game *if there exist functions* $g, h : X \to \mathbb{R}$ *such that* $u_i^S(x_i, x_{-i}) = g(x_i) + h(x_{-i})$*.*

In such a game, a player's utility is the sum of the utility generated by his own action and the utility generated by his opponent's action. This immediately implies that there exists a dominant strategy for each player $x_D = \max_x g(x)$ (when the game is played only once).

The example prisoner's dilemma game in the introduction is an independent-effect game (but not all prisoner's dilemma games are). The *tragedy of the commons* Hardin (2009) models a shared-resource system where, if each player plays solely according to his own self-interest, the result is contrary to the common good of all users, because they overuse the resource with respect to the socially optimal use of it. One possible formulation of the utility function in the tragedy of the commons is $u_i(x_i, x_{-i}) = U(x_i) - (x_i + x_{-i})$ where x_i is the quantity of resource that player i uses. In this game, the dominant strategy for each player is $x_D = \arg\max_x U(x) - x$. However, if they were to play cooperatively to maximize social welfare, they would play a strategy in $\arg\max_x U(x) - 2x$.

In general independent-effect games, the cooperative strategy that maximizes social welfare is $x_C \in \arg\max_x g(x) + h(x)$. Since x_D is dominant, the optimal action to deviate to for a player of type S is always x_D, no matter what the current belief is.[1] Therefore, from the perspective of the selfish player, the only actions in the game that are relevant are $\{x_C, x_D\}$, resulting in a prisoner's dilemma game. Thus, Theorem 3.1 holds in independent-effect games if we let $a = u_i^S(x_C, x_C)$, $b = u_i^S(x_D, x_C)$, $-c = u_i^S(x_C, x_D)$ and assume without loss of generality that $u_i^S(x_D, x_D) = 0$. We next show that the LA types have no incentive to deviate from cooperation—i.e., they are best off playing x_C as long as the other player has done so.

Consider $0 \le \varepsilon' \le \varepsilon'' \le 1$, where ε' represents the belief that the LA player has that the other player is also an LA type (and will therefore cooperate), and $\varepsilon'' - \varepsilon'$ is the belief that the LA player has that the other player is an S type but will nevertheless cooperate in this round—so that ε'' is the total probability that the other player will cooperate in this round. Then, let

$$x^{LA}(\varepsilon', \varepsilon'') \in \arg\max_x \varepsilon' \cdot u_i^{LA}(x, x_C) + (\varepsilon'' - \varepsilon') u_i^S(x, x_C) + (1 - \varepsilon'') u_i^S(x, x_D)$$

That is, $x^{LA}(\varepsilon', \varepsilon'')$ is an action that maximizes an LA type's one-round utility.

In the last round, a player of type S will definitely play x_D. As for a player of type LA, we require he weakly prefers to play x_C. More precisely

$$\begin{aligned} & p_C^T \cdot u_i^{LA}(x_C, x_C) + (1 - p_C^T) u_i^S(x_C, x_D) \\ = \; & p_C^T \cdot u_i^{LA}(x^{LA}(p_C^T, p_C^T), x_C) + (1 - p_C^T) u_i^S(x^{LA}(p_C^T, p_C^T), x_D) \end{aligned} \tag{3}$$

Lemma 4.1. *If $p_C(T)$ is feasible and p_C^T satisfies (3), then for all $1 \le t \le T$ and $x' \ne x_C$, $u_i(LA, h_C^{t-1}) = u_i(x_C, LA, h_C^{t-1}) \ge u_i(x', LA, h_C^{t-1})$.*

The proof is in Appendix. In independent-effect games, we have $c = b - a$ and by Lemma 3.3, we can generate a feasible sequence $p_C(T)$ satisfying the requirement that $p_C^T \ge 1 - \frac{a}{b}$. By Lemma 4.1, given a feasible $p_C(T)$, LA types are willing to play x_C as long as the other player has not deviated if p_C^T satisfies (3). We also have that, if (x_C, x_C) forms a strict Nash equilibrium of u^{LA} (i.e., in the case where it is common knowledge that both players have type LA), and moreover the game is finite, then there exists a sufficiently large $p_C^T < 1$ such that (3) is satisfied. Moreover, by Lemma 3.6, if (x_D, x_D) forms a Nash equilibrium of u^{LA}, then both players will have incentive to keep playing x_D once any player has deviated. Combining these observations, we obtain:

Theorem 4.1. *An LA type is desired for finite independent-effect games if we have that (x_C, x_C) and (x_D, x_D) form Nash equilibria of u^{LA} (and the former is a strict equilibrium).*

[1] In repeated games, this is assuming that deviation results in the same subsequent play, no matter what a player deviated to. This will indeed be the case in our setup due to Assumption 4, which we will still maintain here.

We identify two universal LA types that are desired for independent-effect games. The first one is LA_{min}. Since (x_C, x_C) uniquely maximizes the social welfare, we have

$$\forall x_i \neq x_C, 2 \cdot \min(u_i^S(x_i, x_C), u_{-i}^S(x_i, x_C)) \leq u_i^S(x_i, x_C) + u_{-i}^S(x_i, x_C) < 2 \cdot u_i^S(x_C, x_C)$$

Moreover, since x_D is a dominant strategy, we have

$$\forall x_i \neq x_D, \min(u_i^S(x_i, x_D), u_{-i}^S(x_i, x_D)) \leq u_i^S(x_i, x_D) \leq u_i^S(x_D, x_D)$$

Therefore, for LA_{min}, strategy profile (x_C, x_C) and (x_D, x_D) form Nash equilibria of u^{LA}, and the former is strict.

The second type that we identify is $LA_{\text{avg, positive}}$. $LA_{\text{avg, positive}}$ is a type that cares about the average of two players' utilities, but only if his own utility is larger than 0: $f_{\text{avg, positive}}(a, b) = a$ for all $a \leq 0, b \in \mathbb{R}$ and $f_{\text{avg, positive}}(a, b) = (a + b)/2$ for all $a > 0, b \in \mathbb{R}$. Similar to LA_{min}, it can be verified that for $LA_{\text{avg, positive}}$, strategy profile (x_C, x_C) and (x_D, x_D) form Nash equilibria in u^{LA}.

Corollary 4.1. *LA_{min} and $LA_{\text{avg, positive}}$ are desired for finite independent-effect games.*

5 General Games

We now proceed to general symmetric two-player games (that still satisfy the assumptions from Section 2.5, i.e., the game has a dominant strategy which produces utility 0 when both players use it, and a unique social-welfare maximizing outcome where both players play the same). Unfortunately, with the restriction to universal LA types—i.e., LA types that only consider the payoffs to both players to determine the utility—we run into an impossibility.

Theorem 5.1. *Under Assumption 4, no universal LA type is desired.*

Proof. Consider the modified prisoner's dilemma game in Table 3, with sufficiently small $\delta > 0$.

	C	M	D
C	a, a	$a - \delta/m, a - \delta/m$	$-c, b$
M	$a - \delta/m, a - \delta/m$	$-\infty$,$-\infty$	$-c + \delta$, $-c + \delta$
D	b, $-c$	$-c + \delta$, $-c + \delta$	0, 0

Table 3: An example of a game where any universal LA type fails to induce cooperation.

Assume there exists a desired LA_ϕ. By Assumption 2 and 3, we can assume $u_i^{LA_\phi}(v, v) = \alpha \cdot v$ for all $v > 0$, for some $\alpha \geq 1$. In the last round, in order to ensure a player of type LA_ϕ plays C rather than M, we need $\varepsilon_T \cdot u_i^{LA_\phi}(C, C) + (1 - \varepsilon_T) \cdot u_i^S(C, D) \geq \varepsilon_T \cdot u_i^{LA_\phi}(M, C) + (1 - \varepsilon_T) \cdot u_i^S(M, D)$ That is

$$\varepsilon_T \cdot \alpha \cdot a + (1 - \varepsilon_T) \cdot (-c) \geq \varepsilon_T \cdot \alpha \cdot (a - \delta/m) + (1 - \varepsilon_T) \cdot (-c + \delta) \Leftrightarrow \frac{\varepsilon_T}{1 - \varepsilon_T} \geq \frac{m}{\alpha}$$

As $m \to \infty$, we have $p_C^T = \varepsilon_T \to 1$. Consider the case when $c < b - a$. Let $k > 0$ be the minimum integer such that $\varepsilon_{T-k} < 1$. By Lemma 3.2, we have $p_C^{T-k} = \frac{c}{k \cdot a - (1 - p_C^T) \cdot b + c}$ which, as $\delta \to 0$, approaches a value no greater than $\frac{c}{a+c}$, which is strictly less than $1 - \frac{a}{b}$ by the fact that $c < b - a$. By Corollary 3.2, it follows that for sufficiently small δ, there exists no feasible $p_C(T)$. □

Is there a way around this negative result? In the proof of Theorem 5.1, we create a game (Table 3) such that for any universal LA type, to incentivize that type to play $x_C = C$ in the last round, ε_T must be arbitrarily close to 1. To circumvent this problem, we consider a slight expansion of the set of universal LA types: we allow the function f_ϕ to also take as input whether both players play the same action. Call such types *universal*$^+$. As it turns out, there is in fact a universal$^+$ type that is desired for finite general games. The intuition is that we can design such a type in a way that we do not get $\varepsilon_T \to 1$ when $m \to \infty$.

Definition 5.1 (Universal$^+$ LA type)**.** *For $x_1, x_2 \in X$, let $eq(x_1, x_2) = 1$ if $x_1 = x_2$ and $eq(x_1, x_2) = 0$ otherwise. An LA type LA_ϕ is* universal$^+$ *if there exists a function f_ϕ such that for every possible u_1^S and u_2^S, and all $x_1, x_2 \in X$, $u_i^{LA}(x_1, x_2) = f_\phi(u_i^S(x_1, x_2), u_{-i}^S(x_1, x_2), eq(x_1, x_2))$.*

We need to slightly update Assumptions 1, 2, and 3 because these refer to a function f_ϕ with two arguments, whereas now we have one with three arguments. We require Assumptions 1 and 2 hold for any fixed value of the third argument, and Assumption 3 to hold when the third argument takes value 1.

For the selfish type, the situation is similar as in the case of independent-effect games: only x_C and x_D are relevant to the selfish type. Thus, Theorem 3.1 holds in general games if we let $a = u_i^S(x_C, x_C)$, $b = u_i^S(x_D, x_C)$, $-c = u_i^S(x_C, x_D)$ and assume without loss of generality that $u_i^S(x_D, x_D) = 0$.

The analysis for the LA type, however, needs to be generalized. In the last round of the game, we require LA types to play x_C, so we need:

$$\forall x' \neq x_C, p_C^T \cdot u_i^{LA}(x_C, x_C) + (1 - p_C^T) \cdot u_i^S(x_C, x_D) \geq p_C^T \cdot u_i^{LA}(x', x_C) + (1 - p_C^T) \cdot u_i^S(x', x_D) \quad (4)$$

Recall that by Lemma 3.4, if $\varepsilon_T = \gamma^*$, where γ^* is the positive root of the quadratic equation $f(\gamma) = b \cdot \gamma^2 + (a - b + c) \cdot \gamma - c = 0$, then we can construct a feasible sequence $p_C(T) = (1, \cdots, 1, p_C^{t^*}, \cdots, p_C^T)$ with $p_C^{t'} = \gamma^*$ for all $t^* \leq t' \leq T$. The lemma below establishes that with such a sequence $p_C(T)$, LA types are best off playing x_C as long as nobody has deviated, if $p_C^T = \gamma^*$ satisfies (4) and cooperation is a strict equilibrium when it is common knowledge that both agents are of the LA type.

Lemma 5.1. *If $p_C(T) = (1, \cdots, 1, p_C^{t^*}, \cdots, p_C^T)$ with $p_C^{t'} = \gamma^*$ for all $t^* \leq t' \leq T$, γ^* satisfies (4) and (x_C, x_C) forms a strict Nash equilibrium of u^{LA}, then for all $1 \leq t \leq T$ and $x' \neq x_C$, $u_i(LA, h_C^{t-1}) = u_i(x_C, LA, h_C^{t-1}) \geq u_i(x', LA, h_C^{t-1})$.*

The proof is in Appendix. By Lemma 3.6, if (x_D, x_D) forms a Nash equilibrium, it is a best response for both players to keep playing x_D once any player has deviated. This allows us to conclude:

Theorem 5.2. *An LA type is desired for finite general games if we have that $\gamma^*(G)$ satisfies (4) and both (x_C, x_C) and (x_D, x_D) form Nash equilibria in u^{LA} (and the former is a strict equilibrium).*

We identify three universal$^+$ LA types that are desired: $LA_{\text{coordinate}}$, $LA_{\text{min,coordinate}}$, and $LA_{\text{avg,positive,coordinate}}$. $LA_{\text{coordinate}}$ is a type that only cares about whether two players play the same: $f_{\text{coordinate}}(v, v, 1) = v$ for all $v \in \mathbb{R}$ and $f_{\text{coordinate}}(v, w, 0) = 0$ for all $v, w \in \mathbb{R}$. $LA_{\text{min,coordinate}}$ is a type that is exactly like LA_{min} if players do not play the same ($f_{\text{min,coordinate}}(v, w, 0) = \min(v, w)$) but otherwise gets double the utility ($f_{\text{min,coordinate}}(v, v, 1) = 2v$). Similarly, $LA_{\text{avg, positive, coordinate}}$ is a type that is exactly like $LA_{\text{avg, positive}}$ if players do not play the same, but otherwise gets double the utility.

For $LA_{\text{coordinate}}$, both (x_C, x_C) and (x_D, x_D) form Nash equilibria (where the former is strict) in u^{LA} since for all $x' \neq x_C$, $u_i^{LA}(x', x_C) = 0 < u_i^{LA}(x_C, x_C)$ and for all $x' \neq x_D$, $u_i^{LA}(x', x_D) = 0 = u_i^{LA}(x_D, x_D)$. To satisfy (4), we need $p_C^T \cdot a + (1 - p_C^T) \cdot (-c) \geq p_C^T \cdot 0 + (1 - p_C^T) \cdot 0$, which is equivalent to, $p_C^T \geq \frac{c}{a+c}$. In fact, by Proposition 3.1, we have $p_C^T = \gamma^* \geq \frac{c}{a+c}$.

For $LA_{\text{min,coordinate}}$, it can be verified that (x_D, x_D) forms a Nash equilibrium in u^{LA}. To ensure that $\gamma^* = p_C^T$ satisfies (4), note that $p_C^T \cdot u_i^{LA}(x_C, x_C) + (1 - p_C^T) \cdot u_i^S(x_C, x_D) = p_C^T \cdot 2a + (1 - p_C^T) \cdot (-c)$.

Moreover, for all $x_i' \neq x_C$, since (x_C, x_C) maximizes social welfare, we have

$$u_i^{LA}(x_i', x_C) = \min(u_i^S(x_i', x_C), u_{-i}^S(x_i', x_C)) \leq \frac{1}{2}\left(u_i^S(x_i', x_C) + u_{-i}^S(x_i', x_C)\right) \leq u_i^S(x_C, x_C)$$

Henceforth,

$$\max_{x' \neq x_C} p_C^T \cdot u_i^{LA}(x', x_C) + (1 - p_C^T) \cdot u_i^S(x', x_D) \ \leq p_C^T \cdot a + (1 - p_C^T) \cdot u_i^S(x_D, x_D) \leq p_C^T \cdot a$$

Therefore, we only need $p_C^T \cdot 2a + (1 - p_C^T) \cdot (-c) \geq p_C^T \cdot a$, which is equivalent to, $p_C^T = \gamma^* \geq \frac{c}{a+c}$. Again, by Proposition 3.1, we have $p_C^T = \gamma^* \geq \frac{c}{a+c}$. The case of $LA_{\text{avg,positive,coordinate}}$ is entirely similar to that of $LA_{\text{min,coordinate}}$. We conclude:

Corollary 5.1. *$LA_{\textit{coordinate}}$, $LA_{\textit{min,coordinate}}$, and $LA_{\textit{avg, positive, coordinate}}$ are desired for general games.*

6 Discussion

We have shown that under certain conditions, LA types such as LA_{min} that care directly about the other player's payoff, *if* that player is also an LA type, can be very successful in establishing cooperation in finitely repeated games. We have provided characterizations of the LA types that can achieve this in various classes of games. We believe that this provides useful guidance for the design of agents that are to be introduced into a population in small numbers, for the purpose of establishing an equilibrium where all players almost always cooperate. Unfortunately, we have also shown that in sufficiently general games, such types do not suffice, although allowing them to care about whether both players play the same action circumvents this problem.

All of our results aim to establish that there *exists* a desirable equilibrium, but the bad equilibrium, in which both players play the dominant strategy x_D in every round, still exists in each case. This is a concern if we are not confident that we can establish the desirable equilibrium. Kreps et al. (1982) and Maskin and Fudenberg (1986) demonstrate that by introducing a small fraction of "behavioral" types, the bad equilibrium can be eliminated. It would be valuable to investigate whether we can eliminate the bad equilibrium even with a small fraction of "strategic" types.

More broadly, we may ask whether we can obtain similar results in other classes of games. Stochastic games with a finite number of rounds would constitute a natural next step. We may also ask whether we can somehow unify these insights with results in the routing games literature that sound similar at a high level (though they appear technically quite different). In the limit, we would like to obtain very general insights about how to design agent preferences. Can we do so in a way that even when they are introduced into complex, messy, ambiguous environments, they will establish a more desirable equilibrium? This seems to us a compelling challenge problem for those interested in the design of beneficial game-theoretic AI, and it may yet provide some insights about human societies as well.

References

James Andreoni and John H Miller. 1993. Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The economic journal* 103, 418 (1993), 570–585.

Robert Axelrod and others. 1987. The evolution of strategies in the iterated prisoners dilemma. *The dynamics of norms* (1987), 1–16.

Po-An Chen, Bart De Keijzer, David Kempe, and Guido Schäfer. 2014. Altruism and its impact on the price of anarchy. *ACM Transactions on Economics and Computation* 2, 4 (2014), 17.

Russell Cooper, Douglas V DeJong, Robert Forsythe, and Thomas W Ross. 1996. Cooperation without reputation: experimental evidence from prisoner's dilemma games. *Games and Economic Behavior* 12, 2 (1996), 187–218.

Drew Fudenberg and Jean Tirole. 1991. Perfect Bayesian equilibrium and sequential equilibrium. *journal of Economic Theory* 53, 2 (1991), 236–260.

Garrett Hardin. 2009. The tragedy of the commons. *Journal of Natural Resources Policy Research* 1, 3 (2009), 243–253.

David M Kreps, Paul Milgrom, John Roberts, and Robert Wilson. 1982. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic theory* 27, 2 (1982), 245–252.

David M Kreps and Robert Wilson. 1982. Sequential equilibria. *Econometrica: Journal of the Econometric Society* (1982), 863–894.

Eric Maskin and Drew Fudenberg. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 53, 3 (1986).

Tim Roughgarden. 2004. Stackelberg scheduling strategies. *SIAM J. Comput.* 33, 2 (2004), 332–350.

Tim Roughgarden and Éva Tardos. 2002. How bad is selfish routing? *Journal of the ACM (JACM)* 49, 2 (2002), 236–259.

Appendix

A Omitted Proofs

A.1 Proof of Proposition 3.1

Proof. Note that $f(\cdot)$ is a quadratic function with a positive quadratic coefficient and we have $f(0) = -c < 0$ and $f(\gamma^*) = 0$. Therefore, for $0 < v < 1$, $v \leq \gamma^*$ if and only if $f(v) < 0$. By calculation, we have

$$f(1-\frac{a}{b}) = b \cdot (\frac{b-a}{b})^2 - (a-b+c) \cdot \frac{b-a}{b} - c = -\frac{ac}{b} < 0$$

and

$$f(\frac{c}{a+c}) = \frac{bc^2}{(a+c)^2} + (a-b+c)\frac{c}{a+c} - c = \frac{bc}{a+c} \cdot (\frac{c}{a+c} - 1) \leq 0$$

□

A.2 Proof of Lemma 4.1

Proof. We prove this by induction. Suppose

$$u_i(LA, h_C^{t'-1}) \ = u_i(x_C, LA, h_C^{t'-1}) \geq \max_{x' \neq x_C} u_i(x', LA, h_C^{t'-1})$$

for $t' > t$. The base case is true since p_C^T satisfies (3). As for the t-th round, first note that if a player of type LA plays $x' \neq x_C$, his utility is

$$\begin{aligned} u_i(x', LA, h_C^{t-1}) =& \ \varepsilon_t \cdot u_i^{LA}(x', x_C) + (p_C^t - \varepsilon_t) \cdot u_i^S(x', x_C) \\ &+ (1 - p_C^t) \cdot u_i^S(x', x_D) + (T - t) \cdot u_i^S(x_D, x_D) \\ =& \ \varepsilon_t \cdot u_i^{LA}(x', x_C) + (p_C^t - \varepsilon_t) \cdot u_i^S(x', x_C) + (1 - p_C^t) \cdot u_i^S(x', x_D) \\ =& \ \varepsilon_t \cdot u_i^{LA}(x', x_C) + (1 - \varepsilon_t) \cdot g(x') + (p_C^t - \varepsilon_t) \cdot h(x_C) + (1 - p_C^t) \cdot h(x_D) \end{aligned}$$

The last equality is because the game is an independent-effect game. In such a game, we have $u_i^S(x_D, x_D) = g(x_D) + h(x_D) = 0$, $a = u_i^S(x_C, x_C) = g(x_C) + h(x_C)$ and $b = u_i^S(x_D, x_C) = g(x_D) + h(x_C) = h(x_C) - h(x_D)$. By Corollary 3.2, we have that for all t,

$$p_C^t \geq 1 - \frac{a}{b} \Leftrightarrow p_C^t \geq \frac{g(x_D) - g(x_C)}{g(x_D) + h(x_C)} = \frac{g(x_D) - g(x_C)}{h(x_C) - h(x_D)}$$

For the remainder of the proof, we consider two cases.

Case (1): If $p_C^t = 1$, then we have $\varepsilon_t = \varepsilon_{t+1}$ and

$$\begin{aligned}
u_i(x_C, LA, h_C^{t-1}) =\ & \varepsilon_t \cdot u_i^{LA}(x_C, x_C) + (1-\varepsilon_t) \cdot u_i^S(x_C, x_C) + u_i(LA, h_C^t) \\
\geq\ & a + u_i(LA, h_C^t) \\
\geq\ & a + u_i(x^{LA}(\varepsilon_{t+1}, p_C^{t+1}), LA, h_C^t) \\
\geq\ & a + u_i(x^{LA}(\varepsilon_t, p_C^t), LA, h_C^t) \\
\geq\ & a + \varepsilon_{t+1} \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (p_C^{t+1} - \varepsilon_{t+1}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C) \\
& + (1 - p_C^{t+1}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) \\
\geq\ & a + \varepsilon_t \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (p_C^{t+1} - \varepsilon_t) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C) \\
& + (1 - p_C^{t+1}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) \\
=\ & a + \varepsilon_t \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (1 - \varepsilon_t) \cdot g(x^{LA}(\varepsilon_t, p_C^t)) \\
& + (p_C^{t+1} - \varepsilon_t) \cdot h(x_C) + (1 - p_C^{t+1}) \cdot h(x_D)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
u_i(x_C, LA, h_C^{t-1}) - u_i(x', LA, h_C^{t-1}) & \geq u_i(x_C, LA, h_C^{t-1}) - u_i(x^{LA}(\varepsilon_t, p_C^t), LA, h_C^{t-1}) \\
& \geq a + h(x_D) - h(x_C) + p_C^{t+1} \cdot (h(x_C) - h(x_D)) \\
& \geq a + h(x_D) - h(x_C) + g(x_D) - g(x_C) \\
& = 0
\end{aligned}$$

Case (2): If $p_C^t < 1$, let k be such that $p_C^{t+k} < 1$ and $p_C^{t'} = 1$ for all $t < t' < t + k$. Similarly to the above, we have

$$\begin{aligned}
u_i(x_C, LA, h_C^{t-1}) =\ & \varepsilon_t \cdot k \cdot u_i^{LA}(x_C, x_C) + (p_C^t - \varepsilon_t) \cdot k \cdot u_i^S(x_C, x_C) + p_C^t \cdot u_i(LA, h_C^{t+k-1}) \\
& + (1 - p_C^t) \cdot (u_i^S(x_C, x_D) + (T - t) \cdot u_i^S(x_D, x_D)) \\
\geq\ & p_C^t \cdot (k \cdot a + u_i(LA, h_C^{t+k-1})) + (1 - p_C^t) \cdot ((-c) + u_i^S(x_D, x_D)) \\
\geq\ & p_C^t \cdot (k \cdot a + u_i(x^{LA}(\varepsilon_{t+k}, p_C^{t+k}), LA, h_C^{t+k-1})) + (1 - p_C^t) \cdot ((-c) + u_i^S(x_D, x_D)) \\
\geq\ & p_C^t \cdot (k \cdot a + u_i(x^{LA}(\varepsilon_t, p_C^t), LA, h_C^{t+k-1})) + (1 - p_C^t) \cdot ((-c) + u_i^S(x_D, x_D)) \\
\geq\ & p_C^t \cdot k \cdot a + p_C^t \cdot \varepsilon_{t+k} \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + p_C^t \cdot (p_C^{t+k} - \varepsilon_{t+k}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C) \\
& + p_C^t \cdot (1 - p_C^{t+k}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) + (1 - p_C^t) \cdot ((-c) + u_i^S(x_D, x_D)) \\
=\ & p_C^t \cdot k \cdot a + \varepsilon_t \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (p_C^t \cdot p_C^{t+k} - \varepsilon_t) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C) \\
& + p_C^t \cdot (1 - p_C^{t+k}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) + (1 - p_C^t) \cdot ((-c) + u_i^S(x_D, x_D)) \\
=\ & \varepsilon_t \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (p_C^t \cdot p_C^{t+k} - \varepsilon_t) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C) \\
& + p_C^t \cdot (1 - p_C^{t+k}) \cdot (u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) + u_i^S(x_D, x_C)) + (1 - p_C^t) \cdot u_i^S(x_D, x_D)
\end{aligned}$$

To obtain the inequalities above, we make use of the definition of x^{LA} and Theorem 3.1:

$$p_C^t = \frac{c}{k \cdot a - (1 - p_C^{t+k}) \cdot b + c} \Leftrightarrow p_C^t \cdot k \cdot a + (p_C^t \cdot p_C^{t+k} - p_C^t) \cdot b + (1 - p_C^t) \cdot (-c) = 0$$

Moreover, by $u_i^S(x_D, x_D) \geq u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D)$ and the fact that we have an independent-effect game, we have

$$
\begin{aligned}
u_i(x_C, LA, h_C^{t-1}) \geq\ & \varepsilon_t \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (p_C^t \cdot p_C^{t+k} - \varepsilon_t) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C) \\
& + p_C^t \cdot (1 - p_C^{t+k}) \cdot (u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) + u_i^S(x_D, x_C)) \\
& + (1 - p_C^t) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) \\
=\ & \varepsilon_t \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (1 - \varepsilon_t) \cdot g(x^{LA}(\varepsilon_t, p_C^t)) \\
& + (p_C^t - \varepsilon_t) \cdot h(x_C) + (1 - p_C^t) \cdot h(x_D) \\
=\ & u_i(x^{LA}(\varepsilon_t, p_C^t), LA, h_C^{t-1})
\end{aligned}
$$

Therefore, we have

$$
u_i(x_C, LA, h_C^{t-1}) - u_i(x', LA, h_C^{t-1}) \geq u_i(x_C, LA, h_C^{t-1}) - u_i(x^{LA}(\varepsilon_t, p_C^t), LA, h_C^{t-1}) \geq 0
$$

□

A.3 Proof of Lemma 5.1

Proof. We prove this by induction. Suppose

$$
u_i(LA, h_C^{t'-1}) \ = u_i(x_C, LA, h_C^{t'-1}) \geq \max_{x' \neq x_C} u_i(x', LA, h_C^{t'-1})
$$

for $t' > t$. The base case is true since $p_C^T = \gamma^*$ satisfies (4). As for the t-th round, first note that if a player of type LA plays $x' \neq x_C$, his utility is

$$
\begin{aligned}
u_i(x', LA, h_C^{t-1}) =\ & \varepsilon_t \cdot u_i^{LA}(x', x_C) + (p_C^t - \varepsilon_t) \cdot u_i^S(x', x_C) \\
& + (1 - p_C^t) \cdot u_i^S(x', x_D) + (T - t) \cdot u_i^S(x_D, x_D) \\
=\ & \varepsilon_t \cdot u_i^{LA}(x', x_C) + (p_C^t - \varepsilon_t) \cdot u_i^S(x', x_C) + (1 - p_C^t) \cdot u_i^S(x', x_D)
\end{aligned}
$$

Case (1): If $p_C^t = p_C^{t+1} = \gamma^*$,

$$
\begin{aligned}
u_i(x_C, LA, h_C^{t-1}) =\ & \varepsilon_t \cdot u_i^{LA}(x_C, x_C) + (p_C^t - \varepsilon_t) \cdot u_i^S(x_C, x_C) + p_C^t \cdot u_i(LA, h_C^t) \\
& + (1 - p_C^t) \cdot (u_i^S(x_C, x_D) + (T - t) \cdot u_i^S(x_D, x_D)) \\
\geq\ & p_C^t \cdot (a + u_i(LA, h_C^t)) + (1 - p_C^t) \cdot (-c) \\
\geq\ & p_C^t \cdot (a + u_i(x^{LA}(\varepsilon_{t+1}, p_C^{t+1}), LA, h_C^t)) + (1 - p_C^t) \cdot (-c) \\
\geq\ & p_C^t \cdot (a + u_i(x^{LA}(\varepsilon_t, p_C^t), LA, h_C^t)) + (1 - p_C^t) \cdot (-c) \\
\geq\ & p_C^t \cdot a + p_C^t \cdot \varepsilon_{t+1} \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + p_C^t \cdot (p_C^{t+1} - \varepsilon_{t+1}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C) \\
& + p_C^t \cdot (1 - p_C^{t+1}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) + (1 - p_C^t) \cdot (-c) \\
=\ & p_C^t \cdot a + \varepsilon_t \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (p_C^t \cdot p_C^{t+1} - \varepsilon_t) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C) \\
& + p_C^t \cdot (1 - p_C^{t+1}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) + (1 - p_C^t) \cdot (-c) \\
=\ & \varepsilon_t \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (p_C^t \cdot p_C^{t+1} - \varepsilon_t) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C) \\
& + p_C^t \cdot (1 - p_C^{t+1}) \cdot (u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) + u_i^S(x_D, x_C))
\end{aligned}
$$

To obtain the inequalities above, we make use of the definition of x^{LA} and Theorem 3.1:

$$
p_C^t = \frac{c}{k \cdot a - (1 - p_C^{t+k}) \cdot b + c} \Leftrightarrow p_C^t \cdot k \cdot a + (p_C^t \cdot p_C^{t+k} - p_C^t) \cdot b + (1 - p_C^t) \cdot (-c) = 0
$$

Therefore,

$$\begin{aligned}
& u_i(x_C, LA, h_C^{t-1}) - u_i(x', LA, h_C^{t-1}) \\
\geq\ & u_i(x_C, LA, h_C^{t-1}) - u_i(x^{LA}(\varepsilon_t, p_C^t), LA, h_C^{t-1}) \\
\geq\ & p_C^t \cdot (1 - p_C^{t+1}) \cdot (u_i^S(x_D, x_C) - u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C)) - (1 - 2p_C^t + p_C^t \cdot p_C^{t+1}) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) \\
\geq\ & - (1 - \gamma^*)^2 \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_D) \\
\geq\ & 0
\end{aligned}$$

Case (2): If $p_C^t = 1$ and $p_C^{t+1} = 1$.

$$\begin{aligned}
u_i(x_C, LA, h_C^{t-1}) =\ & \varepsilon_t \cdot u_i^{LA}(x_C, x_C) + (1 - \varepsilon_t) \cdot u_i^S(x_C, x_C) + u_i(LA, h_C^t) \\
\geq\ & a + u_i(LA, h_C^t) \\
\geq\ & a + \varepsilon_{t+1} \cdot u_i^{LA}(x^{LA}(\varepsilon_{t+1}, p_C^{t+1}), x_C) + (1 - \varepsilon_{t+1}) \cdot u_i^S(x^{LA}(\varepsilon_{t+1}, p_C^{t+1}), x_C) \\
\geq\ & a + \varepsilon_t \cdot u_i^{LA}(x^{LA}(\varepsilon_t, p_C^t), x_C) + (1 - \varepsilon_t) \cdot u_i^S(x^{LA}(\varepsilon_t, p_C^t), x_C)
\end{aligned}$$

Therefore,

$$\begin{aligned}
u_i(x_C, LA, h_C^{t-1}) - u_i(x', LA, h_C^{t-1}) & \geq u_i(x_C, LA, h_C^{t-1}) - u_i(x^{LA}(\varepsilon_t, p_C^t), LA, h_C^{t-1}) \\
& \geq a > 0
\end{aligned}$$

Case (3): If $p_C^t = 1$ and $p_C^{t+1} = \gamma^*$, notice that we have $\varepsilon_t = (\gamma^*)^{T-t}$. Moreover, since (x_C, x_C) forms a Nash equilibrium in u^{LA}, we have $u_i^{LA}(x', x_C) \leq u_i^{LA}(x_C, x_C)$ for all $x' \neq x_C$.

$$\begin{aligned}
u_i(x', LA, h_C^{t-1}) & = \varepsilon_t \cdot u_i^{LA}(x', x_C) + (1 - \varepsilon_t) \cdot u_i^S(x', x_C) \\
& = (\gamma^*)^{T-t} \cdot u_i^{LA}(x', x_C) + (1 - (\gamma^*)^{T-t}) \cdot u_i^S(x', x_C) \\
& \leq (\gamma^*)^{T-t} \cdot u_i^{LA}(x_C, x_C) + (1 - (\gamma^*)^{T-t}) \cdot u_i^S(x_D, x_C) \\
& = (\gamma^*)^{T-t} \cdot u_i^{LA}(x_C, x_C) + (1 - (\gamma^*)^{T-t}) \cdot b
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
u_i(x_C, LA, h_C^{t-1}) & = \varepsilon_t \cdot u_i^{LA}(x_C, x_C) + (1 - \varepsilon_t) \cdot u_i^S(x_C, x_C) + u_i(LA, h_C^t) \\
& = (\gamma^*)^{T-t} \cdot u_i^{LA}(x_C, x_C) + (1 - (\gamma^*)^{T-t}) \cdot a + u_i(LA, h_C^t)
\end{aligned}$$

Note that for all $t' > t$, we have $p_C^{t'} = \gamma^*$. We use induction we show that

$$u_i(LA, h_C^{t'}) \geq \sum_{i=0}^{T-t'} (\gamma^*)^i \cdot (\gamma^* \cdot a + (1 - \gamma^*)(-c)) = \frac{1 - (\gamma^*)^{T-t'+1}}{1 - \gamma^*} \cdot (\gamma^* \cdot a + (1 - \gamma^*)(-c))$$

for all $t' > t$. The base case $t' = T$ is true since $\varepsilon_T = \gamma^*$ and we have

$$u_i(LA, h_C^{T-1}) = \gamma^* \cdot u_i^{LA}(x_C, x_C) + (1 - \gamma^*) \cdot u_i^S(x_C, x_D) \geq \gamma^* \cdot a + (1 - \gamma^*)(-c)$$

Assume it is true for $t' = t''$, then for $t' = t'' - 1$, we have

$$\begin{aligned}
u_i(LA, h_C^{t'-1}) & = \varepsilon_{t'} \cdot u_i^{LA}(x_C, x_C) + (\gamma^* - \varepsilon_{t'}) \cdot u_i^S(x_C, x_C) + (1 - \gamma^*) \cdot u_i^S(x_C, x_D) + \gamma^* \cdot u_i(LA, h_C^{t'}) \\
& \geq \gamma^* \cdot a + (1 - \gamma^*)(-c) + \gamma^* \cdot u_i(LA, h_C^{t'}) \\
& \geq \sum_{i=0}^{T-t'} (\gamma^*)^i \cdot (\gamma^* \cdot a + (1 - \gamma^*)(-c))
\end{aligned}$$

Therefore, we have

$$\begin{aligned} u_i(x_C, LA, h_C^{t-1}) &\geq (\gamma^*)^{T-t} \cdot u_i^{LA}(x_C, x_C) + (1-(\gamma^*)^{T-t}) \cdot a + u_i(LA, h_C^t) \\ &\geq (\gamma^*)^{T-t} \cdot u_i^{LA}(x_C, x_C) + (1-(\gamma^*)^{T-t}) \cdot a + \frac{1-(\gamma^*)^{T-t}}{1-\gamma^*} \cdot (\gamma^* \cdot a + (1-\gamma^*)(-c)) \end{aligned}$$

Recall that $b^2 \cdot (\gamma^*)^2 + (a-b+c) \cdot \gamma^* - c = 0$. Henceforth, we have $\gamma^* \cdot a + (1-\gamma^*) \cdot (-c) = b \cdot \gamma^* \cdot (1-\gamma^*)$. Therefore,

$$\begin{aligned} u_i(x_C, LA, h_C^{t-1}) &\geq (\gamma^*)^{T-t} \cdot u_i^{LA}(x_C, x_C) + (1-(\gamma^*)^{T-t}) \cdot a + \frac{1-(\gamma^*)^{T-t}}{1-\gamma^*} \cdot (\gamma^* \cdot a + (1-\gamma^*)(-c)) \\ &= (\gamma^*)^{T-t} \cdot u_i^{LA}(x_C, x_C) + (1-(\gamma^*)^{T-t}) \cdot a + \frac{1-(\gamma^*)^{T-t}}{1-\gamma^*} \cdot b \cdot \gamma^* \cdot (1-\gamma^*) \\ &= (\gamma^*)^{T-t} \cdot u_i^{LA}(x_C, x_C) + (1-(\gamma^*)^{T-t}) \cdot (a + b \cdot \gamma^*) \\ &\geq (\gamma^*)^{T-t} \cdot u_i^{LA}(x_C, x_C) + (1-(\gamma^*)^{T-t}) \cdot b \\ &\geq \max_{x' \neq x_C} u_i(x', LA, h_C^{t-1}) \end{aligned}$$

The last but one inequality is by the fact that $\gamma^* \geq 1 - \frac{a}{b}$ according to Proposition 3.1. □