Evolutionary Stability of Other-Regarding Preferences Under Complexity Costs

Anthony DiGiovanni
Center on Long-Term Risk

Nicolas Macé
Center on Long-Term Risk

Jesse Clifton
Center on Long-Term Risk

Abstract

The evolution of preferences that account for other agents' fitness, or *other-regarding preferences*, has been modeled with the "indirect approach" to evolutionary game theory. Under the indirect evolutionary approach, agents make decisions by optimizing a subjective utility function. Evolution may select for subjective preferences that differ from the fitness function, and in particular, subjective preferences for increasing or reducing other agents' fitness. However, indirect evolutionary models typically artificially restrict the space of strategies that agents might use (assuming that agents always play a Nash equilibrium under their subjective preferences), and dropping this restriction can undermine the finding that other-regarding preferences are selected for. Can the indirect evolutionary approach still be used to explain the apparent existence of other-regarding preferences, like altruism, in humans? We argue that it can, by accounting for the costs associated with the complexity of strategies, giving (to our knowledge) the first account of the relationship between strategy complexity and the evolution of preferences. Our model formalizes the intuition that agents face tradeoffs between the low cognitive cost of simple strategies within a single context, and the ability of more complex (subjective utility-maximizing) strategies to interpolate across contexts. For a single game, these penalties lead to selection for a simple fixed-action strategy, but across games, when there is a sufficiently large penalty on a strategy's number of context-specific parameters, a strategy of maximizing subjective (other-regarding) utility is stable again. Overall, our analysis provides a more nuanced picture of when other-regarding preferences will evolve.

1 Introduction

Under what conditions do agents evolve to maximize a subjective utility function other than their evolutionary fitness? In particular, when is there selection for *other-regarding preferences* [Elster, 1983, Sen, 1986] such as altruism (intrinsically valuing improvements in others agents' fitness) or spite (intrinsically valuing reductions in other agents' fitness)? These questions have been previously studied under the "indirect approach" to evolutionary game theory [Güth and Kliemt, 1998]. Consider a game whose payoffs determine the players' fitness in an evolutionary process, called a base game. The indirect evolutionary approach supposes that selection occurs on agents' subjective preferences (hereafter, "preferences") represented as utility functions, and agents rationally play the base game by optimizing their subjective utility functions. When assessing the evolutionary stability of strategies in the indirect approach, a player's utility function defines their strategy. This is in contrast to the classical "direct" approach where actions in the base game themselves are selected. This indirect approach has been applied in attempts to explain altruism in organisms, especially in contexts where other explanations such as kin selection and reciprocity are inadequate [Bester and Güth, 1998, Janssen, 2008, Konrad and Morath, 2012]. In a simple model of an interaction where two agents' actions have positive externalities for each other — i.e., increasing one's action (represented as a real number) increases the other's marginal payoff from their action — Bester and Güth [1998] find that altruistic preferences are evolutionarily stable. Bolle [2000] and Possajennikov [2000] extended this model to also explain the stability of spiteful preferences in interactions with negative externalities.

However, these models have two key limitations:

1. They assume that agents always play a best response given their preferences and beliefs about the other player's preferences. This precludes agents who commit to following a certain action regardless of their beliefs about the other player. This is important because, as we will show, when such commitments are allowed, a subjective utility-maximizing strategy with other-regarding preferences is no longer the unique evolutionarily stable strategy.
2. They restrict the space of preferences in a way that prevents the use of strategies capable of invading populations of inefficient strategies, called the "secret handshake" in previous work

[Robson, 1990]. As Dekel et al. [2007] show, when the space of preferences is expanded to all possible utility functions, evolutionarily stable strategies in an indirect evolutionary model must be efficient. This is because any population of inefficient strategies can be invaded by mutants who mimic the behavior of the inefficient strategy, and play an efficient action against other mutants.

These two modifications to the original indirect evolutionary models undermine those models' conclusions that other-regarding preferences can be evolutionarily stable, including preferences that lead to inefficient behavior. However, an important feature of the kinds of strategies described in (1) and (2) is that they differ from subjective utility maximization in their *complexity costs*, i.e., the costs an agent must pay to learn and execute strategies [McNamara, 2013]. These costs may play a critical role in evolution; for instance, the tradeoff between the problem-solving benefits and energetic costs of larger brains may explain variation in brain size among primates, and in animal behavior in contests [Isler and Van Schaik, 2014, Reichert and Quinn, 2017]. Previous literature has studied how complexity costs affect the evolutionary stability of strategies [Rubinstein, 1986, Banks and Sundaram, 1990, Binmore and Samuelson, 1992]. The costs of strategy complexity accumulate over the diverse set of environments and interactions an agent faces in its lifetime [Geoffroy and André, 2021]. Thus, instead of using many different strategies that are each simple in isolation, it can be less expensive overall for an agent to use a sophisticated strategy that interpolates well across interactions [Robalino and Robson, 2016, Piccinini and Schulz, 2018]. We will argue that the costs of applying individualized heuristics to each new interaction may be sufficiently high that evolution selects for agents that consistently optimize some (other-regarding) utility function.

Our key contribution is a revised account of the evolution of other-regarding preferences, based on a novel framework accounting for the fitness costs that strategies incur due to their complexity in *multiple* strategic contexts. While existing indirect evolutionary models are inadequate because they artificially restrict the space of strategies, we show that their predictions can be recovered by accounting for how subjective utility-maximizing strategies optimally trade off complexity within and across decision contexts. In particular:

- We characterize the Nash equilibria (and stability thereof) of the space of subjective utility-maximizing strategies from Possajennikov [2000] augmented with strategies that commit to a certain action ("behavioral strategies"), in a general class of symmetric two-player games. In this expanded space, rational strategies with other-regarding preferences that are evolutionarily stable against other rational strategies, as in Bester and Güth [1998] and Possajennikov [2000], are no longer the unique evolutionarily stable strategies. This result motivates the search for an alternative explanation of the evolution of other-regarding preferences.
- While previous work has shown how finite computational costs of strategies in repeated games significantly alter the set of stable strategies, we present two results illustrating a tradeoff between within-game and across-game complexity costs: (1) Suppose that rational strategies are more costly in a single interaction than behavioral strategies, given the greater energetic costs associated with their complex cognition [Abreu and Sethi, 2003]. Then in an *individual* complexity-penalized game, the multiplicity of neutrally stable strategies, including rational strategies with other-regarding preferences, is replaced with a unique evolutionarily stable strategy, the Nash equilibrium of the base game. (2) When agents play multiple games, a sufficiently large penalty on the number of game-specific parameters used by a strategy reproduces the results of Bester and Güth [1998] and Possajennikov [2000] — in numerical experiments, the population converges (under a particular evolutionary dynamic) to a rational strategy with other-regarding preferences. Our experiments also explore how the size of the penalty on game-specific parameters necessary for other-regarding preferences to evolve, and the strength of altruism or spite that evolves, depend on the distribution of games.
- We argue that accounting for complexity costs blocks the secret handshake argument: Mutant strategies that both mimic an inefficient action and play an efficient action against themselves are more complex than behavioral strategies, and thus cannot invade a strategy that always plays the Nash equilibrium of the base game.

2 Related Work

Indirect evolutionary approach Like Bester and Güth [1998], Bolle [2000], and Possajennikov [2000], we model rational players as playing Nash equilibria with respect to utility functions given by their own fitness plus a (possibly negative) multiple of their opponent's fitness. Heifetz et al. [2003] generalize this model to utility functions given by one's own fitness plus some function called a *disposition*. They show that dispositions are not eliminated by selection in a wide variety of games. Generalizing further to the space of all possible utility functions in finite-action games, Dekel et al. [2007] show that any strategy achieving an inefficient payoff against itself — including the kinds of strategies with other-regarding preferences predicted by Possajennikov [2000] — is not evolutionarily stable. We will argue, however, that the invader strategies that make efficiency necessary for stability are more complex than behavioral or rational strategies, and thus when complexity costs are accounted for in an agent's fitness, a stable strategy can lead to inefficiency. Ok and Vega-Redondo [2001] and Ely and Yilankaya [2001] note that in order for utility functions to evolve such that players with those utility functions do not play the base game Nash equilibrium, players must have information about each other's utility functions. We assume utility functions are known, and briefly discuss how players can learn each other's utility functions over repeated interactions, but acknowledge that this is a substantive assumption since players often have incentives to send deceptive signals of their utility functions [Heller and Mohlin, 2019].

Games with complexity costs Rubinstein [1986] characterizes Nash equilibria in repeated games under computational costs. He represents strategies in the repeated Prisoner's Dilemma as finite-state automata (sets of states determining the player's action with rules for transitions between states). Complexity costs are lexicographic: an automaton achieving a strictly higher payoff is always preferred, but when two automata achieve the same average payoff, the automaton with fewer states is preferred. Binmore and Samuelson [1992] show that although no evolutionarily stable strategies exist in the repeated Prisoner's Dilemma without complexity costs, adding these lexicographic costs leads to the existence of some evolutionarily stable strategies. We similarly show that in one-shot games, when we account for the greater complexity of "rational" strategies relative to "behavioral" (fixed-action) strategies, a set of multiple neutrally stable strategies is replaced with a unique evolutionarily stable strategy. Our distinction between the complexity of rational and behavioral strategies follows that of Abreu and Sethi [2003], who show that under an arbitrarily small cost of the complexity of rationality, behavioral strategies are evolutionarily stable in a bargaining game. If automata are also penalized based on the number of different states each state can transition to, the evolutionarily stable strategies are restricted to the Nash equilibria of the (non-repeated) base game [Banks and Sundaram, 1990]. We find an analogous result in one-shot games with a different complexity metric. Lastly, van Veelen and García [2019] find that in the repeated Prisoner's Dilemma, increasing non-lexicographic complexity costs decreases the frequency of cooperation in finite-population stochastic evolutionary simulations. Similarly, in the multi-game setting, we find numerically that as complexity costs on a strategy's number of game-specific parameters increase, there are transitions between more or less efficient stable strategies.

Coevolution of rationality and other-regarding preferences A key theme in our work is that selection may favor the ability of rational agents, which have other-regarding preferences and model other players as optimizing their own utility functions, to solve a variety of strategic problems. Building on Robson [2001]'s analogous results in single-agent problems, Robalino and Robson [2016] model the coevolution of utility maximization and ability to attribute preferences to others. Like us, they show that after accounting for the advantages of interpolation across strategic contexts, selection favors a rational strategy that learns and responds to the preferences of its opponent, as opposed to strategies that do not know how to respond to new games. However, we study selection pressures towards rationality in the context of evolution of preferences. Further, in our analysis, the advantage of rationality comes from avoiding costs that non-rational strategies pay to adapt a response to each separate game, rather than from non-rational strategies' inability to respond to new games. Heller and Mohlin [2019] model the evolution of both preferences and the cognitive capacity necessary to signal false preferences to others. Their argument for the efficiency of stable strategies is vulnerable to the counterargument that we raise to Dekel et al. [2007] above. However, their results are similar to ours in that the set of stable strategies is sensitive to whether the costs of cognitive complexity are sufficiently high, relative to the direct fitness benefits of complex cognition. Like us, Geoffroy and André [2021] model the evolution of strategies that interpolate across different contexts, but their analysis is restricted to cooperation in a certain class of games rather than evolution of other-regarding preferences in general (including uncooperative preferences like spite).

3 Preliminaries and Running Example

We begin with definitions and notation and introduce a well-studied game that will illustrate principles of the indirect evolutionary approach.

Let G be any symmetric two-player game (called the base game) between players i and j, with action space $\mathcal{A}$ and payoff functions $u_i, u_j : \mathcal{A}^2 \to \mathbb{R}$. Players choose actions in the base game as functions of strategies that are selected in an evolutionary process. Suppose players simultaneously play strategies (elements of some abstract space Σ) and observe each other's strategies, then play G with actions determined by the pair of strategies. Then, define the function $r_i : \Sigma^2 \to \mathcal{A}$, where player i's action in G given the players' strategies $\sigma_i, \sigma_j \in \Sigma$ is $r_i(\sigma_i, \sigma_j)$. In standard evolutionary analysis the fitness of a strategy equals its payoff in G, thus we write player i's fitness from a strategy profile as $f_i(\sigma_i, \sigma_j) = u_i(r_i(\sigma_i, \sigma_j), r_j(\sigma_j, \sigma_i))$. (We distinguish fitness from payoffs because once complexity costs are included, as in Section 5.1, this identity no longer holds.) The following definitions classify a strategy based on the robustness to mutations of a population purely consisting of that strategy.

Definition 1. *Relative to a fixed strategy space Σ for G, a strategy $\sigma \in \Sigma$ is:*

- *A* ***Nash equilibrium*** *if, for all $\sigma' \in \Sigma$, $f_i(\sigma, \sigma) \geq f_i(\sigma', \sigma)$.*

- *A **neutrally stable strategy (NSS)** if (1) it is a Nash equilibrium, and (2) for all σ' such that $f_i(\sigma', \sigma) = f_i(\sigma, \sigma)$, then $f_i(\sigma, \sigma') \geq f_i(\sigma', \sigma')$.*
- *An **evolutionarily stable strategy (ESS)** if it is an NSS and the inequality in 2 is always strict.*

The strict inequality $f_i(\sigma, \sigma') > f_i(\sigma', \sigma')$ in the definition of ESS implies a stronger "pull" towards an ESS in evolutionary dynamics (such as the replicator dynamic) than towards an NSS: If a rare mutant that enters a population consisting of an ESS has the same fitness when paired with itself as the ESS has against this mutant, the mutant goes extinct under the replicator dynamic, but this does not necessarily hold for an NSS [van Veelen, 2010].

Our running example is the following symmetric two-player game, which we call the externality game [Bester and Güth, 1998]. Each player i simultaneously chooses $a_i \in \mathbb{R}$, and, for some $m > 0$ and $\kappa \in [-2, 0) \cup (0, 1)$, the players receive payoffs:

$$u_i(a_1, a_2) = a_i(\kappa a_j + m - a_i).$$

Thus, κ represents negative or positive externalities of each player's action for the other's payoff (when $\kappa < 0$ or $\kappa > 0$, respectively). In the original model, players are assumed to have the following *subjective utility functions*, for $\alpha_i \in \mathbb{R}$:

$$V_i^{\alpha_i}(a_1, a_2) = u_i(a_1, a_2) + \alpha_i u_j(a_1, a_2).$$

Players behave rationally with respect to their subjective utility functions, and subjective utility functions are common knowledge. Thus the players play the Nash equilibrium of the game $G(\alpha_i, \alpha_j)$ in which payoffs are given by $V_i^{\alpha_i}, V_j^{\alpha_j}$, denoted $\text{NE}_i(\alpha_i, \alpha_j)$. That is, letting α_i represent player i's strategy, $r_i(\alpha_i, \alpha_j) = \text{NE}_i(\alpha_i, \alpha_j)$.

A player with $\alpha_i > 0$ (respectively, $\alpha_i < 0$) has subjective utility increasing (decreasing) with the other's payoff — these ranges of α_i can be interpreted as altruistic and spiteful, respectively. Generalizing Bester and Güth [1998], Possajennikov [2000] showed that the unique ESS in this strategy space is $\alpha_i = \alpha_j = \frac{\kappa}{2-\kappa}$. Thus, when $\kappa > 0$, this ESS corresponds to players with altruistic preferences, and when $\kappa < 0$, their preferences are spiteful. Players who follow the subjective Nash equilibrium with respect to $V_i^{\alpha_i}$ given by the altruistic ESS both receive a higher payoff than the equilibrium of G, while the payoffs of the spiteful ESS are both lower. Since the Pareto-efficient symmetric subjective Nash equilibrium is at $\alpha_i = \alpha_j = 1$, this means that as $\kappa \nearrow 1$, the ESS approaches efficiency. Intuitively, these other-regarding preferences are stable in Possajennikov [2000]'s model because they serve as commitment devices that elicit favorable responses from the other player [Frank, 1987, Dufwenberg and Güth, 1999]. That is, each agent best-responds under the assumption that the other player will play rationally with respect to their utility function, and as utility functions are selected based on payoffs from the opponent's best response to the action optimizing those utility functions, the population converges to some α.

4 Setup

We now discuss the formal framework on which our results are based. Let $V_i^{\alpha}(a_1, a_2) = u_i(a_1, a_2) + \alpha u_j(a_1, a_2)$ as above. We say that a preference parameter α is *egoistic* if $\alpha = 0$, and *other-regarding* otherwise. In our results we will use the following assumptions, which are satisfied by the externality game:

1. For any α_i, α_j, $\text{NE}_i(\alpha_i, \alpha_j)$ is unique.
2. For any b and α, the function $h_i(a) = V_i^{\alpha}(a, b)$ has a unique global maximum, $\text{BR}_i(b; \alpha)$. (That is, the best response to some action under any subjective utility function is unique.)
3. For any α, the function $g_i(\beta) = \text{NE}_i(\beta, \alpha)$ is surjective on $\mathbb{R}$.[1]

We give some remarks on the typical indirect evolutionary models before presenting our generalized model. Recall our claim that the strategy space assumed by much of the indirect evolutionary game theory literature is too restrictive, due to the assumption that agents always play the Nash equilibrium of $G(\alpha_i, \alpha_j)$. Playing a Nash equilibrium in response to the other player's α_j can be exploitable, in the sense that a player j can "force" another rational agent to play an action that is more favorable to player j (see Section 4.1 for an example). A player may avoid being exploited in this way by committing to some action, independent of opponents' preferences. We will therefore enrich the strategy space in G to relax this assumption (Section 4.1).

Standard indirect evolutionary game theory also assumes players perfectly observe each other's payoff functions and subjective utility functions. This premise has been questioned in previous work, e.g., Heifetz et al. [2007], Gardner and West [2010]. We keep this assumption due to findings by Jordan [1991] and Kalai and Lehrer [1993] that, if players use Bayesian updating in repeated interactions with each other, under certain conditions they converge to accurate beliefs about each other's utility functions and play the Nash equilibrium. Dekel et al. [2007] and Heller and Mohlin [2019] give similar justifications for this assumption in their indirect evolutionary models.

4.1 Strategy space

Our strategy space combines the "direct" and "indirect" approaches to evolutionary game theory [Güth and Kliemt, 1998]. That is, this space includes both fixed actions of the base game and strategies that choose actions

[1] For the externality game, there is no β such that $g_i(\beta) = -\frac{m}{\kappa(1+\alpha)}$. However, one can check directly that $\max_{a_i} u_i(a_i, \text{BR}_j(a_i; \alpha)) = u_i(\text{NE}_i(\alpha, \alpha), \text{NE}_j(\alpha, \alpha))$, which is the only condition for which this assumption is necessary.

as a function of the player's own subjective utility function and the other player's strategy.

First, a *behavioral strategy* plays an action a_i, independent of the other player's strategy. The action a_i is common knowledge to both players before G is played. Second, as in the standard indirect evolutionary approach [Bester and Güth, 1998, Possajennikov, 2000], a *rational strategy* has a commonly known preference parameter α_i, and always plays a best response given α_i to their beliefs about the other player. A rational player believes that another rational player plays the Nash equilibrium of $G(\alpha_i, \alpha_j)$. Thus the best response to another rational player with parameter α_j is $\text{NE}_i(\alpha_i, \alpha_j)$. A rational player i believes behavioral player j plays action a_j, so the rational strategy is $\text{BR}_i(a_j; \alpha_i)$.

To see the reason for including both classes of strategies in one model, consider the externality game with $\kappa < 0$. If a rational player i faces rational player j with $\alpha_j = 0$, and $\alpha_i = \alpha^* := \frac{\kappa}{2-\kappa} < 0$, we can check that the payoff of i increases while that of j decreases: $u_i(\text{NE}_i(\alpha^*, 0), \text{NE}_j(\alpha^*, 0)) > u_i(\text{NE}_i(0, 0), \text{NE}_j(0, 0)) > u_j(\text{NE}_i(\alpha^*, 0), \text{NE}_j(\alpha^*, 0))$. That is, i can exploit the rationality of j by adopting an other-regarding preference parameter as a commitment. We therefore ask what strategies are selected for when we allow players to *ignore* each other's commitments (preferences), in order to avoid exploitation, and instead play some fitness-maximizing action.

In summary, our strategy space $\mathcal{S}$ is the union of these sets:

1. $\mathcal{B} = \{B(a) \mid a \in \mathcal{A}\}$: Behavioral strategy whose action is $r_i(B(a), \sigma_j) = a$ for all σ_j.
2. $\mathcal{R} = \{R(\alpha) \mid \alpha \in \mathbb{R}\}$: Rational strategy whose action is $r_i(R(\alpha), \sigma_j) = \text{BR}_i(a; \alpha)$ if $\sigma_j = B(a)$, or $r_i(R(\alpha), \sigma_j) = \text{NE}_i(\alpha, \alpha')$ if $\sigma_j = R(\alpha')$.

5 Results

We now characterize the Nash equilibria and stable strategies of $\mathcal{S}$. We show that there are multiple neutrally stable strategies, one of which acts according to egoistic preferences, and no evolutionarily stable strategies. This is in contrast to the results of Bester and Güth [1998] and Possajennikov [2000], who showed that without behavioral strategies, a population with other-regarding preferences is the unique ESS in the externality game. All proofs are in Appendix A.

Proposition 1. *Let G be a symmetric two-player game that satisfies assumptions 1- 3. Then a strategy is a Nash equilibrium in $\mathcal{S}$ if and only if it is either $B(NE_i(0, 0))$ or a strategy $R(\alpha)$ that is a Nash equilibrium in $\mathcal{R}$. Further, $B(NE_i(0, 0))$ is an NSS in $\mathcal{S}$, and $R(\alpha)$ is an NSS in $\mathcal{S}$ if and only if it is an NSS in $\mathcal{R}$. There are no ESSes.*

Informally, a population that always plays the base game Nash equilibrium can be invaded by rational players with egoistic preferences, whose fitness against each other matches that of the original population. When the population consists of rational players with other-regarding preferences that are stable against other rational strategies, it can be invaded by agents that always play the Nash equilibrium of the game with payoffs given by those same other-regarding preferences.

5.1 Complexity penalties

Single game Proposition 1 showed that strategies with either egoistic or other-regarding preferences can be neutrally stable, and neither are evolutionarily stable. This suggests that the standard indirect evolutionary approach is insufficient to explain the unique stability of other-regarding preferences. However, our analysis above assumed that players can use arbitrarily complex strategies at no greater cost than simpler ones; fitness is a function only of the payoffs of strategies, not of the cognitive resources required to use them [McNamara, 2013].

We introduce complexity costs as follows. For some complexity function $c : \Sigma \to \mathbb{R}$, we apply the usual evolutionary stability analysis to a modified strategy fitness function:

$$f_i(\sigma_i, \sigma_j) = u_i(r_i(\sigma_i, \sigma_j), r_j(\sigma_j, \sigma_i)) - c(\sigma_i).$$

While behavioral strategies always play a fixed action, rational strategies compute a best response to each given opponent. Within a single game, a behavioral strategy thus requires less computation than a rational strategy (this assumption was also used by Abreu and Sethi [2003]). Given this observation, for some $\epsilon_R > 0$ we define $c(\sigma) = \epsilon_R \mathbb{I}[\sigma \in \mathcal{R}]$ (where the function $\mathbb{I}$ returns 1 if the condition in brackets is true, and 0 otherwise). Once this cost is accounted for, selection favors the behavioral strategy that plays the Nash equilibrium of G (even when assumption 3 does not hold).

Proposition 2. *Let G be a symmetric two-player game that satisfies assumptions 1 and 2. Then for any $\epsilon_R > 0$, the unique Nash equilibrium in $\mathcal{S}$ under penalties is $B(NE_i(0, 0))$, and this strategy is an ESS.*

An arbitrarily small cost of complexity prevents rational strategies from matching the fitness of the Nash equilibrium behavioral strategy.

Multiple games Proposition 2, again, appears inconsistent with the stability of other-regarding preferences. However, this result is based on a metric of complexity that only accounts for costs within one game — the cost of rational optimization versus playing a constant action for any opponent — rather than cumulative costs *across* games. As Piccinini and Schulz [2018] discuss qualitatively, although agents who rely on situation-specific heuristics avoid the fixed cost of explicit optimization paid by rational agents, they do worse in some variable environments than the latter, who can profit from having a general and compact strategy of optimizing utility functions. We formalize this tradeoff in this section.

Suppose that in each generation, the players in a population face a collection of games $\{G_1, \dots, G_K\}$. Each player uses a strategy that (through the function r_i) outputs an action conditional on both the other player's strategy *and* the identity of the game. One can apply the usual evolutionary stability analysis to strategies that play the collection of games, by defining fitness as the sum of fitness from each game minus a multi-game complexity function c_K. If a given strategy has $N(K)$ parameters under selection across K games, c_K should increase with $N(K)$. An ideal definition of this function would be informed by an accurate model of the energetic costs of different kinds of cognition, which is beyond the scope of this work. We can define multi-game complexity in our setting by generalizing the strategy space from Section 4 to multiple games:

1. $\mathcal{B}_K = \{(B(a_k))_{k=1}^K\}$: Plays a_k in game G_k.
2. $\mathcal{R}_K = \{R(\alpha)\}$: Plays the rational strategy with respect to α for each G_k.

The motivation for parameterizing a strategy in $\mathcal{R}_K$ by a single α is that, across a distribution of relevantly similar games (e.g., variants of the externality game with different values of κ), a rational player might be able to perform well by interpolating its other-regarding preferences.[2] Then, for some $\epsilon_P > 0$, define:

$$c_K(\sigma) = \begin{cases} \epsilon_P |\{a_k\}_{k=1}^K|, & \sigma \in \mathcal{B}_K \\ \epsilon_R + \epsilon_P, & \sigma \in \mathcal{R}_K. \end{cases}$$

The set of stable strategies under these multi-game penalties is sensitive to the values of ϵ_R and ϵ_P. Intuitively, a behavioral strategy will be stable when ϵ_P is small, relative to the profits this strategy can make by adapting its response precisely to each game. Conversely, when ϵ_P is sufficiently large, a rational strategy can compensate for applying the same decision rule to every game by avoiding the costs of game-specific heuristics. In the next section, we show these patterns numerically.

5.2 Evolutionary simulations with multi-game complexity penalties

Here, we will use an evolutionary simulation algorithm to see how complexity costs across games influence stable strategies — in particular, which (if any) other-regarding preferences are selected? For simplicity, we consider a set of just two externality games for a fixed m with $\kappa = \kappa_1$ and $\kappa = \kappa_2$, denoted G_{κ_1} and G_{κ_2}. However, to investigate the effects of imbalanced environments (i.e., where G_{κ_1} is played more or less frequently than G_{κ_2}) we suppose that players spend a fraction p of their time in game G_{κ_1} and $1-p$ in G_{κ_2}. Then, with u_i^κ as the externality game payoff function for a given κ, the multi-game penalized fitness of a strategy σ against σ' is:

$$f_i^{\kappa_1,\kappa_2}(\sigma, \sigma') = p u_i^{\kappa_1}(r_i(\sigma, \sigma', G_{\kappa_1}), r_j(\sigma', \sigma, G_{\kappa_1})) \\ + (1-p) u_i^{\kappa_2}(r_i(\sigma, \sigma', G_{\kappa_2}), r_j(\sigma', \sigma, G_{\kappa_2})) - c_2(\sigma).$$

Due to the continuous strategy space, a replicator dynamic simulation is intractable. Instead, we simulate an evolutionary process on $\mathcal{S}$ using the *adaptive learning* algorithm [Young, 1993], implemented as follows (details are in Appendix B). An initial population of size $N = 10$ is randomly sampled from the spaces of rational and behavioral strategies. In each round $t = 1, \dots, 30$ of evolution, each player in the population either (with low probability) switches to a random strategy, or else switches to the best response to a uniformly sampled opponent in the population (with respect to the penalized fitness $f_i^{\kappa_1,\kappa_2}$ above).[3] Note that a best response in the space $\mathcal{B}_K$ might use one action across both games, incurring a complexity cost of ϵ_P instead of $2\epsilon_P$. We fix $\epsilon_R = 10^{-5}$ and $m = 1$. In each experiment, we tune the multi-game complexity penalty ϵ_P (hereafter, "per-parameter penalty") to approximately the smallest value necessary to ensure that the population almost always converges to an element of $\mathcal{R}_K$ (a rational strategy).

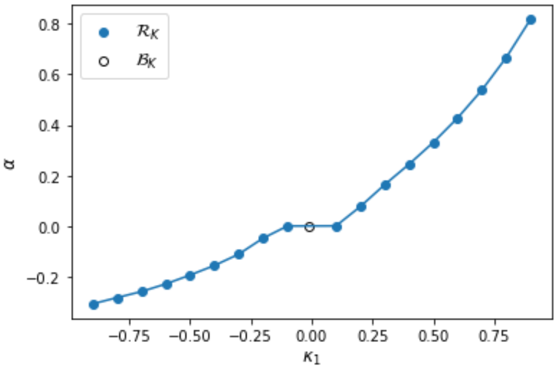

Figure 1: α values of the limit of an evolutionary simulation, in which a proportion 0.5 of games have κ_1 and a proportion 0.5 have $\kappa_2 = 0.001$, as a function of κ_1. All members of the final population are in $\mathcal{R}_K$ except when $\kappa_1 = -0.01$, where the population consists of behavioral strategies (depicted with an open circle). The per-parameter penalty is $\epsilon_P = 0.001$.

Varying strength of negative or positive externalities in one game First, we show that other-regarding preferences evolve under sufficiently strong negative or positive externalities, given a sufficiently high per-parameter penalty. We fix $\kappa_2 = 0.001$, $p = 0.5$, and $\epsilon_P = 0.001$, and vary $\kappa_1 \in$

[2] Compare to Berninghaus et al. [2007]'s model in which players evolve preferences for reciprocity that they apply to both the ultimatum and dictator games.

[3] This algorithm is most appropriate when the evolutionary process is interpreted as agents learning over their lifetimes, updating their responses to each other, rather than as genetic transmission.

$\{-0.9, -0.8, \ldots, -0.1, -0.01, 0.1, \ldots, 0.9\}$. For $\kappa_1 = -0.01$, the population converged to a behavioral strategy that uses only one action, for all values of ϵ_P we tested (see the open circle in Figure 1). This suggests that when both games are sufficiently similar, a behavioral strategy can interpolate across both games at less expense than a rational strategy. Figure 1 shows that, as expected, the sign and magnitude of the stable α value scales with κ_1. For $\kappa_1 \in [-0.1, 0.1]$, the population converges to $\alpha \approx 0$, suggesting that other-regarding preferences only interpolate well across these externality games when the externalities are sufficiently strong in magnitude.

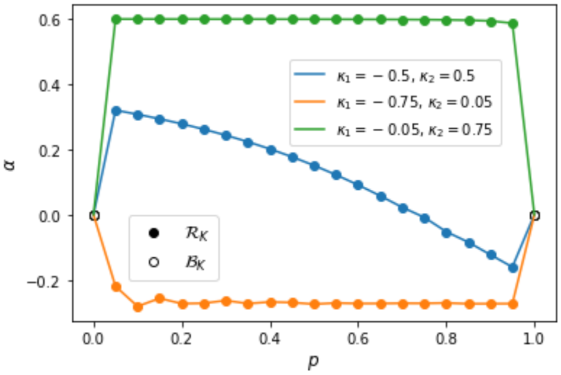

Figure 2: α values of the limit of an evolutionary simulation, in which a proportion p of games have κ_1 and a proportion $1 - p$ have κ_2, as a function of p, with three pairs of κ values. All populations are in $\mathcal{R}_K$ except $p = 0.00005$ and $p = 0.99995$, where the population consists of behavioral strategies (depicted with open circles). The per-parameter penalty is $\epsilon_P = 0.002$.

Varying proportion of games with negative versus positive externalities Next, we show that the strength of altruism versus spite in the limiting population scales nonlinearly with the proportion of games with negative versus positive externalities. With $\epsilon_P = 0.002$, we vary the fraction of games with $\kappa < 0$, over $p \in \{0.00005, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99995\}$, for three pairs of games. For all pairs of κ in this experiment, the values $p = 0.00005$ and $p = 0.99995$ have one-action behavioral strategies in the limiting population (see the open circles in Figure 2). When one game is extremely rare, the rational strategy's gains from interpolation across games do not outweigh the cost ϵ_R of rationality.

First we fix $\kappa_1 = -0.5$ and $\kappa_2 = 0.5$ (blue curve in Figure 2). Again, the trend of decreasing α with greater p is as expected, though there is a bias towards altruism: an equal proportion of positive and negative externalities gives $\alpha > 0$. When $\kappa_1 = -0.75$ and $\kappa_2 = 0.05$ (orange curve), even small proportions of the large-magnitude negative κ_1 are sufficient for the rational population to adopt $\alpha < 0$, and α remains roughly constant above $p \approx 0.1$. That is, in an environment where one game has weak positive externalities and the other has strong negative externalities, most of the effect on the population's other-regarding preferences comes just from having a frequency of strong negative externalities *above some (small) threshold*. The same pattern holds in the opposite direction when $\kappa_1 = -0.05$ and $\kappa_2 = 0.75$ (green curve).

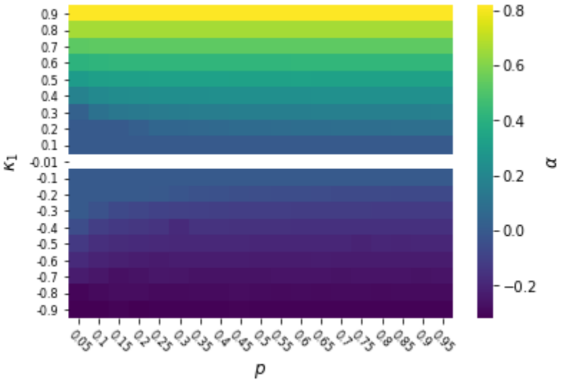

Figure 3: α values of the limit of an evolutionary simulation in which a proportion p of games have κ_1 and a proportion $1-p$ have $\kappa_2 = 0.001$, as a function of κ_1 and p. White cells indicate that the limiting population is not in $\mathcal{R}_K$. The per-parameter penalty is $\epsilon_P = 0.002$.

In Figure 3, we vary both κ_1 and p, keeping $\kappa_2 = 0.001$. For any p, the result from Figure 1 where a rational strategy is not stable for small κ_1 still holds. Likewise, the result that $R(0)$ takes over the population when $\kappa \in [-0.1, 0.1]$ is not sensitive to p. Generalizing the trend from Figure 2, for sufficiently large magnitudes of κ_1, only a minority of games need to have κ far from 0 for strong other-regarding preferences to be stable.

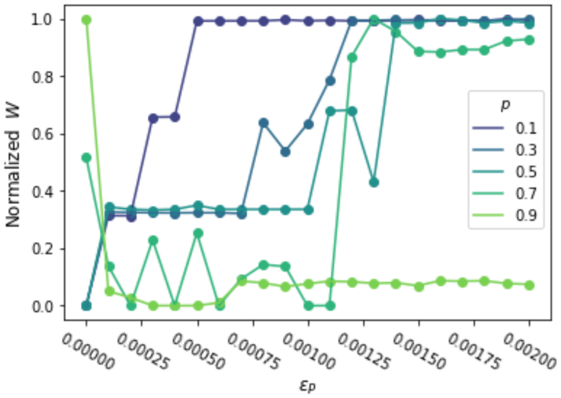

Figure 4: Normalized social welfare of the limit of an evolutionary simulation, in which a proportion p of games have $\kappa_1 = 0.5$ and a proportion $1-p$ have $\kappa_2 = -0.5$, as a function of the per-parameter penalty ϵ_P, for different values of p.

Social welfare in the limiting population as a function of the per-parameter penalty Finally, we show

how the total payoffs of the limiting population vary both with the size of the per-parameter penalty, and with the proportion of games with positive versus negative externalities. Fixing $\kappa_1 = -0.5$ and $\kappa_2 = 0.5$, we vary $\epsilon_P \in \{0, 0.0001, \ldots, 0.0019, 0.002\}$ for each $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. To visualize the transitions between limiting populations of behavioral versus rational strategies, we compute the social welfare $W_{\epsilon_P,p} = u_i + u_j$ averaged over the last two rounds (for some parameter values, the population oscillates) of each evolutionary simulation for penalty ϵ_P and proportion p, shown in Figure 4. [4]

For most values of p, when there is no per-parameter penalty ($\epsilon_P = 0$) the population attains the near-lowest social welfare, where all in the population play the base game Nash equilibrium. The penalty $\epsilon_P = 0.0015$ is sufficient for all populations to converge to an other-regarding rational strategy, which attains the highest social welfare when $p \leq 0.7$ but nearly the lowest when $p = 0.9$, i.e., when most of the games have $\kappa < 0$. For intermediate values of ϵ_P, the population oscillates between $R(0)$ and a behavioral best response to $R(0)$ in each game, usually resulting in social welfare between that of very low or high ϵ_P. The minimum value of ϵ_P necessary for convergence to the rational strategy is largest for values of p closest to 0.5, while only a small penalty is necessary when $p = 0.1$ or 0.9 (see the values of ϵ_P where the curves in Figure 4 plateau). Intuitively, if the large majority of games have the same κ, a behavioral strategy does not profit much from adapting with multiple actions, relative to the complexity costs of playing different actions for two games.

The magnitude of ϵ_P relative to ϵ_R required for other-regarding preferences to be stable might appear unrealistically large, based on these results. We note the distinction between the fixed cognitive costs of developing a rational decision procedure, and the per-use costs of learning heuristics for each context and recognizing when each is appropriate. Cooper [1996] argues that lexicographic, or infinitesimal, complexity costs are appropriate for the former — these start up costs are a tiebreaker between strategies that are otherwise equally capable — while finite non-negligible costs are suitable for the latter. It is therefore plausible that in several evolutionary contexts, the costs of adapting to each interaction from scratch outweigh costs of rationality. Regardless, given the sensitivity of the stable populations in these experiments to ϵ_P, it is important to account for the relative strength of these two factors when predicting the result of an evolutionary process.

5.3 Inefficiency and the secret handshake

Lastly, we discuss the implications of complexity costs for another model that appears to preclude the evolution of certain other-regarding preferences. Recall that we have defined the utility functions of rational strategies as the player's own payoff plus a multiple of the opponent's payoff. Previous work has shown (in finite-action games) that if *all* possible subjective utility functions are permitted, and players observe each other's subjective utility functions, then all stable strategies achieve a Pareto efficient payoff [Dekel et al., 2007, Heller and Mohlin, 2019]. This conclusion follows from the "secret handshake" argument: a player who is indifferent among all action pairs can select an equilibrium that matches any other strategy's action against that strategy, but plays an action achieving an efficient payoff against itself [Robson, 1990]. These results rule out both the base game Nash equilibrium and the ESS in $\mathcal{R}$ of the externality game, which is $R(\alpha^*)$ for $\alpha^* = \frac{\kappa}{2-\kappa} < 1$, while $R(1)$ is the unique efficient rational strategy.

One might suspect, then, that our conclusion from the numerical experiments — i.e., inefficient other-regarding preferences can be stable when agents play multiple games — would not hold after including the strategy classes from Dekel et al. [2007] and Heller and Mohlin [2019]. When we include complexity costs, however, the secret handshake argument does not follow. Let $\mathcal{H}$ be the class of strategies whose subjective utility functions are constant over all action pairs, and which use the equilibrium selection rule described above. Because this strategy requires choosing different Nash equilibria depending on the opponent, we claim that it is more complex than either a behavioral or rational strategy. For $\epsilon_H > \epsilon_R$, let $c(\sigma) = \epsilon_H \mathbb{I}[\sigma \in \mathcal{H}] + \epsilon_R \mathbb{I}[\sigma \in \mathcal{R}]$. Then $B(\text{NE}_i(0,0))$ is still an ESS under the conditions of Proposition 2, with $\mathcal{H}$ added to the strategy space. The proof is straightforward; given a positive penalty, a strategy from $\mathcal{H}$ cannot match the payoff of $B(\text{NE}_i(0,0))$ against itself, by the definition of the base game Nash equilibrium:

$$\begin{aligned}
&\max_{\sigma \in \mathcal{H}} f_i(\sigma, B(\text{NE}_i(0,0))) \\
&\quad = u_i(\text{NE}_i(0,0), \text{NE}_j(0,0)) - \epsilon_H \\
&\quad < f_i(B(\text{NE}_i(0,0)), B(\text{NE}_i(0,0))).
\end{aligned}$$

We conjecture that across multiple games, a sufficiently large penalty ϵ_H would yield similar results to Section 5.2.

6 Discussion

The puzzle that motivated this work was the apparent prevalence of other-regarding preferences, such as altruism and spite, despite the possibility of selection for commitment strategies that ignore the signals of other-

[4] We take the average $\overline{W}_{\epsilon_P,p}$ over 10 runs of each simulation, and given the list $\mathcal{W}_p = \{\overline{W}_{0,p}, \ldots \overline{W}_{0.00225,p}\}$, we normalize each $\tilde{W}_{\epsilon_P,p} = \frac{\overline{W}_{\epsilon_P,p} - \min \mathcal{W}_p}{\max \mathcal{W}_p - \min \mathcal{W}_p}$. (This is for ease of visualization.) Note that the trend in Figure 4 for $p = 0.7$ is exaggerated by normalization; the stable rational value is $\alpha = 0.0004$, so the social welfare does not actually vary significantly.

regarding preferences. Our results suggest that this puzzle stems from a neglect of complexity considerations in previous literature on the evolution of preferences. We considered a class of two-player symmetric games that includes the games used by Bester and Güth [1998] and Possajennikov [2000] to illustrate the stability of altruism and spite. First, via evolutionary stability analysis on a strategy space that combines the direct and indirect approaches, we confirmed that other-regarding preferences are no longer uniquely stable when fixed-action strategies can also evolve. We then showed numerically that, although other-regarding preferences are unstable when agents play a single game under costs of strategy complexity, if the costs of distinct fixed actions across *multiple* games are sufficiently high, other-regarding preferences are stable. These costs also explain why inefficient stable strategies can persist — the flexible "secret handshake" strategy, which has been purported to guarantee that stability implies efficiency, is too complex to invade populations with certain inefficient strategies.

Accounting for the costs of adapting strategies to specific games plausibly sheds light on other phenomena in evolutionary game theory. For example, Boyd and Richerson [1992] argued that a common explanation of cooperation as a product of punishment, e.g., as in tit-for-tat in the repeated Prisoner's Dilemma, proves too much: "Moralistic" strategies, which not only punish noncooperation but also punish those who do not punish noncooperation, can enforce the stability of *any* individually rational behavior. These moralistic strategies require sophisticated recognition of the behaviors that constitute cooperation or punishment in each given game. If some individually rational behavior enforced by a moralistic strategy is only marginally better for the cooperating player than getting punished, another strategy could invade by avoiding the complexity cost of the moralistic strategy, which outweighs the direct fitness cost of being punished. Thus, under complexity costs, the set of evolutionarily stable behaviors may be much smaller. It is also important to note that classes of simple, generalizable utility functions other than those we have considered might evolve. Instead of having utility functions given by their payoff plus a multiple of the other agent's payoff, agents could develop utility functions with an aversion to exploitation or inequity [Huck and Oechssler, 1999, Güth and Napel, 2006].

Besides explaining biological behavior, our model of complexity-penalized preference evolution might also motivate predictions of the behavior of artificial agents, such as reinforcement learning (RL) algorithms. Policies are updated based on reward signals similarly to fitness-based updating of populations in evolutionary models [Börgers and Sarin, 1997]. It is common in RL training to penalize strategies ("policies") according to their complexity, and deep learning researchers have argued that artificial neural networks have an implicit bias towards simple functions [Mingard et al., 2021, Valle-Perez et al., 2019]. Thus, RL agents trained together may develop other-regarding preferences, as far as the assumptions of our model are satisfied by the tasks these agents are trained in. A better understanding of the relationship between complexity costs and the distribution of environments these agents are trained in may help us better understand what kinds of preferences they acquire.

References

Dilip Abreu and Rajiv Sethi. Evolutionary stability in a reputational model of bargaining. *Games and Economic Behavior*, 44(2):195–216, 2003.

Jeffrey S Banks and Rangarajan K Sundaram. Repeated games, finite automata, and complexity. *Games and Economic Behavior*, 2(2):97–117, 1990. ISSN 0899-8256. doi: https://doi.org/10.1016/0899-8256(90)90024-O. URL https://www.sciencedirect.com/science/article/pii/089982569090024O.

Siegfried Berninghaus, Christian Korth, and Stefan Napel. Reciprocity—an indirect evolutionary analysis. *Journal of Evolutionary Economics*, 17:579–603, 02 2007. doi: 10.1007/s00191-006-0053-1.

Helmut Bester and Werner Güth. Is altruism evolutionarily stable? *Journal of Economic Behavior & Organization*, 34(2):193–209, 1998.

Kenneth G Binmore and Larry Samuelson. Evolutionary stability in repeated games played by finite automata. *Journal of Economic Theory*, 57(2):278–305, 1992. ISSN 0022-0531. doi: https://doi.org/10.1016/0022-0531(92)90037-I. URL https://www.sciencedirect.com/science/article/pii/002205319290037I.

Friedel Bolle. Is altruism evolutionarily stable? And envy and malevolence?: Remarks on Bester and Güth. *Journal of Economic Behavior & Organization*, 42(1):131–133, 2000.

Robert Boyd and Peter J. Richerson. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, 13(3):171–195, 1992. ISSN 0162-3095. doi: https://doi.org/10.1016/0162-3095(92)90032-Y. URL https://www.sciencedirect.com/science/article/pii/016230959290032Y.

Tilman Börgers and Rajiv Sarin. Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997. ISSN 0022-0531. doi: https://doi.org/10.1006/jeth.1997.2319. URL https://www.sciencedirect.com/science/article/pii/S002205319792319X.

David J. Cooper. Supergames Played by Finite Automata with Finite Costs of Complexity in an Evolutionary Setting. *Journal of Economic Theory*, 68(1):266–275, 1996. ISSN 0022-0531. doi: https://doi.org/10.1006/jeth.1996.0015. URL https://www.sciencedirect.com/science/article/pii/S0022053196900150.

Eddie Dekel, Jeffrey C. Ely, and Okan Yilankaya. Evolution of Preferences. *The Review of Economic Studies*, 74(3):685–704, 2007. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/4626157.

Martin Dufwenberg and Werner Güth. Indirect evolution vs. strategic delegation: a comparison of two approaches to explaining economic institutions. *European Journal of Political Economy*, 15(2):281–295, 1999. ISSN 0176-2680. doi: https://doi.org/10.1016/S0176-2680(99)00006-3. URL https://www.sciencedirect.com/science/article/pii/S0176268099000063.

Jon Elster. *Rationality*, page 1–42. Cambridge University Press, 1983. doi: 10.1017/CBO9781139171694.002.

Jeffrey C. Ely and Okan Yilankaya. Nash Equilibrium and the Evolution of Preferences. *Journal of Economic Theory*, 97(2):255–272, 2001. ISSN 0022-0531. doi: https://doi.org/10.1006/jeth.2000.2735. URL https://www.sciencedirect.com/science/article/pii/S0022053100927352.

Robert H. Frank. If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience? *The American Economic Review*, 77 (4):593–604, 1987. ISSN 00028282. URL http://www.jstor.org/stable/1814533.

Andy Gardner and Stuart A. West. Greenbeards. *Evolution*, 64(1):25–38, 2010. doi: https://doi.org/10.1111/j.1558-5646.2009.00842.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2009.00842.x.

Félix Geoffroy and Jean-Baptiste André. The emergence of cooperation by evolutionary generalization. *Proc Biol Sci*, 2021.

Werner Güth and Hartmut Kliemt. The indirect evolutionary approach: Bridging the gap between rationality and adaptation. *Rationality and Society*, 10(3):377–399, 1998. doi: 10.1177/104346398010003005. URL https://doi.org/10.1177/104346398010003005.

Werner Güth and Stefan Napel. Inequality aversion in a variety of games: An indirect evolutionary analysis. *The Economic Journal*, 116(514):1037–1056, 2006. ISSN 00130133, 14680297. URL http://www.jstor.org/stable/4121943.

Aviad Heifetz, Chris Shannon, and Yossi Spiegel. What to Maximize If You Must. *Journal of Economic Theory*, pages 31–57, 2003.

Aviad Heifetz, Chris Shannon, and Yossi Spiegel. The Dynamic Evolution of Preferences. *Economic Theory*, 32:251–286, 2007.

Yuval Heller and Erik Mohlin. Coevolution of deception and preferences: Darwin and Nash meet Machiavelli. *Games and Economic Behavior*, 113:223–247, 2019. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2018.09.011. URL https://www.sciencedirect.com/science/article/pii/S0899825618301532.

Steffen Huck and Jörg Oechssler. The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior*, 28(1):13–24, 1999. ISSN 0899-8256. doi: https://doi.org/10.1006/game.1998.0691. URL https://www.sciencedirect.com/science/article/pii/S0899825698906911.

Karin Isler and Carel P. Van Schaik. How humans evolved large brains: Comparative evidence. *Evolutionary Anthropology: Issues, News, and Reviews*, 23 (2):65–75, 2014. doi: https://doi.org/10.1002/evan.21403. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/evan.21403.

Marco A. Janssen. Evolution of cooperation in a one-shot prisoner's dilemma based on recognition of trustworthy and untrustworthy agents. *Journal of Economic Behavior & Organization*, 65(3):458–471, 2008. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2006.02.004. URL https://www.sciencedirect.com/science/article/pii/S0167268106001934.

James Jordan. Bayesian learning in normal form games. *Games and Economic Behavior*, 3(1):60–81, 1991. URL https://EconPapers.repec.org/RePEc:eee:gamebe:v:3:y:1991:i:1:p:60-81.

Ehud Kalai and Ehud Lehrer. Rational Learning Leads to Nash Equilibrium. *Econometrica*, 61(5):1019–1045, 1993. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951492.

Kai A. Konrad and Florian Morath. Evolutionarily stable in-group favoritism and out-group spite in intergroup conflict. *Journal of Theoretical Biology*, 306:61–67, 2012. ISSN 0022-5193. doi: https://doi.org/10.1016/j.jtbi.2012.04.013. URL https://www.sciencedirect.com/science/article/pii/S0022519312001944.

John M McNamara. Towards a richer evolutionary game theory. *Journal of the Royal Society Interface*, 10(88): 20130544, November 2013. ISSN 1742-5689. doi: 10.1098/rsif.2013.0544.

Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A. Louis. Is SGD a Bayesian sampler? Well, almost. *Journal of Machine Learning Research*, 22(79): 1–64, 2021. URL http://jmlr.org/papers/v22/20-676.html.

Efe A. Ok and Fernando Vega-Redondo. On the Evolution of Individualistic Preferences: An Incomplete Information Scenario. *Journal of Economic Theory*, 97(2):231–254, 2001. ISSN 0022-0531. doi: https://doi.org/10.1006/jeth.2000.2668. URL https://www.sciencedirect.com/science/article/pii/S0022053100926681.

Gualtiero Piccinini and Armin W. Schulz. The Ways of Altruism. *Evolutionary Psychological Science*, 5:58–70, 2018.

Alex Possajennikov. On the evolutionary stability of altruistic and spiteful preferences. *Journal of Economic Behavior & Organization*, 42(1):125–129, 2000.

Michael S. Reichert and John L. Quinn. Cognition in contests: Mechanisms, ecology, and evolution. *Trends in Ecology & Evolution*, 32(10):773–785, 2017. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2017.07.003. URL https://www.sciencedirect.com/science/article/pii/S0169534717301799.

Nikolaus Robalino and Arthur Robson. The Evolution of Strategic Sophistication. *The American Economic Review*, 106(4):1046–1072, 2016. ISSN 00028282. URL http://www.jstor.org/stable/43821484.

Arthur J. Robson. Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Theoretical Biology*, 144(3):379–396, 1990. ISSN 0022-5193. doi: https://doi.org/10.1016/S0022-5193(05)80082-7. URL https://www.sciencedirect.com/science/article/pii/S0022519305800827.

Arthur J. Robson. Why Would Nature Give Individuals Utility Functions? *Journal of Political Economy*, 109(4):900–914, 2001. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/10.1086/322083.

Ariel Rubinstein. Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory*, 39(1):83–96, 1986. ISSN 0022-0531. doi: https://doi.org/10.1016/0022-0531(86)90021-9. URL https://www.sciencedirect.com/science/article/pii/0022053186900219.

Amartya Sen. *Foundations of Social Choice Theory: An Epilogue*. Cambridge University Press, Cambridge, 1986.

Guillermo Valle-Perez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rye4g3AqFm.

Matthijs van Veelen. But Some Neutrally Stable Strategies are More Neutrally Stable than Others. Tinbergen Institute Discussion Papers 10-033/1, Tinbergen Institute, March 2010. URL https://ideas.repec.org/p/tin/wpaper/20100033.html.

Matthijs van Veelen and Julián García. In and out of equilibrium II: Evolution in repeated games with discounting and complexity costs. *Games and Economic Behavior*, 115:113–130, 2019. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2019.02.013. URL https://www.sciencedirect.com/science/article/pii/S0899825619300314.

H. Young. The evolution of conventions. *Econometrica*, 61(1):57–84, 1993. URL https://EconPapers.repec.org/RePEc:ecm:emetrp:v:61:y:1993:i:1:p:57-84.

A Proofs

A.1 Proof of Proposition 1

Behavioral Define $u_N = u_i(\text{NE}_i(0,0), \text{NE}_j(0,0))$, the payoff of the Nash equilibrium with egoistic preferences. By the definition of Nash equilibrium of G, since $\max_{a_i} u_i(a_i, \text{NE}_i(0,0)) = u_N$, the strategy $B(\text{NE}_i(0,0))$ is a Nash equilibrium in $\mathcal{S}$. Suppose $a \neq \text{NE}_i(0,0)$. Then we must have $\max_{a_i} u_i(a_i, a) > u_i(a,a)$, because otherwise uniqueness of the Nash equilibrium (assumption 1) would be violated. So $B(a)$ is not a Nash equilibrium in $\mathcal{S}$.

Since the Nash equilibrium of $G(0,0)$ is unique, there is no behavioral strategy $B(a) \neq B(\text{NE}_i(0,0))$ such that $f_i(B(a), B(\text{NE}_i(0,0))) = f_i(B(\text{NE}_i(0,0)), B(\text{NE}_i(0,0)))$. Suppose a rational strategy $R(\alpha)$ satisfies $f_i(R(\alpha), B(\text{NE}_i(0,0))) = f_i(B(\text{NE}_i(0,0)), B(\text{NE}_i(0,0)))$. Then $\arg\max_a V_i^\alpha(a, \text{NE}_i(0,0)) = \arg\max_a \{u_i(a, \text{NE}_i(0,0)) + \alpha u_j(a, \text{NE}_i(0,0))\} = \text{NE}_i(0,0)$. (This is satisfied for $\alpha = 0$.) But this implies that $\text{NE}_i(0,0) = \text{NE}_i(\alpha,\alpha)$. So $f_i(R(\alpha), R(\alpha)) = u_i(\text{NE}_i(\alpha,\alpha), \text{NE}_j(\alpha,\alpha)) = u_N$, and $f_i(B(\text{NE}_i(0,0)), R(\alpha)) = u_N$, therefore $B(\text{NE}_i(0,0))$ is neutrally stable (but not an ESS).

Rational It is immediate that $R(\alpha)$ can only be a Nash equilibrium in $\mathcal{S}$ if it is a Nash equilibrium in $\mathcal{R}$. Let $R(\alpha)$ be such a strategy. $R(\alpha)$ always plays $\text{NE}_i(\alpha,\alpha)$ against itself, so $f_i(R(\alpha), R(\alpha)) = u_i(\text{NE}_i(\alpha,\alpha), \text{NE}_j(\alpha,\alpha))$. Suppose a deviator σ plays a_i. Given assumption 3, for any a_i there exists a β such that $a_i = \text{NE}_i(\beta,\alpha)$. Therefore:

$$\begin{aligned} f_i(\sigma, R(\alpha)) &\leq \max_{a_i} u_i(a_i, \text{BR}_j(a_i;\alpha)) \\ &= \max_\beta u_i(\text{NE}_i(\beta,\alpha), \text{BR}_j(\text{NE}_i(\beta,\alpha);\alpha)) \\ &= \max_\beta u_i(\text{NE}_i(\beta,\alpha), \text{NE}_j(\beta,\alpha)) \\ &= u_i(\text{NE}_i(\alpha,\alpha), \text{NE}_j(\alpha,\alpha)). \end{aligned}$$

Where the last line follows because $R(\alpha)$ is a Nash equilibrium in the space of rational strategies. So $R(\alpha)$ is a Nash equilibrium in $\mathcal{S}$.

Suppose $R(\alpha)$ is neutrally stable in $\mathcal{R}$. By assumption 2, $\text{NE}_i(\alpha,\alpha)$ is the unique action a such that $f_i(B(a), R(\alpha)) = f_i(R(\alpha), R(\alpha))$. Then $f_i(B(\text{NE}_i(\alpha,\alpha)), R(\alpha)) = u_i(\text{NE}_i(\alpha,\alpha), \text{NE}_j(\alpha,\alpha)) = f_i(R(\alpha), R(\alpha))$, and $f_i(B(\text{NE}_i(\alpha,\alpha)), B(\text{NE}_i(\alpha,\alpha))) = u_i(\text{NE}_i(\alpha,\alpha), \text{NE}_j(\alpha,\alpha)) = f_i(R(\alpha), B(\text{NE}_i(\alpha,\alpha)))$. So $R(\alpha)$ is neutrally stable (but not an ESS). On the other hand, if $R(\alpha)$ is not an NSS in $\mathcal{R}$, the same counterexample to neutral stability applies in the expanded space $\mathcal{S}$, thus $R(\alpha)$ is not an NSS in $\mathcal{S}$.

A.2 Proof of Proposition 2

Behavioral The conditions for Nash equilibrium in $\mathcal{B}$ do not change, since this set has the lowest complexity. However, when assessing the stability of $B(\text{NE}_i(0,0))$, it suffices to only consider invader strategies in $\mathcal{B}$, because for a strategy $\sigma \in \mathcal{R}$, $f_i(\sigma, B(\text{NE}_i(0,0))) \leq u_N - \epsilon_R < f_i(B(\text{NE}_i(0,0)), B(\text{NE}_i(0,0)))$. Since the Nash equilibrium of $G(0,0)$ is unique (assumption 1), there is no other behavioral strategy $B(a)$ such that $f_i(B(a), B(\text{NE}_i(0,0))) = f_i(B(\text{NE}_i(0,0)), B(\text{NE}_i(0,0)))$. Thus $B(\text{NE}_i(0,0))$ is an ESS under penalties.

Rational Let $R(\alpha)$ be any rational strategy. Then $f_i(R(\alpha), R(\alpha)) = u_i(\text{NE}_i(\alpha,\alpha), \text{NE}_j(\alpha,\alpha)) - \epsilon_R$. But $f_i(B(\text{NE}_i(\alpha,\alpha)), R(\alpha)) = u_i(\text{NE}_i(\alpha,\alpha), \text{NE}_j(\alpha,\alpha))$, so $R(\alpha)$ cannot be a Nash equilibrium under penalties.

B Details on Numerical Experiments

Each strategy is parameterized by (α, a_1, a_2, n), where the strategy is $R(\alpha)$ if $n = 0$ or $(B(a_1), B(a_2))$ if $n = 1$. A population of size $N = 10$ is initialized with $\alpha, a_1, a_2 \overset{i.i.d.}{\sim} N(0,1)$ and $n \overset{i.i.d.}{\sim} \text{Bern}(0.5)$ for each player in the population. Let $q_1 = 0.01$ and $q_{t+1} = q_t \cdot \frac{t}{t+1}$ if $t \leq 20$, otherwise $q_{t+1} = 0$. The probability of switching to a random strategy from the initialization distribution in round t of evolution is q_t. (We decay q_t to decrease the rate of stochasticity and thus help convergence.) Best responses in the space of $\mathcal{B}_K$ are computed analytically; for $\mathcal{R}_K$, we use gradient ascent.