The Korean Economic Review
Volume 31, Number 2, Winter 2015, 239-266.

Evolution of Altruistic Preferences among Boundedly Rational Agents*

Nayoung Kim** · Sung-Ha Hwang***

We study the co-evolution of social preferences and bounded rationality. In particular, we show that when agents are boundedly rational, altruistic preferences are evolutionarily stable, even in environments that are deemed unfavorable for altruism in the literature. The existing standard result is that when interactions are strategic substitutes and exhibit negative externality, only selfish preferences are evolutionary stable. The key assumption underlying this result is that agents are perfectly rational. Selfish agents are thus able to play the Nash equilibrium, gaining evolutionary advantages over altruists. By relaxing this assumption, we show that altruist preferences can survive among bounded rational agents. The simple intuition is that selfish agents, now with bounded rationality, choose excessive action, which in turn induces altruists to choose an action level closer to the Nash equilibrium–an action level evolutionarily stable in the long run. We combine the level-k model of bounded rationality and the standard evolutionary model of altruistic preferences and characterize for the conditions under which altruism can proliferate in the long run.

JEL Classification: D64, D03
Keywords: Social Preferences, Altruism, Bounded Rationality

I. Introduction

Social preferences such as altruism and reciprocity help us explain why people often cooperate in social dilemmas⁻situations in which unilateral defection gives higher material payoffs (Fehr and Gaechter, 2000; Bowles, 2004). Many researchers have examined whether such social preferences exist and, if so, whether they are

Received: June 29, 2015. Revised: Oct. 13, 2015. Accepted: Nov. 10, 2015.

* We would like to thank two anonymous referees, Jungmin Lee and Kang-Oh Yi for their valuable comments.

** First Author, School of Economics, Sogang University, Seoul, Korea, E-mail: nykim0104@gmail.com

*** Corresponding Author, Associate Professor, School of Economics, Sogang University, Seoul, Korea, E-mail: sunghah@sogang.ac.kr. Sung-Ha Hwang's work was supported by the National Research Foun-dation of Korea Grant funded by the Korean Government(NRF-2013S1A5A8022517).

stable in the long run when preferences evolve over time (Bester and Guth, 1998; Sethi, 2001; Alger and Weibull, 2013). Further, a substantial amount of behavioral and experimental literature reports that agents typically have limited and bounded rationality, following Herbert Simon who proposed bounded rationality as an alternative basis for economic decision making (Simon, 1975; Kahneman, 2003). A natural question then is how do social preferences evolve among agents with limited and bounded rationality? Few studies have addressed this question and hence the relationship between the two traits is poorly understood. In this study, we examine the evolution of altruistic preferences among agents with bounded rationality.

Our study is motivated by various recent empirical and experimental findings of interdependence between preferences and rationality (Brandstater and Guth, 2002; Ben-Ner et al., 2004; Millet and Dewitte, 2007; Oechssler et al., 2009; Liberali et al., 2012; Benjamin et al., 2013; Dittrich and Leipold, 2014). For example, in a laboratory study of Chilean high school students, Benjamin et al. (2013) find elementary school grade point averages predictive of preferences measured at the end of high school. Similarly, some experiments report that cognitive abilities play a prominent role in behavioral bias (see, e.g., Oechssler et al. (2009)). A more direct relation between reasoning and social preferences is found in Dittrich and Leipold (2014), who show through experiments that “subjects who use most steps of reasoning are more selfish, trust less and are less reciprocal than subjects who perform no steps of reasoning.”

In this study, we examine the evolution of altruistic preferences among bounded rational agents in environments where interactions are strategic substitutes and exhibit negative externality. These environments are typically considered unfavorable for the evolution of altruistic preferences. For example, Bester and Guth (1998) show that under these environments, altruistic preferences are not evolutionarily stable and only selfish preferences are stable. In the presence of negative externality and strategic substitutability, the trait adopting the action level of the Nash equilibrium is evolutionarily stable, whereas the trait adopting other action is not. Since selfish agents adopt the Nash equilibrium action, the selfish trait is evolutionarily stable. By contrast, altruistic agents internalizing negative externality choose actions different from the Nash equilibrium action, creating unfavorable evolutionary forces against the selection of altruistic traits. However, the key assumption underlying this argument is that selfish agents know how to play the Nash equilibrium—the assumption of perfect rationality. We show, perhaps surprisingly, that the effect of bounded rationality, by ameliorating the negative effect of unfavorable environments, allows altruism to proliferate and hence altruism and bounded rationality may co-evolve. The simple intuition is that selfish agents, now with bounded rationality, choose excessive action, which in turn induces altruists to choose an action level closer to the Nash equilibrium—an action level evolutionarily stable in the long run.

Specifically, we study a population of agents randomly matched to play a game in which interactions are strategical substitutes and exhibit negative externality—a game commonly used to describe the tragedy of the commons. We then combine the standard evolutionary model of social preferences and the bounded rationality model, namely, the level-k model (see, e.g., Hwang and Bowles (2012) for the model of social preferences and see Stahl and Wilson (1994), Stahl and Wilson (1995), and Nagel (1995) for the level-k model). In particular, we use the so-called indirect evolutionary approach. According to this approach, individual behaviors are based on subjective utility derived from social preferences as well as material payoffs, while the long run success of such behaviors is based only on material payoffs (Guth and Yaari, 1992). Thus, evolutionary success depends only on the material payoffs of agents adopting a certain preference and hence provides a strong case for the evolution of that preference, when it turns out to be evolutionarily successful. The level-k model, introduced by Stahl and Wilson (1994) and Nagel (1995), is also one of the popular bounded rationality models frequently applied in the literature (see Crawford (2013)).

By combining these two elements in our model, we find various conditions for the action levels of altruist and selfish agents as well as the degrees of altruism (Propositions 2, 3) under which altruistic preferences are evolutionarily stable and selfish preferences are not evolutionarily stable. The paper is organized as follows. Section 2 presents the main model, Section 3 provides the main analysis, Section 4 explores some extensions of the main model, and Section 5 concludes the paper.

II. Altruism and Bounded Rationality

Consider a community of a large number of members randomly paired to play a symmetric game whose payoff functions are given by

$$\pi_1(x,y) := xf(x+y) - g(x), \ \ \pi_2(x,y) := yf(x+y) - g(y),$$

where $x, y \geq 0$. Here, $f(x+y)$ specifies the marginal benefit of varying action x, which in turn depends on the joint action of x and y, and function $g(x)$ is the cost of such action. We assume that $f \geq 0, g \geq 0$, and $g' > 0$. Since this is a symmetric game, we write $\pi(x,y) = \pi_1(x,y)$. Then π_2 is given by $\pi_2(x,y) = \pi(y,x)$.

We mainly focus on the case where interactions exhibit negative externality and are strategic substitutes:

$$\frac{\partial \pi}{\partial y}(x,y) = xf'(x+y) < 0 \quad \text{and} \quad \frac{\partial^2 \pi}{\partial x \partial y}(x,y) = f'(x+y) + xf''(x+y) < 0 . \qquad (1)$$

We assume that $f' < 0$ and $f'' \leq 0$, to ensure that the conditions in (1) hold. This game includes the popular Cournot competition model and is also commonly used to study common-pool resource problems such as the tragedy of commons, a social dilemma. We also consider extensions to other cases such as positive externality and strategic complementary interactions. Letting α be the degree of an agent's altruism, we suppose that an individual choosing x against y with the degree of altruism α derives the utility of $u(x,y;\alpha)$:

$$\begin{aligned} u(x,y;\alpha) &:= \pi(x,y) + \alpha\pi(y,x) \\ &= xf(x+y) - g(x) + \alpha[yf(x+y) - g(y)]. \end{aligned} \tag{2}$$

If $\alpha = 0$, we call the agent selfish, because his sole concern is his own material payoff. When $0 < \alpha \leq 1$, the agent is altruistic and, especially when $\alpha = 1$, he cares equally for his own and his partner's payoffs (see Hwang and Bowles (2012)). Function u is sometimes called a behavioral utility function because it determines the behavior of altruists (Bester and Guth, 1998).

Now, we introduce behavioral traits by combining social preferences and the degree of rationality based on a level-k model. First, we suppose that there exists a social norm of action which agents with the least cognitive ability naively adopt. Much behavioral economics literature reports the existence of such norms in social dilemmas (see, e.g., Bowles (2004)). Trait-0 represents the type of agent who chooses an action level x_0 dictated by the social norm and is hence termed a naive social norm adopter. If the action of trait-0 is equal to the expected action when uniformly randomizing over the action space, trait-0 becomes the so-called level-0 type in the level-k literature. To study the co-evolution of altruism and bounded rationality under various circumstances, we define trait-0 as the trait of a naive social norm adopter and examine all possible values of x_0, rather than restricting it to the expected value of the uniform random variable.

To define level-k agents as selfish and altruistic, we first introduce a best response function br :

$$br(y;\alpha) = \arg\max_{x}\{f(y+x)x - g(x) + \alpha[f(x+y)y - g(y)]\},$$

where our assumptions for f and g ensure that "arg max" has a singleton element. Following the standard literature on the level-k model (Stahl and Wilson, 1995; Nagel, 1995), a level-k selfish agent's action $x_S^{(k)}$ is recursively defined as that of an agent who best responds to a level-($k-1$) selfish agent's action $x_S^{(k-1)}$, that is,

$$x_S^{(k)} = br(x_S^{(k-1)};0),$$

where $x_S^{(0)}$ is equal to x_0, the social norm action. We also define a level-k altruist as an agent who best responds to a level-($k-1$) selfish agent *altruistically*. Thus, the action of the level-k altruist, $x_A^{(k)}$, can be given by

$$x_A^{(k)} = br(x_S^{(k-1)};\alpha).$$

By $x_S^{(\infty)}$ we denote the action satisfying $x_S^{(\infty)} = br(x_S^{(\infty)};0)$; i.e., a level-$\infty$ selfish agent with action $x_S^{(\infty)}$ adopts a Nash equilibrium action level, and a level-∞ altruistic agent best responds to the level-∞ selfish agent altruistically: $x_A^{(\infty)} :=$ $br(x_S^{(\infty)};\alpha)$.

Notice that in our definition of a level-k agent, we implicitly assume that a level-k agent, whether selfish or altruistic, best responds to a selfish level-($k-1$). One may define a level-k agent differently. For example, we can imagine that a selfish agent with level-k responds to an altruist with level-($k-1$) by assuming his partner is an altruistic agent. By allowing this possibility, we need to consider 2^k traits in total, whose analysis easily becomes intractable as k increases. Thus, we choose to settle on the current definition to simplify. However, we consider a variation in this assumption in Section 4, where we introduce an alternative assumption that a level-k altruist best responds to a level-($k-1$) altruist. Throughout the paper, we use subscripts S and A to indicate selfish and altruistic agents, respectively.

In the following analysis, we study the co-evolution of altruistic preferences among bounded rational agents by assuming that they adopt some traits from a list of traits consisting of $x_S^{(k)}$'s and $x_A^{(k)}$'s. The following sets are some examples: a trait set consisting of trait-0 and general level-k traits $\mathcal{T}_k = \{x_0, x_S^{(k)}, x_A^{(k)}\}$, and a trait set consisting of trait-0 and selfish and altruistic traits with levels-1 and 2, $\mathcal{T} = \{x_0, x_S^{(1)}, x_A^{(1)}, x_S^{(2)}, x_A^{(2)}\}$.

Now, for a moment, assume three traits with population fraction $p = (p_0, p_S, p_A) \in \Delta$, where p_0, p_S, and p_A are the fractions of trait-0, the selfish trait, and the altruistic trait (for some level), respectively and Δ is the simplex in $\mathbb{R}^n$. When $p_i = 1$ for trait i, the state corresponds to a monomorphic (or homogeneous) state where all agents adopt trait-i. For instance, $p = (0,1,0)$ refers to the monomorphic population of all selfish agents. Thus we indicate by writing $p_i = 1$ the monomorphic population state at which all agents adopt trait-i. Since individuals are randomly matched to play the underlying game, the expected *material* fitness of individuals with each trait at population state p is given as follows:

$$\Pi_0(p) = \pi(x_0, x_0)p_0 + \pi(x_0, x_S)p_S + \pi(x_0, x_A(\alpha))p_A$$
$$\Pi_S(p) = \pi(x_S, x_0)p_0 + \pi(x_S, x_S)p_S + \pi(x_S, x_A(\alpha))p_A$$

$$\prod_A(p) = \pi(x_A(\alpha), x_0)p_0 + \pi(x_A(\alpha), x_S)p_S + \pi(x_A(\alpha), x_A(\alpha))p_A, \tag{3}$$

where $x_0, x_S, x_A(\alpha)$ are the actions chosen by trait-0, selfish, and altruistic agents, respectively.

Note that in equation (3), the expected fitness of altruists ($\prod_A(p)$) depends only on material payoffs (π), not on behavioral utility function u, and the degree of altruism (α) can only "indirectly" affect the fitness of altruists thru $x_A(\alpha)$. That is, evolutionary selection does not account for the behavioral satisfaction derived from altruism (the term $\alpha[yf(x+y) - g(y)]$ in equation (2)) but determines the long-run success of the selfish and altruistic traits, based on the same criteria–only on the material payoffs. This is because evolutionary selection occurs via an indirect effect of preferences on behaviors; the approach adopted in specification (3) is called an indirect evolutionary approach (Guth and Yaari (1992); Bester and Guth (1998)).

Observe that if we interpret $p \in \Delta$ as a mixed strategy of an underlying game, the payoff to strategy $p \in \Delta$ when played against $q \in \Delta$ is given by

$$\prod(q, p) := q_0 \prod_0(p) + q_S \prod_S(p) + q_A \prod_A(p). \tag{4}$$

Standard literature on evolutionary game theory (for example, Weibull (1995)) defines an evolutionarily stable strategy as follow: $p \in \Delta$ is an evolutionarily stable strategy if for every strategy $q \neq p$ there exists some $\bar{\varepsilon} > 0$ such that

$$\prod(p, (1-\varepsilon)p + \varepsilon q) > \prod(q, (1-\varepsilon)p + \varepsilon q) \tag{5}$$

for all $0 < \varepsilon < \bar{\varepsilon}$. The idea behind this definition is as follows. Consider a large population of agents playing an incumbent strategy $p \in \Delta$. Suppose that a small group of mutants (ε) playing some other mutant strategy $q \in \Delta$ enters the population. Since the share of mutants is ε, the post-entry population consists of $(1-\varepsilon)p + \varepsilon p$. If individuals are randomly matched to play the game, the expected postentry payoff to the incumbent strategy p is given by $\prod(p, (1-\varepsilon)p + \varepsilon q)$ and that of the mutant strategy $\prod(q, (1-\varepsilon)p + \varepsilon q)$. Thus, (5) requires that the post-entry payoff to the incumbent strategy is greater than the mutant strategy, as long as the fraction of the mutant population remains small, $\varepsilon \approx 0$.

Since we wish to study the population state indicating the fractions of agents adopting each trait, we introduce the following definition.

Definition 1. (i) A population state $p \in \Delta$ is an *evolutionarily stable state* if for every state $q \neq p$ there exists $\bar{\varepsilon} > 0$ such that

$$\prod(p, (1-\varepsilon)p + \varepsilon q) > \prod(q, (1-\varepsilon)p + \varepsilon q) \tag{6}$$

for all $0 < \varepsilon < \bar{\varepsilon}$.
(ii) A population state $q \in \Delta$ *cannot invade* a population state $p \in \Delta$ if there exists $\bar{\varepsilon} > 0$ such that

$$\prod(p,(1-\varepsilon)p+\varepsilon q) > \prod(q,(1-\varepsilon)p+\varepsilon q) \tag{7}$$

for all $0 < \varepsilon < \bar{\varepsilon}$.
(iii) We say that trait-i is an *evolutionarily stable trait* (**ES** trait) if the monomorphic state ($p_i = 1$) at which all agents adopt action x_i is an evolutionarily stable state and trait-i *cannot invade* trait-j if the monomorphic state ($p_i = 1$) *cannot invade* the monomorphic state ($p_j = 1$).

When we check whether a state p is an evolutionarily stable state or not, we sometimes call p an incumbent population state and an alternative state q a mutant population state (see also Sandholm (2010) for evolutionarily stable states).

Note that the definition of an evolutionarily stable state is based on a static concept. To study the dynamics explicitly, again following the standard literature on evolutionary games (Weibull, 1995), we define the replicator dynamics as follows:

$$\frac{dp_l}{dt} = p_l(\prod_l(p) - \prod(p,p)) \quad \text{for} \quad l \in \{0,S,A\}. \tag{8}$$

In other words, in the replicator dynamics, the growth rate $\frac{dp_l}{dt} / p_l$ of the population fraction of trait-l is given by the difference between the material fitness of trait-l, $\prod_l(p)$, and the average payoff for the population, $\prod(p,p)$. This highlights the Darwinian selection idea that the higher the fitness of trait-l relative to the average fitness of all traits, the more likely it is for trait-l to proliferate in the long run.

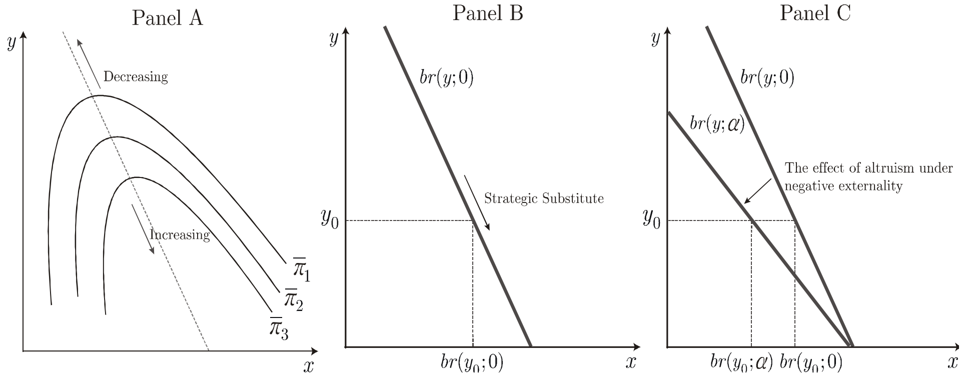
III. Is Altruism with Bounded Rationality Evolutionarily Stable?

A handy tool to study evolutionary stability in our analysis is an iso-(material) payoff locus:

$$\{(x,y) \in \mathbb{R}^2 \mid \pi(x,y) = \bar{\pi}, \text{ for given payoff level } \bar{\pi}\}.$$

Panel A in Figure 1 shows the familiar shapes of iso-payoff loci. Note that negative externality implies that given x, a lower value of y yields a higher payoff for an agent choosing x. Thus, the lower an iso-payoff locus, the higher is the level of payoffs; that is, $\bar{\pi}_1 < \bar{\pi}_2 < \bar{\pi}_3$. Furthermore, because of strategic substitutability, one's best response varies inversely with his opponent's response. Thus, the best response line satisfying $\partial\pi / \partial x = 0$ is negatively sloped (Panel B). In Panel C, we compare the best responses of altruist and selfish agents. In the presence of negative externality, an altruist internalizes the effect of negative externality, reducing his action. Thus, for a given level y_0, the best response of the altruist agent, $br(y_0;\alpha)$, is smaller than the best response of the selfish agent, $br(y_0;0)$ (in Panel C, Figure 1).

[Figure 1] Isoprofit loci, strategic substitutes, and negative externality



Note: Panel A shows iso-(material) payoff loci; every point on the iso-payoff locus gives the same level of payoffs. Further, a lower iso-payoff locus represents a higher payoff level due to negative externality; that is, $\bar{\pi}_1 < \bar{\pi}_2 < \bar{\pi}_3$. Panel B shows the best response function of selfish agents ($br(y;0)$). Panel C gives the best response function of altruistic agents ($br(y;\alpha)$), illustrating how an altruist internalizes negative externality.

Now we derive some sufficient conditions for an ES trait. Let p be an incumbent population state. Following the standard arguments in evolutionary game theory (Weibull, 1995; Hofbauer and Sigmund, 1998), it is easy to show that if the incumbent population state p satisfies

$$\Pi(p,p) > \Pi(q,p) \quad \text{for any state } q, \tag{9}$$

condition (6) in Definition 1 holds (see Appendix A for more details). That is, the incumbent population state p is evolutionary stable. Recall that p satisfying $p_l = 1$ is the monomorphic state consisting of trait-l (i.e., the state at which everyone adopts trait-l). We set $p_l = 1$ and if

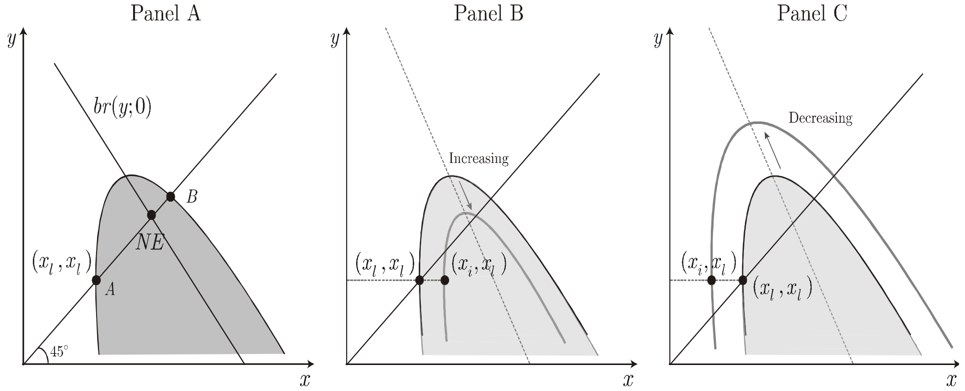
$$\pi(x_l, x_l) > \pi(x_i, x_l) \quad \text{for any other trait } i \tag{10}$$

holds, then condition (10) clearly implies condition (9) (see again Appendix A for details). Thus, condition (10) is a sufficient condition for the monomorphic state $p_l = 1$ to be an evolutionarily stable state. Further, if

$$\pi(x_l, x_l) < \pi(x_i, x_l) \quad \text{for some trait } i \tag{11}$$

holds, condition (6), where $p_l = 1$, cannot be satisfied (see Appendix A). Thus, condition (11) is a sufficient condition to ensure that the monomorphic population state $p_l = 1$ cannot be an evolutionarily stable state. Hereafter, we call $\pi(x_l, x_l)$ the incumbent payoff of trait-l and $\pi(x_i, x_l)$ the mutant trait-i's payoff against incumbent trait-l.

[Figure 2] Illustration of conditions (10) and (11)



Note: The shaded region in Panel A gives the points at which payoffs are higher than the payoff of the incumbent population, $\pi(x_l, x_l)$. The point (x_i, x_l) in Panel B gives a higher payoff than (x_l, x_l), since it is located within the shaded region in Panel A. Panel C shows the point (x_i, x_l) that gives a lower payoff than point (x_l, x_l), since it lies outside the shaded region.

The analytic advantage of sufficient conditions (10) and (11) is that they can be easily checked with iso-payoff loci. To explain this, first note that the payoff for incumbent population $\pi(x_l, x_l)$ in condition (10) or (11) can be represented by an iso-payoff curve passing through a point on the 45 degree line (Panel A of Figure 2). The payoff for mutant trait-i's payoff against incumbent trait-l, $\pi(x_i, x_l)$, can also be identified by the point horizontally away from (x_l, x_l) as in Panels B and C. In Panel A of Figure 2, the shaded region of the points gives a higher payoff than $\pi(x_l, x_l)$. Thus, if point (x_i, x_l) is located in the shaded region, trait-l cannot be an ES trait because of condition (11) (Panel B of Figure 2). If for any other trait-$i \neq l$, point (x_i, x_l) is located outside the shaded region as in Panel C of Figure 2, trait-l is an ES trait because of condition (10).

Intuitively, if an incumbent trait action x_l is small compared to the Nash Equilibrium (NE, point A in Panel A of Figure 2), a mutant trait with an action larger than x_l (but smaller than NE) can invade. In contrast, if an incumbent trait action x_l is large compared to the NE (point B in Panel A of Figure 2), a mutant trait with an action smaller than x_l (but larger than the NE) can invade. Thus, whenever there is a trait with action too small or too large compared to the NE, a trait choosing an action closer to the NE can invade this trait. In other words, the closer a trait's action to an NE, the more likely it is for that trait to sustain as an ES trait. This leads to the following lemma.

Lemma 1. *Let x_{NE} be the NE action level; that is, $x_{NE} = br(x_{NE};0)$. Suppose that either $x < y \le x_{NE}$ or $x > y \ge x_{NE}$ holds. Then,*

$$\pi(y,y) > \pi(x,y) \text{ and } \pi(y,x) > \pi(x,x). \tag{12}$$

Proof. Letting $x < y \le x_{NE}$, we show that (12) holds. The case where $x > y \ge x_{NE}$ follows similarly. Let z be fixed. First, note that from our assumptions of f and g, $\pi(\cdot,z)$ is concave with respect to the first argument and the first order partial derivative of $\pi(\cdot,z)$ with respect to the first argument vanishes at $br(z;0)$. Thus, $\pi(\cdot,z)$ is increasing over $[0,br(z;0)]$. Then, because of strategic substitutability, $y \le x_{NE}$ implies $x_{NE} \le br(y;0)$. Thus, we have

$$x < y \le br(y;0).$$

By taking $z = y$, since $\pi(\cdot,z)$ is increasing over $[0,br(z;0)]$, we find that

$$\pi(x,y) < \pi(y,y) \le \pi(br(y;0),y).$$

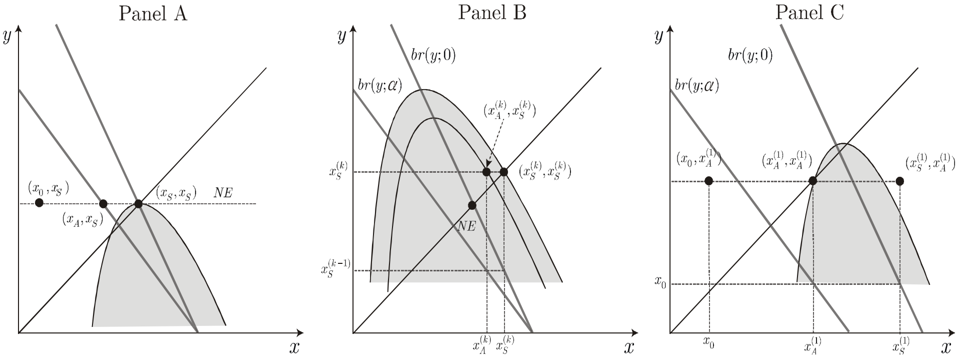
Similarly, from strategic substitutability, we have $x_{NE} < br(x;0)$, and exactly same arguments yield $\pi(x,x) < \pi(y,x) \le \pi(br(x;0),x)$. □

Using these tools, we first study level-∞ traits as a benchmarking case. This corresponds to the study of Bester and Guth (1998) on the evolution of altruism among perfectly rational players. We consider a trait set consisting of trait-0, level-∞ selfish trait, and level-∞ altruistic trait, $\mathcal{T}_\infty = \{x_0, x_A^{(\infty)}, x_S^{(\infty)}\}$. The (almost) immediate consequence of Lemma 1 is that the level-∞ selfish trait is an ES trait, whereas the level-∞ altruist trait cannot be an ES trait. In the case of perfectly rational agents (level-∞ model) with strategic substitutability and negative externality, the selfish agents always adopt the NE action (Panel A in Figure 3). Then, Lemma 1 implies that the payoff of incumbent selfish agents is greater than those of mutant altruistic agents and mutant trait-0 agents $(\pi(x_S^{(\infty)}, x_S^{(\infty)}) > \pi(x_A^{(\infty)}, x_S^{(\infty)})$ and $\pi(x_S^{(\infty)}, x_S^{(\infty)}) > \pi(x_0, x_S^{(\infty)}))$. That is, condition (10) is satisfied and the level-∞ selfish trait is an ES trait. To check whether an altruistic trait is an ES trait or not, note, again from Lemma 1, that the payoff of mutant selfish agents adopting the NE is greater than that of incumbent altruist agents $(\pi(x_S^{(\infty)}, x_A^{(\infty)}) > \pi(x_A^{(\infty)}, x_A^{(\infty)}))$. Thus, from (11), the level-∞ altruistic trait cannot be an ES trait.

Proposition 1. (Level-∞ Selfish and Altruistic Agents). *Consider the evolution of traits among the naive, level-∞ selfish, and level-∞ altruistic trait agents: $\mathcal{T}_\infty = \{x_0, x_S^{(\infty)}, x_A^{(\infty)}\}$. Then, the level-$\infty$ selfish trait is an ES trait, whereas the level-∞ altruistic trait is not an ES trait.*

Proof. See Appendix A. □

[Figure 3] Illustrations of Propositions 1, 2, and 3



Note: In Panel A, the selfish agents choose the NE and the altruist's action is smaller than the NE (Proposition 1). Since the combination of any mutation trait's action, x, and the incumbent selfish traits' action, x_S, (x, x_S), must lie in the dotted line in Panel A, it is clear that no other mutant trait's action can obtain a payoff greater than $\pi(x_S, x_S)$. Panel B illustrates Proposition 2 where altruist trait agents can invade the incumbent selfish population, since the point $(x_A^{(k)}, x_S^{(k)})$ is located within the shaded region. Finally, Panel C illustrates Proposition 3 where neither the selfish trait agents $(x_S^{(1)})$ nor the naive norm adopter (x_0) as mutants can invade the incumbent altruist population, since $(x_0, x_A^{(1)})$ and $(x_S^{(1)}, x_A^{(1)})$ are located outside the shaded region, implying that the altruistic trait is an ES trait.

However, as emphasized earlier, the above argument assumes that selfish agents know how to play the NE. When the agents' rationality is bounded, this need not be the case. Suppose that an altruist with level-k and a selfish agent with level-k best respond to a certain level-($k-1$), $x_S^{(k-1)}$ (see Panel B in Figure 3). Further, assume that $x_S^{(k-1)}$ is smaller than the NE. In this case, the actions of the altruist and selfish agents with level-k $(x_A^{(k)}, x_S^{(k)})$ are larger than the NE. Moreover, the selfish trait's action $(x_S^{(k)})$ is larger than the altruistic trait's action $(x_A^{(k)})$, because the selfish agent does not internalize negative externality. Thus, the selfish trait cannot be an ES trait (Lemma 1 and condition (11)). In short, selfish agents choose excessive action compared to the NE. Panel B in Figure 3 shows that a mutant altruistic trait can invade an incumbent selfish trait, since the point $\pi(x_A^{(k)}, x_S^{(k)})$ is located within the shaded region of points with payoff greater than $\pi(x_S^{(k)}, x_S^{(k)})$.

Proposition 2. (Level-k Selfish Agents). *Let* $1 \le k < \infty$. *Consider the evolution of traits among the naive, level-*k *selfish, and level-*k *altruistic trait agents:* $\mathcal{T}_k = \{x_0, x_S^{(k)}, x_A^{(k)}\}$. *If* $x_{NE} < x_A^{(k)}$*, then the level-*k *selfish trait is not an ES trait.*

Proof. See Appendix A. □

When agents are boundedly rational, an altruistic trait can also be an ES trait. Unlike Proposition 2, this is more likely to occur when the action of level-k altruists is smaller than the NE, but that of level-k selfish agents is larger than the NE. The shaded region in Panel C of Figure 3 shows the points with payoffs greater than the incumbent altruistic trait's payoff, $\pi(x_A^{(1)}, x_A^{(1)})$. In this case, the mutant selfish trait's payoff $(\pi(x_S^{(1)}, x_A^{(1)}))$ is less than the incumbent altruistic trait's payoff $(\pi(x_A^{(1)}, x_A^{(1)}))$ since point $(x_S^{(1)}, x_A^{(1)})$ is located outside the shaded region. If the naive trait action, x_0, gives a lower payoff as a mutant trait (i.e., $\pi(x_0, x_A^{(1)}) < \pi(x_A^{(1)}, x_A^{(1)})$), the altruist trait is an ES trait. When $k = 1$, we can easily show that a level-1 altruist trait is indeed an ES trait (point $(x_0, x_A^{(1)})$ is again located outside the shaded region). For an arbitrary level-k, there exists a level-k altruist trait that cannot be invaded by the level-k selfish trait. This is the content of the following proposition.

Proposition 3. (Level-k Altruistic Agents). *Let* $1 \le k < \infty$. *Consider the evolution of traits among the naive, level-*k *selfish, and level-*k *altruistic trait agents:* $\mathcal{T}_k = \{x_0, x_S^{(k)}, x_A^{(k)}\}$. *There exists* $\underline{x}_A < x_{NE}$ *such that if* $\underline{x}_A < x_A^{(k)} < x_{NE} < x_S^{(k)}$*, then the level-*$k$ *selfish trait cannot invade the level-*k *altruistic trait. Moreover, if* $k = 1$*, the level-1 altruistic trait is an ES trait.*

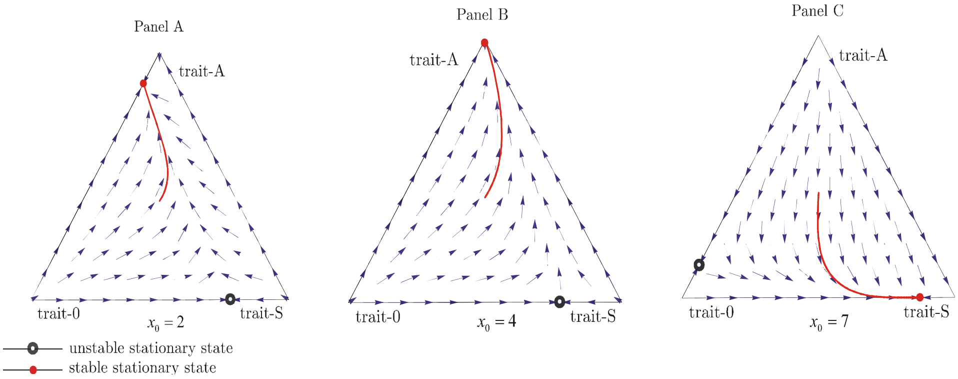
Proof. See Appendix A. □

Recall that we define an evolutionarily stable state as a static concept. How does the population fraction adopting each trait evolve over time? To explore this question, we use simulations of the replicator dynamics in equation (8).

The panels in Figure 4 show the time evolution of the population fraction of each trait under the replicator dynamics. In each simplex, the vertices at (1,0,0), (0,1,0), and (0,0,1) correspond to the monomorphic population states of all altruist (trait A), trait-0, and selfish (trait-*S*) agents, respectively. Further, the closer a point to the given vertex, the more abundant is the corresponding trait in the population.

From Panels A and B of Figure 4, when the social norm action (x_0) is small, the altruistic trait, whether in a mixed or monomorphic population, can survive in the long run under the replicator dynamics. This confirms Propositions 2 and 3. In Panel A, the coexistence of altruistic and trait-0 agents is globally stable, but in Panel B, the monomorphic altruistic trait is globally stable. Note that even if the majority of agents in the population are selfish when the system starts (near the right bottom corner), all agents eventually adopt altruistic traits.

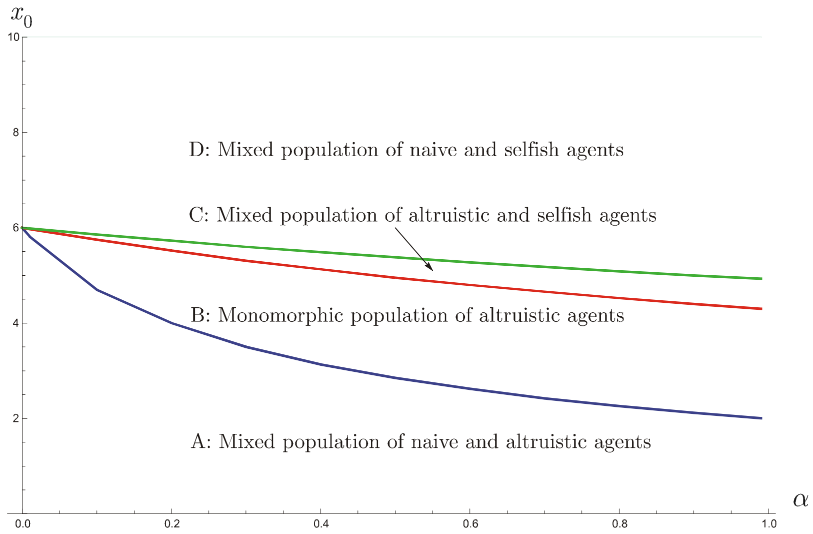
[Figure 4] The solution orbits of the replicator dynamics



Note: Panels show the solution orbits of the replicator dynamics (8). We take $f(x+y) := 300-(x+y)^2$, $g(x) := x^2$, and $\alpha = \frac{1}{2}$. In this case, x_{NE} is 6 and $\underline{x}_A$ is 5.71. Panel A shows the results of Proposition 2; that is, the level-1 selfish trait cannot be an ES trait. In Panel A, when the action of trait 0 is given by 2, the actions of level-1 altruist and selfish agents are 8.07 and 8.41, respectively. Panel B presents the results of Proposition 3. In this case, the level-1 altruistic trait is an ES trait when the actions of the level-1 altruist and selfish agents are 6.47 and 7.18, respectively, in response to the action of trait 0 agents, $x_0 = 4$. Panel C illustrates the coexistence of the trait 0 and level-1 selfish agents, when the trait 0 agent's action level is 7 and the level-1 altruist and selfish agents choose 4.10 and 5.42, respectively.

Can the selfish trait also survive in the long run among boundedly rational agents? Panel C in Figure 4 illustrates this possibility. In Panel C, the coexistence of selfish and trait-0 agents is globally stable. This occurs when the social norm action (x_0) is sufficiently large so that the selfish agent's action (hence the altruistic agent's action) is smaller than the NE action level. In this case, the selfish trait with action larger than the altruistic trait but still smaller than NE chooses an action closer to the NE action level (see Lemma 1). Thus, the selfish trait can do better than altruists. This is how a substantial proportion of selfish agents survive in the long run under the replicator dynamics.

[Figure 5] Stable states under the replicator dynamics in the space of (α, x_0)



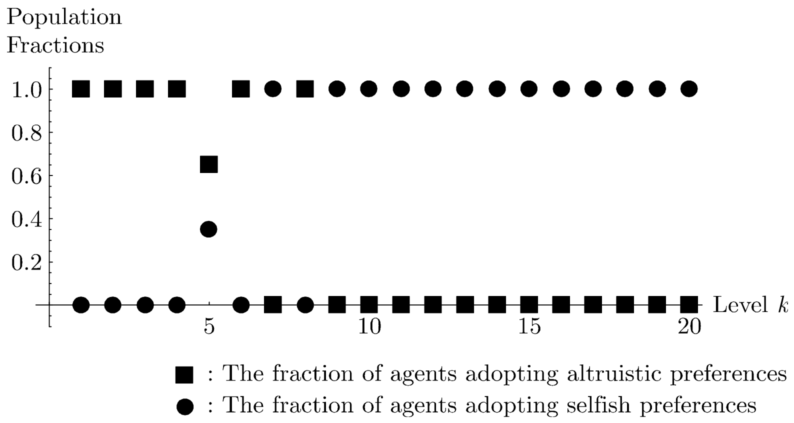
Note: Each region shows the asymptotically stable state under the replicator dynamics of a given combination of degree of altruism (α) and the social norm action (x_0). We take $f(x+y):=300-(x+y)^2$ and $g(x):=x^2$.

Recall that our definitions of the level- k selfish and altruistic traits depend on the social norm action x_0 and the degree of altruism α . To show the effect of the social norm action (x_0) in relation to the degree of altruism (α), we present Figure 5 using simulations. Figure 5 shows the regions of the degree of altruism (α) and the social norm action (x_0) where various traits are stable. For instance, in region A, the mixed population of the naive and altruistic traits is stable under the replicator dynamics.

First, Figure 5 shows that as the degree of altruism increases, the range of x_0 supporting the altruistic trait as a globally stable state (region B) enlarges. Thus, the more altruistic the agents, the more likely are the altruistic preferences to be evolutionary stable. Second, for a given level of α , when x_0 is in the intermediate range, the altruistic trait is stable. Thus, when the social norm action is in the intermediate range (not in the extreme range), the altruistic trait is more likely to be stable. If we interpret the social norm action x_0 as an expected level under uniform randomization, the action level would lie in the middle range. Thus, under this situation, we expect that altruism co-evolves with bounded rationality.

We show that when the degree of rationality k is given by level-1, the altruistic trait is evolutionarily stable (Proposition 3), whereas when the degree of rationality k is given by level- ∞ , the selfish trait is evolutionarily stable (Proposition 1). Thus, there may be a monotonic relation between the stability of the altruistic trait and the degree of rationality k . However, in the presence of strategic substitute interactions, if a level-1 selfish action is larger than the NE, the level-2 selfish agent's action is smaller than the NE, and the level-3 selfish agent's action is larger than the NE and so on. Thus, agents with a level even number k may behave differently from agents with a level odd number k, because the best response is to adopt the opposite action, which could lead to a cyclic relation between the stability of traits and the degree of rationality k. As Figure 6 shows, this is indeed the case around a relatively low degree of rationality ($k = 5,6,7$). However, as the degree of rationality increases, the monomorphic state where every agent adopts the selfish trait is stable. This is because as the degree of rationality increases, the selfish action level converges to the NE while the altruistic action converges to another state, namely a hypothetical NE in which both players are altruists.

[Figure 6] Relations between the degree of rationality (k) and the stability of traits in the replicator dynamics



Note: The horizontal axis and vertical axis represent the degree of rationality k and the population fraction of agents adopting each trait in the replicator dynamics, respectively. We use $f(x+y) = 300-(x+y)$ and $g(x) = 3x$.

IV. Alternative Assumptions about Traits, Externalities and Strategic Interactions

4.1. Alternative Traits

So far, we assumed that a level-k altruistic agent best responds to a level-($k-1$) selfish agent. Alternatively, the level-k altruistic agent might believe that the level-($k-1$) agent to whom he best responds is an *altruist* as well. To examine whether this modification in assumption changes our results, we consider the following variation in the baseline model. Suppose that there are five traits–trait-0, level-1 selfish trait, level-2 selfish trait, level-1 altruistic trait, and level-2 altruistic trait. Except for level-2 altruistic trait, the other traits are defined as in the previous section. A level-2 altruist is assumed to best respond to a level-1 altruist, thus the action of a level-2 altruist, x_{AA}, is defined to be $x_{AA} = br(x_A;\alpha)$. To simplify, we use the following linear value and cost functions:

$$f(x+y) = a - b(x+y)\text{, and } g(x) = cx\,.$$

Under these assumptions, we explicitly compute the actions of a level-1 selfish agent, x_S, and a level-1 altruist agent, x_A, for the social norm action, x_0, as follows:

$$x_S = br(x_0;0) = \frac{a-c-bx_0}{2b},\quad x_A = br(x_0;\alpha) = \frac{a-c-b(1+\alpha)x_0}{2b}\,. \tag{13}$$

Note that if $\alpha = 0$, then $x_A = x_S$ and if $\alpha > 0$, $x_A < x_S$, because of negative externality. The actions of a level-2 selfish agent, x_{SS}, and a level-2 altruist, x_{AA}, are as follows:

$$x_{SS} = br(x_S;0) = \frac{a-c+bx_0}{4b},\quad x_{AA} = br(x_A;\alpha) = \frac{(1+\alpha)(a-c)+b(1+\alpha)^2 x_0}{4b}\,.$$

Again, if $\alpha = 0$, then $x_{SS} = x_{AA}$. However, differing from (13), if $\alpha > 0$, then $x_{AA} > x_{SS}$. This is because when altruistic agents best respond to another altruist, since an altruist partner's action is smaller than a selfish agent's action, the best responding altruistic agents react more because of strategic substitutability.

Recall our basic intuition that an action closer to the NE is more likely to be evolutionarily stable than actions further from the NE (Lemma 1). As Figure 5 illustrates, the action of altruistic preferences is closer to the NE than the action of selfish agents, when the social norm action (x_0) is in the intermediate range. Thus, we expect that when x_0 takes the intermediate value, the altruistic trait is evolutionary stable, whereas when x_0 is high, the selfish trait is evolutionarily stable.

Proposition 4. *Consider the evolution of traits among the naive, level-*$1, 2$ *selfish, and level-*$1, 2$ *altruistic trait agents:* $\mathcal{T} = \{x_0, x_S, x_A, x_{SS}, x_{AA}\}$. *There exists* $\underline{y}, \overline{y}$ *such that*

i) when $\overline{y} < x_0$, *a level-2 selfish trait is an ES trait,*

ii) when $\underline{y} < x_0 < \overline{y}$, *a level-1 altruistic trait is an ES trait, and*

iii) when $x_0 < \underline{y}$, *a mixed population state consisting of level-1 altruist and level-2* *selfish agents is evolutionarily stable.*

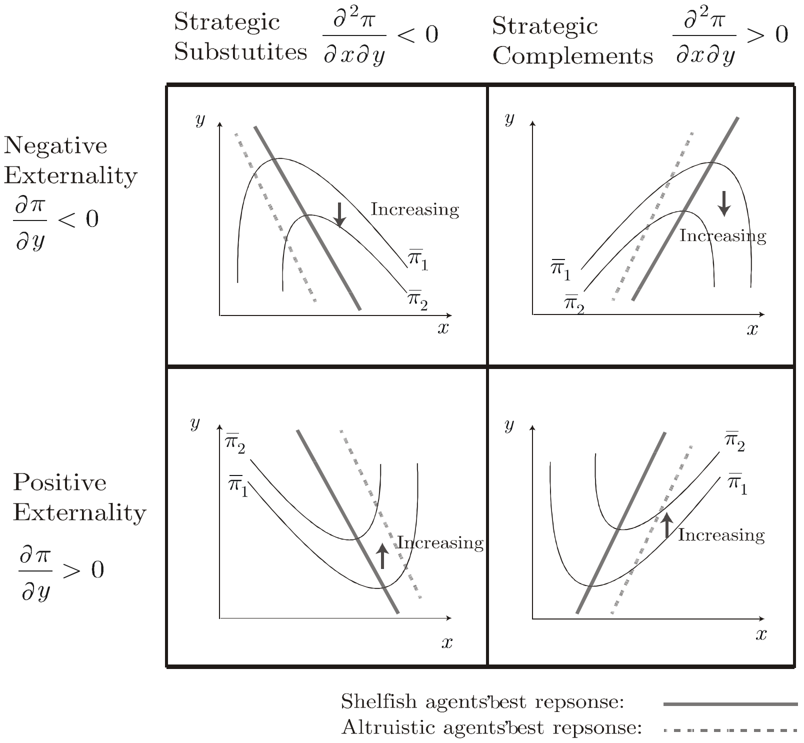
Proof. See Appendix A. □

Thus, Proposition 4, along with the numerical simulations in Figures 4 and 5, corroborates our main results, Propositions 2 and 3.

4.2. Negative Versus Positive Externality and Strategic Complements Versus Substitutes

In this section, we consider alternative assumptions for externality and strategic relations and examine the roles of externality and strategic interactions in facilitating the evolution of altruistic preferences in the long run (see Figure 7). On the one hand, the strategic substitute interaction induces the selfish agents to adopt an *opposite* action to the social norm, while the strategic complement interaction induces the selfish agents to adopt a *similar* action to the social norm. On the other

[Figure 7] Negative vs positive externalities and strategic substitutes vs complements.


Strategic Substutites $\frac{\partial^2 \pi}{\partial x \partial y} < 0$
Strategic Complements $\frac{\partial^2 \pi}{\partial x \partial y} > 0$
Negative Externality $\frac{\partial \pi}{\partial y} < 0$
Positive Externality $\frac{\partial \pi}{\partial y} > 0$
y
x
Increasing
$\bar{\pi}_1$
$\bar{\pi}_2$
Shelfish agents'best repsonse:
Altruistic agents'best repsonse:

Note: These four panels shows four possible combinations of externalities and strategic relations. Arrows show the direction in which the payoffs increase.

hand, negative externality induces the altruistic agents to adopt an action level *smaller* than the selfish agents, while positive externality induces the altruistic agents to adopt an action level *larger* than the selfish agents. By combining these two effects, we will illustrate that our main results may be extended to general settings as follows: in environments where a social norm action induces selfish agents to adopt an extreme action, whether excessive or insufficient, the altruistic preference may be evolutionarily stable in the long run.

More precisely, consider the effect of strategic relations (see Figure 7 again) first:

$$\text{Strategic substitutes: } \frac{\partial^2 \pi}{\partial x \partial y} < 0 \quad \text{Strategic complements: } \frac{\partial^2 \pi}{\partial x \partial y} > 0 .$$

Recall that $\partial^2 \pi / \partial x \partial y$ determines the slope of the best response function of a selfish agent. Thus, the larger social norm action induces a larger (or smaller) action of selfish agents under strategic complementary interactions (or strategic substitutable interactions). Next, in the presence of externality, an altruist internalizes the effect of externality, thus adopting an action level larger (or smaller) than a selfish agent's action level in the presence of positive externality (or negative externality, respectively). Indeed, from $u(x,y,\alpha) = \pi(x,y) + \alpha\pi(y,x)$, we easily verify that

$$\frac{\partial br}{\partial \alpha} > 0 \quad \text{if and only if} \quad \frac{\partial \pi}{\partial y} > 0 .$$

Also, we also assume the following so-called stability condition for the NE (see Glaeser and Scheinkman (2000) and the appendix in Hwang and Bowles (2014))

$$\left| \frac{\partial br}{\partial y} \right| < 1 \tag{14}$$

Equation (14) requires that one's action change does not entail more than a proportional change in the other's action; otherwise, an initial small change in one's action will produce ever increasing chain reactions, destabilizing the NE.

Can the altruistic trait survive in each of the four cases in Figure 7? We have already showed that this is the case under the assumptions of negative externality and strategic substitutes. As we emphasized, the crucial condition for this result is that a trait which adopts an action level closer to the NE is more likely to be evolutionarily successful. In particular, Lemma 1 is used to show the sufficiency of evolutionary stability. The following lemma provides a generalization of Lemma 1.

Lemma 2. *Suppose that condition (14) holds and* $x < y \le x_{NE}$ *or* $x_{NE} \le y < x$ *holds. Then we have*

$$\pi(x,y) < \pi(y,y) \tag{15}$$

Proof. In this proof, we will ignore the dependence of br on α. Let $x < y \le x_{NE}$. The other case follows similarly. We first show that $y \le br(y)$. If $y = x_{NE}$, then this follows since $x_{NE} = br(x_{NE})$. Thus we suppose that $y < x_{NE}$. Then by integrating equation (14) over $[\,y, x_{NE}\,]$, we obtain

$$br(x_{NE}) - br(y) = \int_{y}^{x_{NE}} \frac{\partial br}{\partial z}(z)dz < x_{NE} - y$$

Thus we find that

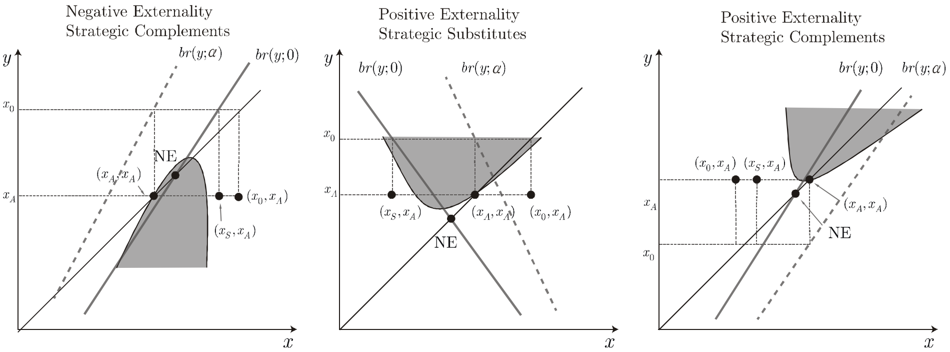
$$\frac{br(x_{NE}) - br(y)}{x_{NE} - y} < 1 = \frac{x_{NE} - y}{x_{NE} - y} = \frac{br(x_{NE}) - y}{x_{NE} - y}$$

and by rearranging this equation, we find $y < br(y)$. So, we have $x < y \le br(y)$. Again from the condition for π, $\pi(\cdot, z)$ is increasing over $[\,0, br(z)\,]$. Thus, we find $\pi(x,y) < \pi(y,y) \le \pi(br(y),y)$ and we obtain the desired results. □

Observe that Lemma 2 holds for all four cases (positive or negative externality and strategic substitute or complement). However, unlike Lemma 1, $\pi(x,x) < \pi(x,y)$ may not hold in the case of strategic complements. Thus we may regard Lemma 2 as a partial generalization of Lemma 1. In principle, one can rigorously study the evolutionary stability of the selfish trait and the altruistic trait combining Lemma 2 with further analysis. However, this task can easily become complicated depending on the specific assumptions for the payoff function and other parameters. We thus illustrate how the altruistic preferences can be evolutionarily stable, replying on the graphical tools developed in Section 3 and the intuition that the trait whose action closer to the NE is more likely to be evolutionarily stable (Lemma 2).

We recall that in the case of strategic substitutes and negative externality, if the social norm action is smaller than the NE, then the selfish agents' action is larger than the NE (strategic substitute) and the altruists' action is smaller than the selfish (negative externality). This opens up the possibility that the altruist's action is closer to the NE. We can use the same intuition in the other three cases to examine the evolutionarily stability of altruistic preferences (see Figure 8).

[Figure 8] Illustration of evolutionary stability of the altruistic preferences



Note: In each panel, the population state in which every agent adopts x_A is evolutionarily stable.

Take the example of a case with positive externality and strategic complements. Suppose that the social norm action, x_0, is smaller than the NE (see the third panel in Figure 8). First, because of strategic complements, the selfish action level is smaller than the NE, too. Second, because of positive externality, the altruist's action level is larger than the selfish agent. This again leads to the possibility that the altruist's action is closer to the NE and the altruistic trait may be evolutionarily stable (see again the third panel in Figure 8). Indeed, if the altruist action level is smaller than the NE (but still larger than that of the selfish agent because of positive externality), Lemma 2 shows that altruist agents as an incumbent population can do better than selfish agents—a necessary condition for the evolutionary stability of the altruistic preference. In Figure 8, we illustrate the possible situations under which the altruistic trait is evolutionarily stable under these varying assumptions. Also, the following table summarizes the underlying causal relations which may lead to the evolutionary success of the altruistic trait.

	Causality
N. Ex. and S. Sub.	low $x_0 \xrightarrow[\text{S. Sub.}]{}$ excessive selfish action $\xrightarrow[\text{N. Ex.}]{}$ smaller altruist action
N. Ex. and S. Com.	high $x_0 \xrightarrow[\text{S. Com.}]{}$ excessive selfish action $\xrightarrow[\text{N. Ex.}]{}$ smaller altruist action
P. Ex. and S. Sub.	high $x_0 \xrightarrow[\text{S. Sub.}]{}$ insufficient selfish action $\xrightarrow[\text{P. Ex.}]{}$ larger altruist action
P. Ex. and S. Com.	low $x_0 \xrightarrow[\text{S. Com.}]{}$ insufficient selfish action $\xrightarrow[\text{P. Ex.}]{}$ larger altruist action

In this section, we illustrated the possibility of evolutionary stability of the altruistic trait. The precise results depend on assumptions for the payoff functions and other parameters and we thus leave complete analysis for future research.

V. Conclusion

In this paper, we examined the evolution of altruistic preferences among bounded rational agents when interactions are strategic substitutes and exhibit negative externality. We combined the indirect evolutionary approaches of preference evolution and the level-k model. Using this model, we identified the conditions under which an altruistic trait is evolutionary stable and a selfish trait is evolutionarily unstable. Few studies have theoretically examined the relationship between social preferences and bounded rationality, despite numerous experimental and empirical findings of interdependence between them. Our results hence fill the theoretical gap in the existing literature under empirically more plausible assumptions for the rationality of agents.

In an extension, we might empirically or experimentally verify the conditions in Propositions 2, 3, and 4. Since the evolutionary stability of altruistic preferences depends on the degree of altruism and the social norm action, as in Figure 5, whether coevolution of preferences and bounded rationality actually occurs in a specific context depends on the specific values of x_0 and α. By combining our theoretical results with an experimental study, we may identify more plausible conditions for co-evolution.

Appendix A. Appendix

Sufficient conditions for an ES trait: conditions in (9), (10), and (11)

These easily follow from the standard arguments in evolutionary game theory (Weibull, 1995; Hofbauer and Sigmund, 1998). We first show that (9) implies (6). If $\prod(p,q) \geq \prod(q,q)$, then we choose any $\bar{\varepsilon} < 1$ and if $\prod(p,q) < \prod(q,q)$, then we choose

$$\bar{\varepsilon} = \frac{\prod(p,p) - \prod(q,p)}{\prod(p,p) - \prod(q,p) + \prod(q,q) - \prod(p,q)} < 1 .$$

Then if $0 < \varepsilon < \bar{\varepsilon}$, (6) is satisfied. Next, we show that (10) implies (9). Recall that p satisfying $p_l = 1$ is the state where all agents choose trait-l. Then

$$\prod(p,p) = \pi(x_l,x_l) = \sum_i q_i \pi(x_l,x_l) > \sum_i q_i \pi(x_i,x_l) = \prod(q,x_l) .$$

Next, suppose that condition (11) holds. Then we need to show that for all $\varepsilon > 0$, there exists $0 < \tilde{\varepsilon} < \varepsilon$ such that

$$\prod(x_l,(1-\tilde{\varepsilon})x_l + \tilde{\varepsilon}x_i) \leq \prod(x_i,(1-\tilde{\varepsilon})x_l + \tilde{\varepsilon}x_i) . \tag{A.1}$$

Let $\varepsilon > 0$ be given. We similarly choose $\bar{\varepsilon}$ as before. That is, if $\pi(x_i,x_i) \geq \pi(x_l,x_i)$ then we choose any $\bar{\varepsilon} < 1$ and if $\pi(x_i,x_i) < \pi(x_l,x_i)$, then we choose

$$\bar{\varepsilon} = \frac{\pi(x_i,x_l) - \pi(x_l,x_l)}{\pi(x_i,x_l) - \pi(x_l,x_l) + \pi(x_l,x_i) - \pi(x_i,x_i)} < 1 .$$

We let $\tilde{\varepsilon} := \min\{\varepsilon, \bar{\varepsilon}\}$. Then it is easy to verify that (A.1) holds. Thus x_l cannot be an ES trait. Finally, suppose that $\pi(x_l,x_l) > \pi(x_i,x_l)$ for some i. Then by following the same argument, it is easy to show that (7) holds. Thus x_i cannot invade x_l.

Proof of Proposition 1. From the first-order condition and our assumptions of f and g, we obtain

$$\frac{\partial br}{\partial \alpha}(y;\alpha) < 0 \quad \text{for any } y \in \mathbb{R} . \tag{A.2}$$

Equation (A.2) shows that the action of a selfish agent is always larger than that of altruists, because of negative externality. Thus, a level-∞ altruist (for any degree α) adopts $x_A^{(\infty)} < x_{NE}$ because a level-∞ selfish agent chooses $x_S^{(\infty)} = x_{NE}$. From Lemma 1, we have

$$\pi(x_S^{(\infty)}, x_S^{(\infty)}) > \pi(x_A^{(\infty)}, x_S^{(\infty)}), \pi(x_S^{(\infty)}, x_S^{(\infty)}) > \pi(x_0, x_S^{(\infty)}) \text{ and}$$
$$\pi(x_S^{(\infty)}, x_A^{(\infty)}) > \pi(x_A^{(\infty)}, x_A^{(\infty)}).$$

Then equations (10) and (11) show that the level-∞ selfish trait is an ES trait and that the level-∞ altruist trait is not an ES trait. □

We next prove Proposition 2.

Proof of Proposition 2. Let $x_{NE} < x_A^{(k)}$. Negative externality again implies that $x_A^{(k)} < x_S^{(k)}$. Thus, we have $x_{NE} < x_A^{(k)} < x_S^{(k)}$. From Lemma 1, we have

$$\pi(x_S^{(k)}, x_S^{(k)}) < \pi(x_A^{(k)}, x_S^{(k)})$$

and equation (11) shows that the selfish trait with level-k cannot be an ES trait.□

To prove Proposition 3, we need the following lemma. To do this we let x_{ANE}, and x_{ASN} be

$$x_{ANE} = br(x_{ANE}; \alpha), \text{ and } x_{NE} = br(x_{ASN}; \alpha). \tag{A.3}$$

Lemma 3. *Suppose that x_{ANE} and x_{ASN} are given by (A.3). Then there exists x_{AAS} such that*

$$\pi(br(x_{AAS}; 0), br(x_{AAS}; \alpha)) = \pi(br(x_{AAS}; \alpha), br(x_{AAS}; \alpha))$$

and

$$x_{ASN} < x_{AAS} < x_{ANE} < x_{NE}.$$

Proof. First, from our assumptions for f and g (i.e., strategic substitutes), we have

$$\frac{\partial br}{\partial y}(y; \alpha) < 0 \quad \text{for any } \alpha \in [0,1]. \tag{A.4}$$

Let $\alpha \in [0,1]$ be fixed. From equations (A.2), (A.3), and (A.4), we have $x_{ANE} < x_{NE}$, whereas equations (A.3) and (A.4) imply $x_{ASN} < x_{ANE}$. Thus, we have $x_{ASN} < x_{ANE} < x_{NE}$. Next, we show that there exists a unique $x_{AAS} \in [x_{ASN}, x_{ANE}]$. We define $\Psi : \mathbb{R} \to \mathbb{R}$ such that

$$\Psi(y) := \pi(br(y;0), br(y;\alpha)) - \pi(br(y;\alpha), br(y;\alpha)).$$

We then check the values of function Ψ at x_{ASN} and x_{ANE}. By the definition of Nash Equilibrium, for any $s \neq x_{NE}$, $\pi(x_{NE}, x_{NE}) > \pi(s, x_{NE})$. This implies that

$$\Psi(x_{ASN}) = \pi(br(x_{ASN};0), x_{NE}) - \pi(x_{NE}, x_{NE}) < 0. \tag{A.5}$$

Moreover, from the definition of best responses, we have

$$\Psi(x_{ANE}) = \pi(br(x_{ANE};0), x_{ANE}) - \pi(x_{ANE}, x_{ANE}) > 0. \tag{A.6}$$

For y such that $x_{ASN} < y < x_{ANE}$, we have

$$y < x_{ANE} < br(y;\alpha) < x_{NE} < br(br(y;\alpha);0) < br(y;0).$$

Below, we denote by s_1 and s_2 the first and second arguments of π to avoid confusion. Since the payoff function π is concave, we have

$$\frac{\partial \pi}{\partial s_1}(br(y;\alpha), br(y;\alpha)) > 0, \quad \frac{\partial \pi}{\partial s_1}(br(y;0), br(y;\alpha)) < 0$$

and from negative externality and strategic substitutes,

$$\frac{\partial \pi}{\partial s_2}(br(y;0), br(y;\alpha)) - \frac{\partial \pi}{\partial s_2}(br(y;\alpha), br(y;\alpha)) < 0$$

Therefore, for y such that $x_{ASN} < y < x_{ANE}$, we have

$$\begin{aligned}\frac{d\Psi}{dy} &= \frac{\partial \pi}{\partial s_1}(br(y;0), br(y;\alpha))\frac{\partial br}{\partial y}(y;0) - \frac{\partial \pi}{\partial s_1}(br(y;\alpha), br(y;\alpha))\frac{\partial br}{\partial y}(y;\alpha) \\ &+ \frac{\partial \pi}{\partial s_2}(br(y;0), br(y;\alpha))\frac{\partial br}{\partial y}(y;\alpha) - \frac{\partial \pi}{\partial s_2}(br(y;\alpha), br(y;\alpha))\frac{\partial br}{\partial y}(y;\alpha) \\ &> 0.\end{aligned}$$

which shows that Ψ is increasing. Thus, from (A.5) and (A.6), we have the unique $x_{AAS} \in [x_{ASN}, x_{ANE}]$ such that $\Psi(x_{AAS}) = 0$. □

Next we prove Proposition 3.

Proof of Proposition 3. From Lemma 3, we first find x_{AAS}. Note that if $x_{ASN} < y < x_{AAS}$, then $\Psi(y) < 0$. Let $\underline{x}_A = br(x_{AAS};\alpha)$. Then, $\underline{x}_A < x_A^{(k)} < x_{NE} < x_S^{(k)}$ implies $br(x_{AAS};\alpha) < br(x_S^{(k-1)};\alpha) < br(x_{ASN};\alpha)$, which in turn implies that

$$x_{ASN} < x_S^{(k-1)} < x_{AAS} \tag{A.7}$$

because of negative externality. Since

$$\begin{aligned}\Psi(x_S^{(k-1)}) &= \pi(br(x_S^{(k-1)};0), br(x_S^{(k-1)};\alpha)) - \pi(br(x_S^{(k-1)};\alpha), br(x_S^{(k-1)};\alpha)) \\ &= \pi(x_S^{(k)}, x_A^{(k)}) - \pi(x_A^{(k)}, x_A^{(k)}) < 0,\end{aligned}$$

we find that

$$\pi(x_S^{(k)}, x_A^{(k)}) < \pi(x_A^{(k)}, x_A^{(k)}). \tag{A.8}$$

and find that the selfish trait with level-k cannot invade the altruist trait with level-k. Now suppose that $k = 1$. Then from equation (A.7) we have $x_0 = x_S^{(0)} < x_{AAS} < x_{ANE}$. Also since $\underline{x}_A = br(x_{AAS};\alpha) > br(x_{ANE};\alpha) = x_{ANE}$ and $x_0 < x_{ANE}$, we find $x_0 < \underline{x}_A < x_A^{(1)} < x_{NE}$. Thus, from Lemma 1, we find $\pi(x_A^{(1)}, x_A^{(1)}) > \pi(x_0, x_A^{(1)})$ and equation (10) implies that the altruistic trait with level-1 is an ES trait. □

Next we prove Proposition 4.

Proof. With the action level of each trait, we can explicitly calculate the payoffs of all agents for a given action of trait-0, x_0. From this, we obtain the following results (see Kim (2013) for more details).

First, when trait-0 chooses smaller than $\frac{a-c}{(3+4\alpha)b}$, trait-0 is strictly dominated, and if no one adopts trait-0, the level-1 altruistic trait dominates the level-1 selfish trait. When no agent adopts trait-0 and the level-1 selfish trait, the level-2 selfish trait dominates the level-2 altruistic trait. For level-1 altruists and level-2 selfish agents, we obtain

$$\pi(x_A, x_A) < \pi(x_{SS}, x_A) \quad \text{and} \quad \pi(x_{SS}, x_A) < \pi(x_A, x_{SS}).$$

Thus, the mixed population of level-1 altruistic and level-2 selfish corresponds to a unique NE.

Second, when the action of trait-0, x_0, satisfies $\frac{a-c}{(3+4\alpha)b} < x_0 < \frac{a-c}{(3+2\alpha)b}$, trait-0 is strictly dominated as well. At the same time, the level-2 selfish trait dominates the level-2 altruistic trait. Without these traits, the level-1 altruistic trait dominates the level-1 selfish trait and the remaining traits obtains

$$\pi(x_{SS}, x_A) < \pi(x_A, x_A) \quad \text{and} \quad \pi(x_{SS}, x_{SS}) < \pi(x_A, x_{SS}) .$$

Thus, level-1 altruistic trait corresponds to a uniquely strict NE.

Finally, when the action of trait-0, x_0, is larger than $\frac{a-c}{(3+2\alpha)b}$, the level-2 selfish trait dominates the level-2 altruistic trait and trait-0. Further, it dominates the level-1 altruistic trait when no one adopts the level-2 altruistic trait and trait-0. Then, level-1 and level-2 selfish traits have

$$\pi(x_S, x_S) < \pi(x_{SS}, x_S) \quad \text{and} \quad \pi(x_S, x_{SS}) = \pi(x_{SS}, x_{SS}) .$$

This implies that level-2 selfish trait corresponds to the NE.

In these three cases, we found each unique NE by iterated deletion of strictly dominated strategies and we can show that the monomorphic state of all gents using an unique NE is an evolutionarily stable state as before. Now we set $\underline{y} := \frac{a-c}{(3+4\alpha)b}$ and $\overline{y} := \frac{a-c}{(3+2\alpha)b}$, to obtain the desired results. □

References

Alger, I. and J. Weibull (2013), "Homo Moralis - Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*, 81(6), 2269-2302.

Ben-Ner, A., F. Kong, and L. Putterman (2004), "Share and Share Alike? Genderpairing, Personality, and Cognitive Ability as Determinants of Giving," *Journal of Economic Psychology*, 25, 581-589.

Benjamin, D. J., S. A. Brown, and J. M. Shapiro (2013), "Who is 'Behavioral'? Cognitive Ability and Anomalous Preferences," *Journal of the European Economic Association*, 11 (6), 1231-1255.

Bester, H. and W. Guth (1998), "Is Altruism Evolutionary Stable?," *Journal of Economic Behavior and Organization*.

Bowles, S. (2004), *Microeconomics*, Princeton University Press.

Brandstater, H. and W. Guth (2002), "Personality in Dictator and Ultimatum Games," *Central European Journal of Operations Research*, 10, 191-215.

Crawford, V. P. (2013), "Boundedly Rational Versus Optimization-based Models of Strategic Thinking and Learning in Games," *Journal of Economic Literature*, 51(2), 512-527.

Dittrich, M. and K. Leipold (2014), "Clever and Selfish? On the Relationship between Strategic Reasoning and Social Preferences," Unpublished.

Fehr, E. and S. Gaechter (2000), "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90(4), 980-94.

Glaeser, E. and J. Scheinkman (2000), "Non-market interactions," *NBER Working Paper* No. 8053.

Guth, W. and M. Yaari (1992), "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game," In *Explaining Process and Change-Approaches to Evolutionary Economics*, University of Michigan Press.

Hofbauer, J. and K. Sigmund (1998), *Evolutionary Games and Population Dynamics*, Cambridge University Press.

Hwang, S.-H. and S. Bowles (2012), "Is altruism Bad for Cooperation?," *Journal of Economic Behavior and Organization*, 83, 330-341.

Hwang, S.-H. and S. Bowles (2014), "Optimal Incentives with State-dependent Preferences," *Journal of Public Economic Theory*, 16(5), 681-705.

Kahneman, D. (2003), "Maps of Bounded Rationality: Psychology for Behavioral Economics," *American Economic Review*, 93(5), 1449-75.

Kim, N. (2013), *Coevolution of Bounded Rationality and Social Preferences*, Master Thesis, Sogang University.

Liberali, J. M., V. F. Reyna, S. Furlan, L. Stein, and S. T. Pardo (2012), "Individual Differences in Numeracy and Cognitive Reection, with Implications for Biases and Fallacies in Probability Judgment," *Journal of Behavioral Decision Making*, 25(4), 361-381.

Millet, K. and S. Dewitte (2007), "Altruistic Behavior as a Costly Signal of General

Intelligence," *Journal of Research in Personality*, 41(2), 316-326.

Nagel, R. (1995), "Unraveling in Guessing Games: An Experimental Study," *American Economic Review*, 85, 1313-1326.

Oechssler, J., A. Roider, and P. W. Schmitz (2009), "Cognitive Abilities and Behavioral Biases," *Journal of Economic Behavior and Organization*, 72(1), 147-152.

Sandholm, W. (2010), *Population Games and Evolutionary Dynamics*, MIT Press.

Sethi, R. (2001), "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 97, 273-297.

Simon, H. (1975), *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting*, Wiley.

Stahl, D. and P. Wilson (1994), "Unraveling in Guessing Games: An Experimental Study," *Journal of Economic Behavior and Organization*, 25, 309-327.

Stahl, D. and P. Wilson (1995), "On Players' Models of other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 218-254.

Weibull, J. (1995), *Evolutionary Game Theory*, MIT Press.