Shared intentions: the evolution of collaboration

Jonathan Newton[a,1]

[a]*School of Economics, University of Sydney.*

Abstract

The ability to share intentions and adjust one's choices in collaboration with others is a fundamental aspect of human nature. We discuss the forces that would have acted for and against the evolution of this ability for a large class of dilemmas and coordination problems that would have been faced by our hominin ancestors. In contrast to altruism and other non-fitness maximizing preferences, the ability to share intentions proliferates when rare without requiring repeated interaction or assortativity in matching.

Keywords: Evolution, shared intentions, norm

JEL Classification Numbers: C73.

> *"Yet how much and how correctly would we think if we did not think, as it were, in community with others to whom we communicate our thoughts, and who communicate theirs with us!"*
>
> – Immanuel Kant (1786)

1. Introduction

Humans are a collaborative species. We collaborate for good and for ill, motivated by love, hate, spite, envy, self-aggrandisement and the basic urges to feed and to reproduce. The understanding of collaboration and cooperation

[1]Come rain or come shine, I can be reached at jonathan.newton@sydney.edu.au, telephone +61293514429. This work was completed while the author was supported by a Discovery Early Career Researcher Award funded by the Australian Research Council (Grant Number: 130101768), and originated in discussions with Sung-Ha Hwang about my work with Simon Angus. I am thankful to both for helping to shape my ideas.

has long been a goal of economics. The current paper models the ability to collaborate as the ability to jointly optimize. That is, can *we* choose what is best for *us* rather than merely making decisions as individuals? It is shown that the ability to *share intentions* and take such joint decisions could have evolved amongst ancient populations who lacked the foresight and reasoning abilities of modern humans, and moreover, that this could happen even in circumstances hostile to the evolution of other behavioral types such as cooperators or altruists who might be expected to behave in ways which appear collaborative.

It has been argued in the philosophical literature that the intentions behind collective acts can be distinct from an aggregation of individual intentions (Bratman, 1992; Searle, 1990; Tuomela and Miller, 1988). This is "shared intentionality", the idea that "we intend to do X" is distinct from "I intend to do X [because I think that she also intends to do X]". There is disagreement amongst philosophers as to what extent shared intentions can be reduced to individual intentions.[2] We take no position on this as our results hold regardless of how agents form shared intentions, whether it be through conversation, pointing and gesturing, or alternative forms of reasoning such as 'team reasoning' (Bacharach, 1999, 2006; Sugden, 2000). To see that shared intentions naturally give rise to joint optimization, consider Alice and Bob who wish to take a drink together at one of two bars, *Grandma's* and *Stitch*. Both Alice and Bob prefer Grandma's to Stitch. Now imagine Alice stating *"I intend to go to Stitch because I think that Bob intends to go to Stitch."* Such an intention is optimal from Alice's perspective, given her beliefs about Bob's intentions, regardless of the Pareto suboptimality of Stitch as a venue. Now, were Alice instead to state *"We intend to go to Stitch,"* then there exists a perfectly valid criticism: given that both Alice and Bob prefer Grandma's to Stitch, and neither has any incentive to deceive the other, it is irrational for them (as a plural entity) to hold such an intention. Economists will recognize this reasoning as similar to that underpinning concepts in game theory such as the Core (Gillies, 1959), Strong Equilibrium

[2]See also Butterfill (2012); Gilbert (1990); Gold and Sugden (2007); Velleman (1997).

(Aumann, 1959), Coalition Proofness (Bernheim, Peleg and Whinston, 1987), Coalitional Rationalizability (Ambrus, 2009), Renegotiation Proofness (Farrell and Maskin, 1989) and Coalitional Stochastic Stability (Newton, 2012).

This paper demonstrates how conditions faced by paleolithic hunter-gatherer societies could have led to the evolution of the ability to collaboratively share intentions. On the one hand, the existence of problems that could be solved by collective action would have spurred the evolution of the ability to form shared intentions. On the other hand, those who could not participate in collaborative acts could sometimes free ride on the successes of others. This free riding would work against the evolution of the ability to share intentions. Note that the sharing of intentions and joint optimization is a *mutualistic* behavior: all participants gain from engaging in it. This does not prevent free riding, as third parties can obtain positive externalities from the collaboration of others, for example if Alice and Bob collaborate in hunting a buffalo, but Colm eats some of the leftovers. The mutualistic nature of jointly intentional behavior can be contrasted with altruistic behavior, in which one party sacrifices fitness for the benefit of another. It has been documented in the anthropology literature that much of the cooperation observed in hunter-gatherer societies is mutualistic. See Smith (2003) for a survey.

A consequence of the mutualistic nature of the sharing of intentions is that such behavior can proliferate when rare. This is in stark contrast to cooperator types or altruists, who become extinct in similar circumstances. Furthermore, even when conditions are adverse to the evolution of the sharing of intentions, for example when there are many opportunities for free riding, some amount of sharing of intentions will persist in the population, a minority behavior that can then spread when conditions become favourable. Note that unlike models of the evolution of altruism and other non-fitness maximizing behaviors, neither repeated interaction (Trivers, 1971), nor kin-selection (Fisher, 1930; Hamilton, 1963), nor assortativity of interaction (Alger and Weibull, 2013; Eshel and Cavalli-Sforza, 1982; Wilson and Dugatkin, 1997), nor group selection (Bowles, 2006; Choi and Bowles, 2007; Haldane, 1932) is required for shared intentions to evolve.

It is hard to overstate the importance of shared intentions to human behavior. Recent work in developmental psychology has shown that from early childhood, human subjects display the ability and desire to engage in collaborative activities. This collaborative urge emerges prior to sophisticated logical inference and the ability to articulate hierarchical beliefs (Tomasello and Rakoczy, 2003, and citations therein). Moreover, the inclination towards collaborative behaviors is considerably weaker in non-human great apes (Tomasello and Carpenter, 2007; Tomasello and Herrmann, 2010).[3] This accumulated evidence has lent support to the hypothesis that human collaborative activity provided a niche in which a uniquely human cognition, replete with sophisticated modes of reasoning, could evolve. This is known as as the shared intentionality hypothesis (Call, 2009) or the Vygotskian intelligence hypothesis (Moll and Tomasello, 2007; Tomasello, 2014; Vygotsky, 1980). The results of the current paper add to the plausibility of this hypothesis, as they show how even in populations of unsophisticated agents, collaborative behavior can evolve.

The author knows of only two other works that deal directly with the topic of the current paper[4,5]: Bacharach (2006, Chapter 3) and the study of Angus and Newton (2015). Bacharach (2006) gives a predominantly non-quantitative argument as to why a group selection mechanism would lead to collaborative 'team reasoning' in coordination problems and social dilemmas. However, in a simulations-based study of coordination games on networks, Angus and Newton (2015) show that group selection is far from sufficient for the evolution of

[3]See also Wobber, Herrmann, Hare, Wrangham and Tomasello (2014); Tomasello, Carpenter, Call, Behne and Moll (2005) and the accompanying critical responses.

[4]We emphasize that we are considering the evolution of a trait - the ability to collaborate and share intentions, *not* the evolution of the play of any specific 'cooperative' action. Alice and Bob may intend to plan a surprise party for Colm, or to rob him of his possessions. Either way, Alice and Bob are collaborating, but to quite different ends.

[5]An alternative approach to understanding collaboration is that of Gavrilets (2014), who models collaborative ability as entering directly into production functions. Groups with high levels of collaborative ability produce more of a public good, giving an advantage in a group selection framework.

collaboration, and that selective pressure *against* the sharing of intentions can arise at a group level due to the possibility of collaborative behavior slowing techno-cultural advance. The cited papers focus on multiple pairwise interactions for which payoffs are given by an underlying two player game. The current paper does not restrict itself to pairwise interaction and gives analytic results for a setting in which members of a population are randomly matched to play m-player public goods games. In contrast to previous work, there is no group selection and selective pressure against the sharing of intentions arises from the possibility of free riding. Finally, it is instructive to compare the evolution of shared intentions to the evolution of preferences (e.g. Dekel, Ely and Yilankaya, 2007; Güth and Kliemt, 1998; Robson, 1996; Samuelson, 2001). In contrast to the evolution of preferences, the ability to collaboratively share intentions does not change individuals' ranking of outcomes. Instead it makes new outcomes available to individuals when they update their strategies as part of a group. Any individual's ranking of menu items does not change, but the variety of items on the menu becomes more appealing.

The paper is organized as follows. Section 2 gives the model and considers the evolution of shared intentions for a large class of public goods games, including the extended example of threshold public goods games. Section 3 analyzes the evolution of shared intentions when collaboration can exert negative externalities on others, such as when two people team up to steal from a third party. Section 4 considers a continuum of types distinguished by different probabilities of an individual of a given type being in a collaborative frame of mind. Section 5 compares and contrasts our results to those for altruism and other behavioral explanations for 'cooperation' found in the literature. Section 6 concludes. All proofs not in the main text are relegated to the appendix.

2. Model and analysis

We shall consider a population of individuals represented by the unit interval. Fitnesses will be determined when randomly formed groups of m individuals

encounter public goods problems. These problems could be opporunities to hunt large prey such as whales (Alvard, 2001; Alvard and Nolin, 2002), or the possibility that coordinated action could bring about a large haul of small prey, such as is the case with fishing (Sosis, Feldstein and Hill, 1998). For a group to at least partially solve the public goods problem it will be required that at least $n \leq m$ of the individuals in the group contribute. This contribution could be of time, effort or resources. The contribution need not be the same for each individual, for example different roles may need to be carried out during a hunt.

2.1. The game

Formally, we represent a problem faced by a group of m individuals by Γ, a symmetric m-player game with player set M and strategy sets $S_i = \{\times, +_1, +_2, \ldots, +_N\}$, $N \in \mathbb{N}$, $i \in M$, where $\times$ represents non-contribution and $+_i$ denote different forms of contribution. Let $s_i \in S_i$ and $s = (s_1, s_2, \ldots, s_m)$ be representative strategies and strategy profiles respectively. Let S be the set of all strategy profiles. Let $\pi_i(s)$ be the payoff of player i at strategy profile s. Payoffs represent reproductive fitness. Let $\underline{\times} := (\times, \ldots, \times)$ be the status quo strategy profile and a Nash equilibrium of the game. We assume that actions other than $\times$ exert (weakly) positive externalities (relative to $\times$) on other individuals. This gives the public goods element to the game: a contribution of any form by i is at least as good for j as is non-contribution by i.

(PG) For all $i, j \in M$, $i \neq j$, $s \in S$, we have $\pi_j(s_i, s_{-i}) \geq \pi_j(\times, s_{-i})$.

Condition (PG) is satisfied by threshold public goods games and by m-player Prisoner's Dilemmas. These are both games in which there is one way to contribute ($N = 1$) and thus no possibility of asymmetry of roles in provision of goods. Figure 1 gives two examples that illustrate how Condition (PG) is also satisfied by more exotic setups. In these examples, group size is two ($m = 2$), there are two ways to contribute ($N = 2$), and both group members are required to contribute for them to gain some net benefit. Figure 1(i) is a stag hunt with two stags, and the hunters must both pursue the same stag in order to be

	$+_1$	$+_2$	$\times$
$+_1$	$b-c$	$-c$	$-c$
$+_2$	$-c$	$b-c$	$-c$
$\times$	0	0	0

(i) Two prey

	$+_1$	$+_2$	$\times$
$+_1$	$-c$	$b-c$	$-c$
$+_2$	$b-c$	$-c$	$-c$
$\times$	0	0	0

(ii) Chase and ambush

Figure 1: Examples for $m = n = 2$, $N = 2$, $b > c > 0$. For each combination of contribution ($+_1$,$+_2$) and non-contribution ($\times$), entries give fitnesses for the row player.

successful. Figure 1(ii) represents a situation where there are two roles required for a successful hunt, such as when one hunter pursues the quarry and a second hunter lies in wait, ready to ambush the quarry when it flees from the first hunter.

From the status quo of no contribution, the assumption that $\underline{\times}$ is a Nash equilibrium implies that no individual acting alone can improve his payoff.[6] However, there exist opportunities for coalitions of players who can share their intentions to collaborate and adjust their actions together in order to obtain higher payoffs. Let the set of collaborative opportunities for a set $T \subseteq M$ be

$$\mathcal{C}(T) = \{s \in S : \; s_i \neq \times \implies i \in T \implies \pi_i(s) > \pi_i(\underline{\times})\}.$$

That is, $\mathcal{C}(T)$ gives the ways in which individuals in T can collaboratively adjust their strategies so that their own payoffs improve, leaving the strategies of individuals outside of T fixed. We assume that the game affords at least some prospect of collaboration. That is, $\mathcal{C}(T) \neq \emptyset$ for at least some $T \subseteq M$. This is equivalent to $\underline{\times}$ not being a Strong Equilibrium in the sense of Aumann

[6] We assume myopia. That is, individuals do not think, or rather act, beyond the implications of a direct adjustment to their strategy or the strategies of those with whom they share intentions. We are modeling early man, not bands of game theorists roving across the savannah. Note that even myopic payoff improvers are considerably more sophisticated than types of players - *cooperators*, *defectors* and so on, who only play a specific action.

(1959). For the example in Figure 1(i), $\mathcal{C}(M) = \{(+_1, +_1), (+_2, +_2)\}$ and for the example in Figure 1(ii), $\mathcal{C}(M) = \{(+_1, +_2), (+_2, +_1)\}$. The size of the smallest coalition that can benefit from collaboration is

$$n = \min_{T:\mathcal{C}(T)\neq\emptyset} |T|$$

Note that our assumption that $\underline{\times}$ is a Nash equilibrium implies than $n \geq 2$ and the existence of at least some collaborative opportunity implies that $n \leq m$.

We shall allow for the possibility that the behavior of individuals outside of a set T will alter as a consequence of T exploiting a collaborative opportunity. With this in mind, define the set of outcomes that could occur following a set of individuals T exploiting a collaborative opportunity.

$$\mathcal{C}^*(T) = \{s^* \in S : \text{ For some } s \in \mathcal{C}(T),\ s_i^* = s_i \text{ for all } i \in T\}.$$

That is, were a set of individuals T to adjust their strategies according to some collaborative opportunity in $\mathcal{C}(T)$, following which the remainder $M \setminus T$ of the individuals were to adjust their strategies in some way, then $\mathcal{C}^*(T)$ is the set of strategy profiles that could be reached.

2.2. Types and behavior

There are two types of individual, those who can share intentions and those who cannot. Those who can share intentions can collaboratively optimize when choosing their action. We refer to such individuals as SI types. Those who lack the cognitive ability to engage in such joint optimization we refer to as N types.

From a status quo at which everybody is playing $\times$, any set of n individuals within a group can gain by adjusting their actions. However, this may not lead to a Nash equilibrium. In fact, it may even be that playing an action other than $\times$ is never individually rational, such as in an m-person Prisoner's Dilemma. However, if nobody is playing anything other than $\times$, then for any set of n SI types not to contribute is not collectively rational. We defer consideration of individuals who are sometimes in the mood for collaboration and are sometimes

not until later in the paper. For now we assume that SI types are always willing to participate in collaborative decisions when opportunities present themselves.[7]

Fix an m-player game Γ as described in Section 2.1. Consider a group M of m individuals who encounter this problem. Let $M_{SI} \subseteq M$ denote the set of SI type individuals within the group. Then we assume the outcome of the game will be given by some strategy profile s^* satisfying

(C) **(i)** If $|M_{SI}| < n$, then $s^* = \underline{\times}$.

(ii) If $|M_{SI}| \geq n$, then select some set of SI type individuals $T \subseteq M_{SI}$, $\mathcal{C}(T) \neq \emptyset$, according to some probability measure $F_{M_{SI},\Gamma}(.)$. Then let $s^* \in \mathcal{C}^*(T)$ be chosen according to some probability measure $G_{T,\Gamma}(.)$. Let $F_{M_{SI},\Gamma}(.)$ and $G_{T,\Gamma}(.)$ be symmetric with respect to the identities of the players in M, except insofar as they are members of M_{SI} or T.

That is, when there are insufficient SI types in the group for them to exploit a collaborative opportunity, the outcome of the game is the status quo $\underline{\times}$. When there exist enough SI type individuals to exploit at least one collaborative opportunity, some set of SI types will exploit some collaborative opportunity. Note that without specifying $F_{M_{SI},\Gamma}(.)$ and $G_{T,\Gamma}(.)$, condition (C) is not a complete description of behavior. In particular, when there are multiple collaborative opportunities any of them could be taken with any probability, and the behavior of the players who are not involved in exploiting the collaborative opportunity is similarly arbitrary. This is important to note as our main results will be independent of $F_{M_{SI},\Gamma}(.)$ and $G_{T,\Gamma}(.)$.

2.3. Matching

We consider a population comprising unit mass of individuals, each of whom may be of SI or N type. Let the share of SI types in the population be x_{SI} and the share of N types be x_N. Let the population state be $x = (x_{SI}, x_N)$.

[7] That is to say, N types would play a Prisoner's Dilemma, whereas SI types would play a Prisoners' Dilemma, the difference being in the positioning of the respective apostrophes.

Each member of the population is matched to play Γ in a group of m individuals. Given a population state x, and an individual, let Z be a random variable denoting the number of SI types amongst the other $m-1$ individuals with whom the given individual is matched. We allow correlation between Z and the type of the individual concerned. Write $Pr_x[Z = k \,|\, SI]$ and $Pr_x[Z = k \,|\, N]$ as the probabilities that there are k SI type individuals amongst the other members of the group, conditional on a given individual being SI type and N type respectively. A *matching protocol* specifies these values for all x and all values of k from 0 to $m-1$. We assume that $Pr_x[Z = k \,|\, SI]$ and $Pr_x[Z = k \,|\, N]$ are continuous in x_{SI}, and strictly positive for $x_{SI}, x_N > 0$.

Given this notation, any SI type has a probability $Pr_x[Z = k-1 \,|\, SI]$ of being in a group that includes exactly k SI types, including himself. Therefore, the mass of SI types in such groups equals $x_{SI} Pr_x[Z = k-1 \,|\, SI]$. Any N type has a probability $Pr_x[Z = k \,|\, N]$ of being in a group that includes exactly k SI types. Therefore, the mass of N types in such groups equals $x_N Pr_x[Z = k \,|\, N]$. Now, any group with k SI types has $m-k$ N types, so the ratio of SI type individuals in such groups to N type individuals in such groups must equal $k/(m-k)$. Noting that $x_N = 1 - x_{SI}$, we have the balance condition (B).[8]

(B) $\frac{x_{SI} Pr_x[Z=k-1 \,|\, SI]}{(1-x_{SI}) Pr_x[Z=k \,|\, N]} = \frac{k}{m-k}$.

An implication of (B) is that as x_{SI} approaches zero, the ratio of $Pr_x[Z = k-1 \,|\, SI]$ to $Pr_x[Z = k \,|\, N]$ approaches infinity. Consequently, SI types find themselves in groups in which collaboration occurs infinitely more often than N types find themselves in such groups. Furthermore, $Pr_x[Z = k \,|\, N]$ must

[8] Given the preceding, one might ask why it is that this is stated as an condition, rather than merely as a consequence of any matching protocol. The reason for this is that although (B) is a logical consequence of matching, it is not a mathematical consequence. Consider groups of size two, with SI types making up a share of a quarter of the population. It is mathematically possible to match this quarter of the population one-to-one to the remaining three quarters of the population via a bijection. This would clearly be against the spirit of the model.

approach zero for $k \geq 1$ and $Pr_x[Z = 0 \,|\, N]$ must approach unity.

2.4. The main theorem

Given a game Γ, some behavioral rule satisfying (C), and some matching protocol, let $f_{SI}(x)$, $f_N(x)$ denote the expected fitnesses of SI and N types respectively at population state x. Specifically, for $|M_{SI}| = k$, denote the expected payoff of $i \in M_{SI}$ by π^k_{SI}, and denote the expected payoff of $i \notin M_{SI}$ by π^k_N. Then

$$f_{SI}(x) = \sum_{k=0}^{m-1} Pr_x[Z = k \,|\, SI]\, \pi^{k+1}_{SI}, \qquad f_N(x) = \sum_{k=0}^{m-1} Pr_x[Z = k \,|\, N]\, \pi^k_N.$$

Note that $f_{SI}(x)$ and $f_N(x)$ depend continuously on the probabilities $Pr_x[Z = k \,|\, SI]$ and $Pr_x[Z = k \,|\, N]$, which in turn are continuous in x_{SI}. Therefore, $f_{SI}(x)$ and $f_N(x)$ are continuous in x_{SI}. For some subpopulation with shares of SI and N types given by $\tilde{x}$, let $f_{\tilde{x}}(x)$ be the average fitness of members of this subpopulation when the population state is x. That is,

$$f_{\tilde{x}}(x) := \tilde{x}_{SI} f_{SI}(x) + \tilde{x}_N f_N(x).$$

We use the concept of an evolutionarily stable state (Taylor and Jonker, 1978). An evolutionarily stable state is a state such that following the invasion of the population by a small population share ε of mutants, the non-mutant share of the population outperforms the invading mutants.

Definition 2.1. A state x^* is an evolutionarily stable state (ESS) if for any other state $\tilde{x}$, defining $x_\varepsilon = (1 - \varepsilon)x^* + \varepsilon\tilde{x}$, there exists $\tilde{\varepsilon}$ such that

$$\text{For all } \varepsilon < \tilde{\varepsilon}, \quad f_{x^*}(x_\varepsilon) > f_{\tilde{x}}(x_\varepsilon).$$

An interior state x^* is an ESS if and only if $f_{SI}(.) - f_N(.)$ is strictly decreasing at x^* and equal to 0. The extremal state $x^*_{SI} = 0$ ($x^*_{SI} = 1$) is an ESS if and only if $f_{SI}(.) - f_N(.)$ is strictly negative (positive) in some open interval bounded below (above) by x^*_{SI}. This implies that, unless there exists some open interval

of x_{SI} on which $f_{SI}(.) - f_N(.) = 0$, at least one ESS must exist. Such examples can be constructed, but will necessarily be special.[9]

We are now in a position to state our main theorem. For any public goods problem and any plausible matching protocol, SI types will make up a positive share of the population at any evolutionarily stable state. Even when conditions are highly adverse to the evolution of shared intentions, SI types will still persist as a small share of the population, ready to expand and take a greater share as soon as conditions become more favorable.

Theorem 1. *If (C),(B),(PG) hold, then $x_{SI} > 0$ in any ESS.*

The reasoning behind the Theorem is as follows. Firstly, collaboration (C) is a mutualistic act, not an altruistic act, therefore when SI types collaborate they improve their payoffs relative to the status quo, holding fixed the strategies of the non-collaborators. Secondly, the public goods condition (PG) ensures that any response to the collaboration by other players can only (weakly) increase the payoffs of the collaborators. Finally, the balance condition (B) implies that when SI types are a small share of the population, any given SI type will find himself in a group in which collaboration occurs much more frequently than any given N type finds himself in such a group. Thus an SI type will enjoy the benefits of collaboration far more often than an N type will get the opportunity to free ride on the collaboration of others.

Note that the possibility of free riding by non-collaborating players disappears when $n = m$. Furthermore, as any collaboration will then always involve every group member, there is no need to consider the responses of non-collaborating players to collaborative activity. Therefore, the result of Theorem 1 holds, regardless of whether or not (B) or (PG) hold.

Corollary 1. *If (C) holds and $n = m$, then a unique ESS x^* exists and $x^*_{SI} = 1$.*

[9] To make specific statements about genericity requires consideration not only of Γ, but also of different behavioral rules satisfying (C), and different matching processes. This is sufficiently involved that it is omitted here.

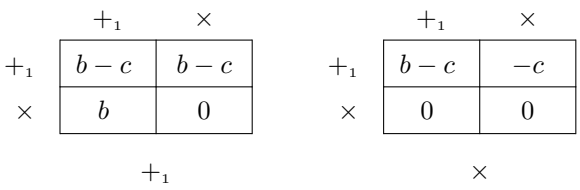

Figure 2: Three player threshold public goods game. Any two players must contribute for the good to be provided. $m = 3$, $n = 2$, $N = 1$, $b > c > 0$. For each combination of contribution ($+_1$) and non-contribution ($\times$), entries give fitnesses for the row player.

An important class of games covered by the Corollary is the class of coordination games for which all symmetric pure strategy profiles are Nash equilibria and $\underline{\times}$ is Pareto dominated by one of the other equilibria.

Finally, as the state x is unidimensional, evolutionarily stable states correspond to strongly uninvadable states in the sense of Bomze (1991) and are thus locally asymptotically stable under the replicator dynamic (Bomze and Weibull, 1995).

2.5. Example: threshold public goods problems

Let Γ be a threshold public goods game with threshold $n \leq m$. There are two strategies, contribute ($+_1$) and don't contribute ($\times$). If at least n individuals contribute then the good is provided, otherwise it is not provided. When the good is provided, every individual in the group obtains a benefit of b. When an individual contributes, he incurs a cost of c. Assume $b > c > 0$. The case of $m = 3$, $n = 2$ is shown in Figure 2. Let matching be binomial so that, given x, from the perspective of any given individual, each of the other individuals in the group will independently be SI type with probability x_{SI}. That is,

(Bin) $Pr_x[Z = k] := Pr_x[Z = k \,|\, SI] = Pr_x[Z = k \,|\, N] = \binom{m-1}{k} x_{SI}^k (1 - x_{SI})^{m-1-k}$.

Assume that when $|M_{SI}| \geq n$, the realized strategy profile, s^*, always has exactly n SI types contributing. This seems plausible, as there is no advantage

to anyone from any additional individual contributing.[10] This implies that from the perspective of an SI type, if the good is provided and there are k other SI types in the group, he will contribute $n/k+1$ of the time. The average fitness of SI types in this setting is

$$f_{SI}(x) = \sum_{k=n-1}^{m-1} Pr_x[Z=k]\left(b - c\frac{n}{k+1}\right). \tag{1}$$

N types will benefit when the good is provided, but are never able to be part of a collaborative effort to provide the good. Therefore, the good will only be provided if at least n of the other $m-1$ players are SI types. When $n=m$, the fitness of an N type is always zero. When $n<m$, the fitness of an N type is

$$f_N(x) = Pr_x[Z \geq n]\, b. \tag{2}$$

When $x_{SI} > 0$, the fitness advantage (disadvantage when negative) of SI types over N types is

$$f_{SI}(x) - f_N(x) = Pr_x[Z=n-1]\left(b - \sum_{k=n-1}^{m-1} \frac{Pr_x[Z=k]}{Pr_x[Z=n-1]} c\frac{n}{k+1}\right) \tag{3}$$

This expression equals $Pr_x[Z=n-1](b-c) > 0$ when $n=m$. When $n<m$, then by showing that the term in brackets is decreasing in x_{SI}, positive for small positive values of x_{SI} and negative for values of x_{SI} close to 1, we prove the following proposition.

Example 1. *For threshold public goods games, when (C),(Bin) hold,*

(i) There is a unique ESS, x^*. If $n<m$, then $x^*_{SI} \in (0,1)$, and if $n=m$, then $x^*_{SI}=1$.

[10] Note that our simple story of when the good is provided is borne out by more complex dynamic processes. If the number of SI types in a group is at least n, then under a coalitional better response dynamic with uniform mistakes (see, for example Newton and Angus, 2015), provision of the good is uniquely stochastically stable in the sense of (Young, 1993). If the number of SI types is strictly less than $n-1$, then non-provision is uniquely stochastically stable. Note that a stochastic stability analysis of such problems under *individualistic* best response dynamics is given by Myatt and Wallace (2008), but results change when coalitional behavior is allowed.

(ii) From any mixed population such that $x_{SI}, x_N > 0$, the replicator dynamic converges to x^*.

(iii) x_{SI} decreases in m, increases in n, increases in b/c.

As (Bin) implies (B), and threshold public goods games satisfy (PG), Theorem 1 applies and tells us that even when conditions are bad for collaboration, a minority of SI types will persist in the population. Example 1 shows that, for threshold public goods problems, such conditions are when m is large relative to n so that there are many free riders whenever the public good is provided, or when the benefit-cost ratio b/c is low. This minority of SI types can then expand when changes in the environment or technology lead to conditions which are more favourable for collaborative behavior. Such a change could be an increase in n caused by an increase in the availability of larger prey due to migration, or changes in the climate when moving between glacial and interglacial periods. Another example would be a reduction in c due to reduced risks from hunting due to improved technology providing better weapons.

3. Collaboration with negative externalities

A plausible sounding conjecture would be that if collaborating players gain fitness from their collaboration following any response by the other players, then SI will evolve. This conjecture is false. When externalities from collaboration are negative and SI types are disproportionately likely to match with other SI types, then negative externalities caused by collaborating SI types on other SI types can outweigh the benefits of collaboration. To see this, first formalize the condition, Profitable Collaboration (PC).

(PC) If $T \subseteq M$, $G_{T,\Gamma}(s^*) > 0$, then $\pi_i(s^*) > \pi_i(\underline{\times})$ for all $i \in T$.

Now, consider the three player game in Figure 3. We call this a three player Hawk-Dove game as any two players can exploit the third (steal his food) and it is never worthwhile for the third player to resist this. Thus there are three asymmetric Nash equilibria (in fact, Strong Equilibria) in which two players

	$+_1$	$\times$
$+_1$	-4	1
$\times$	-3	0

$+_1$

	$+_1$	$\times$
$+_1$	1	-1
$\times$	0	0

$\times$

Figure 3: Three player Hawk-Dove game. Any two players must attack the remaining player for an attack to be successful. $m = 3$, $n = 2$, $N = 1$. For each combination of contribution ($+_1$) and non-contribution ($\times$), entries give fitnesses for the row player.

exploit the remaining player. However, there is also another Nash equilibrium at which all players play $\times$. The only available collaborative opportunity is for two SI type players to exploit the third (eg. $s = (+_1, +_1, \times)$). When a pair of players T take such a collaborative opportunity $s \in \mathcal{C}(T)$, assume that the realized strategy profile $s^* = s$. That is, $G_{T,\Gamma}(s) = 1$. This seems plausible as the exploited player would lose payoff by adjusting his strategy. Note that (PC) is satisfied. Now, the fitness of an N type is

$$f_N(x) = \underbrace{Pr_x[Z = 2 \,|\, N]}_{\substack{\text{Prob. of N type} \\ \text{being exploited} \\ \text{by SI types}}} (-3) \underbrace{\xrightarrow{x_{SI} \to 0} 0}_{\text{by (B)}}.$$

The fitness of an SI type is

$$f_{SI}(x) = Pr_x[Z = 2 \,|\, SI] \underbrace{\left(\frac{2}{3}(1) + \frac{1}{3}(-3)\right)}_{\substack{\text{When all three are SI,} \\ 2/3 \text{ chance of being} \\ \text{an exploiter}}} + Pr_x[Z = 1 \,|\, SI]\,(1),$$

so if $\lim_{x_{SI} \to 0} Pr_x[Z = 2 \,|\, SI] > 3 \lim_{x_{SI} \to 0} Pr_x[Z = 1 \,|\, SI]$, then $\lim_{x_{SI} \to 0} f_{SI}(x)$ is bounded above by a number strictly below zero. That is, positive assortative matching can cause SI types to have lower fitness than N types, even when the share of SI types in the population is small.

Now, consider the case where there is no assortativity in matching; a given individual's type does not affect the type distribution of the remaining $m - 1$

individuals with whom he is matched. This is the no assortativity condition (NA).

(NA) $Pr_x[Z = k] := Pr_x[Z = k \mid SI] = Pr_x[Z = k \mid N]$ for all k, x.

Note that binomial matching satisfies (NA). It turns out that when matching is not assortative and collaboration is always profitable, then SI types will always make up a strictly positive share of the population at any ESS.

Theorem 2. *If (C),(B),(NA),(PC) hold, then* $x_{SI} > 0$ *in any ESS.*

The intuition behind the Theorem is simple. Recall that the balance condition (B) implies that ${}^{Pr_x[Z = k-1 \mid SI]}/_{Pr_x[Z = k \mid N]}$ approaches infinity as x_{SI} approaches zero. Using (NA) and substituting, it must then be true that ${}^{Pr_x[Z = k-1]}/_{Pr_x[Z = k]}$ and thus ${}^{Pr_x[Z = n-1]}/_{Pr_x[Z = k]}$, $k \geq n$, also approach infinity. That is, when collaboration occurs it will usually be when there are exactly n SI types in the group. (PC) implies that these collaborators gain fitness from their collaboration, and there are no other SI types in the group to be affected by any negative externalities. Hence, SI types outperform N types for small, positive values of x_{SI}.

For the Hawk-Dove example of Figure 3, under binomial matching (Bin), we have that $f_N = x_{SI}^2(-3)$, and $f_{SI} = x_{SI}^2\left(\frac{2}{3}(1) + \frac{1}{3}(-3)\right) + 2x_{SI}(1 - x_{SI})1$. Then $f_N < 0$ and $f_{SI} > 0$ for small values of x_{SI}. Once again, SI types proliferate when rare.

4. A continuum of types ordered by likelihood of collaboration

Consider a model where instead of two types, we have a continuum of types, specifically the unit interval. Each time he faces a problem, any given individual of type $\sigma \in [0, 1]$ will be in a *collaborative mood* with probability σ, and in an *individualistic mood* with probability $1 - \sigma$. An individual in an individualistic mood will behave as an N type and an individual in a collaborative mood will behave as an SI type. Let the state, x, be a probability measure on the Borel

sets $\mathcal{B}([0,1])$. This approach, modeling the same individual as sometimes collaborative and sometimes not, is that suggested by Bacharach (2006) for dealing with potential conflicts between individual and collective rationality. Any individual, when facing a problem as part of a group, will sometimes be driven by individual considerations and sometimes by collective considerations.[11] Define $\bar{\sigma}(x) := \int_{[0,1]} \sigma x(d\sigma)$ as the probability that a randomly drawn individual from a population at state x is in a collaborative mood. We can adapt binomial matching to this setting.

(Bin-σ) $Pr_x[Z = k] = \binom{m-1}{k}(\bar{\sigma}(x))^k(1-\bar{\sigma}(x))^{m-1-k}$.

An evolutionarily stable state will not typically exist. The reason for this is that under binomial matching, at any interior ESS, any individual in the population must have equal expected fitness when he collaborates and when he does not collaborate. But then, from any state x such that $x(\{0\}) \neq 1$, $x(\{1\}) \neq 1$, a small mutant subpopulation with type shares $\tilde{x} \neq x$ could emerge such that $\bar{\sigma}(\tilde{x}) = \bar{\sigma}(x)$. That is, some of the mutants have increased σ and some have decreased σ, but the average remains the same as before. Such mutant invasions do not alter the expected fitness of any individual in the population. In particular, the mutants still obtain the same average fitness as non-mutants, so x cannot be evolutionarily stable. Consequently, we use the weaker concept of Neutral Stability (Maynard Smith, 1982). Write $g_\sigma(x)$ for the fitness of type σ at state x. Note that the average fitness of a subpopulation of types distributed according to $\tilde{x}$ when the state is x is now

$$g_{\tilde{x}}(x) := \int_{[0,1]} g_\sigma(x)\tilde{x}(d\sigma).$$

[11] Bacharach (2006) thinks of individuals as sometimes reasoning individualistically and sometimes engaging in 'team reasoning'. Our assumptions relate to behavior and not to reasoning per se, but our model can, should the reader wish, be interpreted as a model of the evolution of team reasoning, specifically what Bacharach refers to as *restricted team reasoning*, where at any given point in time, not every individual can team reason but those that can, recognize one another as such.

A neutrally stable state is then a population state such that following the invasion of the population by a small population share ε of mutants, the invaders do not do better than the non-mutants.

Definition 4.1. A state $\hat{x}$ is a neutrally stable state (NSS) if for any other state $\tilde{x}$, defining $x_\varepsilon = (1-\varepsilon)\hat{x} + \varepsilon\tilde{x}$, there exists $\tilde{\varepsilon}$ such that

$$\text{For all } \varepsilon < \tilde{\varepsilon}, \quad g_{\hat{x}}(x_\varepsilon) \geq g_{\tilde{x}}(x_\varepsilon).$$

Now, by definition of the behavior of type σ, and comparing (Bin) and (Bin-σ), we have

$$g_\sigma(.) = \sigma f_{SI}(\bar{\sigma}(.)) + (1-\sigma) f_N(\bar{\sigma}(.))$$

where f_{SI}, f_N denote fitnesses in the two type model under (Bin), slightly abusing notation to write x_{SI} rather than x as the argument of f_{SI}, f_N. This gives

$$g_x(.) = f_N(\bar{\sigma}(.)) + \bar{\sigma}(x)(f_{SI}(\bar{\sigma}(.)) - f_N(\bar{\sigma}(.))).$$

That is, $g_x(.)$, and specifically $g_x(x_\varepsilon)$, is monotonic in $\bar{\sigma}(x)$. This implies we only need to check robustness of any conjectured NSS to invasions of extreme types $\sigma = 0$ and $\sigma = 1$. These types correspond to N and SI types of the two type model. Now, for the two type model under (Bin), there exists an ESS with a positive share of SI types. Therefore, letting x^* be an ESS of the two type model, we have that $\hat{x}$ such that $\hat{x}(0) = x^*_N$, $\hat{x}(1) = x^*_{SI}$, is an NSS of the continuum type model. Furthermore, as at an NSS under binomial matching, fitness from collaboration and non-collaboration must be the same, the only factor that affects the fitness of any given type is the distribution over how many of his fellow group members are in a collaborative mood. But under (Bin-σ), this distribution is completely determined by $\bar{\sigma}(.)$. Therefore, if x' is vulnerable to an invasion by mutants, and $\bar{\sigma}(x') = \bar{\sigma}(x'')$, then x'' must be vulnerable to the same mutant invasion. That is, the only factor that determines whether a state x is an NSS is the value of $\bar{\sigma}(x)$.

Theorem 3. *If (C),(Bin-σ),(PC) hold, then at least one NSS of the continuum model exists. Under these conditions, a state x of the continuum model is an*

	$+_1$	$\times$
$+_1$	$b-c$	$-c$
$\times$	0	0

(i) Fitnesses

	$+_1$	$\times$
$+_1$	$b-c$	$b-c$
$\times$	0	0

(ii) Magical thinker

	$+_1$	$\times$
$+_1$	$b-c$	$-c/2$
$\times$	$-c/2$	0

(iii) Altruist

Figure 4: For $m = n = 2$, $N = 1$, for each combination of contribution ($+_1$) and non-contribution ($\times$), entries give, for the row player, his (i) fitnesses and his preferences when he is a (ii) magical thinker and (iii) altruist.

NSS if and only if $\bar{\sigma}(x) = x^*_{SI}$ *for some ESS* x^* *of the two type model under (Bin). This implies that a monomorphic NSS* $\hat{x}$ *exists, with* $\hat{x}(\hat{\sigma}) = 1$ *and* $\hat{\sigma} = x^*_{SI}$.

5. Cooperators, altruists and magical thinkers

Here we compare the collaborative sharing of intentions to some other modes of behavior that have been considered in the literature. The crucial distinction is that collaboration involves coordination in *how* actions are chosen rather than in the chosen actions themselves, as we saw in the three player Hawk-Dove game, where an efficient symmetric Nash equilibrium is destroyed by collaboration. Naturally, collaboration will often lead to efficient coordination, but the two things are not the same.[12] This is the reason we have so far avoided the use of the term "cooperation" in describing our model, as practitioners have become accustomed to using this word to describe a state of efficient coordination rather than its attainment.

[12] For more on this point, see Newton and Angus (2015), where it is shown how coordinated action choice by small groups within a population can slow convergence to a globally efficient action profile, even when all players have perfectly common interests.

5.1. Cooperators

There has been much consideration in the academic literature of situations where one symmetric action profile Pareto dominates all other symmetric action profiles. The action corresponding to such a profile is then described as the "cooperative" action. Individuals who always play such an action are called *cooperators* and those who play an action corresponding to some inefficient Nash equilibrium are called *defectors*. In the absence of assortative matching, when there are few cooperators in the population, they will rarely match with one another and will be outperformed by defectors. That is, cooperators do not proliferate when rare and there exists an ESS in which they are absent from the population. For the threshold public goods game of Section 2.5, this has been formally shown by Pacheco, Santos, Souza and Skyrms (2009).

5.2. Magical thinkers

Magical Thinkers erroneously attribute causal powers to their own decisions (Elster, 1979). Consider those magical thinkers that are referred to as 'Kantian' types by Alger and Weibull (2013). These types behave as if their fellow group members will always take the same action as they take. This implies that they will always choose the action corresponding to the most efficient of all symmetric action profiles. For the threshold public goods game of Section 2.5, the fitness of any given individual in the $m = n = 2$ case for each combination of contribution ($+_1$) and non-contribution ($\times$) by the individual and his fellow group member is given in Figure 4(i). However, the magical thinker will act as if his fitness is given by Figure 4(ii). From this we see that if $b - c > 0$, then magical thinkers will behave identically to cooperators, and if $b - c < 0$, then magical thinkers will behave as defectors. Unlike cooperators and defectors, magical thinkers are not automata, but the ordering that determines their choice of action (their preferences) differs from the ranking given by their fitnesses. This is not the case for SI types, whose preferences (which are given at the level of the individual) are unaffected by their SI-ness but who may, in collaboration with other SI

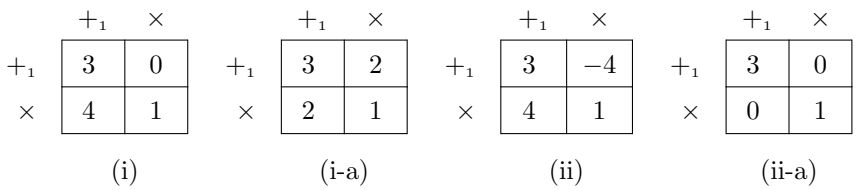

Figure 5: For each combination of contribution ($+_1$) and non-contribution ($\times$), entries give, for the row player in two prisoner's dilemmas, his fitnesses [(i),(ii)] or his preferences when he is an altruist [(i-a),(ii-a) corresponding to (i),(ii) respectively].

types, choose action profiles from a richer set of options. Their preferences are the same, but the menu is larger.

5.3. Altruists

Similarly to those of magical thinkers, the preferences of altruists differ from those of a fitness maximizing individual. Altruists will, when given the opportunity, sacrifice some amount of their own fitness in order to increase the fitness of others. This can sometimes solve coordination problems. Consider utilitarian altruists whose preferences correspond to maximizing the average fitness of those playing a game. For the prisoner's dilemma in Figure 5(i) this gives preferences as in Figure 5(i-a). Given these preferences, an altruist will act as a cooperator and so, as discussed above, altruism will not proliferate when rare in the absence of assortative matching. However, altruism may not even solve the coordination problem to begin with, even for prisoner's dilemmas. For the prisoner's dilemma in Figure 5(ii), a utilitarian altruist will still face the coordination problem of Figure 5(ii-a). A similar comment applies to threshold public goods problems (Figure 4).

5.4. Conditional cooperators

Conditional cooperators identify the type of those with whom they interact and condition their action choice on this information (Hamilton, 1964a,b). This

has been called a *green-beard* effect (Dawkins, 1976), as individuals with some observable characteristic - a "green beard", behave cooperatively towards other individuals with this characteristic. When there is a unique "cooperative action" ($S_i = \{\times, +_1\}$, (PG) holds), conditional cooperators who play $+_1$ if and only if there are least n conditional cooperators in the group are similar to SI types and will proliferate when rare. This is the case for the threshold public goods model of Section 2.5 and for prisoner's dilemmas.

However, it is not clear how conditional cooperation should work when there are multiple ways to contribute ($N \geq 2$). In particular, if collaborative opportunities are asymmetric, such as in the Chase and Ambush game in Figure 1(ii), something more than merely conditioning on the other players being conditional cooperator types is required. One possibility would be for there to exist multiple types of conditional cooperator. For example, there could exist individuals with multiple shades of green beard, with the individual who sports the lighter shade of beard playing $+_1$ and the darker individual playing $+_2$. What this conditioning is of course doing, is to implement jointly profitable behavior. That is, the individuals concerned act as if they can collaboratively share intentions in the game under consideration. The form that conditional cooperation takes would have to differ from game to game. How much better would it be for an individual to have a comprehensive, multipurpose faculty that he could use for all such problems? That faculty is, of course, what we model in our SI types.

6. Conclusion: a modest proposal

It has been shown that for broad classes of games we can expect at least some degree of agency to act at a collective level as if motivated by shared intentions. More specifically, we can expect to observe behavior that accords with some degree of agency being exercised at a collective level. The paper is silent as to how this collective agency is created, which as noted in the introduction, could be via explicit communication, tacit understanding, or team reasoning. Such an approach is not unusual to economics, where concepts such as the 'firm' and

the 'household' are frequently used. It is clear that when decisions at a firm or household level are discussed, some degree of collective agency must be present, although the nature of this collective agency is not usually made explicit.

However, game theory in economics is in a weaker position. The most commonly used solution concept, Nash equilibrium, is habitually used without any explicit justification. Moreover, many of the Nash equilibria that occur in the literature are not Strong Equilibria; they are not robust to coalitional deviation, or to use the language of the current paper, from a status quo of such a Nash equilibrium, there exists an opportunity for collaboration. Therefore, an implication of the current work is that when using the concept of Nash equilibrium, an economist should ask whether the equilibrium is a Strong Equilibrium, and if it is not, should carefully consider the extent to which joint agency might be expected to manifest itself in the problem under consideration. For example, does the problem satisfy (PG), or would a collaborative move away from the Nash equilibrium in question be likely to satisfy (PC)? How large would the gains be for collaborators? What would the externalities of collaboration be? Moreover, this proposal is not just predicated on the work here, but also on rich empirical evidence that the ability to share intentions and engage in goal directed action is a basic human trait that cannot be ignored by any field that purports to scientifically consider human action.

Appendix A. Proofs

Denote the status quo payoff, which by symmetry is identical for all $i \in M$, by

$$\underline{\pi} := \pi_i(\underline{\times}), \quad i \in M.$$

Let $\underline{\underline{\pi}}$ be the lowest payoff in Γ that corresponds to some $s \in S$ and is strictly greater than $\underline{\pi}$.

$$\underline{\underline{\pi}} := \min_{\substack{s \in S \\ \pi_i(s) > \underline{\pi}}} \pi_i(s), \quad i \in M.$$

Let the highest payoff in Γ that corresponds to some $s \in S$ be

$$\pi_{max} := \max_{s \in S} \pi_i(s), \quad i \in M.$$

Proof of Theorem 1. When a set of SI types in a group collaborates, any N type will obtain a payoff of no more than π_{max}. The average fitness of an N type is then bounded above by

$$f_N(x) \leq \underbrace{Pr_x[Z < n \,|\, N]}_{\substack{\text{Prob. too few} \\ \text{SI types for} \\ \text{collaboration}}} \underline{\pi} + \underbrace{Pr_x[Z \geq n \,|\, N]}_{\substack{\text{Prob. some set} \\ \text{of SI types} \\ \text{collaborates}}} \pi_{max}.$$

When collaboration occurs, by (C) and (PG), any SI type within the group will obtain an expected payoff of at least

$$\pi_{min} := \frac{n}{m}\underline{\pi} + \left(1 - \frac{n}{m}\right)\underline{\pi} > \underline{\pi},$$

where n/m is a lower bound on the probability of any given SI type individual being one of the collaborators. The average fitness of an SI type is then bounded below by

$$f_{SI}(x) \geq \underbrace{Pr_x[Z < n-1 \,|\, SI]}_{\substack{\text{Prob. too few} \\ \text{SI types for} \\ \text{collaboration}}} \underline{\pi} + \underbrace{Pr_x[Z \geq n-1 \,|\, SI]}_{\substack{\text{Prob. some set} \\ \text{of SI types} \\ \text{collaborates}}} \pi_{min}.$$

Subtracting,

$$\begin{aligned} f_{SI}(x) - f_N(x) &= (f_{SI}(x) - \underline{\pi}) - (f_N(x) - \underline{\pi}) \\ &\geq Pr_x[Z \geq n-1 \,|\, SI](\pi_{min} - \underline{\pi}) - Pr_x[Z \geq n \,|\, N](\pi_{max} - \underline{\pi}). \end{aligned} \tag{A.1}$$

Now, (B) implies that for small enough x_{SI},

$$Pr_x[Z \geq n-1 \,|\, SI] > Pr_x[Z \geq n \,|\, N] \left(\frac{\pi_{max} - \underline{\pi}}{\pi_{min} - \underline{\pi}}\right),$$

which, from (A.1), implies that $f_{SI}(x) - f_N(x) > 0$ for small enough x_{SI}. That is, $x_{SI} = 0$ cannot be an ESS, so any ESS must have $x_{SI} > 0$. □

Proof of Corollary 1. When $n = m$, unless $M_{SI} = M$, $s^* = \underline{\times}$. Therefore, for any x, the fitness of an N type is

$$f_N(x) = \underline{\pi},$$

but the fitness of an SI type is bounded below by

$$f_{SI}(x) \geq Pr_x[Z < m-1 \,|\, SI]\,\underline{\pi} + Pr_x[Z = m-1 \,|\, SI]\,\pi_{min} > \underline{\pi}.$$

□

Proof of Example 1. The term in brackets in (3), simplified and divided by c equals

$$\frac{b}{c} - \sum_{k=n-1}^{m-1} \frac{n!(m-n)!}{(k+1)!(m-1-k)!} \left(\frac{x_{SI}}{1-x_{SI}}\right)^{k-(n-1)}, \tag{A.2}$$

which is clearly strictly decreasing in x_{SI} when $n < m$, approaches $b/c - 1 > 0$ as $x_{SI} \to 0$, and diverges to $-\infty$ as $x_{SI} \to 1$. Therefore (A.2) equals zero and (3) crosses zero at some unique x^*, is strictly positive for all x such that $x_{SI} \in (0, x^*_{SI})$, and is strictly negative for all x such that $x_{SI} \in (x^*_{SI}, 1)$. Therefore x^* is the unique ESS and the replicator dynamic converges to x^* from all x such that $x_{SI} \in (0,1)$. Now, (A.2) increases in b/c, n and decreases in m, so the value of x_{SI} at which (A.2) equals zero must increase or decrease respectively.

□

Proof of Theorem 2.

$$f_N(x) - \underline{\pi} = \sum_{k=n}^{m-1} Pr_x[Z = k](\pi_N^k - \underline{\pi})$$

and

$$f_{SI}(x) - \underline{\pi} = \sum_{k=n-1}^{m-1} Pr_x[Z = k](\pi_{SI}^{k+1} - \underline{\pi}),$$

giving

$$f_{SI} - f_N = Pr_x[Z = n-1] \left((\pi_{SI}^n - \underline{\pi}) + \sum_{k=n}^{m-1} \frac{Pr_x[Z=k]}{Pr_x[Z=n-1]} (\pi_{SI}^{k+1} - \pi_N^k) \right). \tag{A.3}$$

Now, for a given individual $i \in M_{SI}$, when $|M_{SI}| = n$, all possible collaborative opportunities involve i, so (PC) implies that $(\pi_{SI}^n - \underline{\pi}) > 0$. Furthermore, as

discussed in the main body of the paper, (B) and (NA) imply that for $k \geq n$, ${}^{Pr_x[Z=k]}/_{Pr_x[Z=n-1]} \to 0$ as $x_{SI} \to 0$. This implies that for small enough x_{SI}, the right hand side of (A.3) is strictly positive, so any ESS must have $x_{SI} > 0$. □

Proof of Theorem 3. For $x_\varepsilon = (1-\varepsilon)\hat{x} + \varepsilon\tilde{x}$,

$$g_{\hat{x}}(x_\varepsilon) - g_{\tilde{x}}(x_\varepsilon) = \left(f_{SI}(\bar{\sigma}(x_\varepsilon)) - f_N(\bar{\sigma}(x_\varepsilon))\right)\left(\bar{\sigma}(\hat{x}) - \bar{\sigma}(\tilde{x})\right).$$

If $\bar{\sigma}(\hat{x}) = \bar{\sigma}(\tilde{x})$, we have $g_{\hat{x}}(x_\varepsilon) - g_{\tilde{x}}(x_\varepsilon) = 0$. If $\bar{\sigma}(\hat{x}) > \bar{\sigma}(\tilde{x})$, then $\bar{\sigma}(\hat{x}) > \bar{\sigma}(x_\varepsilon)$, and if $\bar{\sigma}(\hat{x}) < \bar{\sigma}(\tilde{x})$, then $\bar{\sigma}(\hat{x}) < \bar{\sigma}(x_\varepsilon)$. So $\hat{x}$ is a NSS if and only if for small enough $\delta > 0$,

$$f_{SI}(\bar{\sigma}(\hat{x}) - \delta) - f_N(\bar{\sigma}(\hat{x}) - \delta) \geq 0, \qquad f_{SI}(\bar{\sigma}(\hat{x}) + \delta) - f_N(\bar{\sigma}(\hat{x}) + \delta) \leq 0,$$

which, as (Bin) implies that no part of $f_{SI}(.) - f_N(.)$ is linear, is the same condition on $\bar{\sigma}(\hat{x})$ as that placed on x^*_{SI} for an ESS of the two type model. □

References

Alger, I., Weibull, J.W., 2013. Homo moralis–preference evolution under incomplete information and assortative matching. Econometrica 81, 2269–2302.

Alvard, M., 2001. Mutualistic hunting, in: Stanford, C., Bunn, H. (Eds.), The early human diet: The role of meat. Oxford University Press, Oxford, pp. 261–278.

Alvard, M.S., Nolin, D.A., 2002. Rousseau's whale hunt? Current Anthropology 43, 533–559.

Ambrus, A., 2009. Theories of coalitional rationality. Journal of Economic Theory 144, 676 – 695.

Angus, S.D., Newton, J., 2015. Shared intentions and the advance of cumulative culture in hunter-gatherers. arXiv:1503.06522v2 [q-bio.PE] http://arxiv.org/abs/1503.06522.

Aumann, R., 1959. Acceptable points in general cooperative n-person games, in: Tucker, A.W., Luce, R.D. (Eds.), Contributions to the Theory of Games IV. Princeton University Press, pp. 287–324.

Bacharach, M., 1999. Interactive team reasoning: a contribution to the theory of co-operation. Research in economics 53, 117–147.

Bacharach, M., 2006. Beyond individual choice: teams and frames in game theory. Princeton University Press.

Bernheim, B.D., Peleg, B., Whinston, M.D., 1987. Coalition-proof nash equilibria i. concepts. Journal of Economic Theory 42, 1–12.

Bomze, I.M., 1991. Cross entropy minimization in uninvadable states of complex populations. Journal of Mathematical Biology 30, 73–87.

Bomze, I.M., Weibull, J.W., 1995. Does neutral stability imply Lyapunov stability? Games and Economic Behavior 11, 173–192.

Bowles, S., 2006. Group competition, reproductive leveling, and the evolution of human altruism. Science 314, 1569–1572.

Bratman, M.E., 1992. Shared cooperative activity. The Philosophical Review 101, 327–341.

Butterfill, S., 2012. Joint action and development. The Philosophical Quarterly 62, 23–47.

Call, J., 2009. Contrasting the social cognition of humans and nonhuman apes: The shared intentionality hypothesis. Topics in Cognitive Science 1, 368–379.

Choi, J.K., Bowles, S., 2007. The coevolution of parochial altruism and war. Science 318, 636–640.

Dawkins, R., 1976. The selfish gene. revised edn. 1989 Oxford .

Dekel, E., Ely, J.C., Yilankaya, O., 2007. Evolution of preferences. The Review of Economic Studies 74, 685–704.

Elster, J., 1979. Ulysses and the Sirens: Studies in Rationaltiy and Irrationality. Cambridge University Press.

Eshel, I., Cavalli-Sforza, L.L., 1982. Assortment of encounters and evolution of cooperativeness. Proceedings of the National Academy of Sciences 79, 1331–1335.

Farrell, J., Maskin, E., 1989. Renegotiation in repeated games. Games and economic behavior 1, 327–360.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection, ISBN 0198504403, variorum ed.(2000). Oxford University Press, USA.

Gavrilets, S., 2014. Collective action and the collaborative brain. Journal of The Royal Society Interface 12.

Gilbert, M., 1990. Walking together: A paradigmatic social phenomenon. Midwest Studies in Philosophy 15, 1–14.

Gillies, D.B., 1959. Solutions to general non-zero-sum games. Contributions to the Theory of Games 4, 47–85.

Gold, N., Sugden, R., 2007. Collective intentions and team agency. The Journal of Philosophy 104, 109–137.

Güth, W., Kliemt, H., 1998. The indirect evolutionary approach: Bridging the gap between rationality and adaptation. Rationality and Society 10, 377–399.

Haldane, J.B.S., 1932. The causes of evolution. Princeton University Press, 1990 ed.

Hamilton, W., 1964a. The genetical evolution of social behaviour. i. Journal of Theoretical Biology 7, 1 – 16.

Hamilton, W., 1964b. The genetical evolution of social behaviour. {II}. Journal of Theoretical Biology 7, 17 – 52.

Hamilton, W.D., 1963. The evolution of altruistic behavior. American naturalist 97, 354–356.

Kant, I., 1786. What does it mean to orient oneself in thinking?

Maynard Smith, J., 1982. Evolution and the Theory of Games. Cambridge university press.

Moll, H., Tomasello, M., 2007. Cooperation and human cognition: the Vygotskian intelligence hypothesis. Philosophical Transactions of the Royal Society B: Biological Sciences 362, 639–648.

Myatt, D.P., Wallace, C., 2008. When does one bad apple spoil the barrel? an evolutionary analysis of collective action. The Review of Economic Studies 75, 499–527.

Newton, J., 2012. Coalitional stochastic stability. Games and Economic Behavior 75, 842–54.

Newton, J., Angus, S.D., 2015. Coalitions, tipping points and the speed of evolution. Journal of Economic Theory 157, 172 – 187.

Pacheco, J.M., Santos, F.C., Souza, M.O., Skyrms, B., 2009. Evolutionary dynamics of collective action in n-person stag hunt dilemmas. Proceedings of the Royal Society of London B: Biological Sciences 276, 315–321.

Robson, A.J., 1996. The evolution of attitudes to risk: Lottery tickets and relative wealth. Games and economic behavior 14, 190–207.

Samuelson, L., 2001. Introduction to the evolution of preferences. Journal of Economic Theory 97, 225–230.

Searle, J., 1990. Collective intentions and actions, in: Cohen, P.R., Morgan, J., Pollack, M. (Eds.), Intentions in communication. MIT Press, pp. 401–15.

Smith, E.A., 2003. Human cooperation: Perspectives from behavioral ecology, in: P.Hammerstein (Ed.), Genetic and cultural evolution of cooperation. MIT Press, pp. 401–427.

Sosis, R., Feldstein, S., Hill, K., 1998. Bargaining theory and cooperative fishing participation on Ifaluk atoll. Human Nature 9, 163–203.

Sugden, R., 2000. Team preferences. Economics and Philosophy 16, 175–204.

Taylor, P.D., Jonker, L.B., 1978. Evolutionary stable strategies and game dynamics. Mathematical biosciences 40, 145–156.

Tomasello, M., 2014. A natural history of human thinking. Harvard University Press.

Tomasello, M., Carpenter, M., 2007. Shared intentionality. Developmental science 10, 121–125.

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., 2005. Understanding and sharing intentions: The origins of cultural cognition. Behavioral and brain sciences 28, 675–691.

Tomasello, M., Herrmann, E., 2010. Ape and human cognition what's the difference? Current Directions in Psychological Science 19, 3–8.

Tomasello, M., Rakoczy, H., 2003. What makes human cognition unique? from individual to shared to collective intentionality. Mind & Language 18, 121–147.

Trivers, R.L., 1971. The evolution of reciprocal altruism. Quarterly review of biology 46, 35–57.

Tuomela, R., Miller, K., 1988. We-intentions. Philosophical Studies 53, 367–389.

Velleman, J.D., 1997. How to share an intention. Philosophy and Phenomenological Research: A Quarterly Journal 57, 29–50.

Vygotsky, L.S., 1980. Mind in society: The development of higher psychological processes. Harvard University Press.

Wilson, D.S., Dugatkin, L.A., 1997. Group selection and assortative interactions. American Naturalist 149, 336–351.

Wobber, V., Herrmann, E., Hare, B., Wrangham, R., Tomasello, M., 2014. Differences in the early cognitive development of children and great apes. Developmental Psychobiology 56, 547–573.

Young, H.P., 1993. The evolution of conventions. Econometrica 61, 57–84.