Games and Economic Behavior 104 (2017) 517–534

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Shared intentions: The evolution of collaboration ☆

Jonathan Newton [1]

ARTICLE INFO

Article history:
Received 13 September 2016
Available online xxxx

JEL classification:
C73

Keywords:
Shared intentions
Collaboration
Evolution
Game theory

ABSTRACT

The ability to form shared intentions and adjust one's choices in collaboration with others is a fundamental aspect of human nature. We discuss the forces that act for and against the evolution of this ability. In contrast to altruism and other non-fitness maximizing preferences, for large classes of games the ability to form shared intentions proliferates when rare without requiring group selection or assortativity in matching.

"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 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 form *shared 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 (Tuomela and Miller, 1988; Searle, 1990; Bratman, 1992). That is, "we intend to do X" is distinct from "I intend to do X [because I think that she also intends to do X]".[2] Through conversation, pointing and

☆ Follow the latest research at http://sharedintentions.net.

E-mail address: jcsnewton@gmail.com.

[1] This work originated in discussions with Sung-Ha Hwang about my work with Simon Angus. In addition, my work on this topic has benefited from the comments and criticisms of Ingela Alger, Herb Gintis, Natalie Gold, Toby Handfield, Atsushi Kajii, Kohei Kawamura, Christoph Kuzmics, Erik Mohlin, Heinrich Nax, Bary Pradelski, Larry Samuelson, William Sandholm, Ryoji Sawa, Peyton Young and anonymous referees; as well as seminar audiences at Edinburgh, Oxford, Cologne, Bielefeld, ANU, Sydney, Monash, Aizu, Kyoto, Keio, Tokyo, Hitotsubashi, UTS, HKUST and Surrey. Many thanks to Xiaotong Sun for help with diagrams.

[2] There is disagreement amongst philosophers as to what extent shared intentions can be reduced to individual intentions. See also Butterfill (2012), Gilbert (1990), Gold and Sugden (2007), Velleman (1997). We take no position on this. Our results hold regardless of how agents form shared intentions.

http://dx.doi.org/10.1016/j.geb.2017.06.001
0899-8256/

gesturing, or alternative forms of reasoning such as 'team reasoning' (Bacharach, 1999, 2006; Sugden, 2000) people form shared intentions. When combined with notions of optimization, shared intentions naturally give rise to collective agency. To see this, 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 (Aumann, 1959), Coalition Proofness (Bernheim et al., 1987), Coalitional Rationalizability (Ambrus, 2009), Renegotiation Proofness (Farrell and Maskin, 1989) and Coalitional Stochastic Stability (Newton, 2012a).

This paper demonstrates how conditions faced by primitive 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 across a large variety of games, including threshold public goods games, m-player prisoner's dilemmas, trust games, the centipede game, Nash demand games, Bertrand oligopoly, minimum effort games and, under an additional condition, finitely or infinitely repeated versions of all of the above. 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; Newton, 2016), nor group selection (Haldane, 1932; Choi and Bowles, 2007; Bowles, 2006) is required for shared intentions to evolve. This is in stark contrast to cooperator types or altruists, who in similar circumstances fail to proliferate when rare in games such as prisoner's dilemmas and threshold public goods games (see, e.g. Pacheco et al., 2009; Hauert et al., 2006).

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 Herrmann, 2010; Tomasello and Carpenter, 2007).[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 the shared intentionality hypothesis (Call, 2009) or the Vygotskian intelligence hypothesis (Vygotsky, 1980; Tomasello, 2014; Moll and Tomasello, 2007).[4] The results of the current paper show how in populations of unsophisticated agents, collaborative behavior can evolve or fail to evolve, raising the prospect of rigorous, game theoretic analysis of the factors that led to differing degrees of collaborative behavior across species.

The author knows of only two other works that deal directly with the topic of the current paper[5,6]: 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 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 games. In contrast to previous work, there is no group selection and selective pressure against the sharing of intentions arises from either (i) free riding on the positive externalities of collaboration by others, or (ii) negative externalities of collaboration on other potential collaborators. Finally, it is instructive to compare the

[3] See also Wobber et al. (2014), Tomasello et al. (2005) and the accompanying critical responses.

[4] One specific possibility is that the ability to form shared intentional states was fundamental to the evolution of language. For a discussion of the former in the context of a broad review of the latter, see Fitch (2010).

[5] 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.

[6] 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.