Games and Economic Behavior 87 (2014) 122–135

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Preferences and beliefs in a sequential social dilemma: a within-subjects analysis ☆

Mariana Blanco [a], Dirk Engelmann [b,c,d,e], Alexander K. Koch [f], Hans-Theo Normann [g,h,*]

[a] Universidad del Rosario, Economics Department, Calle 14 No. 4-80, Bogotá, Colombia
[b] Department of Economics, University of Mannheim, L7, 3-5, D-68131 Mannheim, Germany
[c] Centre for Experimental Economics, University of Copenhagen, Denmark
[d] CERGE-EI, Prague, Czech Republic
[e] CESIfo, Munich, Germany
[f] School of Economics and Management, Aarhus University, Building 1322, 8000 Aarhus C, Denmark
[g] Duesseldorf Institute for Competition Economics (DICE), Duesseldorf University, 40225 Duesseldorf, Germany
[h] Max-Planck Institute for Research on Collective Goods, Germany

ARTICLE INFO

Article history:
Received 1 February 2011
Available online 17 May 2014

JEL classification:
C72
C90

Keywords:
Beliefs
Consensus effect
Social dilemma
Experimental economics

ABSTRACT

In empirical analyses of games, preferences and beliefs are typically treated as independent. However, if beliefs and preferences interact, this may have implications for the interpretation of observed behavior. Our sequential social dilemma experiment allows us to separate different interaction channels. When subjects play both roles in such experiments, a positive correlation between first- and second-mover behavior is frequently reported. We find that the observed correlation primarily originates via an indirect channel, where second-mover decisions influence beliefs through a consensus effect, and the first-mover decision is a best response to these beliefs. Specifically, beliefs about second-mover cooperation are biased toward own second-mover behavior, and most subjects best respond to stated beliefs. However, we also find evidence for a direct, preference-based channel. When first movers know the true probability of second-mover cooperation, subjects' own second moves still have predictive power regarding their first moves.

1. Introduction

Behavioral economic theory offers a wide range of models that predict how actions in social dilemmas will vary for people with different types of (social) preferences and what an individual's best response is for a given set of beliefs.[1] While these models broaden the spectrum of preferences that people may hold, they typically stick to the standard assumption

☆ Financial support from the Nuffield Foundation, grant No. SGS/34070, is gratefully acknowledged. We thank Steffen Altmann, Maria Bigoni, Steve Burks, Ernst Fehr, Simon Gächter, Sebastian Kranz, Louis Levy-Garboua, Michael Naef, Daniele Nosenzo, Matthias Wibral as well as the anonymous advisory editor and referee for helpful comments.

* Corresponding author.
E-mail addresses: mariana.blanco@urosario.edu.co (M. Blanco), dirk.engelmann@uni-mannheim.de (D. Engelmann), akoch@econ.au.dk (A.K. Koch), normann@dice.hhu.de (H.-T. Normann).

[1] Frequently, beliefs are considered to be distributions that merely rationalize revealed preference orders. We follow here the interpretation that beliefs are real, meaning that they are an independent part of decision making, implying that they can be elicited in experiments. See also Costa-Gomes et al. (2010).

http://dx.doi.org/10.1016/j.geb.2014.05.005
0899-8256/

that people hold correct beliefs (in equilibrium).[2] The downside with this approach is to miss a crucial point: how likely a person thinks it is that others will defect in a social dilemma may well depend on her own attitude toward cooperation. As such an interaction of preferences and beliefs is of general importance for decision making in games, the topic appears to be strangely underdeveloped in the economic literature.

The significance of this issue is underlined by recent findings from sequential social dilemma experiments using a within-subjects design.[3] The data show that subjects who defect as first movers are more likely to exploit first-mover cooperation in their second-mover choice than those who cooperate as first movers. Blanco et al. (2011) document this for the sequential-move prisoners' dilemma.[4] Altmann et al. (2008) and Gächter et al. (2012) have a similar result for the trust game and for a sequential voluntary contribution game, respectively.

The observed within-subjects correlation of the first and the second move is provocative in several ways. First, as noted by Altmann et al. (2008) and Blanco et al. (2011), the finding is at odds with prominent social preference models that are frequently invoked for explaining behavior in social dilemma games. Both *inequality aversion* (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) and *reciprocal preferences* (Dufwenberg and Kirchsteiger, 2004) – under standard assumptions, including that beliefs are not correlated with preferences – would predict a negative correlation of first- and second-mover choices, and not the positive correlation observed. While altruism could rationalize the correlation pattern it would also predict unconditional cooperation, which, however, is at odds with the data (see Bolle and Ockenfels, 1990; Clark and Sefton, 2001, and Blanco et al., 2011).

Second, for simultaneous-move prisoners' dilemma experiments, it has been argued that "fear" and "greed" are main driving forces of behavior (Ahn et al., 2001; Simpson, 2003). Fear refers to the risk of being exploited by the other player when cooperating. Greed describes a player's willingness to defect if the other player cooperates. The sequential-move prisoners' dilemma separates the two motives: fear applies to the first move and greed to the second move. Thus, the correlation of first and second moves suggests that fear and greed are correlated at the individual level. But it does not seem evident why fearful people should be more greedy.

Third and more fundamentally, following standard game-theoretic arguments, first-mover choices should follow a "best respond to your beliefs" principle,[5] and hence reflect the natural variation in beliefs across subjects in an experiment. Second-mover choices, in contrast, are simple decision problems and should depend on players' preferences only. Thus, one would not expect the choices of a person in the role of first and second mover to be strongly related to each other – unless beliefs and preferences are correlated.

A correlation between preferences and beliefs may, however, be exactly what drives the correlation between first-mover and second-mover decisions. The so-called *consensus effect*, according to which players' beliefs are biased toward their own type, would suggest that those subjects who cooperate as second movers will expect a higher second-mover cooperation rate among others than those subjects who defect as second movers (Mullen et al., 1985; Engelmann and Strobel, 2000). Second-move cooperators hence will perceive a higher expected payoff from cooperating as first mover than second-move defectors. So, all else equal (that is, if there is no relationship between preferences for cooperation in the role of first and second mover), second-move cooperators should be more likely than second-move defectors to cooperate as first mover.

Another response to the above issues raised by the experimental data is to turn to alternative social preference models that are consistent with the observed correlation of choices without assuming systematic differences in beliefs across players. A combination of efficiency concerns with maximin preferences (Charness and Rabin, 2002) and reciprocal altruism (Levine, 1998; Cox et al., 2008) are among the alternatives that can explain why first-mover decisions differ between second-mover cooperators and second-mover defectors, even if they hold the same beliefs.

The aforementioned theories presume a *direct, preference-based channel* that influences both first- and second-mover behavior. The consensus effect, in contrast, suggests an *indirect channel* that links preferences (as reflected in a person's second-mover decision) to the first-mover decision via beliefs. But what is the right approach?

The issue of indirect versus direct channel seems particularly relevant because the consensus effect has emerged already in other settings as a plausible alternative to preference-based explanations in rationalizing certain patterns of behavior. For instance, dictator- and trust-game studies where participants report what they believe their counterpart expects in the game, show significant correlations between these second-order beliefs and actions. An explanation for this pattern is that some people are guilt averse. That is, they experience a utility loss if they believe to let someone down (Charness and Dufwenberg, 2006). But Ellingsen et al. (2010) conclude from their own experiments that the correlation can almost

[2] Osborne (2009, p. 379) presents this as the standard approach. Some approaches within behavioral economics relax the assumption of correct beliefs. For example, the level-k literature is explicitly based on assuming very different (non-equilibrium) beliefs. These models have, however, typically not been applied to explaining behavior in social dilemmas.

[3] Earlier experimental analyses of sequential social dilemmas include the sequential-move prisoners' dilemma (Bolle and Ockenfels, 1990; Clark and Sefton, 2001), the gift-exchange game (Fehr et al., 1993), the trust or investment game (Berg et al., 1995), the lost wallet game (Dufwenberg and Gneezy, 2000), and public-good games with a front runner (Potters et al., 2007).

[4] Blanco et al. (2011) check for the within-subjects correlation of six different moves in four different games. The correlation of the first and the second move (given first-mover cooperation) in the sequential-move prisoners' dilemma was the strongest among all 15 correlations.

[5] For recent experiments investigating this issue see, for example, Dhaene and Bouckaert (2010), Costa-Gomes and Weizsäcker (2008), Rey-Biel (2009), and Koch et al. (2009). On the fundamental question whether beliefs are causal for behavior see Costa-Gomes et al. (2010). In a trust game they exogenously shift the trustee's repayment and use this shift to instrument the trustor's beliefs. Their results provide evidence of a causal effect of beliefs on actions.