A short note on the rationality of the false consensus effect

Christoph Vanberg*

Department of Economics, University of Heidelberg

Bergheimer Str. 58, 69115 Heidelberg, Germany

Abstract

A number of experimental studies have documented systematic correlations between subjects' behavior and their higher-order beliefs. Such evidence is often interpreted as indicating a causal relationship between beliefs and behavior, as predicted by models in the Psychological Game Theory framework. An alternative explanation attributes them to what psychologists refer to as a 'false consensus effect.' The latter explanation is often discounted on the grounds that it is based on an implausible psychological bias. The goal of this note is to show that the false consensus effect does not rely on such a bias. I demonstrate that rational belief formation implies a correlation of behavior and beliefs of all orders whenever behaviorally relevant traits are drawn from an unknown common distribution.

Keywords: Beliefs, false consensus effect, guilt aversion, experimental economics, behavioral economics, psychological game theory

JEL Codes: C7, C9, D8

*Email: vanberg@uni-hd.de, Tel: +49 6221 54 2947

1 Introduction

A number of experimental studies have documented systematic correlations between subjects' behavior and their higher-order beliefs, e.g. beliefs about what other players expect them to do. Such evidence is often interpreted as evidence of a causal relationship between beliefs and behavior, as predicted by Psychological Game Theory models such as Guilt aversion (Bacharach et al., 2007; Charness and Dufwenberg, 2006; Dufwenberg and Gneezy, 2000; Guerra and Zizzo, 2004).

An alternative explanation attributes belief-behavior correlations to what psychologists refer to as a 'false consensus effect' (Ross et al., 1977). According to this explanation, subjects may systematically overestimate the extent to which others think and behave as they do. Usually, this is taken to mean that "a person believes others would act similarly rather than that a person believes others believe he or she would make a certain choice" (Charness and Dufwenberg, 2006). In the context of testing psychological game theory models, various authors have raised the concern that subjects may overestimate the extent to which others expect them to behave as they actually do (Ellingsen et al., 2009; Vanberg, 2008). It is my impression that this concern is often discounted because it is thought to be based on an implausible psychological bias. The goal of this note is to show that the false consensus effect does not rely on such a bias. I demonstrate that rational belief formation implies a correlation of behavior and beliefs of all orders whenever behaviorally relevant traits are drawn from unknown but common distribution.

Section 2 presents a simple model to illustrate the argument. Section 3 extends the model to show how experimental treatments can simultaneously affect beliefs and behavior. Section 4 concludes and discusses the underlying assumption that psychological traits are drawn from a common distribution. [1]

[1] It has been brought to my attention that an argument similar to the one I am formalizing here has previously been made by Dawes (1989). I hope that the simple formalization I provide has some added value, especially in relating the false consensus effect to higher order beliefs relevant to Psychological Game Theory.

2 Model

Consider a world with two states labeled $\theta \in \{\theta_L, \theta_H\}$, both equally likely. There are $N \geq 2$ players, who each have two available actions, a_L and a_H. Each player i receives a private signal $s_i \in \{s_L, s_H\}$. In state θ_K, the probability that $s_i = s_K$ is equal to $p > \frac{1}{2}$. Thus, each agent's private signal is correlated with the state of the world, which is common to all agents. Assume that behavior is entirely determined by an agent's signal. Specifically, when $s_i = s_K$, agent i takes action a_K.[2]

By construction, behavior in this example is not a function of an agent's *beliefs*. None the less, it is easy to show that second order beliefs will be perfectly correlated with behavior. To see this, note first that a player who receives signal s_K attaches probability $\frac{\frac{1}{2}p}{\frac{1}{2}p+\frac{1}{2}(1-p)} = p > \frac{1}{2}$ to state θ_K. Thus, the posterior probability that another agent $j \neq i$ receives the *same* signal s_K is given by $q = p^2 + (1-p)^2 > \frac{1}{2}$.

It follows that an agent who receives signal s_K *first order believes* that another agent will take action a_K with probability $\mu_1(a_K|s_K) = q > \frac{1}{2}$. Now consider agent i's *second order* beliefs after receiving signal s_K. With probability q, agent $j \neq i$ receives the same signal s_K and (first order) believes that agent i will take action a_K with probability $\mu_1(a_K|s_K) = q$. With probability $(1-q)$, agent j receives signal s_{-K} and believes that i will take action a_K with probability $\mu_1(a_K|s_{-K}) = (1-q)$. Thus i *second order believes* that j attaches, *in expectation*, a probability $\overline{\mu}_2(a_K|s_K) = q^2+(1-q)^2 > \frac{1}{2}$ to her (i) choosing action a_K. Similarly, i believes that, in expectation, j attaches probability $\overline{\mu}_2(a_{-K}|s_K) = 2 \cdot q \cdot (1-q) = 1 - \overline{\mu}_2(a_K|s_K)$ to her choosing action a_{-K}.[3]

Suppose that we can observe behavior as well as second order beliefs concerning the (expected) probability of choosing action a_K. Without loss of generality, consider

[2]The signal can be interpreted in any number of ways. It may reflect a player's *type* in terms of intrinsic motivations to choose an action, or it may reflect information concerning the state of the world, on which action preferences depend. What's important is that the signal *causes* the agent to behave in one way or the other.

[3]With probability q, agent j believes that i will choose a_{-K} with probability $(1-q)$. With probability $(1-q)$, j attaches probability q to this event.

what will happen in state θ_K. Clearly, an expected fraction p of all agents will choose action a_K and have second order beliefs $\overline{\mu}_2(a_K|s_K) > \frac{1}{2}$. Conversely, an expected fraction $(1-p)$ of all agents will choose action a_{-K} and have second order beliefs $\overline{\mu}_2(a_K|s_{-K}) < \frac{1}{2}$. Thus, behavior will be perfectly correlated with second order beliefs even though it is *causally* determined only by the s_i.

This example shows that a rational agent's second order beliefs will tend to be correlated with her behavior if private factors relevant to choice (e.g. preferences, information) are correlated across agents. Thus, if experimental subjects believe that other subjects' preferences or information are similar to their own, we should expect to see a correlation of second order beliefs and behavior in *any* experimental setting, even if behavior is driven by other factors.

3 Extension: Treatment effects

The basic example can be extended to discuss the effects of a *treatment* variable on beliefs and behavior. In addition to the private signals s_i, all agents now observe a *public* signal $t \in \{0,1\}$. Suppose that this signal *directly* affects the behavior of some subjects. If $t = 0$, behavior is determined as before. If $t = 1$, a fraction $r \in (0,1]$ of all agents prefers action a_H, irrespective of their private signal. The remaining 'flexible' agents behave as before.

When $t = 0$, beliefs are determined as above. What happens to beliefs when $t = 1$? An agent that receives signal s_H will *first order believe* that others will choose action a_H with probability $\tilde{\mu}_1(a_H|s_H) = r + (1-r) \cdot q > q = \mu_1(a_H|s_H)$. An agent who receives signal s_L will *first order believe* that others will choose action a_H with probability $\tilde{\mu}_1(a_H|s_L) = r + (1-r) \cdot (1-q) > (1-q) = \mu_1(a_H|s_L)$. An agent who receives signal s_K will *second order believe* that another agent's first order belief is given by $\tilde{\mu}_1(a_H|s_K)$ with probability q, and $\tilde{\mu}_1(a_H|s_{-K})$ with probability $(1-q)$. In expectation, she believes that another agent attaches probability $\tilde{\mu}_2(a_H|s_K) = q \cdot \tilde{\mu}_1(a_H|s_K) + (1-q) \cdot \tilde{\mu}_1(a_H|s_{-K}) > \mu_2(a_H|s_K)$ to the event that she will choose

action a_H. Thus, both first and second order beliefs of all agents will put more weight on action a_H under the treatment condition.

Again, consider what would happen if we were to observe behavior and beliefs in this setting. Nothing changes relative to the previous example when $t = 0$. When $t = 1$ and $\theta = \theta_H$, an expected fraction $p + r \cdot (1 - p)$ of all agents will choose action a_H. (All those who receive signal s_H, plus those who receive signal s_L, but are sensitive to the treatment.) Among these agents, the mean second order belief will be $\beta(a_H, \theta_H) = \frac{p \cdot \tilde{\mu}_2(a_H|s_H) + r \cdot (1-p) \cdot \tilde{\mu}_2(a_H|s_L)}{p + r \cdot (1-p)}$. When $\theta = \theta_L$, an expected fraction $(1 - p) + r \cdot p$ will choose action a_H, and the mean second order belief among these agents will be $\beta(a_H, \theta_L) = \frac{(1-p) \cdot \tilde{\mu}_2(a_H|s_H) + r \cdot p \cdot \tilde{\mu}_2(a_H|s_L)}{(1-p) + r \cdot p}$. Among those choosing a_L, the mean second order belief associated with action a_H is equal to $\beta(a_L, \theta_K) = \tilde{\mu}_2(a_H|s_L)$.

Relative to the baseline condition $t = 0$, the treatment condition $t = 1$ causes the expected fraction of subjects choosing action a_H to increase by $r \cdot (1 - p)$ when $\theta = \theta_H$, and by $r \cdot p$ when $\theta = \theta_L$. Further, $\beta(a_H, \theta_K) > \beta(a_L, \theta_K)$ for $K = L, H$. That is, subjects choosing action a_H will have 'higher' second order beliefs than those choosing action a_L.

Thus, the data will have the following features: (1) beliefs and behavior are correlated *within* each of the treatment conditions (2) second order beliefs are correlated with the treatment condition, and (3) behavior is correlated with the treatment condition. Despite the fact that behavior is directly affected only by the treatment signal t, features (2) and (3) are consistent with the false hypothesis that behavior is causally driven by second order beliefs.

4 Conclusion

Using a simple example, I have shown that behavior and second order beliefs will be perfectly correlated whenever subjects believe that privately known factors relevant to their decisions are drawn from a common but unknown distribution. It is straightforward to extend this argument to any order of beliefs. If accepted, this argument

poses a significant challenge to researchers interested in testing Psychological Game Theory models such as Guilt or Let-down aversion.

Before concluding, it will be useful to consider the assumption that drives the result - namely, that factors relevant to subjects' choices are drawn from an unknown but common distribution. There are two ways to interpret these factors. The first is that they represent "tastes" for certain kinds of behavior. The second is that they represent subjective theories about what constitutes appropriate behavior in a given situation. The intuition underlying our main assumption is that (a) individuals share both tastes and theories with other members of their species and cultural group. Thus, when I am asked to guess whether others will enjoy, say, the taste of a previously unfamiliar food, I can regard my own reaction as an experiment that provides (albeit limited) data as to the likely reaction of other members of my culture or species. Similarly, when faced with an experimental task, I may use my own assessment of what is "fair" or "appropriate" as a predictor of what others will believe.

A potential objection to this argument is that subjects with minority preferences or opinions should eventually stop using them to predict those of others. However, this objection seems to apply only to situations encountered repeatedly. Faced with a novel situation, it seems natural that even "minority" subjects would initially assume that their own reaction is modal. This leads to the falsifiable prediction that the consensus effect should disappear if a game is repeated and information about others' behavior is provided. Testing this prediction may be worthwhile for future research.

References

Bacharach, M., Guerra, G., and Zizzo, D. J. (2007). The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 63:349–388.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, 25(11):1–17.

Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, 30:163–182.

Ellingsen, T., Johannesson, M., Tjotta, S., and Torsvik, G. (2009). Testing guilt aversion. *Games and Economic Behavior*, in press.

Guerra, G. and Zizzo, D. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior and Organization*, 55:25–30.

Ross, L., Greene, D., and House, P. (1977). The false consensus phenomenon: An attributional bias in self-perception and social perception processes. *Journal of Experimental Social Psychology*, 13(3):279–301.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations. *Econometrica*, 76:1467–1480.