University of Heidelberg

Department of Economics

Discussion Paper Series | No. 662

A short note on the rationality of the false consensus effect

Christoph Vanberg

May 2019

A short note on the rationality of the false consensus effect

Christoph Vanberg*

Abstract

In experiments which measure subjects' beliefs, both beliefs about others' behavior and beliefs about others' beliefs, are often correlated with a subject's own choices. Such phenomena have been interpreted as evidence of a causal relationship between beliefs and behavior. An alternative explanation attributes them to what psychologists refer to as a 'false consensus effect.' It is my impression that the latter explanation is often prematurely dismissed 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 an unknown common distribution. Thus, if we assume that subjects rationally update beliefs, correlations of beliefs and behavior cannot support a causal relationship.

KEYWORDS: Beliefs, behavioral economics, experimental economics

[1]University of Heidelberg, Department of Economics. Contact: vanberg@uni-hd.de

1 Introduction

My aim in this short note is to demonstrate that experimental economists should be careful when lending a causal interpretation to observed correlations of beliefs and behavior. An alternative interpretation of such a correlation is what psychologists refer to as the 'false consensus effect': People may systematically over-estimate the extent to which others behave and think as they do.

Most papers investigating ostensibly causal relationships between beliefs and behavior typically mention the false consensus effect as a potentially confounding factor. However, it is my impression that many authors do not consider it a major concern. A possible reason is that many experimental and behavioral economists think that a false consensus effect would have to be based on a rare psychological bias, and therefore the problem may be safe to ignore.[1] This short note demonstrates, using a simple model, that this is wrong. The correlations of beliefs and behavior that are conventionally referred to as a false consensus effect do not require any kind of psychological bias.

The basic argument is the following. If traits relevant to behavior are shared (formally, correlated), a rational agent should use his own inclinations to predict the behavior of others, his own beliefs to predict the beliefs of others, and so forth, up to arbitrary orders of belief. Absent additional information, it makes sense for people to hypothesize that others behave and think as they do, and (as a consequence) that others expect them to behave exactly as they do.

To illustrate the argument informally, imagine two tourists in an exotic country being offered the choice between two previously unfamiliar foods A and B. After inspecting the choices, the first tourist feels inclined to choose A. Now imagine asking him 'what do you think the other tourist will choose?' Then it is perfectly rational for the first tourist to think 'Well, I think A looks better, and he's probably similar to me, so I guess he will feel the same.' And indeed it is rational to conclude

[1] Readers familiar with the experimental literature on Psychological Game Theory will recognize what I am talking about. However I will deliberately refrain from citing specific sources here.

that the other tourist probably expects him to choose A as well, and to believe that the other tourist believes that he (tourist 1) expects him (tourist 2) to choose A, etc. ad infinitum.

It has been brought to my attention that part of the argument I am developing here was already presented in Dawes (1989). Indeed Dawes argued in essentially the same way that it is rational for an individual to use her own (binary) response to a task as an estimator of the average response in a population of which she herself is a member. While this is the essential ingredient in the arguments made below, my analysis will go slightly further in that I will show that rational belief formation will lead to correlations of behavior and beliefs of *any* order. In addition, I will explicitly consider how this phenomenon affects our ability to experimentally test theories using treament parameters. In particular, I will show that the experimenter may falsely attribute treatment effects to changes in (higher order) beliefs when if fact they are *directly* related to the treatment parameter.

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 second order 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

[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.

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]

Now, consider what will happen in state θ_K. Suppose that we can observe behavior as well as (mean) second order beliefs concerning the probability of choosing action a_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) are correlated across agents. Thus, if experimental subjects believe that other subjects' private preferences and inclinations 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. It is immediately obvious that the argument can be extended to yield the same conclusion for beliefs of *any order*.

[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.

3 Extension: Treatment effects

This example can be expanded 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, we can consider what would happen if we were to observe behavior and beliefs in this setting. Clearly, 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 $\beta\left(a_H, \theta_H\right) = \frac{p \cdot \tilde{\mu}^2(a_H|s_H) + r \cdot (1-p) \cdot \tilde{\mu}^2(a_H|s_L)}{p + r \cdot (1-p)}$. 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 $\beta\left(a_H, \theta_L\right) = \frac{(1-p) \cdot \tilde{\mu}^2(a_H|s_H) + r \cdot p \cdot \tilde{\mu}^2(a_H|s_L)}{(1-p) + r \cdot p}$. Among those choosing a_L, the mean second order belief associated with action a_H is equal to $\beta\left(a_L, \theta_K\right) = \tilde{\mu}^2(a_H|s_L)$.

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 by the treatment signal t, these features are consistent with the *false* hypothesis that behavior is causally driven by second order beliefs. If follows that data of this type cannot be used to support that hypothesis.

4 Conclusion

The simple model presented in this short note suggests that a rational agent's behavior may be perfectly correlated with his beliefs (of any order) even in a setting where beliefs do not causally affect behavior. The essential feature of the setting considered is that the behavior of agents belonging to a relevant reference group is determined by some individual characteristic which is drawn from the same (unknown) distribution. Substantively, this means that the members of the reference group are expected to be similar.

This assumption is natural and plausible in almost any application, including experimental games. When faced with an experimental decision task, an individual participant will feels a *disposition* to choose a certain option. This disposition reflects genetic, cultural, and other factors that make certain choices appear attractive or appropriate. Although these factors are likely to vary between individuals, it is reasonable for a subject to assume that they will be correlated within a reference group (typically, students of the same university).

If I sample an exotic food and find it delicious, it is reasonable for me to think that other members of my reference group are similarly disposed, and therefore I should expect that others will also find the food delicious. And indeed this logic can be extended to higher order beliefs, for example I should expect others to predict that I will find the food delicious, and to believe that I will predict the same about them, etc. ad infinitum

Since all of this is true when agents update their beliefs rationally, the phenomenon conventionally referred to as the 'false consensus effect' does *not* represent a psychological bias. This suggests that it should be taken seriously. If so, it represents a serious challenge to researchers attempting to test theories that stipulate direct effects of (higher order) beliefs on motivation and behavior. In particular, it is not the case that such theories can be supported by data that demonstrates a correlation of beliefs and behavior, be it within or between treatments (or both).

One way to test such theories would be to induce transparently *exogenous* variation in beliefs using treatment variables that affect only beliefs but not other factors relevant to a subject's choices. And a way to test theories stipulating a direct effect of a treatment condition (not via beliefs) is to induce exogenous variation in the treamtent condition while holding beliefs constant.[4]

References

Dawes, R. (1989). Statistical Criteria for Establishing a Truly False Consensus Effect, *Journal of Experimental Social Psychology*, 25, 1(17).

[4]Both of these strategies have been employed in practice, but as mentioned above I will deliberately refrain from referencing specific studies here.