Journal of Economic Behavior & Organization 106 (2014) 281–297

Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

Signs and (counter)signals of trustworthiness☆

Diego Gambetta[a,b], Áron Székely[c,d,*]

[a] European University Institute, Via dei Roccettini 9, San Domenico, Firenze, Italy
[b] Nuffield College, New Road, Oxford OX1 1NF, UK
[c] University of Oxford, Department of Sociology, Manor Road, Oxford OX1 3UQ, UK
[d] Institute of Cognitive Sciences and Technologies, CNR, Via San Martino della Battaglia 44, Rome, Italy

ARTICLE INFO

Article history:
Received 4 October 2013
Received in revised form 17 July 2014
Accepted 20 July 2014
Available online 28 July 2014

JEL classification:
D83
C72
C91

Keywords:
Information
Costly signals
Generosity
Trust
Laboratory experiments

ABSTRACT

When we become aware that our past actions carry information about qualities that we possess or lack, which others use to decide how to deal with us, are we unconcerned, content to rely on what we have done, or do we take action to alter this information? We study this question experimentally using generosity as a sign and a signal of trustworthiness, and a trust game. Subjects play a dictator game unaware that later they will play a trust game and that their level of generosity in the dictator game will be revealed to trusters, with some inaccuracy, before trusters decide whether to trust or not. Once made aware of what follows, trustees have the option to play a second dictator game, from which their choice will be accurately conveyed to trusters in addition to their decision in the initial game. Consistent with 'countersignalling theory', those who, in the first dictator game, were either miserly or generous do not play the second dictator game, resigned or content with the information conveyed by their past actions. Those neither miserly nor generous in the first dictator game, an intermediate generous group, are likeliest to use the second dictator game; many of them for the purpose of signalling, so that they are not confused with the miserly.

1. Introduction

We investigate a common case in which people realise that some past action of theirs, which they performed for motives other than informing others of their qualities, becomes known to others in situations in which this could have consequences for them.

Going about their lives, people act in myriad ways, good and bad. Many of these actions are trivial and go unnoticed, forever lost. But some actions are perceived by other people, other actions are recorded on media whether agents are aware of it or not, others still leave enduring marks on them, perceivable well after the event that produced them occurred. The information potential contained in these past actions can remain dormant. However, out of this baggage with which everyone travels, people often extract *signs* of other people's qualities, good or bad, and use this information in deciding how to deal with them.

☆ We thank Ozan Aksoy, Dominic Burbidge, Giorgia Gon, Dan Honig, Wojtek Przepiorka, Dingeman Wiertz, Nan Zhang, and two anonymous referees for their helpful comments. This work was supported by the Economic and Social Research Council [Grant number ES/I022066/1]. Data are available upon request from the authors.

* Corresponding author at: Institute of Cognitive Sciences and Technologies, CNR, Via San Martino della Battaglia 44, Rome, Italy. Tel.: +39 06 44362370x1; fax: +39 06445 952 43.
E-mail addresses: diego.gambetta@eui.eu (D. Gambetta), aron.szekely@sociology.ox.ac.uk (Á. Székely).

http://dx.doi.org/10.1016/j.jebo.2014.07.009
0167-2681/

Our question is: what do people do when they realise that this is happening? Under what conditions do they take further action or do nothing about it? If the sign is of a quality of theirs that can help them in the situation in which they find themselves, should they still invest resources in signalling to further strengthen the receiver's belief that they truly have that quality? If an inmate enters jail preceded by the information that he has received a 20-year sentence for armed robbery, does he need to, as it were, flex his muscles publicly and vigorously to ensure that other inmates will leave him alone, or can he afford to sit back and just bank on that indirect sign of his toughness?[1] And if the sign is of a quality that could be detrimental, should people engage in some new deed to make up for it? Can an inmate rumoured to be a 'snitch' do anything to persuade other inmates that it is all a big misunderstanding, and that he really is a most loyal fellow, or would further actions make his situation even worse?

2. Theoretical background

Observable features of an agent that are displayed *intentionally* by that agent for the purpose of changing the beliefs of others about a quality of that agent are called signals. Like signals, signs carry information about people's qualities, but unlike signals signs are not produced with the intention of informing anyone. Signs, in our context, are perceivable features of an agent which *unintentionally* convey information about qualities of that agent (Gambetta, 2009, p. 170). They are by-products of our actions rather than being produced by us in order to send a particular message. On the one hand, the lack of intention to inform makes signs a source of reliable information—for instance, small gestures of generosity when it is clear that they are not carried out instrumentally are signs of genuine generosity (Gambetta and Przepiorka, 2014). On the other hand, the information that signs convey tends to be *noisy*: while raising the probability that an agent has or lacks a certain quality, they often fall short of providing full reassurance. This is either because some 'noise' in the transmission affects their reception or because the actions that created the signs are imperfectly linked to the quality under scrutiny.

Signalling theory—arguably the best instrument we have to understand intentional communication when agents' interests do not fully overlap and information is asymmetric—is silent about signs.[2] In the theory-stylised landscape, the receiver can only use the signalling action taken (or avoided) by the sender to update his beliefs about that sender. Until the sender takes an action, the receiver has to rely on prior beliefs about the base-rate distribution of types to identify the sender.[3,4] In real life, however, this simplification—especially outside of anonymous markets—is almost always violated: before receiving the signal, receivers already know something relevant about the qualities of the sender they are interacting with.

An extension of signalling theory, known as 'countersignalling theory' (CS) (Feltovich et al., 2002), can help us to deal with these richer situations. This theory discards the assumption of a single source of information and admits, realistically, three conditions:

- the existence of prior information independent of signalling but dependent on a sender's quality—sometimes called 'exogenous' information—which the receiver has about the sender and which the sender knows the receiver has;
- the noisiness of the exogenous information[5]; and
- the existence of three levels of sender quality, or types, instead of the customary two types—countersignalling theory includes not only those who have or do not have the quality of concern, known in the theory jargon as 'high types' and 'low types' respectively, but also those who fall somewhere in between, the 'medium types'.

When these features are introduced, the main finding[6] of CS theory is that in certain conditions medium types will signal more intensely than both low types and, more surprisingly, high types too.[7]

---

[1] What, exactly, committing armed robbery conveys depends to some extent on the context. In America, it may be less convincing as a sign of toughness than in Britain. Anticipating that civilians will be armed, robbers in America arm themselves too, making armed robbery the standard form of robbery. In Britain, armed robbery counts as an extreme action that only those strongly committed to the possibility of using violence take.

[2] We refer here only to the signalling literature developed in economics. The theory in biology differentiates between signals and signs—which are called cues (Diggle et al., 2007, p. 1242; Scott-Phillips, 2008).

[3] In some economic renditions of the theory, agents are conceived as strategic to the hilt, ready to always take into account the signalling value of anything they do before they do it, so nothing is ever a sign. This may be true in some extreme cases in which the peril of mutual aggression is permanently present as among Mafiosi or inmates. But this seems hardly the case for most ordinary people in most daily activities.

[4] In his seminal paper (1973) and subsequent book (1974) Michael Spence touches on this issue. In his paper, his second model pp. 368–374 (and Model 3a in his 1974 book), addresses the impact that an uninformative—evenly distributed according to quality—non-alterable characteristic (which he calls an *index*) has on signalling equilibria. Also, in Model 3b in his book (pp. 38–46), he looks at the effect that an index has when it is linked to signalling costs. However, Spence's indexes are different to what we are concerned with here for they are not correlated with quality.

[5] Specifically, four features typify the exogenous information included in Feltovich et al.'s model. First, senders know the probability distribution that their exogenous information can take, but not its specific realised outcome that receivers observe. Second, receivers observe the specific point outcome of the exogenous information that is drawn from the sender's probability distribution. Third, the probability of sending a higher level of exogenous information correlates positively with quality. Fourth, senders have no control over it: it is exogenous to sender choice.

[6] Although there are multiple equilibria as with other signalling games, the countersignalling equilibrium is robust to common out of equilibrium belief refinements, namely the Intuitive Criterion, D1, and D2.

[7] Like signalling theory, CS theory assumes that there is a negative relationship between quality and signalling cost, and that as the signalling level increases the signalling cost increases at an increasing rate. In other words, cost is decreasing in quality and convex in signalling level. The cost of signalling is assumed to satisfy the single-crossing property, thus the marginal cost to signalling is lower for higher type senders.