Trust and Deception in Virtual Societies

Edited by

Cristiano Castelfranchi

National Research Council,
Institute of Psychology,
Division of 'Artificial Intelligence, Cognitive and Interaction Modelling',
Roma, Italy and
University of Siena,
Communication Sciences 'Cognitive Science'

and

Yao-Hua Tan

Erasmus University Research Institute for Decision and
Information Sciences (EURIDIS) and
Erasmus Center for Electronic Commerce (ECEC),
Rotterdam, The Netherlands

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

Chapter 1

TRUST AS TYPE DETECTION

Michael Bacharach*
Department of Economics
University of Oxford
michael.bacharach@economics.oxford.ac.uk

Diego Gambetta
Department of Sociology
University of Oxford
diego.gambetta@all-souls.oxford.ac.uk

1. INTRODUCTION

In this article we embark on a re-orientation of the theory of trust which contains three steps[1].

In section 2, we first define 'trust' as a particular belief, which arises in games with a certain payoff structure. We then identify the source of the 'primary problem of trust' - the problem someone faces in answering the question "Can I trust this person to do X?" - in her uncertainty about the payoffs of the trustee. Departing from standard rational choice treatments, we allow that a trustee's 'all-in' payoffs in a trust game may differ from his 'raw' payoffs. Whether they do so depends on non-observable qualities of his. Some of these - virtues, internalised norms, certain character features - are trustworthy-making qualities, qualities which induce all-in payoffs which motivate him to resist the pull of the raw payoffs and instead to do X. The primary problem of trust lies in the truster's uncertainty as to whether the trustee will be governed

*Several colleagues sent us comments on an earlier version of this paper, some of which have led to changes and improvements. Not all the points that deserved attention, however, receive it here, mostly because of lack of space. We are particularly grateful to Tyler Cowen, Jon Elster, Russell Hardin, Susan Hurley, David Laitin, and Gerry Mackie.

C. Castelfranchi and Y.-H. Tan (eds.), Trust and Deception in Virtual Societies, 1–26.

by raw payoffs, or by all-in payoffs induced by such trustworthy-making qualities[2].

In section 3, we first note that in virtually all games of this type which occur naturally, the truster sees or otherwise observes the trustee before deciding. She therefore can, and should, use these observations as evidence for the trustee's having, or lacking, trustworthy-making qualities. We next note that, since the truster will be proceeding in this way, and given the payoff structure, there is a motive for an opportunistic trustee to 'mimic' - to emit signs of trustworthy qualities when he lacks them. This complicates the truster's problem: she must judge whether apparent signs of trustworthiness are themselves to be trusted. Thus the problem of trust from which we started is transformed into a 'secondary problem of trust'.

In the third step (section 4) we show how the secondary problem of trust can be recast as a particular class of signalling games. This move has two advantages. First, the theory of signalling games is a well-developed part of game theory, and the understanding of such games which game theory provides serves to illuminate the present problem. Secondly, signalling-game theory allows precise predictions about the conditions under which mimicry can occur and how much of it we can expect.

Does it follow from our re-orientation of the theory that the problem of trust is not, in its essentials, *sui generis*? As we argue in section 5, with respect to secondary trust, the answer is largely positive. We say 'largely' rather than 'fully' because we do not exclude the possibility that the act of trusting itself can act as a signal of trustworthiness in so far as trusting persons are believed to be more trustworthy than untrusting ones. Were it to exist, as some empirical evidence suggests (Orbell and Dawes [13]), this mechanism would be *sui generis.*

With respect to the mechanisms which produce the emergence of trust-warranting properties the answer is also only partially positive. The explanation of many such properties can be explained only by broader theories of character, norms, virtues, abilities etc. There is however a family of mechanisms which may be trust-specific. These concern those cases in which trusting itself makes people positively responsive and more trustworthy than they would otherwise be. In the last section we discuss this problem and provide a game theoretic treatment of the self-fulfilling case.

2. PRIMARY TRUST

The notion of trust whose conditions of existence we investigate in this paper we shall call 'trusting someone to do something'. This notion captures a vast range of trust cases, even though it is not intended as an analysis of all varieties of trust. Its essence can be conveyed by an example. Suppose a person extends a loan to another because she expects him to do his best to repay it, when it is clear that she would do better to refuse the loan if he will make no effort to repay it, and when it is also clear that his selfish interest is to make no effort. Then, we shall say, she trusts him to try to repay the loan. In general, we say that a person 'trusts someone to do X' if she acts on the expectation that he will do X when both know that two conditions obtain: if he fails to do X she would have done better to act otherwise, and her acting in the way she does gives him a selfish reason not to do X.

2.1 THE BASIC TRUST GAME

We can make the notion more precise by describing the interaction as a two-player strategic-form noncooperative game, which we shall hereafter refer to as a **basic trust game**. This game has two players, a potential trusteR R ('she') and a potential trusteE E ('he'). Each has two strategies: R's are labelled A and A', and E's are labelled B and B'. Fig. 1.1a shows the payoffs of the truster, and Fig. 1.1b the 'raw' payoffs of the trustee; that is, the payoffs that would be his were he motivated by simple self-interest. In Fig. 1.1a, the numbers 3 and -3 are purely illustrative, and x and y are any numbers that satisfy $x < 3$ and $y > -3$: thus, if the trustee chooses B, A is better than A' for the truster, and if the trustee chooses B', A' is better than A for the truster. Finally, if for whatever reason R chooses A, B' would best serve E's selfish interest. (The trustee's preferences if the truster does A' are irrelevant.)

	B	B'
A	3	-3
A'	x	y

Figure 1.1a: *R's Payoffs*

	B	B'
A	1	4
A'	..	..

Figure 1.1b: *E's Raw Payoffs*

In the loan request example, the strategies A and A′ are simply the actions 'lend' and 'refuse'; B and B′ are the conditional actions 'try to repay if R lends' and 'make no effort to repay if R lends'. This conditionality comes from the dynamic structure of the example, in which E's action (trying or not trying) succeeds R's action (lending or refusing), which E observes. This is a very common pattern in trust situations, but not invariable, and not part of the definition of a basic trust game, which is in terms of strategies not actions.

However, for all the truster knows, there is no need for the trustee's payoffs - that is, his all-things-considered or 'all-in' payoffs - to be identical with his raw payoffs. Perhaps the borrower's upbringing has instilled in him a horror of not repaying debts, so that his payoffs are as in Fig. 1.1c. If we suppose simplifying that this is the only other possible case, the primary problem of trust consists, in effect, in identifying which of the two payoff structures really governs E's actions.

	B	B′
A	2	−4
A′	..	..

Figure 1.1c: *Possible Payoffs of E*

The remainder of the definition of basic trust games concerns what the players know. We assume, first of all, that there is mutual knowledge between R and E[3] that R can do either A or A′, and E either B or B′. Secondly, there is mutual knowledge of R's payoffs, but in general only E knows E's payoffs: a basic trust game is thus a 'game of asymmetric information' concerning E's payoffs. However, we assume that there is mutual knowledge of E's raw payoffs. What only E knows, however, is whether or not his all-in payoffs coincide with, or differ from, his raw payoffs[4].

Basic trust games are paradigms of primary problems of trust, and by no means exhaust the range of such problems. They simplify in several ways, which we will record *seriatim*. Here we draw attention to two. For there to be a problem for R whether to trust E, it is not necessary that the forces pushing E towards B′ be E's raw payoffs (selfish interests); the same problem arises for R if she is concerned that E may be guided by unselfish, even noble or spiritual, reasons for choosing B′. For example, R has a trust problem if she thinks it possible that E has no intention of repaying her because he plans to give all his future unspent earnings to cancer relief. We have chosen to characterise the bad-case payoff structure, which would lead E to do B′, as selfish only for the sake of concreteness and simplicity.

Basic trust games abstract from 'multiple peril'. In a basic trust game there is only one way (B′) in which trust can be disappointed, and there is a single bad-case payoff structure, which puts R in peril of B′, that of Fig. 1.1c. Often, however, there may be several such perils, each favoured by a different 'raw' type. A traveller in an unknown city with a plane to catch, seeing a waiting cab, may wonder "Can I trust that driver?" She may be concerned about more than one scenario. He might be driven by greed, and take a roundabout route (B′) for the sake of money; he might be libidinous and molest her (B″). Such cases, and yet more elaborate ones, can easily be handled by exactly the same methods we develop for basic trust games, at the cost only of some increase in complexity. Corresponding to perils B′ and B″ there are types of trustee whose raw payoffs are as shown in Figure 1.1d.

	B	B′	B″
A	1	4	0
A′	..	..	..

GREEDY TYPE

	B	B′	B″
A	1	0	4
A′	..	..	..

LIBIDINOUS TYPE

Figure 1.1d: *E's Raw Payoffs in Multiple Peril Game*

We shall say that, in a basic trust game, the truster R **trusts** the trustee E if *R expects E to choose B*. We assume that all-in payoffs govern choice; hence if R trusts E, R will do A. We therefore call A the **trusting act** and B the **trusted-in act**. The trustee E will be said to be **trustworthy** if the following obtains: *if E believes R trusts him, E chooses B*. Since all-in preference guides choice, for E to be trustworthy his all-in payoffs must diverge from his raw payoffs.

If R thinks E is trustworthy it now follows - almost - that R will trust E and do A. What follows, more precisely, is that there is an equilibrium in which this is so. There is a consistent state of affairs in which R trusts E and so does A, and E thinks R trusts him and so does B. This is an

equilibrium because E's doing B fulfils R's expectation, and R's doing A fulfils E's expectation.

There are cases that are often regarded as canonical cases of trusting which are not covered by the notion we just have defined. It is often thought that in Prisoner's Dilemmas and games of the family of Rousseau's Stag Hunt mutual trust can produce cooperative outcomes. Because such trust is symmetric, with each player both a truster and a trustee, our asymmetric notion cannot be directly applied. There is, however, in Prisoner's Dilemmas and kindred symmetric games, a close bilateral analogue of our unilateral notion, and there is a bilateral trusting equilibrium that goes with it[5].

In a moment we shall describe a type of participant in basic trust games we dub 'opportunists'. An opportunist is defined by a bundle of features of which one (not the most unsavoury) is that he maximises raw payoff. It follows immediately from this that if E is an opportunist, and believes that R trusts him, he does B′. But something else follows, provided E's raw betrayal payoff (4 in Figure 1.1b) exceeds his raw *status quo* payoff. Given that R will do A if she expects E to do B, and that E knows this, it will *pay* an opportunist to cause R to think that he will do B (if he can do this cheaply), then do B′. This 'scope for opportunism', which is a consequence of the structure of payoffs and information which defines a basic trust game, is crucial in converting the primary trust problem for R into a sign-reading problem.

2.2 TRUST-WARRANTING PROPERTIES

We have characterised what it is to trust someone to do something in terms of certain expectations and actions which might possibly arise within a certain abstract structure of payoffs. To make progress with the question of when we can, or cannot, expect trust to occur in the world, we need to fill out this bare structure. In this section we address the dispositions of the trustee which determine his all-in payoffs. In the next, we address the truster's information about these dispositions.

There are many possible properties, and even kinds of property, which can make a trustee trustworthy. Some of these take the form of reasons the trustee has to choose B. One such reason, is his very belief that the truster is trusting him (Dasgupta [3], Gambetta [5], Hausman [8], Hirschman [9], Pettit [16])[6]. Another class of reasons are generated by values held by the trustee which make him put a high value on an (A, B) outcome in itself, and override the raw preference he has for B′ given that the truster will do A. These include general moral principles supporting altruistic behaviour, and context-specific norms, such as those which govern teamwork. Other properties making for trustworthiness are individual character traits and evolved or cultural dispositions. These properties may make E trustworthy either through some unreflective mechanism of habit or internalised norm, or by making E see the force of reasons to be trustworthy. Another important sort of consideration that may lead E to do B is his susceptibility to possible rewards and punishments; this we take up later in this section.

From a game-theoretic standpoint, all these properties have the effect of transforming the trustee's raw payoffs. They replace them with 'all-in' payoffs, or evaluations, on which the trustees place a higher value on B than on B′ (given that he thinks R expects him to do B). The particular case of altruism has often been modelled thus, as a value which transforms a person's first-order, raw payoffs into higher-order, all-in payoffs (Kitcher [10], Nagel [12]).

We shall call any property (or combination of properties) of a trustee in a basic trust game which suffices for him to be trustworthy in it a **trust-warranting** property for that game[7]. We shall assume that in a basic trust game both parties know perfectly well which properties of trustees are trust-warranting in this game; and indeed that this is mutual knowledge. We have also assumed that, in a basic trust game, R and E know that E's raw payoffs are as in Figure 1.1a, and that R's payoffs are as in Figure 1.1b. These are conceptual necessities: these elements in the structure of payoffs, and mutual knowledge of them, are parts of our defined notion of trusting someone to do something because without them the question "Can I trust him to do B?" could not arise.

In contrast, for all that has been said to this point, R may know much or little about whether E has trust-warranting properties. Every degree of informedness is conceptually compatible with R's asking herself the primary question. However, across an extremely broad range of empirical cases, the way in which R acquires whatever information he has about this matter is much the same. In the rest of this paper, we shall focus on this information process.

Before turning to it, we must address one important point. We have said that, in a basic trust game, there are always properties of E which R must know E has if trusting E is to be warranted, and that these are 'payoff-transforming' properties of E. At first blush this claim seems to be false in an important subclass of basic trust games. In this subclass trust is usually thought to be warranted in virtue of features of the situation which R knows about other than such properties of E. It is the subclass in which the encounter whose payoffs are given in Figure 1.1 has a sequel in which R will have opportunities to punish a B′ choice, or reward a B choice, if she has trusted. The trustee will therefore be

concerned that R will punish him if he does B′, or that the reputation he earns with R will adversely affect his dealings in the sequel with her or others who hear of it. Both in classical economics (Hume, Smith) and in much contemporary rational choice writings on trust and reputation, this forward payoff structure is said to provide E with strong enough reasons, in the form of incentives, to be trustworthy.

It may seem that this solution to the primary trust problem eliminates the need for R to be concerned about what we have called E's trust-warranting properties. However, what matters is not these raw future payoffs alone, but the way in which E's knowledge of them transforms the raw payoffs to B and B′. Into this transformation enter dispositions of E to be swayed by these possible sequels, including his time preference and his imagination (Selten [17]). What is true is that, given E's dispositions (for example given his insensitivity to the fact of being trusted), a known system of sufficiently great punishments and rewards will usually incline E towards B′. But it does not follow from this that for a given system of known punishment and rewards his dispositions become irrelevant. On the contrary, for any plausible such system there are usually values of dispositional variables that make E untrustworthy, as well as values of them that make him trustworthy. The way such variables can tip the balance is just the same in the case of raw payoffs spread over-time as it is in the case of raw payoffs in a one-shot encounter.

It was exactly for reasons of this kind that American mafiosi, according to the Mafia boss Vincent Teresa, decided not to have dealings with black people. There was no punishment severe enough for the prospect of it to keep them under control. Once that was clear, the sign 'black' was enough to banish trust. Blackness, in the Mafia world, had become a sign of the absence a non-observable trust-warranting property, namely of not responding to the prospect of punishment.

3. SECONDARY TRUST

3.1 KRYPTA AND MANIFESTA

The truster seldom knows the trustee's trust-relevant properties directly from observation. True, one may say, "I could see at once that he was an honest fellow". But it is in signs that one sees it. One observes, for instance, physiognomic features - the set of the eyes, a firm chin - and behavioural features - a steady look, relaxed shoulders - and treats them as evidence of an internal disposition. Trust-warranting properties - honesty, benevolence, love of children, time preference, sect membership - may come variably close to being observable. But, except in limiting cases[8], they are unobservable and signs mediate the knowledge of them.

Let us call unobservable properties of a person **krypta**. We shall call those krypta of a person which are his trust-warranting properties in a basic trust game his **t-krypta** in that game[9].

The truster in a basic trust game may have sources of knowledge of the trustee's inner properties, and in particular his t-krypta, other than direct observation. Nothing in the notion of a basic trust game excludes background knowledge or knowledge got from others; similarly, nothing excludes direct observation of features of the trustee that are directly observable. These various sources may combine to allow the truster to know, or at least form considered beliefs about, the trustee's t-krypta. A truster may observe the tweed jacket and shooting stick of a woman in a departure lounge, and may infer that she is to be trusted to look after the truster's bag while he posts a last-minute letter. Or, a truster may recognise her trustee, by her face, or signature, as someone who has always repaid loans in the past, and may infer that she has now, as then, those properties of character and liquidity which make her credit-worthy. Or, finally, a truster, having heard of the trust-warranting properties of a person of a certain description, may try to match that description with the person in front of her: "Is this the person picked out by the signs of identification I heard of?" Sometimes, signs from different sources may give contrasting evidence: the credit card details look fine, the signature looks fine, but - as one of the authors of this paper knows from experience - if the man speaks with an Italian accent, R may still be reluctant to accept credit card payment. Sign reading is a fundamental part of deciding whether to trust. Correspondingly, the deliberate use of signs, or signalling, is a fundamental part of making oneself appear trustworthy.

By a person's **manifesta** we shall mean any observable features of him. 'Features' of a person include parts or aspects of his body, pieces of behaviour by him, and his appurtenances. Whether a feature of one of these kinds is observable depends on the situation. In face-to-face encounters someone's face is a manifestum. At the immigration desk the passport someone carries is a manifestum. In the street the jacket a woman wears is a manifestum. Depending on the medium of interaction, a feature can be a manifestum or remain hidden. The telephone reveals the voice but not the face; the Internet reveals neither the voice nor the face, but still reveals some electronic address, a language and a style of writing.

Manifesta may be evidence relevant to, signs of, krypta, and in particular of t-krypta. An open look is a sign of honesty. A business suit is a sign of respectability. An accent is a sign of one's social and ethnic origins. A signature is a sign of personal identity. A skullcap is a sign

of piety. A declaration is a sign of good intentions. A hesitation may be a sign of dishonesty. Excessive protestation is a sign of insincerity. Such manifesta can be read by R, allowing her to make inferences about E's krypta and hence his trustworthiness. They may reveal that E is motivated or otherwise disposed to do B if he thinks you trust him to, or that he is not.

Inferences from *prima facie* manifesta are sometimes sound; the observable features (the tweeds, the signature, the name, the accent, the club membership) are sometimes good enough signs. Sometimes, however, it is not certain that signs of trust-warranting properties are to be trusted. Herein lies what we shall call the 'secondary problem of trust'.

A manifestum may be a sign of a krypton without being a reliable sign of it. We shall meet many kinds of case in which signs can be misleading[10]. (Sometimes indeed a manifestum offered as a sign of a krypton may even reduce the probability it is rational to assign to the krypton.) The significance of manifesta is captured by the phrase "He seemed so trustworthy", uttered after trust has been misplaced. Most signs are less than fully reliable: but we do not have in mind this commonplace. In the case of signs of t-krypta there are special forces at work, which threaten reliability.

3.2 OPPORTUNISTS

We now define a type of interactant in a basic trust game, the **opportunist**, who is not only untrustworthy but also manipulative. An opportunist has these two properties: (O1) he is rawly motivated: if an opportunist E knew he was trusted to do B, he would do B′, for the sake of the extra raw payoff; (O2) he is manipulative: if he could obtain trust at low enough cost, he would do so, then betray it. O1 is just the property of being guided by raw payoffs. O2 goes beyond O1: it means the opportunist is not just lacking in trust-warranting properties (love of babies, strength of will, devoutness, responsiveness to being trusted, diagnostic skill), but is proactively deceptive.

An opportunist is a type of interactant in a basic trust game, not necessarily a type of person. A person might be trustworthy in one encounter but an opportunist in another. This is because some trust-warranting properties may vary over time in a given person: people change. Some trust-warranting properties are stable, others not. "I trusted him, but he must have changed since I knew him years ago." He had become an opportunist. Finally, some trust-warranting properties are local rather than global. Someone may be solidly trustworthy as a father and yet an opportunist with his colleagues or the Inland Revenue. Love of children may be a krypton which sustains trustworthiness only in specific relationships. Honesty, in contrast, is a property which warrants trust in a wide range of encounters.

Now, suppose there is a manifestum, m, which people take to be evidence of a krypton, k. Then deliberately displaying m is a possible strategy for convincing someone that you have k. In the language of signalling theory, displaying m is a way of signalling that you have k. Both k's and non-k's may have a motive for signalling that they have k. Rich people sometimes wear expensive clothes to show that they are rich; poor people sometimes wear expensive clothes to seem to be rich. Benevolent uncles smile to show they are benevolently disposed; wicked uncles smile to seem to be benevolently disposed. The deceptive instances of the strategy we call 'mimicking'. More precisely, a **mimic of k through m** is a person who does not have k and deliberately displays m in order to be taken to have k by another[11].

Mimicry may be positive or negative - 'camouflage'. There are often manifesta of a trustee which are likely to be interpreted by the truster, rightly or wrongly, as indicating *not-k*, and so untrustworthiness. Both an honest k who expects to be unjustly perceived if he displays such a manifestum m, and an opportunist non-k[12] who is afraid of being detected if he does, have a reason to camouflage, that is, to take steps *not* to show m. For the purposes of this paper, we may consider deceptive camouflaging as a special case of mimicking, since the strategy of camouflaging non-k-ness by suppressing m is just that of mimicking k through displaying the notional manifestum 'no m'.

3.3 MIMIC-BESET TRUST GAMES

The possible presence in the world of opportunists means that a truster must beware of mimicry in interpreting manifesta of the trustee. Suppose that k is some t-krypton and m is a manifestum which is usually taken to be a sign of k. Then the truster should argue as follows: "An opportunist would wish to make me trust him to do B, then betray me by doing B′. He would know that to get me to trust him it would help to get me to believe he has k. There may be opportunists around. Therefore, this m-bearer may not have k, but may instead be an opportunist mimicking k through m. The m I observe, though it may be produced by a k-possessing and therefore trustworthy trustee, cannot be taken at face value." Apparent guarantors of trust might be the manipulations of a wolf disguised as an old lady.

We shall call an encounter a **mimic-beset trust game** if it is a basic trust game with the following two features: (i) there is some positive prob-

a manifestum or manifesta of the trustee. More precisely: before the truster and trustee have to choose their actions, there is an observational episode, in which the trustee may, if he wishes, display a certain manifestum, and the truster observes it if he does. We suppose that the vast majority of basic trust games, and indeed of small-number problems of trust more generally, include some observation by trusters of manifesta of their trustees. Recall that any piece of observable behaviour, including a message, counts as a manifestum; whence it follows, for example, that every trust-involving interaction which is initiated by a trustee contains an episode of the kind described.

Here is a simple example: the truster is approached by a stranger who tells her a hard-luck story and asks her for money. The truster values a world in which she gives money and the story is true, and values negatively a world in which she gives money and the story is false. As the stranger speaks he has the opportunity to look the truster in the face, or else he may look down at the ground. The t-krypton is honesty, the manifestum is the action of looking in the face. If we like, we can complicate this story to make it more realistic, without changing the essential structure of the trust game. For example, it might be that there are two displayable manifesta: looking in the face, and offering an address; or indeed many. Similarly, the t-krypta may be many and various. It is clear that very many basic trust games are mimic-beset trust games.

3.4 SECONDARY TRUST

If a truster is playing a mimic-beset trust game, she should give her trust only with caution. She should exercise care in this example not only if the stranger looks down, but also if he looks her in the face. The assumptions of mimic-beset trust games imply that whenever there is a question "Can I trust this person to do B?", a truster must first, and need only, answer another question: "Is the manifestum m of the trust-warranting cryptic property k which I observe a reliable sign of k?" - in brief: "Can I trust this sign of k?" The conditions in which manifesta of t-krypta may be trusted or distrusted is a special topic in trust. It is the 'problem of secondary trust'.

If we are right that typically problems of trust are mimic-beset trust games, then the problem of secondary trust is the key to answering the two questions which normative trust theory should be concerned to answer: (i) "When is it rational to trust someone to do something?" (ii) "How can I, who am trustworthy, convince the truster that I am?". who has trust-warranting properties. A theoretical understanding of the secondary problem of trust directly answers the truster's question. It also indirectly answers the trustworthy trustee's question: he should, if he wishes to be trusted and if he thinks the truster rational, display those manifesta which can be trusted as signs.

4. SECONDARY TRUST AND SIGNALLING

In the first part we have suggested that secondary trust almost always accompanies, and is often the key to solving, problems of primary trust. In this part we show how the theory of signalling which developed in economics, biology, and game theory (Spence [18], Zahavi [19]), when combined with the notion of a mimic-beset trust game, provides a clear analytical framework for understanding secondary trust. Mimic-beset trust games can be recast as a class of signalling games. Rational trust can therefore be explained, up to a point, as the beliefs that a receiver has in an equilibrium of a signalling game.

4.1 SIGNALLING THEORY

A signal is an action by a player (the 'signaller') whose purpose is to raise the probability another player (the 'receiver') assigns to a certain state of affairs or 'event'. This event could be anything. As it happens, in many cases studied in the literature the event is that the signaller herself is of a certain 'type'; for example, in a famous early study (Spence [18]) the event is that the signaller is a worker of high rather than low productivity. In these cases, what is being signalled is what we have called a krypton of the signaller.

Within signalling theory most attention has been paid to a class of games in which there are three kinds of agents: k's, non-k's, both of whom can send some signal from a specified list, and receivers. There is something the receiver can do which benefits a signaller, whether or not he is k; but it benefits the receiver only if the signaller is k, and otherwise hurts her. The benefits to k and non-k signallers need not be equal. Thus k's and receivers share an interest in the truth, but the interests of non-k's and receivers are opposed: non-k's would like to deceive receivers into thinking they have k, in order to receive the benefit, while receivers have an interest in not being deceived. The interests of k's and non-k's are also usually opposed because the activity of the latter damages the credibility of the signals of the former.

The main result in signalling theory is that in a game of this kind[13] there is an equilibrium in which at least some truth is transmitted, pro-

vided that among the possible signals is one, s, which is cheap enough to emit, relatively to the benefit, for signallers who have k, but costly enough to emit, relatively to the benefit, for those who do not.

If the cost relationships are such that all and only k's can afford to emit s, the equilibrium in which they do so is called 'separating' or 'sorting'. In such an equilibrium signals are quite unambiguous, and the receiver is perfectly informed. No poisoner seeks to demonstrate his honesty by drinking from the poisoned chalice.

But the cost conditions may also give rise to dirtier equilibria, so-called 'semi-sorting' ones. In a semi-sorting equilibrium (SSE), there is a signal s which is emitted by all k's but not only k's; a certain proportion of non-k's emit it too. Here, although observing s is evidence for the receiver in favour of k, it is not conclusive evidence; it makes it more likely that the signaller has k, but does not imply that he does. A determined mimic can even try to deceive an audience that he is an operatic singer, by lip synching. The higher the proportion of non-k's who use this signal the less conclusive is the evidence.

The conditions on the parameters of a signalling game for it to have a sorting equilibrium or a semi-sorting equilibrium are well known. It has a sorting equilibrium, in which a k always emits s and a non-k never emits s, if two conditions hold: (i) the benefit to a k of being treated as a k exceeds the cost for him of emitting s; and (ii) the benefit to a non-k of being treated as a k is less than the cost for him of emitting s. In brief, a k can afford to emit s, and a non-k cannot. We shall call these the Can and the Cannot conditions (they are called the 'incentive compatibility condition' and the 'non-pooling condition' in game theory). And the signalling game has an SSE, in which all k's emit s and a certain proportion p of non-k's, less than one, also do so, if these two conditions hold: (i) the Can condition; and (ii) the benefit to a non-k of being treated as a k is exactly equal to the cost for him of emitting s: that is, a non-k can just afford to. We shall call the latter the Can Just condition. We can, further, calculate p in terms of the cost and benefit parameters.

There is evidence that many - perhaps most - cases of krypton signalling in animal life are not sorting but only semi-sorting equilibria (Guilford and Dawkins [7]). The same appears true of human life. For example, verbal reports of inner states such as beliefs and intentions may be regarded as a very large family of krypton signals. Telling lies is commonplace; all such reports are suspect when the signaller has a motive to deceive. Yet whenever a hearer doubts a verbal claim, it remains likely that the speaker's making it is on balance evidence for the krypton (raising, not lowering, the probability the hearer assigns to the krypton) since otherwise the speaker acted counterproductively. If so, then the utterance is a semi-sorting signal of the krypton. If a man with a UPS uniform rings your doorbell, the chances are that he is a UPS delivery man.

The relevance of signalling theory to the problem of trust lies in a simple connection: mimic-beset trust games are instances of signalling games. In a mimic-beset trust game, there is a trustee who may or may not have k and who can, if he so decides, display a manifestum m to a truster. Both k trustees and non-k trustees are interested in getting the truster to believe that they have k[14]. To complete the demonstration of the correspondence between the two kinds of game, it remains only to note that the *display of* m may be regarded as a signal s (and displaying m as emitting the signal s).

Our definition of a mimic-beset trust game is silent about whether or not the opportunist will in fact mimic. For a basic trust game to be mimic-beset, as we have defined this term, is enough that the trustee has an opportunity to display a manifestum and that he may be an opportunist, and so prepared to mimic, if it suits him. We can now use the correspondence with signalling games to give precise conditions in which mimicry will occur, assuming that the truster and trustee are 'subjects of game theory'. More than this, we can say how much will occur, when any does: that is, what proportion of opportunists who have a chance to mimic does so. The nub of the matter is this: *opportunistic trustees in a mimic-beset trust game mimic at least sometimes if the associated signalling game has a semi-sorting equilibrium*[15]. We may conclude that opportunists sometimes mimic if and only if the Can condition is satisfied for a k and the Can Just condition is satisfied for a non-k. And we can determine the proportion of opportunists who do.

Signalling theory is thus a powerful tool: it delivers specific answers to the questions when, and how much, false signalling will occur. However, crucial for the existence of the solutions which it offers are strong assumptions about the players' background knowledge: they know the sizes of the benefits, and of the cost of emitting s, for both k's and non-k's; and a receiver knows the base-rate probabilities that a signaller is k and non-k. When the truster peers through the Judas hole and sees a stranger in a UPS uniform, as she wonders whether to confer the benefit of opening the door she has in mind the general incidence of criminality, the ease or difficulty of acquiring a disguise, and the value to a criminal of what he would get from securing entry.

One of us recently filled the truster role in a mimic-beset trust game which has an SSE. Bacharach was approached by an apparently (m) desperate young man in a public park. He turned out to be an oppor-

tunistic non-k mimicking k through m. Bacharach's donation to him was made despite his being uncomfortably conscious that there was a chance of this, and because it seemed to be outweighed by the chance that he was genuine and the awfulness of his plight if he was. The former chance depends, in the theory, on the base-rate incidence of opportunism, on the size of the donation, and on the difficulty of 'putting on' the m that Bacharach observed, which was a detailed and convincing performance. He and the young man both played their parts just right in an SSE of a mimic-beset trust game[16].

Signalling theory provides a framework for analysing, in a wide if somewhat idealised class of cases, rational krypton signalling both informative and deliberately misleading. It thus serves as a basis for an analytical account of the secondary problem of trust. But it is only a framework that it provides. The explanations it gives remain abstract and incomplete. They are cast in terms of a benefit (from being thought k) and costs (of emitting each possible signal) which are given no structure in the theory and themselves receive no explanation[17].

5. IS TRUST SUI GENERIS?

One upshot of our re-orientation of the theory of trust is a sceptical conclusion: the problem of trust is not, in its essentials, *sui generis*. We have showed what several scholars suspected. We do not wish to imply, however, that the problem of trust thereby evaporates. Rather, we have reformulated it, within a more general yet robustly analytical framework. Trust is a complex phenomenon, but one whose main elements are met in other domains of decision-making. We have emphasised here those elements, which have to do with how to read signals of hidden qualities, when these signals may be meant to mislead. We have pointed out that this 'secondary problem of trust' is almost always present, and stressed its analogy with the problem of how to read quality signals elsewhere. The problem of the truster with respect to trustworthiness is the problem of the employer with respect to productivity, of the dog with respect to the belligerence of the hissing cat, and of the predator with respect to the toxicity of the striped worm.

However, we have said comparatively little about the hidden qualities, the t-krypta themselves. We have noted that the hidden qualities that are relevant for trustworthiness vary greatly with what the trustee is being trusted to do, and with cultural and other determinants of preferences. Moreover, given that our definition of trust as 'trusting someone to do something' is orthogonal to moral considerations, qualities like 'bloodthirstiness' could in certain conditions qualify as trust-warranting properties. The qualities which make a person trustworthy derive from a variety of sources (character, norms, values, innate dispositions, cognitive abilities and many more). This makes it impossible to have a theory specifically targeted on this key aspect of the primary problem. Our understanding of trustworthiness is unavoidably parasitical on the vast array of theories which can explain the myriad hidden qualities which make for persons' trustworthiness, and virtually none of these is likely to be *sui generis*.

It is possible, however, that there are high-level human dispositions which render people trustworthy in general, or at least over a very wide range of basic trust games. To this extent the full analysis of trust involves attitudes which are indeed peculiar to games of trust. So far in the paper we have been agnostic about the existence of dispositions of this kind, of attitudes primed specifically in trust cases. We have also been generally neutral as to whether such dispositions, if there are any, are themselves explicable by methods of rational choice theory.

Here we want to focus on one such case, the case in which the very fact of being trusted induces a positive responsiveness and 'makes' people trustworthy.

5.1 POSITIVE RESPONSIVENESS

Several writers have suggested that someone may be motivated to do something by the awareness that he is being trusted to do it. If this suggestion is true, then a feature of a player which standard game theory takes to be exogenously given – her payoff function – is in fact endogenous. More precisely, since R's trusting E to do something is R's having a certain expectation about what E will do, E's payoffs over the outcomes of action-pairs are dependent on his beliefs about R's *expectations*. It is suggested that they may depend on them positively in the sense that if E thinks R expects him to do B then R is motivated indeed to do B. We shall call this effect positive responsiveness (to being trusted). In a similar vein, it might be suggested that someone's preferences over outcomes might be affected by the belief that her interactant had certain preferences over outcomes: for example, an actor's believing that her interactant has benevolent feelings towards her might make her feel benevolent towards him.

Although such phenomena have been (almost) wholly neglected in game theory, Geanokoplos *et al*'s ([6]) theory of 'psychological games' shows that it is possible to accommodate them within it. Geanokoplos *et al* study games in which a player's payoffs depends on her beliefs about other players' expectations about her action. These they call

'psychological games'. In this subsection we shall model as a psychological game a basic trust game in which E may be a positively responsive type, which we call the Responsive Trust Game.

Our purpose is twofold: first, to make sure that our theory is not misunderstood as claiming that there is nothing *sui generis* about trust. So far from asserting this, our theory, and the game theoretical framework in which it is cast, are quite flexible enough to accommodate motivational effects which are highly distinctive of trust. Second, we want to illustrate that the existence of *sui generis* motivational mechanisms does nothing to remove the secondary problem of trust.

In a 'psychological equilibrium' of a psychological game in the Genaokoplos *et al* theory not only does each player make correct estimates of the others' action-probabilities, but each player also correctly estimates the others' estimates. The conclusion of our analysis is that there are equilibria of this kind in which positive responsiveness induces equilibria of fulfilled trust.

We suppose that a truster R and a trustee E play a basic trust game whose raw payoffs are as in Figure 1.1 with the further (unrestrictive) gloss that R has a payoff of zero from choosing A′, regardless of what E does.

We suppose next that with some probability, π, the trustee, instead of having the raw payoffs, is positively responsive, and that this *incidence* π of positive responsiveness is common knowledge in the community. We model the positively responsive (POS) type thus. Although we may expect a POS type to choose B in circumstances in which a 'raw' type does not, we do not wish to presume that a POS type does so. Instead, we write ρ for the probability with which the POS type plays B, and let the game-theoretic analysis determine ρ. We call ρ the reliability of a POS type. Assume that R believes (with certainty) that the reliablity ρ has a certain value, $\overline{\rho}$ say, and that E believes (with certainty) that R's estimate $\overline{\rho}$ has a certain value, $\overline{\overline{\rho}}$. In general neither the estimate of ρ, $\overline{\rho}$, nor the estimate of $\overline{\rho}$, $\overline{\overline{\rho}}$, need be correct, but in a psychological equilibrium they are, by definition.

	B	B′
A	3, 1	−3, 4
A′	0, ..	0, ..

Figure 1.2: *R's Payoffs with E's Raw Payoffs*

The POS type we characterize as follows. If E is of POS type, her payoffs from doing B and B′ respectively, when R does A, are given by:

$$u_{\mathrm{E}}(\mathrm{A},\mathrm{B}) = 1 + s\overline{\overline{\rho}}, \qquad u_{\mathrm{E}}(\mathrm{A},\mathrm{B}') = 1 - s\overline{\overline{\rho}}.$$

Thus if, for example, E believes that R trusts him completely (the case $\overline{\overline{\rho}} = 1$), fulfilling that trust gives E a payoff s higher than his raw payoff, while betraying trust gives him a payoff s lower than his raw payoff. The increment and the decrement are both reduced, however, if E thinks that R is not completely counting on him (that is, if $\overline{\overline{\rho}} < 1$). A high value of the coefficient s means a strong positive responsiveness. We shall call s the sensitivity of a POS type.

So far in this paper we have abstracted from partial trusting and partial trustworthiness, focussing on cases in which R has a definite expectation of B and in which E definitely chooses B. But in reality it may often be that R expects B only with limited confidence, or that E only tends to choose B with some probability or frequency. In such cases we might want to say that R trusts E up to a point, and that E is fairly trustworthy. It turns out that the above game may have equilibria in which positive responsiveness induces the trusted-in action (B) only with some probability, that is, in which it E's positive responsiveness makes him only fairly trustworthy.

Let us now derive the psychological equilibria of this game which feature trust and trustworthiness, in the extended senses of trust and trustworthiness in which these may be only partial. First consider R's expected payoff from choosing A. If R estimates that a POS type chooses B with probability $\overline{\rho}$, R's expected payoff from A given that the trustee is a POS type is $3\overline{\rho} - 3(1 - \overline{\rho})$, and given that he is not a POS type is -3. Hence since R knows that her trustee is type POS with probability π, her overall expected payoff from A is

$$(1-\pi)(-3) + \pi(3\overline{\rho} - 3[1-\overline{\rho}]) = 6\overline{\rho}\pi - 3.$$

Recalling that in a psychological equilibrium R's estimate $\overline{\rho}$ of ρ is correct, we have that R's expected payoff from A in an equilibrium is $6\rho\pi - 3$. Hence, since the payoff from A′ is 0:

In a psychological equilibrium, R chooses A iff $\rho\pi \geq 1/2$. (1.1)

(1.1) tells us that R trusts the trustee enough to choose the trusting act if and only if the joint probability of the trustee's being a responsive type *and* a responsive type's choosing B is sufficiently great.

Now consider E's expected payoffs from choosing B and B′ respectively, when he believes that R will choose A, and his estimate of R's

estimate $\overline{\rho}$ is $\overline{\rho}$. A POS type is pulled two ways: choosing B is more humanly rewarding, but choosing B′ is more convenient. The two net payoffs are $1 + s\overline{\overline{\rho}}$ and $4 - s\overline{\overline{\rho}}$. Since in a psychological equilibrium $\overline{\overline{\rho}} = \overline{\rho} = \rho$, in such an equilibrium the two payoffs are $1 + s\rho$ and $4 - s\rho$. Hence there is a psychological equilibrium in which a POS type plays B (that is, $\rho = 1$) only if $1 + s \geq 4 - s$, that is, only if $s \geq 3/2$. This condition and (1.1) are together sufficient. Call an equilibrium of this kind, in which R plays A and E plays B if he is of type POS, a **pure responsive** equilibrium of the Responsive Trust Game. Thus we have:

There is pure responsive equilibrium iff $s/3 \geq 1/2$ and $\pi \geq 1/2$. (1.2)

Call a psychological equilibrium in which R plays A and E plays B with probability ρ positive but less than 1 if he is of type POS, a **mixed responsive** equilibrium of the Responsive Trust Game. The two pulls on a responsive type just balance if $\overline{\overline{\rho}} = \frac{3}{2s}$. By a basic theorem, there can be an equilibrium with ρ positive and less than 1 only in this case. Since in a psychological equilibrium $\overline{\overline{\rho}} = \rho$, there can be a mixed responsive equilibrium only if $\rho = \frac{3}{2s}$. On the other hand, this condition and (1.1) are together sufficient. Notice that the balancing condition can be met if and only if $s/3 \geq 1/2$. Notice also that, given the balancing condition, (1) is equivalent to $(\frac{3}{2s})\pi \geq 1/2$, or $\pi \geq s/3$. Thus we have:

There is a mixed responsive equilibrium iff
$s/3 > 1/2$ and $\pi \geq s/3$, and in it $\rho = 3/2s$. (1.3)

Putting (1.2) and (1.3) together gives us:

THEOREM. *In the Responsive Trust Game in which the positively responsive type occurs with probability π and has responsiveness coefficient s, there is at least one trusting psychological equilibrium if and only if $\pi \geq 1/2$ and $s \geq 3/2$. If $\pi < s/3$ the unique such equilibrium is the pure responsive equilibrium; if $\pi \geq s/3$ there are the pure responsive equilibrium and a mixed responsive equilibrium with $\rho = 3/2s$.*

It is not hard to recognize the mixed responsive equilibrium as a real possibility. As the POS type's reliability ρ rises (with R aware of its value, and E knowing that R is), there comes a point, depending on the POS type's sensitivity s, at which B and B′ are equally worthwhile for E, all things considered. Hence, at this point, being reliable to degree ρ is also as good as anything else. This balancing value of ρ is $3/2s$. This mixed behaviour by the POS type forms part of a trust equilibrium if and only if it is reliable enough to motivate R to play E. This happens if and only if the incidence of the POS type, π, is sufficiently high – concretely, high enough to make the probability of B at least $1/2$. This probability is $\rho\pi$, so the condition is that $3\pi/2s \geq 1/2$, or $\pi \geq s/3$. The more responsive types there are, the less reliable need they be for R to be trusting enough to choose A.

Paradoxically, the more *sensitive* POS types are, the less reliable they are in the mixed responsive equilibrium. If s is high, the balancing value of ρ is correspondingly low, because for a sensitive E knowing that R has even a weak expectation of B suffices to make fulfilling this expectation compensate for E's inconvenience. However, as we have noted, this low ρ must be offset by a high π if R is to judge it better to choose A.

In an example of Hausman's a trusting R, off for the weekend, pins a note to her neighbour E's door asking if he could possibly feed the cat, and E feeds the cat because he knows that R is counting on him to. The above analysis suggests a more modulated world: some people are nice (responsive), and in E's position these people feed cats some percentage of the time, but not always. They do as much of this sort of thing as they do because, knowing that weekending cat-owners and the like hope, cautiously, that their neighbour is both responsive by nature and that on this occasion he will actually feed, induces in them actual feeding on just that percentage of occasions.

What conclusions may be drawn? First, our analysis has confirmed that if the '*sui generis*' property of positive responsiveness exists, it can generate warranted trust: it can be part of a complex of attitudes and dispositions in trusters and trustees which mutually support each other in an equilibrium. For good measure, we have shown conditions in which it can do so, as a function of its strength and of its incidence in the population.

Second, the main lines of our theory are unaffected by the existence of this possibility. Having the positive responsiveness property is clearly a trust warranting property: we may model a case in which R knows E has it by putting $\pi = 1$, and the theorem shows that in this case there is a pure responsive equilibrium as long as $s \geq 3/2$. But it is equally clear that it is not a transparent one; even if it were the case that everyone is POS, and that this had become commonly known, there would surely remain unknown variation in the sensitivity s.

As voters, we may elect a politician to do B, and thereby overtly put our trust in him to do B. But we know as citiziens and voters how difficult it is to judge how much this trust we put in our politicians will weigh with them once elected. Part of the politician's success rests on his ability to persuade us that it will. His doing so will consist in the exercise of an appropriate repertoire of electioneering signals. ("The British people has put its trust in me, and I for my part will never betray

it ... .") In short, POS is a t-krypton. Furthermore, if R and E have opportunities to observe and display manifesta, the questions with which we have been concerned in this paper arise in just the same way as for any other t-krypton.

6. CONCLUSION

The presence of trustworthy-making properties is not sufficient to induce trust. Trust has two enemies, not just one: bad character and poor information. The incidence of good character sets an upper bound on the amount of trust in a community. Actual trust, however, may fall below that threshold because of the difficulties incurred communicating the presence of these qualities.

The fundamental problem for the truster is therefore independent of the existence or nature of trustworthy-making dispositions. It is the decoding of signs including deliberate signals. The key quality which, in the circumstances, determines trustworthiness may be as unmysterious as the size of a bank balance or the membership of a profession. Even where deep t-krypta are at work, it is often not the deep t-krypton itself that determines trustworthiness, but a mundane quality known to be correlated with it, such as an identity. The question for the truster is whether the trustee has qualities which she believes, in virtue of experience or common sense or on some other basis, to make for trustworthiness; not the sometimes deeper question of why these properties do so. Yet answering this shallower question is a complex business, because in games of trust there is a motive for the opportunist to mimic trustworthiness by displaying its signs.

In the central sections of this paper we have investigated the communication deficit, what causes it and by implication how it might be narrowed. In particular, since communication is through signs, the prevalence of trust increases with the significance of signs of trustworthiness. Here we have shown how signalling theory can provide the basic principles upon which the significance of such signs rests. In doing so we made use of tools which allow one to identify precise conditions under which trustworthiness is communicated and trust thereby justified. In this way we hope to help foster an approach to trust research which questions the belief that trust is something mysteriously different from anything else and aspires to the exact measurement of the conditions under which trust is rationally given. Much of what signalling theory can contribute to the understanding of trust, however, lies in the future.

Notes

1. A longer and partially different version of this paper can be found in [2]. We are very grateful to the editors for giving their consent to the publication of this version of the text in this collection.

2. Furthermore, the particular assumptions about people's motivations that are characteristic of the rational choice approach to trust which is currently dominant play no role in the theory we develop here. It is, for example, not essential to our theory that trust is the product of some system of rewards and penalties which act as incentives to the trustee in repeated interactions (such as repeated PDs). We treat trust as the product of underlying trustworthy-making character features - one of which can be, of course, the degree of susceptibility to rewards and punishments.

3. By this we mean that each knows that the other knows it, each knows this, and so on, up to some finite degree.

4. In the language of games of incomplete information ([4], [14]), there is more than one 'type' of E, of which one is characterized by the raw payoffs and one or more others by other payoffs; and in general only E knows E's type. In this respect, but not in all, basic trust games resemble 'principal-agent' problems with adverse selection and R as principal.

5. If we interpret the standard Prisoner's Dilemma matrix as showing raw payoffs then both have the same raw payoffs, and both are uncertain about the relation between the raw and all-in payoffs of the other. The present notion extends easily. To illustrate, consider the following simple expression of the idea that mutual trust can yield (C, C) in the Dilemma. Say that a player 'trusts' the other if she expects him to do k, and that a player is 'trustworthy' if the following is true of her: if she trusts the other player, this makes her prefer to do k herself, despite the fact that she rawly-prefers to defect. Then there is an equilibrium in which both are trustworthy, both trust, and both do k.

6. This reason for preferring B to B′ takes us outside the standard framework of game theory, because part of it consists of a belief E has about R's expectations. In standard game theory, R's act choice (here A) typically affects E's preference over B and B′, but R's reason for this act-choice does not: the primitive objects of preference are acts given others' acts, not acts given others' reasons for acting. Despite the non-standard character of such preferences, however, there is no difficulty in describing an equilibrium in which E has them. The equilibrium described at the end of section 2 in which R trusts a trustworthy E is an example. By definition, a trustworthy E chooses B if he thinks R trusts him. E's preference for B when he thinks R trusts him might be produced either by E's derived belief that R will choose A and/or non-standardly, by E's belief that R trusts him. And the argument that the described situation is an equilibrium is independent of which of these beliefs produced it. We spell out a mechanism of the second kind in section 5.

7. We have assumed here that if the trustee has a trust-warranting property k then (given the belief that she is trusted) she definitely will do B. But this oversimplifies. Mormons are valued as trustworthy baby-sitters, but there may be some bad Mormons. Here, then, k only makes doing B probable. The whole analysis can easily be generalized to this case, which is perhaps the only one we can expect ever to find in reality. Similarly, we have talked so far as if trust-warranting properties are always dichotomous. In reality, they often come in degrees: someone is more or less loyal, considerate, or whatever. In section 5 we will explicitly model such a case.

8. A limiting case is 'perceiving as': for example, a smile may be seen as a friendly smile: here the disposition is in the content of the perception and not inferred (see e.g. [11]).

9. We have assumed that there is no plasticity about what R can observe: some features she observes costlessly and automatically, in the base interaction; others she can not observe, but can only infer from the observables and background knowledge. Other features are not like that: get closer to a crocodile on a polo shirt and you can see whether it is a Lacoste one or not. If the getting closer or whatever is costly and chosen, then this choice variable should be modelled as part of the description of the game that is played. In the basic trust game and the variants of it which we consider in this paper we abstract from scrutiny by R of this kind, for the sake of simplicity and without affecting our essential argument.

10. Thus, in our (stylized) treatment a t-krypton k is sure evidence of trustworthiness, but manifesta are in general unsure evidence of k.

11. The 'mimicry' defined here is deceptive mimicry; there are also varieties of non-deceptive mimicry. I may mimic to amuse. I may also mimic a type of person (Xs) who are known to have k, by copying some manifestum m of X-ness, to convey that I am k, when I am k; although this behaviour is deceptive, it is so only incidentally, and there is no reason in general why Xs should lose by it; indeed they may benefit from it, if they too display m in order to convey they are k, by the reinforcement of the association between m and k. This is a form of 'Mullerian' mimicry ([15]).

12. We shall often speak of 'k's' and 'non-k's' rather than of people who have and do not have k. A krypton has been defined as a property, so one should perhaps speak of an agent's 'having' k rather than of agents who 'are' k or of 'k's'. But sometimes the latter is smoother, and it does no harm, since corresponding to every property there is a type, the possessors of the property.

13. Other scenarios besides this one are treated in signalling theory, but we shall not consider them.

14. So far we have suggested, about the trustworthy trustee's preferences, only that, in the circumstances in which the truster expects B and so chooses A, the trustee prefers (A, B) to (A, B′). Given that he is trusted, he prefers to fulfil than to let down trust. But nothing has been said about his preference for being trusted or not trusted in the first place. Both cases are possible. Let us say that in a basic trust game the trustee is 'willingly trustworthy' in the first case, that is, if he is trustworthy and prefers to be trusted. Suppose someone is trustworthy because doing B, though onerous, is felt to be an obligation. Then he is also willingly trustworthy if he wishes, all in all, to take on such onerous obligations. If a krypton k makes someone trustworthy but not willingly trustworthy, the mimic-beset trust game is an only slightly different signalling problem, whose signal-theoretic solutions are essentially the same.

15. We here take it for granted that if there is a semi-sorting equilibrium (SSE), players conform to it. Some of the games in question also have 'pooling' equilibria (in which k's and non k's behave in the same way); these generally appear implausible even though, notoriously, principles for excluding them involve unsettled questions about ratio0nal belief revision in multi-stage games. To establish the statement in the text we need only show: (i) in an SSE of the signalling game a non-k signaller sometimes emits s (that is, displays m); (ii) this emitting of s is a case of mimicking k via m. (i) is direct from the definition of an SSE. To see (ii), recall that for E to mimic k via m is for E to display m in order to raise the probability R attaches to E's being k. In an SSE, using s always raises the probability the receiver attaches to your having k. Say that in the equilibrium a non-k uses s with probability p. By definition of an SSE, $0 < p < 1$. Let the population fraction of k's be f. If the receiver observes the non-k's s, she assigns the probability $Pr(\mathrm{k} \mid \mathrm{s})$ to his being k, where $Pr(\mathrm{k} \mid \mathrm{s}) = f/[f + p(1-f)]$, by Bayes' theorem, since $Pr(\mathrm{k} \mid \mathrm{s}) = 1$. This ranges between f (when $p = 1$) and 1 (when $p = 0$). On the other hand, if the receiver observes no s, she assigns the probability $Pr(\mathrm{k} \mid \mathrm{not} - \mathrm{s}) = 0$. Hence using s raises the probability the receiver attaches to k, from 0 to some positive number, as long as $f > 0$ (there are some k's in the population).

16. There is more than one plausible way of modelling the situation. If the size of a donation in a begging game is conventionally given at D, and the truster's preferences are like Bacharach's, we have a mimic-beset trust game as defined in this paper, and it has an SSE in which the truster is just prepared to pay D to a mendicant displaying m if a non-k's payoff from D is just equal to the cost of mimicry (he can just afford to mimic) and the proportion of non-k's who mimic is equal to KZ, where K is the odds that an arbitrary beggar is k, and Z is the ratio of the utility the truster gets from paying D to a k and the regret she suffers from paying D to a non-k. Such an equilibrium might arise over time through adjustments in D and/or in the level of the performance.

17. Elsewhere ([2]), we have made a start in filling the gap that is left.

References

Bacharach, M. 1997. Showing what you are by showing who you are. Russell Sage Foundation Working Paper no. 125

Bacharach, M. and Gambetta, D. 2000. Trust in signs. In Karen Cook (ed) *Trust and Social Structure.* New York: Russell Sage Foundation, forthcoming

Dasgupta, P. 1988. Trust as a commodity. In D. Gambetta (ed), *Trust: Making and Breaking Cooperative Relations.* Oxford: Basil Blackwell, 49 - 72

Fudenberg, D. and Tirole, J. 1991. Game Theory. Cambridge, Mass.: M.I.T. Press

Gambetta, D. 1988. Can we trust trust? In D. Gambetta (ed), *Trust: Making and Breaking Cooperative Relations.* Oxford: Basil Blackwell, 213 - 37

Geanakoplos, J., Pearce, D. and Stacchetti, E. 1990. Psychological games and sequential rationality. *Games and Economic Behavior* 1, 60 - 79

Guilford, T. and Dawkins, M. 1991. Receiver psychology and the evolution of animal signals. *Animal Behaviour* 42, 1 - 14

Hausman, J. 1997. Trust in game theory. Working paper, London School of Economics

Hirschman, A.O. 1984. Against parsimony: Three easy way of complicating some categories of economic discourse. *Amer. Economic Rev. Papers and Proc.* 74, 88 - 96

Kitcher, P. 1993. The evolution of human altruism. *J. Philosophy* 90, 497 - 516

McDowell, J. 1994. *Mind and World.* Cambridge, Mass.: Harvard University Press

Nagel, T. 1978. *The Possibility of Altruism.* Princeton: Princeton University Press

Orbell, J. and Dawes, Robyn M. 1991. A 'cognitive miser' theory of cooperators' advantage. *Amer. Pol. Sci. Rev.* 85, 516 - 528

Ordeshook, P.C. 1986. *Game Theory: An Introduction.* Cambridge: Cambridge University Press

Pasteur, G. 1982. A classificatory review of mimicry systems. *Ann. Rev. Ecological Systems* 13, 169 - 99

Pettit, P. 1995. The cunning of trust. *Philosophy and Public Affairs* 24, 202 - 25

127 - 58

Spence, M.A. 1974. *Market Signaling: Informational Transfer in Hiring and Related Screening Processes.* Cambridge, Mass.: Harvard University Press

Zahavi, A. 1975. Mate selection—A selection for a handicap. *J. Theor. Biology* 53, 205 - 214

Chapter 2

A FORMAL SPECIFICATION OF AUTOMATED AUDITING OF TRUSTWORTHY TRADE PROCEDURES FOR OPEN ELECTRONIC COMMERCE

Roger W. H. Bons*
Corporate IT Department, Philips International B.V., Eindhoven, the Netherlands.
roger.bons@philips.com

Frank Dignum
Faculty of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands.
dignum@win.tue.nl

Ronald M. Lee
Erasmus University Research Institute for Decision and Information Systems (EURIDIS), Erasmus University Rotterdam, The Netherlands.
rlee@fac.fbk.eur.nl

Yao-Hua Tan†
EURIDIS and Erasmus Center for Electronic Commerce (ECEC), Erasmus University Rotterdam, The Netherlands.
ytan@fac.fbk.eur.nl

*The points of view expressed in this publication represent the opinion of the author and do not necessarily always correspond with the current official standpoints of Philips.

†The research of this author was carried out when he was the Harry W. Reynolds Visiting professor at the Department of Operations and Information Management of the Wharton School of the University of Pennsylvania. The financial support from the Harry W. Reynolds fund is gratefully acknowledged.

C. Castelfranchi and Y.-H. Tan (eds.), Trust and Deception in Virtual Societies, 27–54.