TRUST IN SIGNS[1]

Published in Karen Cook (ed.) *Trust in Society*. New York: Russell Sage Foundation, 2001, pp. 148-184

NB: This version of the paper does not coincide in all details with the published version

[1] Several colleagues have 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.

INTRODUCTION

In this article we embark on a re-orientation of the theory of trust which contains five steps. The first four establish a new theoretical framework for determining when trust and its fulfilment are to be expected. The fifth lies partly in the future, and will implement this theoretical framework by setting out the detailed structure of the semiotics of trust, which we initiates in the last section of this essay.

In section I 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 by raw payoffs, or by all-in payoffs induced by such trustworthy-making qualities.

In the second step (section II), 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 has been transformed into a `secondary problem of trust'.

In the third step (section III) 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.

We then take a side step (section IV) and develop signalling theory in a new direction. We extend it to the case in which what is signalled by a signaller is his identity. This is of crucial importance for understanding how trust decisions are taken, because the secondary problem of trust is often soluble by assessing the reliability of identifying marks. This is the case every time we can determine whether a trustee has trustworthy-making qualities indirectly, by establishing whether he is the same person who has proved trustworthy in the past. The opportunists of this case are impersonators, who

mimic a possessor of trustworthy-making qualities by imitating that person's identifying marks.

In step five we describe some salient concrete features of the complex semiotic structure of secondary problems of trust. These features, some conceptual and some empirical, must be studied and specified in order to turn our framework into a tool for the empirical analysis of real-life trust problems. Here we initiate this study and this specification. The semiotic structure involves a range of different genera of signs, a range of different kinds of costs of emitting signs, and a range of strategies for the protection of signs against the threat of mimicry.

Our paper carries a number of implications for the methodological issue of whether rational choice theory is an appropriate tool for the analysis of trust. Our analysis suggests that it is. In particular, it suggests that the subtlety and richness of the judgements a truster makes when she trusts, are due only to the complexity of the game that she finds herself playing. The apparently ineffable nature of these judgements is an illusion, which we hope to dispel by describing the detail of the structure of such games. However, 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.

However, 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. In this paper we treat this incidence as given. Actual trust, however, may fall below that threshold because of the difficulties incurred communicating the presence of these qualities. We investigate only 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 explore the determinants of the significance of such signs.

In doing so we make 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 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.

I

PRIMARY TRUST

The notion of trust 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.

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. 1a shows the payoffs of the truster together with the 'raw' payoffs of the trustee: that is, the payoffs that would be his were he motivated by simple self-interest. 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, 1	-3, 4
A'	x, ..	y, ..

Fig1a: R's payoffs and 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. 1b. 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	3, 2	-3, -4
A'	x, ..	y, ..

Fig1b: R's payoffs and E's Possible All-In Payoffs

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 that R can do either A or A', and E either B or B'.[2] 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.[3]

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, some of which we will record later in the paper. 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.

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

[3] In the language of games of incomplete information (Fudenberg and Tirole 1991, Ordeshook 1986), 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.

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.1b. 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 be handled by exactly the same methods we develop for basic trust games, at the cost only of some increase in complexity.[4]

We shall say that, in a basic trust game, the truster R **trusts**[5] 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.[6]

---

[4] Corresponding to perils B' and B" there are types of trustee whose raw payoffs are as shown in Fig.1c.

	E		
R	B	B'	B"
A	1	4	0
A'	-	-	-

Greedy Type

	E		
R	B	B'	B"
A	1	0	4
A'	-	-	-

Libidinous type

Fig.1c: E's Raw Payoffs in Multiple Peril Game

[5] We use bold characters throughout the text to mark the terms that we define. See the Glossary at the end of the article.

[6] There are cases, 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. 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

Trust-warranting properties

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 1988, Gambetta 1988, Hausman 1997, Pettit 1995, Hirschman 1984).[7] Another class of reasons are generated by values held by the trustee that 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 1993, Nagel 1978).

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.[8] 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.

---

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.

[7] 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 1 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.

[8] 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.

By contrast, 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 Fig. 1a 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 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 1978). 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.

II

SECONDARY TRUST

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,[9] 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.[10]

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.

---

[9] 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. McDowell 1998).

[10] 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 that 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.

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 is 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.[11] (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.

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

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

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, by contrast, is a property that 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.[12] 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'.[13] 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.[14]

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

[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] 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 (Pasteur 1982).

[14] Mimicry may also be 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 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'.

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 probability that the trustee is an opportunist; and (ii) the truster observes 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.

Imagine 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.

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?". The first is asked by the truster; the second by the trustworthy trustee 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.

There is, moreover, one secondary problem of trust that pervades all interactions between any two people which are spread over time. Present behaviour can have no effect on future payoffs, and so future payoffs can provide no present incentives, unless a fundamental condition is present in the future encounters. This is the condition that the truster knows that her interactant is ... who he is! This condition is necessary whatever the specific nature of the successive interactions (whether they are simple repetitions of a fixed basic trust game, opportunities for reciprocation, for punishment, or whatever). For, as we shall show in section IV, identity itself is a krypton.

III

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 do two things: first, we show how the theory of signalling which developed in economics, biology, and game theory (Spence 1974, Zahavi 1975), 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. Second, we expand signalling theory to treat `signalling via identity', a mode of signalling which is of key importance when the basic trust game is repeated over time and reputation is invoked in solving problems of trust. Signalling via identity has until now received little attention (but see Bacharach 1997, Tadelis 1996).

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 1974) 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[15] there is an equilibrium in which at least some truth is transmitted, provided 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.

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

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 1991). 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.[16] 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*.[17] 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

---

[16] 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 fulfill 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.

[17] 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 rational 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 a 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(c; s) to his being k, where $Pr(c; s) = f/[f + p(1 - f)]$, by Bayes' theorem, since $Pr(s; c) = 1$. 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 $Pr(c; \text{not-s}) = 0$. 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).

k's and non-k's; and a receiver knows the base-rate probabilities that a signaller is k and non-k. When she 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 opportunistic 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.[18]

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. In the remainder of the paper we shall make a start in filling the gap that is left.

IV

IDENTITY SIGNALLING

Signalling via individual identity

This section concerns an extremely widespread strategy for using manifesta to give evidence of a special kind of krypton, identity. This strategy gives rise to a special problem of secondary trust - trust in signs of identity. When it is used, the truster employs, instead of the three-layered inferential structure

[18] 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.

m ---> k ---> t

Fig. 2a

the four-layered structure

g ---> i ---> k ---> t

Fig. 2b

where `t' denotes trustworthiness, `k' denotes k-ness, `i' denotes identity, ` m' denotes presence or absence of a manifestum of k, and `g' denotes presence or absence of a manifestum which is a sign of identity called a `signature'. The threat of mimicry of k through m is replaced by the threat of mimicry of i through g.

Signalling via identity (which for the sake of concision we shall call 'identity signalling') arises in dynamic contexts. In particular, it does so when a basic trust game is played repeatedly. Unless players are locked in a room and monitor each other constantly, there is the possibility that partners may change. Whenever this is so, the repeated basic trust game is a basic trust game with random rematching. In such a game, the trustee may wish to signal his identity so as reassure R that he is the same person that R encountered before.

Identity signalling pays, when it does, by enabling the signaller to exploit a reputation. The latter is not always straightforward. This is because, frequently, the fact that someone is the bearer of a certain reputation is a krypton of that person! For example, Armani has a reputation for selling well-designed clothes, but to exploit this reputation a seller must convince customers that he is Armani. Islamic Jihad has a reputation for carrying out its threats against hostages, but to exploit this reputation a group of kidnappers must convince governments that they belong to Jihad. The IRA have devised a special secret code - known to the British police - which they use when they issue warnings about bombs they have placed. (Not only does the code make the information credible because it re-identifies a group which has acquired a reputation for planting bombs but, so long as it remains secret, it also makes the telephoned information credible because it creates an obstacle for would-be IRA impersonators who would like either to play nasty pranks or to plant a real bomb to discredit the IRA.)

Identity signalling is a strategy for signalling a krypton which works by giving evidence of another krypton, that of being the reputation bearer. It can be an efficient means of krypton signalling, partly because identity signals can be very cheap, and partly because it capitalises on certain pre-existing beliefs of the receiver. The signaller does not need to send a signal to induce these beliefs. They are of two kinds: the signaller's reputation with the receiver, and `trait laws' believed by the receiver. We first consider trait laws.

Common-sense belief systems contain trait laws. These are laws of constancy, of two kinds: individual trait laws, of the form "once a k always a c"; and categorial trait laws, of the form "one X a k, all Xs k's". A categorial trait law is a schema in which both `k' and `X' are variables. It is not of the form "all Swedes are fair-haired" (a specific reputation of Swedes), but rather "all people of a given nationality have the same hair colour" or "hair colour is a national characteristic". Similarly, an individual law is not of the form "Alderman Brown is a stuffy so-and-so" but rather "someone's degree of stuffiness is the sort of thing that doesn't change". There is no need for a trait law to be as rigid as this: k may be not a sure property but a merely probabilistic or tendential one, as in "within each nationality, the *probabilities of various hair colours* are the same for all persons", or "a given individual's tendency to cheat is constant over time". In this paper we deal only with sure k's, but this is for simplicity, not because the logic of our argument requires it.

A reputation is formed in someone's mind by, as it were, experiencing a krypton. An **experience krypton** - a generalisation of the economic notion of an `experience good' (Nelson 1970) - is a krypton which is revealed by interacting in a certain way with its possessor. Someone's being the seller of a good wine is revealed by buying wine from him; someone's being honest is revealed by trusting him to repay a loan; an organisation’s readiness to execute hostages if its demands are not met is revealed by not meeting them. The idea is that the outcome of the interaction reveals to one interactant whether or not the other has the krypton.

We can put this in terms of a basic trust game: a t-krypton k is an experience t-krypton if the truster discovers whether or not the trustee has k by trusting her. The interpretation is that the outcome of the trusting act and the trustee's act reveals to the truster whether or not the trustee has k: more precisely, the outcome of (A, B) reveals that he has k, while the outcome of (A, B') reveals that he does not. [19]

This discovery gives birth to a reputation for k-ness or non-k-ness and so either for trustworthiness or untrustworthiness. If the trustee were anonymous, however, reputation would be stillborn. For a reputation to issue in advantage, for its value to be realised, its owner must reveal himself to those who believe in it. That is, he must engage in signalling behaviour in which he signals his identity by displaying an appropriate manifestum.

Manifesta used to signal identity are usually of a kind we shall call `signatures'. A **signature** is a manifestum which is drawn from some fixed family, called a stock, which are all alike in some respect: one stock is the set of all possible English proper

[19] Evidently, act-outcomes often reveal t-krypta less definitively than this. The hustler may feel the truster a few B's, though lacking the t-krypton; an honest man by contrast may occasionally succumb to the lure of B', though possessing it. Our definition of 'experience krypton' defines only an ideal case, which may be relatively rare.

names, another the set of all possible fingerprints, another the set of all possible uniforms, another the set of all possible embroidered emblems.

Suppose that a trust game with rematching is enriched as follows. Before a trustee plays for the first time, she acquires a signature from a certain stock, by a mechanism which ensures that it differs from all other signatures which are allocated (the mechanism is `heteronymous'), thus avoiding the possibility of 'mistaken identity'. Genetic variety does this for fingerprints and faces (and other features too which still keep being discovered, such as odours); the size of the set of common names almost does it for the standard mechanism of name-allocation (help yourself). On any occasion on which this trustee plays a game, she is able to display her signature if she wishes.

Here is how the identity signalling mechanism works to allow a trustee who has k to show that he has. Suppose that on some occasion this trustee, chooses to display his signature. His truster on this occasion, R, if she trusts him, soon finds out whether or not he is k (since k is an experience krypton). Thereafter, whenever E encounters R again, E is in a position to provide evidence that he is E, simply by displaying his signature. Provided that R remembers it, she can infer (from heteronymity) that he is the previously encountered trustee who turned out to have k, and she can conclude (from the individual trait law `once a k always a c') that he still has k.

The simplest example of such a repeated trust game with random rematching and identity signalling is a sequence of face-to-face encounters with random rematching and facial display. It has been suggested that human facial recognition evolved because this capacity allowed interactants in collective action problems to keep track of interactants who had proved cooperative (Cosmides and Tooby 1992). If so, then what took place was identity signalling in which the signature stock was the set of possible human faces.

Face-to-face interactions are (still!) the kind that first spring to mind. They have the properties, key advantages, as we shall explain in Part 3, of being costless to display and almost impossible to reproduce. It is perhaps because it is easy to forget how lucky we are to be able to count on facial recognition in so many of our dealings that we may overlook that individual identity is a krypton, and that even when people repeatedly interact the problem of signalling who you are is often far from trivial. In particular, the further that interaction departs from ideal face-to-face conditions, the less safe are signallers of personal identity from the corruption of their signals by the mimicking activities of impersonators.

Identity signalling often provides a highly efficient way of signalling a krypton, for example a t-krypton which implies trustworthiness. One reason for this is that it is easy to contrive stocks which have heteronymous allocation mechanisms and whose signatures are very cheap to display. Both of these things tend to be true of stocks of symbolic signatures. The ultimate source of the efficiency of identity signalling, however, is the presence in the receiver population - in our case, the population of

trusters - of pre-existing beliefs in trait laws. The identity signalling strategy capitalises efficiently on these beliefs.

Indeed, we can regard the identity signalling mechanism as a prime example of applying the following very general maxim for signallers: seek background beliefs which link the krypton with another more efficiently signalled one. Reputation works as a means of signalling krypta when, but only when, both halves of this recipe are present. The general maxim has wide application in the world. As Figure 2 shows, a truster's inferences contain two sorts of links: from manifesta to krypta, and from krypta to krypta. Optimising a signalling strategy may often involve inducing the receiver to follow a quite roundabout route, going through several krypton-krypton links leading from a signallable krypton to the one that counts; many krypton-signalling mechanisms (for example, Spence's) are indirect in this way.[20] However, it must be remembered that we have been idealising away from uncertainty by considering only rigid trait laws; when beliefs connecting krypta are only partial or probabilistic, increasing the roundaboutness of the mechanism decreases its power.

Categorial identity signalling

So far we have focused on the individual case. Consider now a class or category of people X all of whom are k's. X's may be able to mark themselves with a group signature g (i.e. a signature which is displayed by all X's, but by no non-X's). Suppose they do. Suppose further that a truster R discovers as a result of an interaction with some individual X, x say, that x has k. Thereafter, if R has an encounter with any X, x' say, and the latter displays g, R can conclude (by heteronymity) that x' is an X, and therefore (by the categorial trait law "one X a k, all X's k's") that x' has k.

Categories have both advantages and disadvantages over individuals as units of identity for purposes of signalling trustworthiness. Because they can be big, a reputation for k-ness can be built, and travel, much more rapidly. It may be that, with respect to certain kinds of krypton, people are more inclined to believe in categorial than individual trait laws (people may indeed believe unduly in national characters and ethnic traits), partly because the former provide simplified, more cognitively tractable views of the world (Tajfel 1959). Categories are more powerful in various ways; for example, social groups may be able to secure observance of trait laws, through social influence on their members, while individuals may lack analogous means to keep themselves in line over time.[21]

---

[20] In Spence's model, k is a property (high productivity) which is correlated with another (academic prowess). The k's strategy for demonstrating the former is to signal the latter, which he credibly achieves by displaying a manifestum (a certificate) which he can afford to get and a low-prowess job applicant cannot. He leaves it to the employer to draw the inference from prowess to productivity.

[21] There are two ways in which the presence of non-k members can undermine categorial identity signalling: it undermines social belief in the trait law; and these members are able to mimic k-ness by using the group signature: for as long as the cost of the latter is the same for all group members the Cannot condition fails if the Can condition is met. An important example of this mechanism occurs when group membership (i-ness) is taken as evidence

On the other hand, in order for the identity signalling mechanism for showing that you have k to work, a signature must be fixed over occasions of interaction. For an individual this just means coordinating her display behaviour over time. Above, we merely assumed that she could do this effortlessly, and this is indeed often so. It is not always: for example, in the case of PIN's and distinctive clothing, to display her PIN, or what she wore, again the would-be self-identifier must recall it. But this is often not hard. By contrast, a category of people must, to get identity signals to work, coordinate across members to adopt matching signatures. So the most likely categories to use categorial identity signalling are not just sets of people who share any old k, but sets which already have enough cohesion as a group to facilitate this coordination. Consider an example reported by Frank (1994:207): many New York couples advertise for a governess in Salt Lake City, having learned that people brought up in the Mormon tradition are trustworthy to a degree exceeding that of the average New Yorker. Mormons have the right kind of cohesion and can coordinate their signatures; people who love children have not and cannot. Hence, if k is the property of being diligent in looking after a child left in your care, the krypton Mormon which is a sign of k can be signalled via identity, while the krypton children-loving which is also a sign of k cannot be.

---

against k, so that i members who are k have an honest motive to camouflage i-ness by suppressing a manifestum m which acts as an involuntary signature of i-membership. It is, for example, difficult for k members of an ethnic minority which is reputed to be non-k to signal, successfully, that they are k, by merely camouflaging their ethnicity, because it is just as easy for opportunistic non-k members to camouflage it. On the other hand, such camouflage may be an important method of escaping from the trap of their group identification (their i-ness) in cases where this acts only as an initial filter. Those who are not excluded summarily by their i-ness may have an opportunity to signal their k-ness by some further signal which they can afford and their non-k fellow-i's cannot.

V

THE MANAGEMENT OF MANIFESTA

In section III and IV we have argued that signalling theory provides a general analytical framework for understanding the phenomenon of secondary trust. Signalling theory, however, is abstract and incomplete. It does not arm our imagination with the tentacles it needs to grasp the infinite variety of signs which can be emitted by trustees and processed by trusters in mimic-beset trust games. Its abstractness makes it difficult even to see many of the trust problems we regularly face as what they are, signalling games which involve trust in signs.

In particular, signalling theory lacks a concrete semiotic structure. Yet this is indispensable for introducing order into the variety of signs of trustworthiness, and for accounting for the varied strategies of `sign management' that people use. There is scope, above all in humans, for creating new signs, for discovering latent ones, and for protecting signs against mimics. Protective measures are in turn threatened by stratagems to get round them, giving rise to a relentless semiotic warfare in which technology plays a major part. We see bank notes, forgeries, watermarks, forged watermarks; we see smiles, false smiles, scrutiny, and dark glasses. In section IV we made a beginning in the task of delineating the concrete structure of signs and their management, by describing one important general strategy for signalling krypta, identity signalling. In this part we pursue this enterprise. In section 1 we briefly discuss the routes by which a manifestum may come to signify a particular krypton. In section 2 we give a taxonomy of manifesta, and discuss the relative advantages for the signaller of manifesta of the different taxa. In section 3, we describe some techniques of `sign management', including the `protection' and `multiplication' of signs. In section 4 we recount a pair of real incidents which set primary problems of trust: we show that the theory we have developed provides an analytical reconstruction of the truster's intuitions on that day. By the same token, the truster's intuitions illustrate how rich and subtle the reasoning is which sustains our everyday decisions as trusters about whether to trust, and our everyday moves as trustees in managing signs.

“Model comes before mimic”

The existing literature on mimicry, which has developed in biology, includes some applications of signalling theory, but is mainly concerned with the variety of forms and functions of animal signalling (Hauser 1996). Almost nothing has been written about the human case.

In biology, the standard dynamics are that in a first phase there emerges a mutant of some k-possessing type of organism. This mutant bears a manifestum m. This gives it a selective advantage over other k-possessors because a certain type of receiver learns to associate k with m. For example, k is toxicity, m is a bright marking, receivers are predators. In a second phase, a mutant of non-k (say, non-toxic) organisms emerges which also bears m. It too is selectively advantaged, in this case over non-k's without

m. As this variety (`mimics') multiplies, both its advantage and that of m-bearing k's (`models') diminish, in both cases because the correlation of m with k becomes weaker as mimics multiply.[22]

We have seen that, in any SSE of a signalling game in which the signal is a display of a manifestum m, each non-k who displays m is a mimic of k through m. But SSE's fail to capture other aspects of the notion of mimicry. In particular, the signalling-game model describes no real-time process through which m comes to signify k. Indeed, for all that this model says, the significance of m can spring into existence at the same moment as the deceitful use of the same m. But almost always, in animal and human affairs, `model comes before mimic'. The standard way in which an m comes to be evidence of a krypton k for receivers at time t is the growth of an associative belief: before time t there has been a history of experiencing krypta associated with independent evidence of k-ness.[23] Signalling theory does not deny this explanation of the evidential force of s; it only fails to model it. This is because it is an equilibrium theory with no dynamics. All it says about the evidential force of a manifestum in an SSE is that in such a state the probabilities with which k's and non-k's use s are known to the receiver; it says nothing about how the receiver comes to know them. In one interpretation she works them out a priori from various data including those on benefits and costs; on this interpretation, the equilibrium, complete with the significance and rate of deceptive use of the signal, springs into existence. But, equally, there is nothing to stop us from interpreting the SSE as the outcome of a learning process like the one sketched above.

Kinds of manifesta

The main result of signalling theory implies that a manifestum m is secure against mimicry if and only if it is cheap enough for a k to display and too expensive for an opportunist to display. These two conditions we have called, in section III, the Can and the Cannot conditions. The study of the reliability of different kinds of signs of krypta in general and t-krypta in particular can usefully be organised in terms of these conditions.

---

[22] Since in general there is some fitness cost in displaying m, in general an equilibrium is eventually reached in which, for mimics, the fitness benefit from the now imperfect correlation just covers fitness cost (the Can Just condition is met); while for models it still more than covers the fitness cost (the Can condition is met). This equilibrium is an SSE of an appropriately specified game. Notice that for this process to take place it must be that the cost of mimicry is not too high: it must be lower than the fitness benefit of being taken for a k when there are not yet any mimics. But the cost of mimicry must be higher than the cost of displaying m for a model.

[23] This describes how a well-founded associative belief, between an m and a certain k, forms in the mind of a *rational* truster. In this article we do not consider variously *biased* or e*rroneous* associative beliefs of which there are plenty in the real world: from the ancient Greek belief that beauty means goodness (*kalos k'agathos*), to the many extravagant folk injunctions still around today about whom one should trust: – people who use words sparingly, people who look straight in the eyes – or not trust: redheads, people with thin lips, women, people from the city, people from the country.

Cues

One favourable case is that m is a 'cue' of k. A 'cue of k' is a manifestum whose display is costless for k-possessors.[24] An example is an honest look, or, in identity signalling, one's handwriting or voice.[25] Since in a signalling game the benefit from being thought k is positive for a k, any cue of k satisfies the Can condition. There is no guarantee that cues satisfy the Cannot condition: if assuming an honest look were easy enough, the Cannot condition would fail and honesty would be mimickable through an honest look. However, cues usually have at least some positive cost for non-k's. If so, signalling theory implies that sometimes k's can rely on the cue of k to convince: namely, when mimics have little to gain. Often, though not always, the cue is `there anyway' and the k need take no action to manifest it: these are `automatic cues'. When cues are automatic, a k can take it easy. Indeed he need hardly be aware of the cue's effect to benefit from it, as we shall see.[26]

Evolution has equipped us with many cues. These may be categorial, such as signs of gender, or individual, such as signs of age. Cues of this kind are often costly, sometimes impossible, to mimic. Some, like the face, could have evolved, together with a remarkable ability to discriminate those of other humans, because they are advantageous in that they sustain cooperation by making identity signalling cheap and safe from mimics. By contrast, other biological cues of identity, some of which are still being discovered, may have evolved for reasons unrelated to cooperation, or be just random individual differences which become observable with the right technology; in so far as they are heteronymous, these manifesta can be employed for re-identification. British banks will soon introduce new devices at their cash dispensers, `palm readers' or `finger readers', which trade on the fact that in no two individual is the shape of the hand, or even a single finger, the same. They will replace PIN's, which are kept safe from imitation by mimics only by secrecy.

---

[24] It is the marginal cost of display which is zero, not necessarily the historic cost of developing the capacity to display it.

[25] This sense of `cue' resembles Hauser's (1996). Although k's who display a manifestum m may do so as a signal of k, they may also display m for some other reason, and indeed in some cases without any purpose. Rich people often wear expensive clothes with no thought of conveying anything about their wealth, but merely to make a *bella figura*; as an unintended by-product they give evidence of their wealth. In such cases m is a cue of k even though it is costly to produce, because it is not a costly input *into the activity of inducing a belief in k-ness*.

[26] It may be that m is a cue of k but is also costless to display for *some* non-k's. In this case, the truster ought to be concerned about taking m at face value (i.e. as indicating k and so trustworthiness) even if she is sure there are no opportunists (say because it is common knowledge that the penalty for opportunist behaviour is certain death). For she may think that non-opportunistic egoism possible. Suppose for instance that all k's have wide-set eyes, and most non-k's have narrow-set eyes, but some non-k's (say 10%) have wide-set eyes. Say the population is 50% k and 50% non-k. Then if R sees that E has wide-set eyes she should assign the probability 1/11 to E's being non-k. According to the payoffs it may or may not be wise to trust E.

An interesting class of cues comes as a by-product of the life each individual lives. As we grow older, our diet, occupation and lifestyle shape our limbs, looks and quirks. Cues of this kind also apply to categories of people, who share a language, a pronunciation and practices of many sorts.

Symbolic manifesta

An unfavourable case is that m is symbolic, in the sense that it consists in a configuration of characters however these may be physically realised. It is exemplified by names, logos and oaths. What makes this case unfavourable is that among the physical realisations there are usually some which are very cheap for anyone, non-k's included, to produce. The efficient production cost of a verbal claim or a false signature is virtually zero. Symbolic manifesta are attractive for signallers because the signaller effortlessly meets the Can condition, but since they violate the Cannot condition their evidential value is under threat. The expansion of the scope for ultra-cheap transmission of symbol-strings is indeed a major cause of the growth of mimicry in our time. However, the Cannot condition is not necessarily violated, even if the costs of producing the manifestum are zero, for the costs of producing m may not be, as we shall see, the only costs of displaying it.

Symbols are characteristic of actual identity signalling. Symbolic signatures, individual or categorial, abound. Like other symbolic manifesta, they are vulnerable to the mimic, if the production element in cost dominates. We do indeed find much mimicry of identities through symbolic signatures, both of a personal kind (impersonation), and of a categorial kind (posing). However, it may sometimes be that mimicking the trustee E, or a category of trustworthy Xs (and hence indirectly k) through some g is more costly than mimicking k directly through some m. For even though the Cannot condition fails to be satisfied on the production side, there are ways specific to identity signals in which E, or Xs, can often raise the cost of mimicry to would-be mimics.

Fakeable manifesta

A second important unfavourable case is that m can be faked. Faking is not the same thing as mimicry: what is mimicked is a krypton, what is faked is a certain kind of object. Mimicry uses all sorts of techniques, from lying to plastic surgery, from make-up to the imitation of bodily movements. Fakery is one technique among others. If I mimic a rich man by wearing an expensive suit, I do not fake or forge the suit. If I mimic a devout person by wearing a skullcap, I do not fake the skullcap. If, however, I mimic a rich man by wearing what looks like but is not an expensive suit, then I employ fakery to execute mimicry. If I write your name on a cheque so that it is taken for yours, I forge something to execute mimicry. And if, to convince customers that my roadside restaurant has a good kitchen, I place cardboard mock-ups of container lorries in the parking area, I fake lorries as part of a strategy to mimic an establishment where one eats well.

The objects that are faked or forged in these examples are not true manifesta, but `quasi-manifesta'. By definition manifesta are observables: if m is a manifestum, a

truster can tell by looking (smelling, hearing) whether or not a trustee is displaying an m. But if a thing can be faked, then *ipso facto* trusters cannot tell whether or not what is displayed is that thing. A fakeable object o (or type of object o) is one which can be simulated by another, o'. For faking to be successful, it must be possible for an observer to mistake o' for o.

The following definition of fakery will do for our purposes. First, generalise the notion of manifesta to `exhibits'. An exhibit is a displayable feature of a person (for example, a part of his body, a piece of behaviour, an object attached to him) of which one aspect is observable (that is, is a manifestum), and the other is not: call this the `latent component'. Consider any exhibit (m, n) (where m denotes the manifestum and n the latent component). To fake (m, n) is to display an exhibit (m, n'), where n' differs from n, with the object of convincing an interactant that it is (m, n). To do this successfully, n' must be observationally indistinguishable from n. An opportunist, to convince a trustee that he is to be trusted, may display a fake testimonial, or a fake smile. The latent component of the testimonial is its authorship; the latent aspect of the smile the emotion which it expresses or fails to express.

An important class of cases in which false signalling by faking occurs is that in which displaying a real m satisfies the Cannot condition. For example, m is a certificate which can only be got if you can prove to the issuing office that you are k. In this class of cases, it may be that the best hope of falsely signalling k is by faking m. For it may be that, although displaying m is prohibitively costly for a non-k (the signal display-m satisfies the Cannot condition), displaying a fake m is not, because k's have no particular advantage in producing the manifest component of m.[27]

Sign management

So far, we have described the characteristics which, by helping or hindering the satisfaction of the Can and Cannot conditions, militate for or against the effectiveness of *given* signals. The question naturally arises how we may expect signallers, and in particular trustworthy and opportunist trustees, to optimise over alternative signals and over variables, which may affect the costs to them and others of using given signals. We call the ensemble of such optimising activities `sign management'. In this section we discuss two important principles of sign management, protection and the multiplication of signs.

Humans are often faced with situations in which the Cannot condition fails, and indeed fails badly, so that even quite high rates of mimicry still leave mimics with a profit. Trusters and trustworthy trustees have developed many strategies to combat those situations. One type of such strategy is protection.

---

[27] In real cases of Spence's scenario, the signalling strategy for demonstrating high productivity (demonstrating academic prowess) contains a third step, in which fakery plays a part. Employers observe not the quality of one's performance at college, but a certificate or transcript. The piece of paper you show is a pair whose second element is non-manifest, and could be such that the certificate is no evidence of prowess: it is if you have counterfeited it.

To `protect' a manifestum m of a t-kryptum k is to implement the Cannot condition by deterring would-be mimics either from producing or displaying. But to do its job, the strategy must do this without endangering the Can condition: the cost of, say, punishing mimics, must be affordable by trustworthy trustees (k's). There may be alternative manifesta which would convince trusters without any protection, but these may be so expensive to produce or display as to violate the Can condition. What counts is the sum of costs of production, display, and protection: it is this which must be too high for mimics but affordable by k's. As it happens, cheap manifesta combined with strong protection is a combination that is often found.

Protection strategies are of several kinds: they may be directed at detecting mimics in the act of faking or fraudulently displaying manifesta, or at demonstrating at a later stage that a mimicry has taken place. The mimic is then punished. Such strategies may be put in place either by trustworthy trustees, trusters, or some coalition. Thanks to efficient legal systems, firms can practise effective identity signalling of quality by incurring the trivial cost of displaying a name or a logo. Even if there are manifesta which support tolerable equilibria unprotected, through their production and display costs alone, these are frequently more costly *in toto* than such symbolic signatures protected.

It may be that groups successful in category signalling tend to have members who are not cryptic but relatively transparent to each other. Mormons are no doubt better than non-Mormons both at recognising Mormons and at recognising non-Mormons. In particular, we may expect successful category signallers to have good negative as well as positive recognitional capacities, for this allows them to detect mimics and so to operate protection strategies. For this reason, signalling via identity of X-membership will tend to be more feasible the more familiar to each other are members of the category X. (In the individual case this reaches an extreme degree: people are very expert at telling who is, and who is not, themselves!)

A type of sign-management activity as important as protection is the choice of which signals to use. This may involve searching for or designing new signals. The `model' is often in a position to raise the cost of mimicking her through her manifestum, by modifying it to make this more expensive, or by choosing or sometimes inventing a new type of manifestum. For example, fakeable manifesta may be rendered harder to forge by introducing some costly device in their production, such as digital watermarks which can identify photographs transmitted on the Internet. Several episodes in the history of fashion are of the latter type: when a garment becomes more affordable it stops being a reliable sign of opulence, and those who intend to signal their opulence by the way they dress must switch to costlier ones. If the trustworthy trustee can pursue such a strategy at a low enough cost to herself, he can banish the threat of mimicry. Generally, a failure of the Cannot condition to obtain through mimicry costs in one category can be made good by raising those in another.

A feature of a krypton which militates against mimicry is multiplicity of its manifesta. Say that good eating-places are frequented by many patrons who are mostly local

people. When I see that a place is crowded I may think "Ah! that looks like a good place for lunch." But when I enter I may notice that they are all tourists. This will lead me to conclude that this cannot be an echt establishment. The problem for the mimic is not just that the technical difficulty of the simulation accelerates with the dimension of the manifestum, but that it is very easy to forget one and give the game away. A good mimic must be thorough and display all the characteristic signs of a k. How multiple the manifesta of a krypton are depends on the 'bandwidth' of the signalling stage of the game. It is often thought to be easier to mislead someone about yourself when nothing is exchanged but strings of ASCII symbols, as in Internet communication, than in face-to-face encounters. When a customer opens a bank account in New York he is given a PIN and asked to specify three further identifiers: his mother's maiden name, the name of his elementary school, and a password of his choice. The base-rate incidence of mimicry in New York is probably higher than in Oxford, where only two identifiers are requested.

Ethnic groups with a long common history are usually very robust because the constellation of manifesta that identify them is extensive. Art Spiegelman reports that, during the German occupation of Poland, his father used to travel to town by the tram. It had two cars: "One was only Germans and officials. The second, it was only the Poles. He always went straight to the official car" where a simple salute, 'Heil Hitler', was enough not to call attention, whereas "in the Polish car they could smell if a Polish Jew came in". It was harder for a Polish Jew to mimic the nuanced multiple signs of a Polish gentile than the fewer superficial manifesta of a pro-Nazi. (Spiegelman 1991: 142).

Two incidents

One of us was recently stopped by a young man at the door of the College library. He said: "I am a student and I've forgotten the code to get into the library. Could you please let me in?" His non-verbal manifesta matched his claim. He looked a plausible Oxford undergraduate; he also had what to Gambetta seemed an honest look. Still, the library is full of precious books (raw benefit), while his verbal claim was costless; his scruffy looks are socially approved among today's students, and thus cheap not only to produce but also to display (low cost of mimicry). Before making a decision, Gambetta therefore looked at him with suspicion (closely observed his manifesta), and probed him with questions designed to establish the krypton in question, to establish whether he really was what he claimed to be. Hardly a minute later, Gambetta ran into a group of about ten people chatting in the cold at the entrance of a College seminar room. One of the group said: "There's going to be a seminar in the Wharton room but we're locked out. Could you let us in please?" That room is full of precious paintings and furniture (raw benefit), but Gambetta did not think twice before letting the whole group into the room.

Why was he more suspicious of one person than of many? Had we not been writing about the problem we would hardly have noticed the curious difference between the two incidents in terms of the truster's reaction. In the first, Gambetta sought further evidence of the claim before believing in the student's trustworthiness, in the second

he did not. Even the second incident might have been an instantiation of an SSE: the group of people might have been a mimicking team planning to rob the room clean once Gambetta had let them in and gone. Yet, he took the claim to be honest and believed that the members of the group had the krypton in question, were genuine seminar participants stuck in the cold. We tend to perceive our reactions of this kind as guided by intuition, and as therefore not governed by complex computations. But this is only because we are so good at this activity; it comes naturally to us, even though it entails reading a rich web of signs, and the rapid computation of sometimes complex cost-benefit balances.[28] The key manifesta Gambetta observed were that the trustee was a group and that the group members looked like graduate students and academics. Unreflectingly, Gambetta must have formed a rough idea of the cost relative to the benefit of assembling a whole group of people, and of supplying them all with the right looks for the staging of an elaborate act of concerted mimicry. Given these estimates, the conclusion that it was too expensive followed easily. Had the group asked to be let into the room where the College's silver is kept (very high raw benefit), Gambetta's conclusion would have been different.

Reflecting on the behaviour of the trustees in these examples from the point of view of the trustees shows the potential of signalling theory for predicting fine details of people's sign-management strategies. Neither the student nor the group of academics intentionally chose in the morning to wear their looks as manifesta for displaying to Gambetta later that day. Had they been untrustworthy opportunists they would have prepared for such encounters, and done so; but the trustworthy trustees did not expect to run into the difficulties they did, and made no preparations. By simply keeping to their normal style, and so conforming to the conventions of their categories, they convinced the truster. A substantial part of what it is to look like an academic or a student is a set of automatic cues: pale complexions, drooping shoulders, an absent air. Such cues give the true k-possessor two distinct advantages over her would-be mimics. Because they are cues, the Can condition is necessarily met. Because they are automatic, the k-possessor does not have to be on her toes, planning, her strategic faculties activated. (One needs to worry about cues only when they are counter-indications.) Cues need to be natural to be automatic. Just as the student unreflectingly wears his pale complexion, so too she unreflectingly dons her tracksuit top or her Doc Martens. Automatic cues are information manna which supports all human intercourse whose success depends on the correct identification of krypta and, in particular, does much to engenders trust just when it is warranted.[29]

The fact that, in displaying such non-natural automatic cues as Doc Martens, no conscious intentionality is involved, is consistent with our generally tailoring our

---

[28] Intuitive judgements can involve complex computations. There is a substantial literature in cognitive psychology, describing processes for framing impressions of people from pieces of their behaviour; much of this literature emphasizes that these processes are often `automatic' and without conscious awareness. Yet in some of it (e.g. Kunda and Thagard 1996) the automatic, unaccessed processes have considerable complexity.

[29] When cues are symbolic signs, like the styles of dress in the example, they usually give less security, other things being equal, because they are cheaper for the mimic to reproduce.

appearances in a way which deals with incidents of these kinds. If the same signalling situation arises day after day, what is at first a conscious intention tends to become habitual and internalised. But it remains an intention for all that, on this simple test: were one to be made aware of the issue - for example, by losing a suitcase containing all one's clothes - one would choose to re-outfit oneself with similar manifesta - for example, with new clothes in the same style - with the explicit intention of signalling one's t-krypta. In short, the intention does not have to be `occurrent', but only `triggerable'. So conforming to the dress codes and outward practices of one's group can be explained as a broadly intentional act, with the purpose of signalling membership of that group.

This example illustrates that the secondary problem of trust, 'trust in signs', is present in many daily encounters. It illustrates too that a krypton is often indicated by several manifesta which, in our truster role, we need to disentangle before we can roughly compute their cost and decide on their reliability. As trusters, we do not usually need to bother about the Can condition; typically, we already know from experience that k-possessors can afford to display m.[30] But we do need to judge whether the Cannot condition is satisfied: given the raw benefit, could a mimic afford this manifestum? This is often a complex task, since in analysing the affordability of signs, we face a very wide array of sources of sign cost. But we are often good at it owing to the richness of our background knowledge. It matters for the Cannot condition that the absence of a norm against down-dressing reduces the total cost of a scruffy look; but we come to the problem knowing what the norms are and easily marshalling this knowledge. Sometimes the analysis of costs of signs is less obvious. The longer it takes to acquire a certain look or accent, the more expensive it is for a mimic to assume it: to look like an academic costs nothing extra to academics, and something to non-academics; by contrast, to look like a student costs little to both students and non-students of the right age group. The analysis can produce surprises: while one person is generally less menacing than a group, this is not so when the menace is of mimicry: the number of people involved increases the cost of mimicking because this demands coordination and consistency. Confidence artists do better as loners.

[30] For example, we know that students can afford to look like students, for what it is to `look like a student' is to look the way that, in our experience, they usually do. This reason we have for not having to make calculations about the Can condition is due to the process of associative learning described in section III.

CONCLUSION

Several scholars have expressed scepticism about the specificity of problems of trust. One upshot of our re-orientation of the theory of trust is such a sceptical conclusion: the problem of trust is not, in its essentials, *sui generis*. This conclusion, however, is not nihilistic; we do not suggest that the problem of trust thereby evaporates. Rather, we reformulate 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.

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. It is certainly possible, and we do not wish to deny, 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. They may, for example, include a positive responsiveness to the very fact of being trusted. To this extent the full analysis of trust may involve attitudes which are indeed peculiar to games of trust. We are agnostic about the existence of dispositions of this kind, of attitudes primed specifically in trust cases. We are also neutral as to whether such dispositions, if there are any, are themselves explicable by methods of rational choice theory.

The fundamental problem for the truster, however, is independent of the existence or nature of such 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 unscrupulous to mimic trustworthiness by displaying its signs.

The ceaseless semiotic warfare between mimics on the one hand and their trustworthy models and potential dupes on the other, rather than being the proper target of a single article, defines an entire field of research which, for the case of humans, has barely been opened. Its overall outcome, at any one time, yields the amount of trust society

enjoys. In so far as people can observe, and read, reliable signs of trustworthiness, we can conclude that mimics have backed away or been fought off. But even in these happy circumstances models and dupes should not rest on their laurels. Mimics are always in the offing.

GLOSSARY

Raw payoffs are those that motivate an agent who pursues only his self-interest in the narrowest sense

All-in payoffs are those that motivate an agent once self-interest and all other considerations are taken into account

The **primary problem of trust** consists in identifying which of the two payoff structures really governs a trustee's actions

A **basic trust game** is a two-player – a potential truster and a potential trustee – strategic-form non-cooperative game. Each player has two strategies: in the loan example, the strategies of the truster are simply the actions `lend' and `refuse'; and those of the trustee are to `repay' and `not repay' if the loan is made. There is mutual knowledge of the truster's payoffs, but only the trustee knows his all-in payoffs: a basic trust game is thus a `game of asymmetric information' concerning the trustee's all-in payoffs. However, there is mutual knowledge of the trustee's *raw* payoffs. What only the trustee knows is whether or not his all-in payoffs coincide with, or differ from, his raw payoffs

A person **trusts** someone 'to do X' if she acts on the expectation that he will do X, e.g. in the loan example 'repay', when there is mutual knowledge between them 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

A trustee is **trustworthy** in a basic trust game just if: if he believes the truster trusts him, he chooses X

A **trusted-in act** in a basic trust game is the trustee's response expected by a truster when he believes a trustee is trustworthy in that game.

A **trusting act** in a basic trust game is the act chosen by the truster when he expects the trustee to be trustworthy in that game.

A **trust-warranting property** is any property (or combination of properties) of a trustee in a basic trust game that suffices for him to be trustworthy in that game

Krypta are unobservable properties of a person. Those krypta of a person which are his trust-warranting properties in a basic trust game are his t-krypta in that game

Manifesta are any observable features of a person. `Features' include parts or aspects of the person's body, pieces of behaviour by him, and his appurtenances

An **opportunist** in a basic trust game is an interactant who has two properties: he is motivated by raw payoffs and if he could obtain trust at low enough cost, he would do so, then betray it. An opportunist is not just lacking in trust-warranting properties but is proactively deceptive

A **mimic** is an opportunist who deceives by displaying a manifestum that is a sign of a krypton he does not possess

A **mimic-beset trust game** is a basic trust game where there is some positive probability that the trustee is an opportunist; and the truster observes a manifestum or manifesta of the trustee

The **secondary problem of trust** concerns the conditions in which manifesta of t-krypta may be trusted or distrusted. In order to answer the question "Can I trust this person to do X?", a truster must first, and need only, answer another question: "Is the manifestum of the trust-warranting krypton k which I observe a reliable sign of k?" In brief: "Can I trust this sign of k?"

A **signature** is a manifestum used to signal identity, which is drawn from some fixed stock of manifesta, which are all alike in some respect: one stock is the set of all possible English proper names, another the set of all possible fingerprints, another the set of all possible uniforms, another the set of all possible embroidered emblems.

Bibliography

Bacharach, M.1997. Showing What You Are By Showing Who You Are. Russell Sage Foundation Working Paper

Cosmides L. and Tooby 1992. *The adapted mind: evolutionary psychology and the generation of culture*. Oxford: Oxford University Press

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

Frank, R. 1994. *Microeconomics and behavior*. New York: McGraw-Hill, 2nd edition

Fudenberg, D. and Tirole, J. 1991. *Game theory*. MIT Press

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

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

Hauser M.D. 1996. *The evolution of communication*. Cambridge, Mss.: M.I.T. Press

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

Hirschman A.O. 1984. Against parsimony. Three easy way of complicating some categories ofeconomic discourse. *American Economic Review Proceedings*, 74, 88-96

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

Kunda, Z. and Thagard, P. 1996. Forming impressions from stereotypes, traits, and behaviors: a parallel constraint-satisfaction theory. *Psychological Review*, 103, 284-308.

McDowell, J. 1994. *Mind and world*. Harvard University Press

Nagel, Thomas 1978. *The possibility of altruism*. Princeton University Press

Nelson, P. 1970. Information and consumer behavior. *J. Political Economy*, 78, 311-29.

Ordeshook, P.C. 1986. *Game theory and political theory: An introduction*. Cambridge University Press

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

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

Selten, Reinhard 1978. The chain store paradox. *Theory and Decision*, 9, 127-58

Spence, M.A. 1974. *Market signaling: informational Transfer in Hiring and related screening processes*. Cambridge, Mass.: Harvard University Press

Spiegelman A. 1991. Maus. A survivor's tale. New York: Pantheon Books

Tadelis, Steven 1996. What's in a name? Reputation as a tradeable asset. Working paper, Dept of Economics, Harvard Univ.

Tajfel H. 1959. The anchoring effects of value in a scale of judgments. *British J. of Psych*., 50, 294-304.

Zahavi A. 1975. Mate selection-A selection for a Handicap. *J. of Theoretical Biology*, 53, 205-214