5 Under trusting eyes: the responsive nature of trust

Vittorio Pelligra

(Razumov) Have I by a single word, look, or gesture given him reason to suppose that I accepted his trust in me?


1. Introduction

In his *Under Western Eyes*, first published in 1909, Joseph Conrad tells the story of Razumov, a solitary but well-respected young student, and of his charismatic colleague Victor Victorovitch Haldin. While Razumov is considered by his fellow students as reserved but reliable, Haldin has been classified as restless and unstable even by the local authorities. Razumov has a good reputation: a man who is always willing to help others, even at personal cost. Knowing this, one day Haldin knocks desperately on his door. Razumov lets him in. Haldin looks distraught; he immediately confesses his secret to Razumov: it was he who was responsible, that morning, for the act of terrorism against the carriage of Mr de P., the feared and brutal president of the notorious ‘repressive commission’. Razumov is horrified, first for the gravity of what has happened, but also because, knowing Haldin’s secret, he too now is part of that tragic and dangerous conspiracy. Still shocked, he replies to Haldin’s confession: ‘But pardon me, Victor Victorovitch. We know each other so little. . . I don’t see why you. . . ’ ‘Trust,’ replies Haldin. ‘This word’ – Conrad tells us – ‘sealed Razumov’s lips as if a hand had been clapped on his mouth. His brain seethed with arguments’ (p. 5). These are brief excerpts from the novel’s prologue; the rest of the story narrates the details of Razumov’s reaction to having been so heavily trusted, his inner struggle and, finally, the consequences of his behaviour.

This story also introduces us to the so-called ‘problem of trust’, the main topic of this chapter. We develop a theoretical framework to reveal the reasons and feelings that such an episode probably elicited in Razumov, taking this episode as emblematic of a generic trusting relationship. How can we describe and analyse such a strategic interaction
between a trustor (Haldin) and a trustee (Razumov)? What motivational process does Haldin’s trust activate in Razumov? If we consider, together with John Stuart Mill, how trust ‘penetrates into every crevice and cranny of human life’ (1848, p. 131), it is not difficult to appreciate the relevance of these questions. And, in fact, the problem of trust has recently gained importance and centrality in many areas of the social sciences: economics (Dasgupta, 1988; Fehr and Gächter, 1998a; Harvey, 2002), sociology (Coleman, 1990; Misztal, 1996; Szompka, 1999), political science (Hardin, 2001; Braithwaite and Levi, 2002) and organisational sciences (Kramer and Tyler, 1995; Lane and Bachmann, 1998).

In the economic domain, trust is perceived as playing a crucial role in inter- and intra-organisational relationships, contract theory, labour economics, in the area of socio-economic development, and in the huge literature focused on social capital, to quote only a few relevant fields.

Several explanations have been developed to account for the ‘trust phenomenon’. Here, in particular, I focus on game theoretical explanations. Many strategies have been developed, some more conservative, others more heterodox (see Pelligrá, 2002), to account for trusting behaviours. Which of these strategies best explain the wider class of fiduciary relationships is eminently an empirical matter. Experimental data have recently been produced to test alternative theories. Evidence seems to support the principle of trust responsiveness as the most pristine explanation of trustful and trustworthy behaviour (Dufwenberg and Gneezy, 2000; Bacharach, Guerra and Zizzo, 2001; Pelligrá, 2003, 2004). This chapter focuses in particular on the understanding of the hypothesis of trust responsiveness, its functioning, historical roots and philosophical foundations.

2. ‘Telling a secret’ formalised!

Let us now describe in formal terms the interaction between Haldin and Razumov, and consider the game in extensive form, as depicted in figure 5.1a. Haldin first has the choice between telling Razumov his secret or keeping it to himself. If he decides to tell his secret, then Razumov will have the choice between remaining silent about it or telling the truth by reporting Haldin to the police, and eventually receiving a reward. The numbers represent, ordinally, the preferences of the two subjects. If Haldin says nothing to Razumov, we have a status quo (0,0). If he decides to share his secret and Razumov keeps it to himself, Haldin will somehow be relieved by not bearing the weight of his act alone, and therefore better off with respect to the status quo, whereas Razumov will be in the status quo (1,0). But, if Razumov decides to opt for the reward,
then Haldin will be sent to prison and Razumov will end up richer ($-100, 10$).

This interaction has three basic elements:

(i) **potential gain**: if the trustee fulfils the trustor's trust, the latter will end up better off;

(ii) **potential loss**: if the trustee is trusted and does not fulfil the trustor's trust, the latter will be worse off;

(iii) **temptation**: if the trustee is trusted, he will be materially better off by letting the trustor down.

In more general terms, the situation can be formalised as a ‘basic trust game’ (Figure 5.1b), in which the first element is described by imposing $c > a$, the second $b < a$, and the third $e > f$. It must be noted, however, that the ‘telling-a-secret game’ represents a particular instance of the basic trust game, since it is characterised by the additional condition $d = f$. This condition, together with the others, describes what may be defined as a ‘gratuitous trust game’ (Pelligra, 2003), which formalises a situation in which the trustee’s payoff from not having been trusted and the payoff from having repaid the trustor’s trust are equal.
relationship between Razumov and Haldin is one in which the latter’s trust is ‘gratuitous’ in the sense that the former’s trustworthiness does not yield mutual gains.

The central question now is: how would a rational Razumov and a rational Haldin behave in such a situation? Game theory provides an answer through the so-called Zermelo’s algorithm, which suggests to a rational Haldin to reason backwards before deciding: he should imagine what a rational Razumov would do if called upon to play, and then the consequences of their joint actions. Haldin’s reasoning goes backwards as follows: ‘If I tell him my secret, he will choose the reward (10) and I will be sent to prison, which is, from my viewpoint, the worst of all possible outcomes (−100). Therefore, “continues Haldin,” in keeping my secret, if not 1, at least I can get 0, which is better than −100.’ Thus, classical game theory suggests that player A should not play R (right), and at the same time that player B should play L (left).

But, if – contrary to this normative advice – we observe a pair of strategies such as (R,R) (i.e. share the secret, keep the secret), then common sense would say that A trusted B and B repaid his trust by behaving trustworthy.

Observationally, we define trustful and trustworthy behaviour as follows.

A strategy is trustful when:
1. in a situation that can be modelled as a trust game;
2. player A plays R.

Correspondingly, B’s behaviour is trustworthy when:
1. and (2) apply; and
2. player B plays R.

Was, then, Haldin irrational by playing R? He himself explains his reasons to Razumov: it is not a matter of irrationality, but of trust.

On the empirical side, the prediction that emerges from the backward induction argument is often falsified by robust experimental findings (Camerer, 2003, ch. 2; Ostrom, 2003) that show that people are more trustful and trustworthy than theory would suggest. How can we theoretically account for these anomalous behaviours? Let us find some good reasons for Haldin to behave as he actually did.

3. Tautologism

The most radical response to the divergence between theoretical prediction and observed behaviour implies the interpretation of the payoffs in the game as ex post indexes of player’s preferences – that is, as utility in the theory of revealed preferences: we attach the highest index to the
chosen action (Binmore, 1998). According to this view, having observed actual behaviour that seems to disconfirm the theory’s predictions only means that we are considering an ill-specified game and that, consequently, it should be re-described to fit the data correctly. As already mentioned, this is a radical position that tends to ‘throw the baby out with the bath water’, in the sense that, while, on the one hand, it saves the empirical validity of the theory, on the other it dramatically impairs its practical applicability.

4. **Enlightened self-interest**

A second, less radical though still conservative, reaction is based on the idea of *repetition of the interaction*. If the players know that there is a positive though small probability that there will be another round after that actually being played, they may be motivated to forgo an immediate gain in order to foster a long-standing and more remunerative relation. The main limitation of this theory is that most of the experimental evidence we have on trust has been produced using one-shot games, where long-term considerations do not apply.

Another explanatory strategy aims at developing models in which the players’ utility functions contain some additional element to account for behavioural principles other than self-interest. In some models, these additional factors are introduced into an extended utility function that the players aim at maximising in the usual way. These models may be defined as ‘forward-looking’, since, in fact, players’ actions are motivated only by their consequences. Other models develop new solution concepts based on an idea that an agent is actuated not only by the outcomes of his choices but also by the way such consequences are attained. These models may thus be defined as ‘procedural’.

5. **Altruism**

The first model I present was proposed to formalise the idea of altruism, which implies that the *ego*’s own welfare is directly affected by the *alter*’s welfare. In game theoretical terms, an altruist is defined as a subject whose utility increases (decreases) as the other’s payoff increases (decreases). Since this assumption is still self-centred, the variations are weighted in a way that *ego* attributes more importance to his own utility than to *alter*’s (Margolis, 1982). It is immediately clear that a subject who is sufficiently motivated by altruism would behave trustworthily in the trust game.

Was Haldin assuming that Razumov was an altruist?
6. **Inequity aversion**

A second principle that has recently been incorporated in game theoretical models to explain non-selfish behaviour is that of inequity aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). An agent is considered inequity-averse when he aims both at maximising his payoff and minimising the difference between his own payoff and those of other agents. The basic idea is twofold: first, people dislike being part of an unequal distribution of wealth; second, in such unequal distributions they dislike being in the disadvantageous position even more than in the advantageous one. In a trust game, trustworthiness is consistent with player B being motivated by distributional concerns. Should we now ask: was Halding counting on Razumov’s taste for equity when he decided to tell him his secret?

It is worth noting that the two previous models are based on a purely forward-looking logic, since player B is motivated exclusively by the consequences that his actions would produce and not at all by A’s choice. A’s trustful action, in fact, does not affect B’s motivations. In this respect, all the other theories described from now on are basically different.

7. **Team thinking**

The strategy underlying the two above theories was focused on a modification of the subject’s preference orderings by means of some additional behavioural principle – although, in the end, agents continue to be individually instrumental as they aim at reaching their most individually preferred outcome. On the contrary, theories of team thinking (Sugden, 1993b, 2000a; Bacharach, 1999) postulate a connection between preferences and action that is not only different from that usually assumed by the standard framework but also is by no means reducible to it. When an agent comes to perceive himself as a member of a team, his reasoning style switches to a mode that no longer responds to the question ‘What should I do optimally to attain my goal?’ but to ‘What should I do to play my part in accomplishing the team’s plan?’ From this perspective, a team thinker is considered to be rational when he chooses actions that are part of the team’s plan, whether they lead to optimal outcomes or not. For a team thinker, good outcomes are reasons for action by the team, but not reasons for action by individual agents; for individual agents, they are contingent consequences of good plans. Individual optimality is no longer a criterion by which to judge the rationality of a certain course of action.
While altruism and other theories affect the preference formation process, team thinking postulates a different way of satisfying the team’s preferences. Given this criterion of rationality, we can expect that, if in the trust game the two players A and B perceive themselves as belonging to the same group, and if we assume the team’s goal to be to gain as much as possible, joint strategy R,R will be preferred to R,L or L. Should we, then, be induced to think that Haldin was considering Razumov as a member of his own team?

8. **The motivating power of expectations**

Another strategy that may be followed to account for the trust phenomenon is based on the hypothesis that *alter’s* expectation of *ego’s* behaviour, in given circumstances, may have a motivating power that makes *ego* act in ways different from those suggested by mere material self-interest. Sugden (1998, 2000b) has developed a theoretical framework to analyse situations in which a norm determines the formation of subjects’ expectations. Subjects’ extended utility is broken down into two different parts: material and psychological. To fulfil or to frustrate *alter’s* expectations about *ego’s* choices causes an increase or a decrease in *ego’s* psychological utility. In Sugden’s model, *alter’s* expectations come from the existence of a norm that generally applies to that given situation, and that gives rise to generalities of behaviour. Which norm has to be applied to any given situation is not established a priori, but it is inductively inferred by the agents, from their past experiences.

The crucial assumption in this theory is the so-called ‘resentment hypothesis’. ‘Resentment’ is a feeling ‘which compounds disappointment at the frustration of one’s expectations with anger and hostility directed at the person who is frustrating (or has frustrated) them’. Similarly, ‘aversion’ stands for an emotion that is the negative of desire, triggered by a sense of fear or unease about being the focus of another person’s resentment (Sugden, 2000b, p. 113). It would not be difficult to show that the pair of trustful and trustworthy strategies constitutes an example of such an equilibrium. Did Haldin suspect that?

9. **Reciprocity**

Theories of reciprocity incorporate the idea that an agent would be willing to sacrifice part of his material wealth in order to be kind to someone who has been kind, or is expected to be, kind to him and to punish anyone who has been, or is expected to be, unkind to him. This
principle has been formalised in various ways. Here I focus on the well-known model of reciprocating fairness, proposed by Matthew Rabin (1993). In his model, payoffs depend not only on players’ actions, as in the traditional theory, but also on players’ intentions. Intentions can be formalised by considering not only what the players do but also what they could have chosen to do and did not. As a benchmark to assess the degree of kindness or unkindness incorporated in a given action, Rabin introduces the so-called equitable payoff: if player A expects player B to choose a strategy that leads A to a payoff larger than the equitable one, then B’s expected choice is kind; otherwise it is unkind. If B expects A to choose a similarly kind action and A knows B’s expectation and B knows A’s expectation, both psychological utilities increase, and if, in this way, each player’s overall utility (material and non-material) is maximised then there exists a fairness equilibrium, which allows two rational players to coordinate, for example, the Pareto superior cooperative outcome in a prisoner’s dilemma. The player’s psychological utility crucially depends on the degree of perceived kindness in the way that A’s being kind (unkind) to B when he expects A to be kind (unkind) positively contributes to both A’s and B’s utility; while mixed situations are a source of disutility. This type of reasoning implies that players form expectations not only about each other’s actions but also about each other’s expectations; these are called ‘second-order expectations’.

Let us now consider a trust game and presume that A expects B to play R, to be trustworthy. To be trustful is the only Pareto optimal strategy for A, so this choice conveys neither kindness nor unkindness to B. Let us now presume that B expects A to be trustful by playing R. Logically, replying trustworthily by playing R would be considered as kind, while playing opportunistic strategy L would be considered as unkind. But, in Rabin’s fairness equilibrium, we cannot have one player being kind and the other not. Therefore, although outcome R,R seems coherent with the logic of reciprocating fairness, it is not a formal implication of the model; it is not, in fact, a fairness equilibrium.

To apply Rabin’s model to a trust game, the model must be amended in various respects. Some of these emendations have recently been suggested by Daniel Hausman (1998). In particular, he substitutes the value of the payoff deriving from the predicted Nash equilibrium to the equitable payoff. The intuition behind this substitution is: ‘If you

---

1 Rabin’s model applies only to normal-form games, so we should re-describe the trust game in normal form and consider the players playing simultaneously. This can be done without substantially modifying the underlying argument.

2 Assuming that the game has a single equilibrium.
provide a benefit to me in playing your materially self-interested equilibrium strategy, then you are not being kind to me, and there is nothing unfair if I pursue my own material self-interest’ (p. 10). Genuine kindness needs intentionality. In this way, the expectation of reciprocal behaviour may well be considered as the rationale for trustful and trustworthy behaviour in the trust game. Would Haldin agree?

10. **Trust responsiveness**

While trust tends to appear at first glance as a candidly clear concept, its essential nature turns out to be rather elusive. One consequence is the proliferation of theoretical characterisations that the concept has received in past years. Trust has been defined as a *personality trait* (Baker, 1987; Jones, 1996), an eminently *probabilistic phenomenon* (Baier, 1986, 1994; Gambetta, 1988) or as a matter of *encapsulated interest* (Hardin, 1993, 2001). Among all these conceptions, however, the characterisation that best seems to account for the primary quality, the essential feature of the trust phenomenon, is the idea of trust as *responsive* behaviour (Horsburgh, 1960; Jussim, 1986; Pettit, 1995; Pelligra 2003).

The main idea of the responsive conception of trust is that trust is basically a matter of interpersonal relationships and that the relational factor should play a central part in its understanding. An act of trust takes place within an (often personalised) relationship between two subjects. It is extremely unlikely that a theory that considers the reasons to behave trustfully and trustworthy as external to that relationship will be able to give a satisfactory account of what trust is. Nevertheless, at least theories based on (enlightened) self-interest, altruism, inequity aversion and team thinking consider the reasons to be trustworthy as exogenous. This means that, at a given node of the interaction, whether or not alter decides to behave trustworthy does not depend on ego’s particular behaviour in previous nodes. A more satisfactory theory of trust should be able to account for the influences that alter’s observed choices exert on ego’s preferences and choices. In the trust responsiveness hypothesis, a trusting move induces trustworthiness through an endogenous modification of ego’s preference structure. A single act of genuine trust may provide *additional* reasons to behave trustworthy.

The remainder of this chapter is devoted to exploring the main features of this hypothesis.

Assume two subjects, A and B: when A behaves trustfully, he overtly manifests his expectations about B’s behaviour. The idea of trust responsiveness assumes that this manifestation induces in B a tendency to fulfil A’s expectations, even at some material cost. In this respect, trust is said
to be self-fulfilling. A similar process of inducement, defined as *therapeutic trust*, was first described by H. J. N. Horsburgh (1960), who claims that, in a situation similar to the trust game, ‘[O]ne of the reasons for [A’s] willingness to risk the loss of his money is a belief that this may induce [B] to act more honourably than he originally intended’ (p. 345).

Therapeutic trust is also defined as ‘reliance which aims at increasing the trustworthiness of the person in whom it is placed’ (p. 346). A peculiar aspect of ‘therapeutic trust’ is its purposiveness, namely that it ‘requires that the person trusted should be aware of the reasons for the trust which is placed in him’ (p. 346) and that the trustful action explicitly aims at increasing the trustee’s trustworthiness.

Other forms of ‘responsive trust’, less extreme or purposive, have been described in the philosophical literature. Philip Pettit (1995), for an explanation of the self-fulfilling nature of trust, has suggested the idea of *interactive reliance*. We observe interactive reliance when, relying on B, A thinks that his manifest reliance ‘will strengthen or reinforce [B’s] existing reasons to do that which [A] relies on him to do’ (p. 206). This is because A believes that, once his reliance has been manifested, ‘the utility [B] gets [from fulfilling A’s] increases with the recognition that [doing so] will serve [A’s] purposes’ (p. 206).

The rationale of this process is to be found, according to Pettit, both in exogenous traits displayed by the trustee and in an endogenous process of belief formation. The first element is related to broadly conceived self-interested factors (i.e. individual taste for loyalty, virtue and prudence). The second element is more interesting, as it is based on an original desire for the good opinion of the others. However, in Pettit’s interpretation this desire is not ‘a trait that many will be proud to acknowledge in themselves. . . .[I]t counts by most peoples’ lights, not as a desirable feature for which they need to strive, but rather a disposition – a neutral or even shameful disposition – that it is hard to shed’ (p. 203). The existence of this desire implies that, by manifesting him reliance, A implicitly manifests his belief that B is trustworthy; that belief represents for B a precious good that, however, can be enjoyed only by actually behaving trustworthily. The enjoyment derived from that good is nothing but the satisfaction deriving from having confirmed A’s good opinion.

Despite the centrality of such an interactive process to the understanding of trust and of many other relational phenomena, as far as I know, apart from Pettit’s proposal, it has received very little formal exploration.

At an intuitive level, many scholars emphasise the importance of something similar to trust responsiveness. Psychologist Jonathan Baron, for instance, suggests that ‘[f]ollowing the norm of trust has an effect on
both the beliefs and the norms of others. It creates a virtuous circle...if we act as if we expect the best from the others, they will often behave better as a result’ (Baron, 1998, p. 411). From the same psychological perspective, Lee Jussim (1986) describes and analyses ‘situations in which one person’s expectations about a second person lead the second person to act in ways that confirm the first person’s original expectation’ (p. 265).

Among economists, Partha Dasgupta (1988) notes how ‘the mere fact that someone has placed his trust in us makes us feel obligated, and this makes it harder to betray that trust’ (p. 54). Rabin himself suggests an integration of his model to take account of the fact that, in a sequential non-repeated game, if the first mover trusts the second, the latter may feel motivated to behave fairly even if he has the last move. ‘If player 1 plays ‘trust’, rather than ‘split’, he is showing he trusts player 2. If player 2 feels kindly towards player 1 as a result of this trust, then he might not grab all the profits’ (1993, p. 1297). A similar hypothesis was also advanced by David Good (1988), who claims: ‘There are probably many reasons why these relationships between trusting, being trustworthy, and psychological well-being exist, but the fact that trust at a basic personal level is psychologically rewarding is unsurprising’ (p. 33).

Surprisingly enough, despite having noted the phenomenon, most social scientists and, notably, economists remain silent about its origin and functioning. In the following sections I shall try to fill this gap.

11. **Relational motivation: the evolution of an idea**

As we have seen, one of the most complete accounts of trustworthiness, that provided by Pettit, is grounded on the agent’s desire to be well regarded by his peers. However, Pettit attaches to this desire, which he calls a ‘shameful disposition’ (1995, p. 202), a negative moral status. My position on this maintains, firstly, that – although crucial – this motive should not be considered the ultimate source of motivation and, secondly, that having a desire for the good opinion of others should not always be considered ‘shameful’. This is because there is a wide range of motives, going from vanity to the genuine desire of being praiseworthy, and all these motives are, to different degrees, related to others’ opinion. We cannot fully understand trust responsiveness without exploring these varieties.

In the following, I briefly explore the development of the idea of ‘self-love’, broadly conceived as ‘desire for the good opinion of others’. This is necessary to create a context and to isolate the historical roots of some of the elements that are crucial in explaining the mechanism of
trust responsiveness: in particular, that self-love is ‘relational’ and that self-love motivates us to be praiseworthy, not just eager for praise.

On the one hand, there is the position developed mainly by philosophers such as Thomas Hobbes and Bernard Mandeville, according to which the mere pursuit of self-love is necessarily a source of conflict; on the other hand, there are the champions of the idea of self-love as a source of love of others. There is also a third position, personified mainly by David Hume and Adam Smith, that aims at the reduction of this dichotomy through a purification of self-love’s moral status. Here, I focus mainly on the latter. Aristotle’s (see Aristotle, 1980) theory of sociality constitutes the starting point of Hume’s and Smith’s systems. Its pivotal element is the concept of philautia (literally, ‘self-love’). Philautia is the source of philia, or friendship, which, in Aristotle’s system, is the source of self-knowledge and self-consciousness. And it is by virtue of this mirroring process that, in a friendly relation, one has the opportunity to develop all one’s virtues. According to Aristotle, therefore, self-love represents the bedrock of one’s moral development and flourishing.

Hume’s position (1978/1739, 2001/1751) goes explicitly against both cold cynicism (Hobbes) and candid optimism (Francis Hutcheson, 1971/1725), as he maintains that behaviour should be explained through a mix of coexisting self-interest and benevolence. His crucial assumption is that of an innate sense of sympathy that generates the basic human propensity to sociality. It is through sympathy that we assess whether our actions are ‘useful’ or ‘pernicious’ for others. This assessment is, in turn, the basis for establishing a sense of justice and honesty. This is why virtue is considered ‘desirable on its own account, without fee and reward, merely for the immediate satisfaction which it conveys’ (apps. 1, 5). Man is generally moved by the desire for fame and reputation, a kind of desire that is not radically different from vanity. Here, Hume introduces a distinction, which was further clarified by Smith, between aiming at being praised and approved and aiming at being worthy of praise and approval. This shift is based on what Hume defines as the ‘reverberating’ or ‘reflective’ nature of sympathy (Treatise, para. 365), a quality that implies that we aim at gaining our own approval at least as much as we desire to obtain others’. The reverberating nature of sympathy leads us to internalise other people’s moral judgements and induces us to see ourselves as we appear to others, being pleased with our virtues and disliking our own vices, although they may be of great material benefit.

Smith builds his Theory of Moral Sentiments (1759/1976) around two main empirical assumptions. The first is that the most basic motive for
social action is the desire to be loved and approved: ‘There is a satisfaction in the consciousness of being beloved, which, to a person of delicacy and sensibility, is of more importance to happiness, than all the advantage which he can expect to derive from it’ (part 3, chap. 2, sect. 1).

But vanity and self-love alone are not enough to explain the genesis of social sentiments; Smith needs a second assumption that refers to individuals’ ‘separateness’. We do not have direct experience of what others feel, but we have the natural ability of ‘feeling with others’ – that is, of imagining ourselves as the subject of others’ experiences. To sympathise with others implies not to imagine what I would feel in a given situation but what the subject I am sympathising with would feel in that situation. This imaginative ability is the basis for our self-consciousness, produced by our natural inclination to view ourselves as others see us. Therefore, the ultimate consequence of reciprocal sympathy is that the subject becomes capable of self-reflection. The logic of self-evaluation develops through our imaginative process, which, fuelled by sympathy, leads us to imagine others’ reactions and sentiments in a given situation. Through such exercise we first imagine what they feel and then decide whether or not we would conform to them, in that particular situation. Conformity of sentiment suggests ‘approbation’, the contrary ‘disapprobation’. Once established, this ability for self-approbation and self-disapprobation is supplemented with a morally objective reference point that is provided by the ‘man within’, a ‘cool and impartial spectator’, who, according to the logic of sympathy, does not feel the agent’s emotion and sentiments to the same degree as the agent does, so that a certain degree of detachment between the agent and the spectator is reached, to guarantee the required degree of objectivity.

At this point we have an agent who, because of his innate sociality and sentiment of sympathy, is naturally inclined to self-evaluation. The impartiality, although not the absolute objectivity, of his judgement is guaranteed by the action of the impartial spectator. Therefore, our actions turn out to be motivated by the desire for others’ approbation, which we may represent to ourselves even in the physical absence of others. But what, at this stage, is the difference between the desire for praise and mere vanity? According to Smith’s original view, we are moved by something more than vanity as we aspire not only to be praised but also, and more fundamentally, to be praiseworthy. ‘Man naturally desires, not only to be loved, but to be lovely; or to be that thing which is the natural and proper object of love. . . He desires, not only praise, but praiseworthiness’ (part 3, chap. 2, sect. 1).

What emerges from the previous discussion is the centrality of the desire for being loved and approved, as a primary, if not the only, source
of motivation. We observe that at the end of this itinerary, from Aristotle’s philautia to Smith’s praiseworthiness, this concept emerges as morally purified from any residue of selfishness. According to Smith, this desire cannot be considered either selfish or self-centred, as it is based on the imaginative act of leaving one’s self in order to enter others’ contingencies. I think that it is precisely this imaginative leap that may be considered as the rationale for trustworthiness, as it is depicted in the trust responsiveness mechanism.

12. Self-reflection and trustworthiness

A subject’s ability with regard to self-reflection turns out to be the pivotal element in the functioning of the mechanism of trust responsiveness. This ability cannot be entirely generated internally, since it arises within the relationship in its mirroring function. The trust responsiveness hypothesis suggests that some of the reasons for being trustworthy derive from the mere fact of having been made an object of trust. In my interpretation, Smith’s idea of self-reflection constitutes the starting point for providing a potential justification for that proposition.3

Consider two subjects, A and B; let them interact in a situation like the trust game. Subject A moves first and, according to game theory, he should end the game there by being prudently distrustful. Presume, instead, that he decides to opt for the trustful strategy. Now game theory tells us that B, with certainty, will behave opportunistically. Presume further that, contrary to such a prediction, he decides to play trustworthily. We would observe a pair of trustful and trustworthy strategies. How can we rationalise the reasoning process that motivates B to resist the temptation of an opportunistic move?

First, assume that B is interested in the material payoff and, second, that he is also interested in A’s opinion. We know that subjects have an innate desire for the good opinion of others, and that they also have the ability, through imagination, of changing their point of view to that of others. Having observed A’s move, although he may not know the real reason, B conjectures that A has an expectation of a trustworthy response. In other words, he knows that, because he played trustfully, A must think B is trustworthy. Now, B knows that A believes him to be trustworthy.

3 It must be noted that Smith’s argument is only the starting point for my tentative rationalisation of trustworthy behaviour. With respect to his original construction, I go somewhat further by assuming that the desire for being praised and the desire for being worthy of praise are two distinct, albeit related, sources of motivation, and although this is a philologically sound interpretation of Smith it is still disputed. Elsewhere (Pelligra, 2003), I have provided a more detailed justification for this interpretation.
trustworthy. B, then, has two options: to confirm that expectation by behaving trustworthily, or to let A down by being opportunistic. In thinking about such options, B considers two different orders of judgements: first, he tries to imagine A’s reaction to both responses. Given his prior expectation, A will be satisfied with the former choice and frustrated by the latter. B knows that such reactions will have, respectively, a positive or a negative impact on A’s opinion of him and consequently on his ‘vanity’. That first order of judgements affects what we may understand as B’s desire for praise. But there is also a second order of judgements, that of our own self-appraisal. Nature has, in fact, made us desire not only to be praised but to deserve praise. This is what makes us concerned about what an impartial and well-informed spectator may think and feel about what we are doing, irrespective of the reactions of the actual spectators. I suggest that this second-order judgement affects B’s desire for praiseworthiness – that is, his self-esteem, and not his vanity, as in the previous case. At this point, then, B must balance the effect of the material gains and psychological losses attached to his available options: the material gain and the psychological loss from having been (materially) self-interested, or the material loss and psychological gain from having been trustworthy. What is important to note at this point is that the psychological impact has two distinct sources: it derives not only from the idea that A would form of B, following his actions, but also, and I think principally, from the idea that B himself would form of his own actions, as seen from the external standpoint of the impartial spectator.

The mechanism of self-evaluation works in two stages. In the first, the impartial spectator determines the balance between the actor’s (B) interest as agent and that of the trustor (A). In the second, the impartial spectator determines B’s conscience towards the action, according to what was determined in the first stage. This mechanism provides what may be termed internal psychological reasons for the action. At the same time, B perceives or imagines the kind of reaction that A may express towards B’s action. These are the external psychological reasons for B’s action. Thus, B is in relation with two subjects: the external spectator, trustor A, and the ‘man within’, Smith’s impartial spectator, and he is influenced by the reactions of both.4

In this sense, B’s action is determined by three orders of motives: first, his direct material interest; second, A’s approbation or disapprobation; third, his own self-approbation or disapprobation as derived from the perspective of the impartial spectator.

4 The former is real, the latter is only metaphorical.
This motivational structure is neither narrowly self-interested, because it takes into account A’s interests, nor narrowly self-centred, because it is partially determined by B’s expected reaction. That is, it is relational, because alongside the two self-centred and other-regarding motives there is also a third source of motivation – i.e. what derives from the internalised judgement of others, as described in the metaphor of the impartial spectator. According to this mechanism, our actions are assessed and determined by the consequences that they produce on us, both internally and externally. The internal reasons are those related to our sense of worth, the external reasons are those related to our vanity. These two sources of motivation account for the difference that exists between having a desire for conformity to other’s expectations because of the fear of others’ reactions and having the same desire because of intrinsic reasons related to one’s own sense of worth. Trust responsiveness is based on both. And this composite nature is able to explain why, for instance, we often observe trustworthiness even in anonymous interactions – that is, when A’s reaction cannot be observed directly. Consider, for instance, what happens when someone finds a lost wallet in a street and takes it to the nearest police station. Although this behaviour cannot be aimed at gaining praise from the owner of the wallet, it may be driven by a desire for praiseworthiness. In such a situation there is no room for vanity, but there is still room for the working of the internal reasons, which may, in fact, counterbalance the subject’s material interest – that is, his desire to keep the wallet. In such a situation, B does not directly perceive A’s reaction, but he may still be able to imagine it. By the same means, B may imagine A’s reaction even before materially having done a certain action.

Thus, the actual choice is the compound effect of material self-interest, others’ approbation or disapprobation, and a personal sense of self-worthiness. Self-worthiness is a second-order construct, deriving from the good opinion of one’s self, which the self, in turn, derives from the impartial spectator’s approval.

13. How trust responsiveness is different

Having considered in the previous sections several alternative theories of trust and trustworthiness and having described the basic elements of the trust responsiveness hypothesis, it is now natural to ask where the peculiarities of the latter theory lie, and how they render it different from the others.

As far as the theories of altruism and inequity aversion are concerned, the main divergence can be found in their being eminently
forward-looking. In these theories, subjects aim at achieving their most preferred, although not narrowly self-interested, outcome. It is as if they were maximising a social welfare function. Agents are by no means responsive to others’ behaviour, and their actions or preference orderings are unaffected by interactions with others.

If we consider theories of team thinking, we realise that they differ radically from trust responsiveness in the kind of rationality they imply. For a team thinker, a rational action finds its explanation in the fact that it is functional to the plan of the team. Thus, to play trustworthily in a trust game is rational only if we first assume that the outcome R,R is beneficial for the team, which is not always the case.\(^5\) Within the framework designed by trust responsiveness, being trustworthy may be rational even though the agent, from a purely material perspective, would prefer the outcome to be achieved by opportunistic play L,L. Team thinkers are rational but, because of their team preferences, the two types of rationality cannot be compared.

Yet the idea of normative expectations differs from that of trust responsiveness in the sense that, while the trustee actually responds to the trustor’s expectations, those expectations are grounded on a convention shared by the population as a whole. The trustee’s behaviour does not depend on the character of the relationship in which he is interacting.

Lastly, the idea of reciprocity, as embodied in Rabin’s model, is the most similar to that of trust responsiveness. Both reciprocity and trust responsiveness are principles of norm-guided behaviour. These norms, because of their effect on subjects’ psychological utility, may, in certain conditions, offset the effect of the material payoff. Although both principles may lead subjects to act in ways that appear to be contrary to their material self-interest, they differ regarding the mechanism that elicits individuals’ psychological utility. In Rabin’s theory, the trustee’s motivation is a response to the trustor’s intentions, but in the sense that the trustee assumes the position of a judge, evaluating the trustor’s action and deciding whether to reward or punish it. The trustee considers the trustor as forward-looking, infers his degree of altruism or selfishness from his intentions, and then decides to react consequently. It is noteworthy that the degree of altruism or selfishness is inferred by comparing the outcome that the expected action will lead to with the equitable payoff. In this sense, therefore, the perceived kindness (unkindness) is

\(^5\) The outcome (R,R) may well be mutually beneficial for subjects, but that is not a strict implication of the conditions defining a trust game. The ‘gratuitous trust game’, for example, is an instance of a trust game in which the cooperative outcome is beneficial for the trustor but not for the trustee.
a measure of material gain (loss). The idea of reciprocity is ultimately based on the joint effects of material and psychological incentives. The perceived kindness that elicits reciprocal behaviour is a measure of the material benefit that one agent’s choice attributes to another. Instead, in trust responsiveness the potential material benefit plays no role in motivating the trustee.

Relative to Pettit’s (1995) account, one of the differences between his trust mechanism and the idea of trust responsiveness is that Pettit grounds trustworthiness ultimately on ‘a shameful disposition’, self-love or vanity. Trust responsiveness, as I have described it, has two main sources of motivation. The first is the desire for the good opinion of others, or ‘vanity’, which maintains the negative moral flavour Pettit assigns to it, but there is also a second source of motivation, which derives from our sense of self-worth. Following Smith’s Theory of Moral Sentiments in this, I think that the latter, ‘the tribunal of [our] own consciences’, must be considered as the ultimate ‘judge and arbiter of [our] conduct’. It is perfectly understandable that we follow our own desire for praiseworthiness even when there is no possibility of being praised; it would be more problematic, at best pathological, if we were merely to follow others’ praise with no respect for our own sense of worth.

14. Trustworthiness as relational good

From the above, it emerges that the nature of trustworthiness is neither self-centred, because it is based on others’ imagined reactions, nor completely other-regarding, because it is not based exclusively on others’ opinion. I think the best way of describing the nature of trustworthiness is by using the term ‘relational’. Trustworthiness comes about as a product of the action of self-reflection, which, in turn, arises from the relation with others – others as a mirror of the self. There then arises the similarity between such a conception of trustworthiness and the idea of ‘relational good’.

A ‘relational good’ is, essentially, a kind of good that is produced and consumed within a specific relation. On the production side, this good emerges thanks to a technology that is embedded in ‘encounters’ (Gui, 2000a, chapter 2 in this book), in which the motivations of the interactors are considered the essential, although not unique, inputs in the production process. If we consider trust as a three-way relation – A trusts B to do C – trustworthiness can be manifested only within a relation. I cannot reveal that I am trustworthy if I am not trusted. Furthermore, trustworthiness, as the trust responsiveness hypothesis considers it,
emerges because I have been trusted. While one may have a certain a
priori degree of trustworthiness, it is in the trusting relation that trust-
worthiness is elicited. In this sense, trustworthiness is a genuine product
of the trusting relation. Because it is grounded in our desire for self-
approval, which in turn is based on our sense of sociality, trustworthi-
ness can be neither produced nor enjoyed alone, in isolation from a
trusting relation. Trustworthiness, ontologically, needs to be externally
recognised.

15. Concluding remarks: what happened to Haldin?

This chapter has discussed the phenomenon of trust and some of the
theories that have been developed to explain trustful and trustworthy
behaviour. In particular, a novel explanatory principle is proposed,
grounded on the relational nature of any trusting interaction. The hy-
pothesis of trust responsiveness is unravelled in its philosophical and
historical roots, functioning, and underlying psychological structure.

Our discussion began, however, with the story of Haldin and
Razumov, to the point at which Razumov’s trustworthiness was being
tested. What happened next? Did Razumov turn out to be trustworthy,
or did he betray Haldin? After having explored the cunning of trust, it
would be interesting now to reread that story under trusting eyes and ask
ourselves what we would have done in Razumov’s place. Although it is
impossible to provide a generally valid answer to the question, we may
say that a genuine and somehow desperate act of trust such as Haldin’s
necessarily produces a sense of obligation to maintain the secret. Razu-
mov’s actual behaviour, however, emerges as the result of a balance
between the material gain from having helped the notorious oppressive
regime and the psychological losses due to his betrayal of a desperate
friend. In Conrad’s book, Razumov cannot foresee that cost and decides
to report Haldin to the police. The rest of the novel tells us how
Razumov spends the rest of his life repenting his untrustworthy behav-
iour. Through literary fiction, what emerges clearly is the nature of the
psychological cost that most of us wish to avoid by being trustworthy
when someone trusts us. The very essence of that cost derives from our

6 We know that trustworthiness may arise from a process of self-valuation that is triggered
by the relationship with others. At the centre of this process there is a desire for being
praiseworthy and, therefore, for being pleased with oneself. It is interesting to note that
the Italian word for ‘to be pleased’, compiacere, comes from the Latin cum placere, which
may be translated as ‘to be pleased together’. This provides etymological support for the
idea that our own sense of worth emerges, at least partially, as a reflex of others’
judgement.
own social nature, our own need for mutual recognition. Our identity is shaped in a social environment and we acquire a good amount of self-knowledge through the mirroring effect of others. I am not completely sure that ‘[a] man’s real life’ – as Conrad suggests – ‘is that accorded to him in the thoughts of other men by reason of respect or natural love’ (p. 21). However, what I hope I have shown is that, if economics aims at providing a descriptively adequate picture of economic interactions, we cannot avoid understanding concepts such as trust, and that trust cannot be fully understood if the role of our relational egos continues to be underestimated.

ACKNOWLEDGEMENTS

The ideas expressed here owe a great deal to constant and fruitful exchanges with Michael Bacharach, Luigino Bruni and Benedetto Gui, and especially Shaun Hargreaves Heap and Robert Sugden. This chapter would not have been as it is without their stimulating and attentive support. Philip Pettit provided useful comments on a previous version of the work. The usual attribution of responsibility applies.