Adaptive Human Behavior and Physiology (2015) 1:325–340
DOI 10.1007/s40750-015-0026-4

ORIGINAL ARTICLE

A Model of Smiling as a Costly Signal of Cooperation Opportunities

Samuele Centorrino[1] · **Elodie Djemai**[2] · **Astrid Hopfensitz**[4] · **Manfred Milinski**[3] · **Paul Seabright**[4]

Received: 28 November 2014 / Revised: 6 April 2015 / Accepted: 9 April 2015 /
Published online: 23 April 2015

Abstract We develop a theoretical model under which "genuine" or "convincing" smiling is a costly signal that has evolved to induce cooperation in situations requiring mutual trust. Prior to a trust interaction involving a decision by a sender to send money to a recipient, the recipient can emit a signal to induce the sender to trust them. The signal takes the form of a smile that may be perceived as more or less convincing, and that can be made more convincing with the investment of greater effort. Individuals differ in their degree of altruism and in their tendency to display reciprocity. The model generates three testable predictions. First, the perceived quality of the recipient's smile is increasing in the size of the stake. Secondly, the amount sent by the sender is increasing in the perceived quality of the recipient's smile. Thirdly, the expected gain to senders from sending money to the recipient is increasing in the perceived quality of the recipient's smile.

Keywords Smiling · Costly signaling · Experiment · Trust game

JEL Classifications D03 · D85 · D87 · Z13

Introduction

> The man who indulges us in this natural passion, who invites us into his heart, who, as it were, sets open the gates of his breast to us, seems to exercise a species of hospitality more delightful than any other. No man, who is in ordinary good

✉ Paul Seabright
Paul.Seabright@tse-fr.eu

[1] Economics Department, Stony Brook University, Stony Brook, NY, USA

[2] PSL, Université Paris-Dauphine, LEDa, UMR DIAL, Paris, France

[3] Max Planck Institute of Evolutionary Biology, Plön, Germany

[4] Toulouse School of Economics, Institute for Advanced Study in Toulouse, Toulouse, France

temper, can fail of pleasing, if he has the courage to utter his real sentiments as he feels them, and because he feels them.

Adam Smith—The Theory of Moral Sentiments[1]

This paper develops a model of smiling as a form of signaling behavior whose purpose is to facilitate economic exchange in situations requiring mutual trust. Smiling is a form of behavior that exists in all human societies (see Darwin 1872; Ekman 1982; Niedenthal et al. 2010). It appears to be more elaborate and more central to communication in humans than in any other species, and to play an important part in judgments of individuals about the character and general trustworthiness of others. Yet there is no scientific consensus as to why smiling has evolved to be like this, nor about what it is in smiling that makes it an appropriate basis for judgments of others. Here we develop the idea that in early human history, when there were few formal institutions to regulate social actions and behaviors, human beings needed to find signals that could reliably help them in determining whom to trust and when to do so. In these circumstances individuals might have focused on the facial expressions of others, particularly the expressions of the eyes and mouth.

There is, nevertheless, consensus about a number of the characteristics of smiling behavior. First, viewers perceive smiles as varying in their degree of "genuineness" or "convincingness". Since the work of Duchenne (1862) and Darwin (1872) in the 19th century it has been known that smiles perceived as genuine (known as enjoyment or "Duchenne" smiles) are characterized by use of the *orbicularis oculi* (which surrounds the eyes) in combination with the *zygomatic major* (which raises the corners of the mouth).[2] Honest smiling which involves a contraction of the *orbicularis oculi* further has the obvious cost of reducing the smiler's visual field, which both reduces informational inputs and makes the smiler potentially more vulnerable to enemies and predators. Doing so would have implied fitness costs for the smiler during our evolutionary past, and thus should be undertaken only sparingly, which explains why smiling convincingly may be costly to the smiler.

Second, Duchenne smiles are difficult to fake and are not under straightforward voluntary control. Duchenne smiles are easier to produce in certain affecting states, including when the person is in a positive and sharing mood (Mehu et al. 2007). Third, smiles induce mimicry, both in the sense that individuals viewing smiles by others have an increased tendency to smile themselves (Niedenthal et al. 2010), and in the sense that individuals trying to make a good impression on others (as when posing for photographs) make an effort to smile well. Smiling seems to be a form of communication. But if so, what do people reveal when smiling, and why have we evolved to adopt and interpret a form of communication behavior that is under such imperfect voluntary control?

In this paper we set out a framework in which smiling is modeled as a form of costly communication (costly in a sense we make precise below) that induces cooperation between individuals in situations requiring mutual trust. According to this view, the

[1] Smith (2000), p.497.

[2] Other important characteristics of Duchenne smiles are symmetry and temporal dynamics such as smile onset, apex, and offset durations for perceived genuineness (Krumhuber et al. 2007).

necessary costliness of smiling is precisely the reason why it is under such imperfect conscious control. In a nutshell, smiling may be costly because otherwise it would be easy to fake, and would not reliably be associated with trustworthiness. This does not necessarily imply that smiling has evolved to be difficult to fake in order to act as a signal, but more plausibly that natural selection has recruited as a signal a form of behavior that is already difficult to fake.

This hypothesis is not original to us (it was advanced by Owren and Bachorowski 2001), but to our knowledge it has not previously been formalized in a way that would make it capable of being subjected to a comprehensive experimental test. In the biological and economic literature on signaling, a costly signal is a detectable trait (which may be physiological or behavioral) that in itself has direct fitness costs[3] but conveys indirect advantages because it is correlated with an underlying invisible trait which has fitness consequences for another interested individual (such as a mating partner, a rival or a potential cooperator) whose behavior might be influenced by the signal. The nature of that correlation could be that individuals with the underlying invisible trait feel the signal to be less subjectively costly to emit than do the individuals without the invisible trait, or it could simply be that those with the invisible trait are more likely to possess the visible trait: it does not have to feel psychologically costly to be adaptively costly. What matters is the correlation of the invisible and visible traits and the fact that the visible trait considered purely by itself has adaptive costs.

As will become clear in what follows, whether or not a trait is a costly signal cannot be simply observed directly. It requires verifying independently four different elements: the fact that there is a cost to the sender; the fact that the cost is greater for senders who do not bear the trait; the resulting correlation of the signal with the trait that has fitness consequences to the receiver; and the causal efficacy of the signal in inducing the receiver to behave in a way that brings benefits to the sender to compensate for the fitness costs (see Számadó 2012). We set out precisely how these four elements might interact and what kind of empirical test might be possible for them. We also acknowledge that the empirical evidence for smiling to be a costly signal in social exchanges is weak and, to the best our knowledge, there is no existing direct evidence that smiling is differentially costly; rather, this differential cost is one possible reason for the correlation between the signal and the trait. There are certainly alternative possible theories of smiling: it might, for example, be a pure coordination device or as a simple expression of an inner emotional state (see Schmidt and Cohn 2001). Among other things, our model serves the purpose of making the theory of smiling as a costly signal testable against these alternative views. It also clarifies that being difficult to fake is not a sufficient condition for something to be a costly signal (many aspects of our physiology are difficult to fake); it also has to be reliably correlated with the appropriate adaptive trait.

It is worth explaining briefly the purpose of a formal model in an argument such as ours. A model should be considered as a simplified representation that helps us understand a complex empirical phenomenon, in much the same way that a map is a

[3] The economic literature on costly signalling only rarely considers fitness costs explicitly, and costly signalling may occur when costs are psychologically costly regardless of whether or not these are, in current environments, adaptively costly. Nevertheless we retain this terminology since the best explanation for why certain behaviors are psychologically costly is that they were adaptively costly in evolutionary environments.

simplified representation that helps us to navigate a complex terrain. The purpose of this model is to help us understand why individuals trying to trust each other might use a signal that is cognitively costly. In order to explain this we have to suppose that individuals differ in their degree of trustworthiness, so we use a very simplified representation of this difference in order to show why a costly signal could be a useful thing to transmit and to interpret. Our model assumes a certain set of payoff functions (and does not seek to show how these payoff functions evolved); conditional on these payoff functions it shows that the use of the signal would be rational (that is, fitness-enhancing) for the individuals concerned. It abstracts away from other aspects of individual payoffs that, while doubtless present in reality, do not have to be represented in the model in order for it to illustrate the functioning of the signaling mechanism.

In Section 2 we briefly review the relevant literature. Section 3 describes our model in outline. Section 4 derives the equilibrium behavior of the players in the trust interaction, as a function of the signals already sent and received. Section 5 analyzes the signals sent. Section 6 discusses empirical implications that have been tested by us in a companion paper (Centorrino et al. 2015). Section 7 concludes.

Literature Review

Costly signaling has been widely studied since the work of Spence (1973) in economics and Zahavi (1975) in biology. A signal is any observable trait that is costly to bear for the sender but which reliably indicates the presence of some advantageous hidden trait. While the cost can be a pecuniary or non-pecuniary effort cost in economics, or a fitness cost in biology, the benefit from signaling the hidden trait is that it attracts partners in mating or in some other mutually beneficial cooperative activity. The benefit to the signaler of doing so must exceed the cost of the signal, and the partner must also benefit from responding to the signal. The correlation of the signal and the hidden trait comes about because the cost of sending the signal is greater for those individuals that do not possess the advantageous hidden trait than for those who do (Grafen 1990); in economics this is known as the "single-crossing property". Our use below of the assumption that the cost is greater for those who do not bear the trait corresponds therefore to standard practice in this literature. In economic exchange, the hidden trait signaled by smiling could be any characteristic of the smiler (such as her degree of altruism or tendency to display reciprocity as in Gintis et al. 2003), or a characteristic of the situation in which the smiler finds herself (such as the size of the stake to be shared, as in the context of this paper).

Apart from the paper of Owren and Bachorowski (2001), which suggested the hypothesis, there has not been to our knowledge a significant application of the costly signaling approach to understanding smiling. However, smiles have been considered as a coordination device in Manzini et al. (2009), and more broadly, recent studies in economics and psychology have investigated the importance of emotions in games. Considering that emotions are not just some random noise but essential in the decision making process (Damasio 1994), theoretical and experimental work has investigated how different emotions enter into decision making (Elster 1998; Loewenstein 2000; Kahneman 2003; Frijda et al. 2004). While the focus has been mostly on negative social emotions as anger and guilt (Bosman and van Winden 2002; Sanfey et al. 2003;

de Quervain et al. 2004; Hopfensitz and Reuben 2009), increasing attention has been given to the use of rewards and the experience of happiness (e.g., Kahneman et al. 1999; Frey 2008; Frey and Neckermann 2009).

Altruism and cheater detection in social dilemmas have been investigated both in economics and biology (Cosmides and Tooby 1992; Gintis et al. 2001). If signals that can be used to identify altruists can quickly be imitated by non-altruists, they are not reliable signals anymore (Fehr and Fischbacher 2005). It has been shown that observable altruism can serve as a reliable signal of trustworthiness in Smith and Bliege Bird (2000), Gintis et al. (2001) and Lotem et al. (2003). However, in many situations, altruism cannot be observed, and more broadly the behavior of the interaction partner is unobserved. To detect whether an interaction partner can be trusted we can either rely on third party information regarding the target individual's reputation (Sommerfeld et al. 2008) or use visual signals concerning the individual's character (Frank 1988). In the absence of third party information reputation requires a track record, which is not possible in one-shot interactions. In order to detect trustworthy partners with some degree of reliability in these circumstances, it is necessary to make use of the verbal or non-verbal signals sent by the partner.

Brown and Moore (2002) stress that honest signals with a reliable emotional basis may be needed to guarantee the positive intentions of a counterpart. This leads to the importance of 'emotional expressivity' i.e., the ability to accurately communicate your internal feeling state (Boone and Buck 2003). To be reliable, these signals must be costly and therefore difficult to mimic. Smiles, and especially honest smiles, might be just that. Brown et al. (2003) were the first to observe that videos from self-reported altruists are rated differently by neutral observers than videos of non-altruists. Further, an analysis of video recordings from altruists and non-altruists showed that self-reported altruists showed more *orbicularis oculi* activity and more symmetric smiles (see also Oda et al. 2009). However, the existing empirical evidence in favor of the view that smiles reliably indicate altruistic intent is at best mixed. This is true when players are directly confronted with their partners in dyadic or triadic conversations or are shown pictures and sliced clips of their opponent.

Several different types of interaction have been used to study this question. The preferred setting appears to be the Prisoner's Dilemma game. In this context, Vogt et al. (2013) report an experiment in which they use short video clips of subjects. These 'thin sliced' clips were recorded while participants discussed a variety of topics not related to the game itself. Vogt et al. (2013) find that other informed experimental subjects were not able to make use of these clips to infer trustworthiness. Manson et al. (2013) provide evidence that defectors cannot be detected with better-than-chance accuracy either by their game partners or by informed or uninformed external observers. Lyons and Aitken (2008) show that the frequency of smiles does not significantly affect the ability of a player to identify a cooperative partner; and their duration is negatively correlated with the prediction of cooperation. In contrast to these three studies, Johnston et al. (2010) use video clips and test cooperation on the basis of comparison of enjoyment and non-enjoyment smiles. In their experiment, participants evaluated individuals displaying enjoyment smiles more positively than those displaying non-enjoyment smiles and had higher rates of cooperation with those displaying enjoyment smiles. Finally, Reed et al. (2012) report that players are able to use facial actions as credible signals of cooperative intent in dyadic interactions.

An alternative setting for studying this question is the ultimatum game. Schug et al. (2010) find that individuals who are prone to cooperate as proposers are more emotionally expressive when facing unfair treatment by others than those who do not, including in the tendency to emit Duchenne as opposed to non-Duchenne smiles. Mussel et al. (2014) instead study the reaction of the receiver when she is shown a still picture of the proposer before deciding whether to accept or reject the offer. They find that when proposals are accompanied by smiles this reduces the negative emotions felt by recipients in the face of disappointing offers and increases the probability of acceptance for a given offer.

Finally, the trust game (Berg et al. 1995) has been used by Scharlemann et al. (2001) and in our companion paper (Centorrino et al. 2015). The trust game differs from the Prisoners' Dilemma in two ways: first, the two players play sequentially and the decision of the second is conditional on the decision of the first. Secondly, because of this sequential structure, non-cooperation by the first player is not a dominant strategy, unlike in the Prisoner's Dilemma where non-cooperation is dominant for both. On the contrary, the first player has an incentive to cooperate if she believes that the second will do so, notwithstanding the dominant nature of the second player's non-cooperation strategy.

Scharlemann et al. (2001) use still pictures and find that smiles positively affect trust among strangers. Facial features can also affect cooperation. Centorrino et al. (2015) use instead video clips recorded by trustees in which they incite their partner to play with them the trust game. Trustors decide whether to send the game stake after watching these videos. We find that Duchenne smiles are causally effective in inducing the target to cooperate with the trustee.

The variety of settings used to elicit smiles as a signal of cooperation explain part of the difference in the current experimental literature. However, the informational feature of the game seems also to play an important role. In Lyons and Aitken (2008), Manson et al. (2013), and Vogt et al. (2013), whether in dyadic conversations or in recorded video-clips, players were not informed about the game they were going to play or were explicitly asked not to discuss it. In contrast, the studies by Johnston et al. (2010) and Reed et al. (2012) give full information to participants before they interact with their opponent. This seems to point towards the fact that smiles might signal different things in different contexts and that the information about the task is essential in order to produce the "right" smile. This view is also consistent with the result by Mehu et al. (2007), that honest smiles are produced when the person feels in a particular sharing mood.

The study by Mussel et al. (2014) shows an impact of smiles on the behavior of those who observe them, but (as the title "Smiling faces, sometimes they don't tell the truth" implies) is skeptical of the value of information conveyed by a smile. However, the authors do not test for that value. In their experiment, smiling and non-smiling faces are in fact randomly associated to each offer made by the proposer. The focus is therefore on the psychology of the receiver, who does not seem to adjust her expectations upward when seeing a smiling face, but instead becomes more willing to accept an unfair offer. Smiling, in this case, plays the role of easing the disappointment of the receiver after an unfair behavior from the proposer. Furthermore, a one-shot ultimatum game like the one designed by these authors seems unfit to test for cooperation. The ultimatum game is based on a take-it-or-leave-it offer, which seems more suited for eliciting fairness or inequality aversion.

The existing empirical results only confirm that there may be a plethora of ways in which smiles can be used as a communication device in strategic interactions. This partly depends on the task used and partly on its informational features.

In the present paper, we have decided to use the trust game as framework for our model. The advantages of the trust game are several. First of all, to the best of our knowledge, the existing experimental literature has found that smiles in the trust game can effectively induce cooperation. Secondly, in the trust game, the role of the sender and the receiver of the signal are very well defined. This allows to test whether smiles act through the psychology of the sender or of the receiver or both. This is an impossible task in a simultaneous game, such as the Prisoner's Dilemma, where both players play the role of both sender and receiver. Moreover, unlike in the ultimatum game (which is also sequential and therefore distinguishes the roles of sender and receiver), the trustee also makes a decision and it is therefore possible to test whether the signal of cooperation truthfully reveals the action of the trustee.

The model we develop hypothesizes that trustees may be motivated, to a greater or lesser degree, both by reciprocity and by altruism. In keeping with our approach of using the simplest model necessary to illustrate the mechanism, we do not need to hypothesize that senders are altruistic, which would complicate the algebra for no additional insight. There is a large literature addressing ways of incorporating social preferences in individual utility functions (see Sobel 2005, for a survey). It is safe to say that there is no consensus as to the appropriate way of modelling such motivations, and it is emphatically not our intention to propose a general theory here. For instance, in many models of behavior in public goods games, individuals are considered to be motivated either by reciprocity or by altruism but not both (Fehr et al. 2003); this is a useful device for focusing on the distinction between unconditional contributors and conditional contributors. Other papers (Hwang and Bowles 2010; Brülhart and Usunier 2004) hypothesize that individuals may have both motivations simultaneously to different degrees, and that is the approach we adopt here. This is a plausible and parsimonious way to capture the phenomenon, clearly present in many experimental studies including our own (Centorrino et al. 2015), that individuals vary in their degree of trustworthiness. It is not just that some are trustworthy while others are not, but also that among individuals who are trustworthy, some are more generously or fully so than others. The combination of reciprocity and altruism in our model captures this difference, but we make no claim that it is the only modeling strategy that would do so.

Outline of the Model

In our model players engage in a trust interaction. For ease of exposition, we give this the rather specific structure of an experimental trust game, but it is also a schematic representation of a much wider range of circumstances in which individuals take a chance on the trustworthiness of others in the hope of a subsequent gain from cooperation. They are able to engage in a signaling interaction before they do so. Thus, although our model represents a rather particular type of interaction between the players, the general conclusion of the analysis applies to a much more general class of economic situations, in which the parties interact without expectation of an extended relationship, one of the parties must make a commitment before the other, and the other will therefore try to signal trustworthiness in order to induce that commitment to be made.

There are two players, A and B. To avoid confusion we shall refer to A as "he" and to B as "she", though there is no intrinsic gender difference in the roles.

Player A receives a stake of value s and must decide whether or not to send it to player B. For pedagogical purposes we shall compare two situations: one with $s = 1$ and the other with $s = 2$; once again this is for expository clarity but more general cases could be considered without changing the underlying intuition. If the stake is sent it is multiplied by three, and player B may choose to (i) keep the new enlarged stake for herself; (ii) send back to A the original stake and keep twice the stake; (iii) divide the new enlarged stake equally between herself and player A.

There is nothing special about the number three except that it is significantly greater than two, indicating that if the parties are willing to trust each other they can each gain significantly more than the original stake. The analysis below could be undertaken for any multiplicative factor greater than two without affecting the qualitative results, but we use the number three both to keep the exposition intuitive and because this corresponds to the factor that has typically been used in experimental settings, including that in our companion paper.

Player A's decision will be influenced by his beliefs about player B along two dimensions – how much player B cares about strong reciprocity,[4] and how altruistic she is (we make these terms precise below). With respect to strong reciprocity, Player B may be one of two types $\theta \in (L,H)$; for simplicity we assume there are equal proportions of the two types in the population, though nothing of importance turns on this. H types have stronger preferences for reciprocity than L types (we can call these High Reciprocators and Low Reciprocators respectively). With respect to altruism, player B has a component of her utility that is a stochastic function of the amount she sends back to A. Player B knows her own type at the start of the game, and notably when she makes a video clip in order to persuade player A to send her his stake.

If player A sends the stake, player B must decide to send back to player A a multiple m of the original stake. In principle that multiple could be chosen from a continuous interval, but to aid intuition we are interested in the choice between three types of reply, which we can call "selfish", "reciprocating" and "generous", and which we represent by $m \in (0,1,1.5)$. Since the stake has been multiplied by 3, this means player B has a choice between keeping all the stake (the selfish strategy), keeping two-thirds of it (the reciprocating strategy), and keeping half of it (the generous strategy). Note that since the generous strategy involves the parties splitting the gains equally, it could be motivated by a desire for equality rather than altruism; the latter is the motivation we shall employ in our model. For our purposes nothing of importance turns on this point, though in other contexts it might matter which of these motivations was at work.

Prior to this interaction, player B communicates with player A, sending him a costly signal which can take the form of a smile. Then A forms a belief about B's type based on the signal. If A chooses not to send the stake the game ends, A keeps the stake and B receives a zero monetary payoff (and a total payoff that may include a cost of effort involved in sending the signal). If A chooses to send the stake then B finally chooses what multiple of the stake to return to A, and the game ends.

[4] In games with costly punishment, strong reciprocators are also willing to engage in such punishment, but that is not at issue here.

As is standard we solve the game backwards from the end, finding a perfect Bayesian equilibrium.

Moves in the Trust Game

Player B's Move

We model player B's motivation for returning a multiple of A's original stake using a random utility function. It is separable in money and in two types of social preference. The first social preference is for strong reciprocity, which we model as a fixed utility derived from sending back at least the original stake to player A, but not otherwise varying according to the amount sent. This utility, which differs between types, is given by α_θ, where $1>\alpha_H>\alpha_L>0.5$.

The second motivation is altruism, which is increasing in the amount sent back by B to A (it can be thought of as reflecting B's pleasure at knowing that she is increasing A's payoff). We model this as a utility that is a multiple β of the amount returned, plus a random error term ε. The coefficient β is itself random and may be greater or less than one (capturing the fact that, of players who return at least some money, some return only the original stake while others return a larger amount). Specifically, $\beta \in \{0.5, 1.5\}$ with probability $(1-p_\theta, p_\theta)$. This is without loss of generality. Every value of β greater than 1 would imply that player B derives positive net utility by choosing the generous option. Also, every value of β lower than 1 implies that player B suffers a net loss when transferring the stake to A. She will thus choose the selfish or the reciprocating option, depending on other parameters of the game. We assume that $p_H>p_L$ to reflect the fact that individuals with a greater propensity for reciprocity are also likely to be more altruistic.

We therefore model player B's utility function as follows:

$$U_B = 3s - ms + \alpha_\theta + \beta ms + \varepsilon \quad \text{if} m > 0 \tag{1}$$

$$U_B = 3s \quad \text{if} m = 0 \tag{2}$$

where s is the amount sent, m is the amount returned, the error term ε has a zero mean, and is uniformly distributed between −0.5 and +0.5.

As noted above, if β=1.5, player B will always choose m=1.5, since his utility is always strictly increasing in m. Thus either type of player will choose m=1.5 with probability p_θ.

If β=0.5 on the other hand, player B's utility is strictly decreasing in m once m is positive. Thus B will either choose m=0 or m=1. The probability of choosing m=1 is therefore the probability that:

$$\alpha_\theta > \frac{s}{2} - \varepsilon \tag{3}$$

If s=1, $\frac{s}{2} - \varepsilon$ is distributed uniformly on [0,1], so the probability that m=1 is just $(1-p_\theta)\alpha_\theta$.

If $s=2$, $\frac{s}{2}-\varepsilon$ is distributed uniformly on [0.5,1.5], so the probability that $m=1$ is just $(1-p_\theta)(\alpha_\theta-0.5)$.

We can write this probability as a function of s, namely as

$$(1-p_\theta)\left(\alpha_\theta+\frac{(1-s)}{2}\right).$$

We therefore summarize in Table 1 the probabilities of choosing different values of m according to whether the player is of high or low type and whether the stakes are high or low, as follows:

Player A's Move

Player A will send the money if the expected value of doing so is greater than the sure value of keeping it.

We also model player A's decision using a random utility function. We consider his utility as given by his expected payoff plus an error term η which is uniformly distributed between –e and 0 (we can consider this as a way of allowing for risk aversion while keeping the advantages of linear utility: $\eta=0$ corresponds to risk neutrality, while $\eta=-e$ is the highest risk aversion in the population). We ignore altruism and/or inequality aversion on the part of player A. While these traits can be important to explain the behavior of players in experimental interactions, they are not contingent on the characteristics of player B. Hence, since the focus of our model is on how A reacts to the signal sent by B, we believe that this simplification does not affect our main results.

Player A's decision then depends on γ, his subjective probability of facing a High Reciprocator type. He will send the money if the gain from receiving a net profit of half the original stake, multiplied by the probability that B chooses $m=1.5$, exceeds the loss of the whole original stake, multiplied by the probability that B chooses $m=0$. Formally, A sends the money iff:

$$\begin{aligned} 0.5(\gamma p_H+(1-\gamma)p_L)+\eta > \ & \gamma(1-p_H)\left(\frac{(1+s)}{2}-\alpha_H\right) \\ & +\ (1-\gamma)(1-p_L)\left(\frac{(1+s)}{2}-\alpha_L\right) \end{aligned} \tag{4}$$

Table 1 Probabilities that player B chooses various values of m

	$m=0$	$m=1$	$m=1.5$
High Reciprocator type (θ=H)	$(1-p_H)\left(\frac{(1+s)}{2}-\alpha_H\right)$	$(1-p_H)\left(\alpha_H+\frac{(1-s)}{2}\right)$	p_H
Low Reciprocator type (θ=L)	$(1-p_L)\left(\frac{(1+s)}{2}-\alpha_L\right)$	$(1-p_L)\left(\alpha_L+\frac{(1-s)}{2}\right)$	p_L

Notice that the right hand side of Eq. (4) is strictly increasing in s. This means that, for given γ, player A is less likely to send the money when the stakes are high than when they are low. Thus if we observe a higher probability of sending the money when the stakes are high, this must mean that A players have higher levels of γ.

Because of the uniform distribution of η, we can write the probability that an A player sends the money, given the value of γ, as q_γ, where

$$q_\gamma = \frac{[0.5(\gamma p_H + (1-\gamma)p_L) - \gamma(1-p_H)((1+s)/2-\alpha_H) - (1-\gamma)(1-p_L)((1+s)/2-\alpha_L)\,]}{e} \tag{5}$$

Differentiating (5) with respect to γ yields:

$$\frac{\partial q_\gamma}{\partial \gamma} = \frac{[0.5(p_H-p_L) - (1-p_H)((1+s)/2-\alpha_H) + (1-p_L)((1+s)/2-\alpha_L)\,]}{e} > 0 \tag{6}$$

Differentiating (6) with respect to s yields

$$\frac{\partial^2 q_\gamma}{\partial \gamma \partial s} = \frac{(p_H-p_L)}{2e} > 0 \tag{7}$$

which shows that a given increase in γ will result in a larger increase in q_γ when s=2 than when s=1. So higher stakes make the probability of sending the money more sensitive to player A's subjective probability that player B is the High Reciprocator type.

Smiling as the Signal

Now consider the sending of the signal. Player B invests effort e, which has an increasing convex cost $c_\theta(e)$, where $c_H(e)<c_L(e)$ for all positive values of e. This inequality can be interpreted as a simple incentive compatibility constraint and it is at the core of our model: if player B of L type could successfully mimic an H type at the same (or lower) cost, H would not have any advantage in trying to signal her type.

This effort produces a smile whose quality is related to the effort exerted via an increasing function $g(e,\tau)$, where τ is a random variable, and the probability distribution function $f(g|e)$ has the Monotone Likelihood Ratio Property.

We begin by assuming that this smile has a predictable positive effect on player A's subjective probability γ that player B is the High Reciprocator type. Without such an effect neither player would have any incentive to exert any effort at all. This effect can be represented by the "smile function" $\gamma=\gamma(g)$, where $\gamma'>0$. The function $\gamma(g)$ need not be concave but if not $c_\theta(e)$ must be sufficiently convex to yield a unique interior solution.

We next go on to show that if player B knows this, and if the quality of the smile responds to her effort, she has reason to invest effort in smiling in such a way that the

smile will indeed be a positive signal not just of her effort but also of the probability that she is the High Reciprocator type. Thus A's tendency to display greater trust in individuals who have more convincing smiles is one that could be expected to evolve under natural selection since it would correspond to a real empirical regularity.

To see this, write $V_{s\theta}$ for the expected utility B will receive if player A sends the money and note that $V_{sH} \geq V_{sL}$.[5] Writing $e^*_{s\theta}$ for the optimal choice of effort by a player B who is playing for stake s and is of type θ, since $c_H(e) < c_L(e)$ it follows that

$$e^*_{sH} > e^*_{sL} \tag{8}$$

It is also straightforward that $V_{2\theta} > V_{1\theta}$, and therefore that

$$e^*_{2\theta} > e^*_{1\theta} \tag{9}$$

Any function $g(e,\tau)$ that has the Monotone Likelihood Ratio Property will imply that the conditional probability that player B is the High Reciprocator type is increasing in the value of $g(e,\tau)$. To see this note that Bayes' Law with a uniform prior implies that

$$prob(\theta = H | g(e,\tau)) = \frac{1}{1 + f\left(g|e^*_{sL}\right)/f\left(g|e^*_{sH}\right)} \tag{10}$$

which is monotonically increasing in g by Eq. (8) and the Monotone Likelihood Ratio Property. This means that an increasing smile function $\gamma(g)$ is indeed consistent with natural selection and therefore we can predict, substituting the smile function into Eq. (6), that

$$\frac{\partial q_\gamma}{\partial g} > 0 \tag{11}$$

Finally, given that the convincingness of smiles is the result of effort in the way described in Eq. (10), we can calculate how the expected gain to A from sending money is related to smile quality. We write the expected gain to A from sending the money, conditional on smile quality as follows

$$\begin{aligned} E(U_A|g,s) \;&= (pr(\theta = H|g))\left[1.5sp_H + s(1-p_H)\left(\alpha_H + \frac{(1-s)}{2}\right)\right] \\ &+ \left(1 - pr\left(\theta = H \middle| g\right)\right)\left[1.5sp_L + s(1-p_L)\left(\alpha_L + \frac{(1-s)}{2}\right)\right] - s \end{aligned} \tag{12}$$

[5] The reason why the expected utility for B players of type H is higher than the utility for those that are L is that they have more altruism payoff than L players do. They could choose to return the same amount as L players do and would get at least as much utility as L players from doing so. In fact they choose to return more (in expected terms) than L players do, so their expected utility must be higher.

We can rewrite (12) as

$$E\left(U_A|g,s\right) = (pr(\theta = H|g))\left[1.5s(p_H - p_L) + s(\alpha_H - \alpha_L)\frac{s(1-s)}{2}(p_H - p_L) + s(\alpha_L p_L - \alpha_H p_H)\right] + \left[1.5sp_L + s(1-p_L)\left(\alpha_L + \frac{(1-s)}{2}\right)\right] \tag{13}$$

and therefore we can write the derivative of $E\left(U_A|g,s\right)$ with respect to $pr(\theta=H|g)$ as

$$\frac{\partial E\left(U_A|g,s\right)}{\partial pr(\theta = H|g)} = s[(s+0.5)(p_H - p_L) + (\alpha_H - \alpha_L) + (p_L\alpha_L - p_H\alpha_H)\,] \tag{14}$$

which is strictly positive because

$$[(s+0.5)(p_H - p_L) + (\alpha_H - \alpha_L) + (p_L\alpha_L - p_H\alpha_H)] > (s+0.5)(p_H - p_L) + \alpha_H(1-p_H) - \alpha_H(1-p_L) = [(s+0.5)(p_H - p_L) - \alpha_H(p_H - p_L)\,] \tag{15}$$

and the expression on the RHS is positive for any $s \geq 0.5$.

From this it follows, given (10), that

$$\frac{\partial E\left(U_A|g,s\right)}{\partial g} > 0 \quad \forall g,s \tag{16}$$

which is just the statement that the expected gain to player A from sending the stake to player B is increasing in the perceived convincingness of player B's signal.

Testable Implications

Our hypothesis that smiling in a way that is perceived as convincing is a costly signal has the following testable implications:

H1: The perceived quality of player B's smile is increasing in the size of the stake: this follows from inequality (9) given that g(.) is an increasing function;
H2: The amount sent by player A is increasing in the perceived quality of the smile: this follows from inequality (11);
H3: The expected gain to player A from sending the stake to player B is increasing in the perceived quality of player B's smile: this follows from inequality (16).

H1 is necessary in order to distinguish this hypothesis from two alternative views: first, that smiling is a form of costless communication that solves pure coordination

problems (like "cheap talk"), and secondly, that it is not communication at all but merely an outward sign of an inner emotional state (like blushing, say). H2 is necessary to explain why human beings should have evolved the habit of communicating in this costly way. H3 is necessary to explain why human beings should also have evolved the tendency to be influenced by the smiles of others.

In our companion paper (Centorrino et al. 2015) we subject these three predictions to an experimental test that significantly supports all three. The experiment consisted in a two person trust game where senders observed short video clips of trustees before taking their decisions. Potential trustees (84 participants from Toulouse, France) made two video clips averaging around 15 seconds for viewing by potential senders before the latter decided whether to send them money. Senders (198 participants from Lyon, France) made trust decisions with respect to the recorded clips. Clips were further rated concerning the genuineness of the displayed smiles. We have observed that smiles rated as more genuine by senders strongly predict judgments about the trustworthiness of trustees, and willingness to send them money. We observe a relation between costs and benefits: smiles from trustees playing for higher stakes are rated as significantly more genuine. Finally, we show that those rated as smiling genuinely return more money on average to senders.

In addition there exists some corroborating evidence for H2 and H3 elsewhere in the literature. H2 is the only one of the three predictions that has been tested directly in the existing literature. It has received significant support in Scharlemann et al. (2001) and in Johnston et al. (2010). While Scharlemann et al. (2001) use still pictures to detect trustworthy partners, Johnston et al. (2010) uses video clips and tests cooperation in a prisoners' dilemma on the basis of comparison of two clips.

Schug et al. (2010) find that individuals who are prone to cooperate as proposers in an ultimatum game are more emotionally expressive when facing unfair treatment by others than those who do not, including in the tendency to emit Duchenne as opposed to non-Duchenne smiles. This finding is consistent with H3 though not directly implied by it.

Conclusions

We have developed a model of smiling convincingly as a costly signal that has evolved to induce cooperation in situations requiring mutual trust. Individuals differ both in their willingness to engage in reciprocity and in their degree of altruism, and it is in their interest to signal this to others. In order to do so they must smile convincingly, but to do so involves costly effort. The model generates three testable predictions. First, the perceived quality of player B's smile is increasing in the size of the stake. Secondly, the amount sent by player A is increasing in the perceived quality of the smile. Thirdly, the expected gain to player A from sending the stake to player B is increasing in the perceived convincingness of player B's smile. We test, and find support for, these three predictions in our companion paper.

Our model suggests that individuals who have payoff functions of the kind we hypothesize would be able to signal trustworthiness to each other. It does not consider by what precise process such payoff functions might have evolved. This seems to us – as in the case of other costly signals – an interesting and important subject for future research.

Acknowledgments The authors are grateful to the Max Planck Institute for Evolutionary Biology in Plön for financial support. Support through the ANR - Labex IAST is gratefully acknowledged.

References

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*, 122–142.

Boone, R. T., & Buck, R. (2003). Emotional Expressivity and Trustworthiness: the role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior, 27*(3), 163–182.

Bosman, R., & van Winden, F. (2002). Emotional hazard in a power to take *experiment. The Economic Journal, 112*, 147–169.

Brown, W. M., & Moore, C. (2002). Smile asymmetries and reputation as reliable indicators of likelihood to cooperate: An evolutionary analysis. In S. P. Shobov (Ed.), *Advances in psychology research* (Vol. 11, pp. 59–78). New York: Nova.

Brown, W. M., Palameta, B., & Moore, C. (2003). Are there nonverbal cues to commitment? An exploratory study using the zero-acquaintance video presentation paradigm. *Evolutionary Psychology, 1*, 42–69.

Brülhart, M., & Usunier, J.-C. (2004). Verified trust: Reciprocity, altruism, and noise in trust games. Working paper 04.15, Université de Lausanne.

Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., & Seabright, P. (2015). Honest signalling in trust interactions: smiles rated as genuine induce trust and signal higher earnings opportunities. *Evolution and Human Behavior, 36*(1), 8–16.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Damasio, A. (1994). *Descartes' error: Emotion, reason and the human brain*. New York: Avon Books.

Darwin, C. R. (1872). *The expression of the emotions in man and animals*. London: John Murray.

de Quervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science, 305*(5688), 1254–1258.

Duchenne de Boulogne, C._B. (1862). *The mechanism of human facial expression*. Paris: Jules Renard.

Ekman, P. (1982). *Emotion in the human face* (2nd ed.). New York: Cambridge University Press.

Elster, J. (1998). Emotions and economic theory. *Journal of Economic Literature, 36*(1), 47–74.

Fehr, E., & Fischbacher, U. (2005). Altruists with green beards. *Analyse and Kritik, 27*(2005), 73–84.

Fehr, E., Fischbacher, U., & Gächter, S. S. (2003). Strong reciprocity, human cooperation and the enforcement of social norms. *Human Nature, 13*(1), 1–25.

Frank, R. H. (1988). *Passions within Reason. The strategic role of the emotions*. New York: Norton.

Frey, B. (2008). *Happiness: A revolution in economics*. Cambridge: MIT Press.

Frey, B., & Neckermann, S. (2009). Awards: A disregarded source of motivation. In M. Baurmann & B. Lahno (Eds.), *Perspectives in moral science – contributions from philosophy, economics, and politics in honour of Hartmut Kliemt* (pp. 177–182). Frankfurt: Frankfurt School Verlag.

Frijda, N., Manstead, A., & Fischer, A. (2004). Feelings and emotions: Where do we stand? In A. Manstead, N. Frijda, & A. Fischer (Eds.), *Feelings and emotions: The Amsterdam symposium*. Cambridge: Cambridge University Press.

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology, 213*(1), 103–119.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in Humans. *Evolution and Human Behavior, 24*, 153–172.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*, 517–546.

Hopfensitz, A., & Reuben, E. (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal, 119*, 1534–1559.

Hwang, S.-H., & Bowles, S. (2010). Is Altruism bad for cooperation? Working paper, University of Massachusetts.

Johnston, L., Miles, L., & Macrae, C. N. (2010). Why are you smiling at me? Social functions of enjoyment and non-enjoyment smiles. *British Journal of Social Psychology, 49*, 107–127.

Kahneman, D. (2003). A psychological perspective on economics. *The American Economic Review, 93*(2), 162–168.

Kahneman, D., Diener, E., & Schwarz, N. (Eds.). (1999). *Well-being: The foundations of hedonic psychology*. New York: Russell Sage.

Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behaviour. *Emotion, 7*, 730–735.

Loewenstein, G. (2000). Emotions in economic theory and economic behaviour. *American Economic Review, 90*(2), 426–432.

Lotem, A., Fishman, M. A., & Stone, L. (2003). From reciprocity to unconditional altruism through signalling benefits. *Proceedings of the Royal Society B, 270*, 199–205.

Lyons, M. T., & Aitken, S. J. (2008). Machiavellianism in strangers affects cooperation. *Journal of Evolutionary Psychology, 6*(3), 173–185.

Manson, J. H., Gervais, M. M., & Kline, M. A. (2013). Defectors cannot be detected during "small talk" with strangers. *PLoS ONE, 8*(12), e82531.

Manzini, P., Sadrieh, A., & Vriend, N. J. (2009). On smiles, winks and handshakes as coordination devices. *The Economic Journal, 119*, 826–854.

Mehu, M., Grammer, K., & Dunbar, R. (2007). Smiles when sharing. *Evolution and Human Behavior, 28*, 415–422.

Mussel, P., Hewig, J., Allen, J. J. B., Coles, M. G. H., & Wolfgang, M. (2014). Smiling faces, sometimes they don't tell the truth: facial expression in the ultimatum game impacts decision making and event-related potentials. *Psychophysiology, 51*, 358–363.

Niedenthal, P., Mermillod, M., Maringer, M., & Hess, U. (2010). The Simulation of Smiles (SIMS) model: embodied simulation and the meaning of facial expression. *Behavioral and Brain Sciences, 33*(06), 417–433.

Oda, R., Yamagata, N., Yabiku, Y., & Matsumoto-Oda, A. (2009). Altruism can be assessed correctly based on impression. *Human Nature, 20*(3), 331–341.

Owren, M. J., & Bachorowski, J.-A. (2001). The evolution of emotional expression: A "selfish-gene" account of smiling and laughter in early hominids and humans. In T. J. Mayne & G. A. Bonnano (Eds.), *Emotions: Current issues and future directions* (pp. 152–191). New York: Guilford Press.

Reed, L. I., Zenglen, K. N., & Schmidt, K. L. (2012). Facial expressions as honest signals of cooperative intent in a one-shot anonymous prisoner's dilemma game. *Evolution and Human Behavior, 33*, 200–209.

Sanfey, A., Rilling, J., Aronson, J., Nystrom, L., & Cohen, J. J. (2003). The neural basis of economic decision making in the ultimatum game. *Science, 300*(5626), 1755–1758.

Scharlemann, J. P. W., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: game theory with a human face. *Journal of Economic Psychology, 22*, 617–640.

Schmidt, K. L., & Cohn, J. F. (2001). Human facial expressions as adaptations: evolutionary questions in facial expression research. *American Journal of Physical Anthropology, Suppl 33*, 3–24.

Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T., & Bonnet, K. (2010). Emotional expressivity as a signal of cooperation. *Evolution and Human Behavior, 31*(2), 87–94.

Smith, A. (2000). *The theory of moral sentiments*. New York: Prometheus Books.

Smith, E. A., & Bliege Bird, R. (2000). Turtle hunting and tombstone opening: public generosity as costly signalling. *Evolution and Human Behavior, 21*(4), 245–261.

Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of Economic Literature, XLIII*(June), 392–436.

Sommerfeld, R. D., Krambeck, H.-J., & Milinski, M. (2008). Multiple gossip statements and their effect on reputation and trustworthiness. *Proceedings of the Royal Society B, 275*, 2529–2536.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics, 87*(3), 355–374.

Számadó, S. (2012). The rise and fall of handicap principle: a commentary on the "Modelling and the fall and rise of the handicap principle". *Biological Philosophy, 27*, 279–286.

Vogt, S., Efferson, C., & Fehr, E. (2013). Can we see inside? Predicting strategic behavior given limited information. *Evolution and Human Behavior, 34*, 258–264.

Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology, 53*(1), 205–214.