Evolution and Human Behavior 36 (2015) 44–51

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

Regulatory adaptations for delivering information: the case of confession

Daniel Sznycer [a,*], Eric Schniter [b], John Tooby [a], Leda Cosmides [a]

[a] Center for Evolutionary Psychology, University of California, Santa Barbara
[b] Economic Science Institute, Chapman University

ARTICLE INFO

Article history:
Initial receipt 15 May 2014
Final revision received 27 August 2014

Keywords:
Communication
Decision-making
Motivation
Welfare tradeoff ratio
Shame

ABSTRACT

Prior to, or concurrent with, the encoding of concepts into speech, the individual faces decisions about whether, what, when, how, and with whom to communicate. Compared to the existing wealth of linguistic knowledge however, we know little of the mechanisms that govern the delivery and accrual of information. Here we focus on a fundamental issue of communication: The decision *whether* to deliver information. Specifically, we study spontaneous confession to a victim. Given the costs of social devaluation, offenders are hypothesized to refrain from confessing unless the expected benefits of confession (e.g. enabling the victim to remedially modify their course of action) outweigh its marginal costs—the victim's reaction, discounted by the likelihood that information about the offense has not leaked. The logic of welfare tradeoffs indicates that the victim's reaction will be less severe and, therefore, less costly to the offender, with decreases in the cost of the offense to the victim and, counter-intuitively, with *increases* in the benefit of the offense to the offender. Data from naturalistic offenses and experimental studies supported these predictions. Offenders are more willing to confess when the benefit of the offense to them is high, the cost to the victim is low, and the probability of information leakage is high. This suggests a conflict of interests between senders and receivers: Often, offenders are more willing to confess when confessions are *less* beneficial to the victims. An evolutionary–computational framework is a fruitful approach to understand the factors that regulate communication.

1. Introduction: The puzzle of confession

A popular moral view of confession is that confessing is the right thing to do following wrongdoing and that it positively reflects on the moral character of the offender. Some researchers echo this view arguing that spontaneous confessions are internally motivated, guilt-driven attempts to undo the harm done and to restore the relationship (Weiner, Graham, Peter, & Zmuidinas, 1991; see Tangney, Miller, Flicker, & Barlow, 1996). From an evolutionary perspective, however, this analysis does not provide a theory of why natural selection would have favored this behavior. Consequently, the phenomenon of spontaneous confession remains puzzling.

Undoing the harm done to someone else – or claiming to be the cause of it – is not in any obvious way fitness-promoting in the general case. Indeed, the decision to commit the self-interested, other-harming act in the first place already involved the evaluation that it was of net advantage to the perpetrator. More importantly, disclosing to a victim that you were the one who harmed her could very well damage or destroy a pre-existing cooperative relationship. This is presumably the reason why the perpetrator kept the information from the victim to begin with. Absent compensatory benefits, victims of harm will at the very least revise their valuation of the perpetrator downward. Moreover, the victim may retaliate, making further offenses costlier to the offender. Furthermore, information about the offender's uncooperative inclinations may diffuse into the community, yielding reputational costs. So, regardless of whether the offender up-regulates his valuation of the victim after inflicting harm, why would a brain shaped by natural selection be designed to spontaneously communicate such self-damaging information? Here we subject confession to an evolutionary–functional analysis to understand its logic and to highlight the advantages of analyzing issues in human communication in evolutionary–functional terms.

1.1. Adaptationist approaches to communication, social valuation, and social negotiation

Confession – disclosing to someone that you have harmed them – is a subtype of communication. Systems that signal a given kind of information should only evolve if they yield net benefits to the sender, on average and under ancestral-like conditions (Maynard-Smith & Harper, 2003; Tooby & Cosmides, 1990; Williams, 1966). The adaptive problem for the signaler – sending information when it is likely to be self-beneficial – is framed by the design of information-processing machinery in the receiver. In particular, what effect will the signal have on the receiver's behavior, and what fitness consequences will this behavioral response have on the sender? It is this evolved architecture in signal receivers that shapes the adaptations for communication in senders (Krebs & Dawkins, 1984)—including whether to send a given kind of

* Corresponding author. Center for Evolutionary Psychology, University of California, Santa Barbara, CA 93106-3210.
E-mail address: sznycer@psych.ucsb.edu (D. Sznycer).

http://dx.doi.org/10.1016/j.evolhumbehav.2014.08.008
1090-5138/

signal at all. Since disclosure of an offense triggers victim behavior that adversely impacts the sender, the default strategy is to not send the clarifying signal. This links the puzzle of confession to the evolved emotion program of shame.

By hypothesis, shame is a neurocomputational program whose evolved function is to orchestrate best-bet responses in an individual to the adaptive problem of information potentially or actually spreading to others which would have a negative effect on how those others value and treat that individual (Sznycer, 2010). Shame acts to limit the likelihood and the costs of being socially devalued. Our evolved psychology is packed with devaluation-limiting features (De Hooge, Zeelenberg, & Breugelmans, 2011; Fessler, 2001; Gilbert, 1998; Sznycer et al., 2012). For instance, actions thought to be interpreted as reprehensible by an audience trigger shame even when people know the action caused no harm to the audience or third parties (Sznycer, Tooby, & Cosmides, 2009)—what one would expect from a system designed to tamper with adverse representations about the self in others' minds. On this view, shame prompts non-disclosure. Thus, for a confession to be made, there must be other motivational forces overriding this default.

1.2. Welfare tradeoff functions and their regulation

For ancestral humans, actions taken by one individual would often have had positive or negative effects on the welfare of others. Selection pressures such as kin selection (Hamilton, 1964) and the asymmetric war of attrition (Hammerstein & Parker, 1982) specify strategies that conditionally trade off the welfare of the actor against the welfare of another organism (the target). If these selection pressures acted on our ancestors, then circuits are required to compute the effects of acts on self and others, and deploy a weighting function indexing the extent to which the self will trade off their own welfare to enhance the welfare of another—a welfare tradeoff ratio between the self (i) and individual j: WTRij (Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008).

Recent studies support the hypothesis that the human mind is equipped with a welfare tradeoff architecture. Consistent with the WTR hypothesis, people allocate resources between self and others as if a target-specific threshold divided those tradeoffs that are worth making from those that are not (Delton et al., 2008; see Jones & Rachlin, 2006; Sell, Tooby, & Cosmides, 2009).

An individual is benefited when others recalibrate their welfare trade-off ratios toward her upwards, and is harmed when others recalibrate their WTRs toward her downwards. Reciprocally, it costs more to place a higher weight on someone else's welfare. So there are a variety of social strategies that evolved to minimize losses by not putting too great a weight on others, and to make gains by inducing others to place a higher weight on the actor's welfare. Welfare trade-off ratios as internal magnitudes regulating choices constituted adaptive problems for socially interacting humans, and gave rise to a suite of what have been called recalibrational emotions (Tooby et al., 2008). Anger is triggered when another person places too low a weight on the welfare of the actor (compared to what the actor implicitly computes she is entitled to). Once triggered, the function of anger is to bargain for a higher WTR through conditional threats of increasing harm or decreasing benefits to the "offender," or at worst to curtail the costs of making sacrifices for another that are not redeemed by reciprocal valuation by that person.

In contrast, guilt is a recalibrational emotion that is triggered when one discovers that one has placed too little weight on another's welfare from one's own point of view. It is a recalibrational process that operates even in the absence of knowledge by the victim of the guilt-producing act (Smith, Webster, & Eyre, 2002; Sznycer, 2010). Guilt is triggered by acts that harm a valued target to an unexpectedly large extent. The functional product of guilt (in contrast to shame) is to increase one's WTR toward the harmed individual, to bring it to the level of the equilibrium long-run valuation the system places on the welfare of the other, and in the shorter run to mitigate or remedy the negative effects of the harmful act, when the earlier valuation turned out not to be in the interests of the offender.

In sum, an act entailing too low a WTR toward another will elicit anger in the victim and shame, guilt, or both in the offender. Selfish acts often involve both reputational damage and unwarranted cost imposition from the offender's perspective, and so shame and guilt tend to co-activate. However, the distinct functional signatures of each emotion can be discerned. For instance, publicity exacerbates shame more than guilt (Scarnier, Schmader, & Lickel, 2009; Smith et al., 2002; Sznycer et al., 2009). And cooperative motivations have a robust link to guilt, but only a contingent link to shame (De Hooge, Breugelmans, & Zeelenberg, 2008; Tangney, Stuewig, & Martinez, 2014; Wicker, Payne, & Morgan, 1983).

1.3. Welfare tradeoffs, the infliction of harm, and confession

When a person inflicts harm on another to accrue a benefit for himself, this reflects a WTR the offender behaviorally expressed toward the victim. Regarding selfish transactions – of relevance to confession situations – the WTR indexes the minimum benefit the self requires before imposing a given cost on another. The higher the WTR, the higher the valuation of the other, and the less frequent the selfishness; and vice versa. An individual will be motivated to benefit at the expense of another, but only when:

$$\text{Bof}\,f_O \geq \text{Cof}\,f_V \text{WTR}_{OV} \quad (1)$$

Where WTR_{OV} is the WTR of the offender (O) toward the victim (V), Boff_O is the offender's estimate of the benefit he or she will derive from the offense, and Coff_V is the offender's estimate of the cost of the offense to the victim.

Following the principle of receiver-targeted communication, we advance a model where offenders target their confessions to the social valuation psychology of victims with the aim of minimizing the devaluing and punitive reactions of victims.

This *devaluation-minimizing model of confession* applies to situations where an offender benefits at the expense of a victim and the victim is unaware of the offense or ignorant about the offender's identity—at least in the estimation of the offender. The devaluation-minimizing model is limited to voluntary, spontaneous, truthful confession to the victim. Confession by force or to third parties is beyond its scope (for other types of confession, see Kassin & Gudjonsson, 2004; Schelling, 1960). The model is also limited to cooperative or potentially cooperative relationships (see below). Under the devaluation-minimizing model, spontaneous confession is a decision with one binary choice (confess or not) and three possible outcomes, each of which is associated with a different payoff.

Outcome 1. *Offender confesses*. Confessing entails costs to the offender. By hypothesis, these lie not in the production of the signal (where the cost is trivial) but in the impact on the offender of the victim's reaction to the information (Lachmann, Számadó, & Bergstrom, 2001). This reaction is driven by anger (McCullough, 2008; Sell, 2005), and may include retaliation, demands for WTR up-regulation and restitution, victim WTR down-regulation, and/or withdrawal from the cooperative relationship. Other emotions such as sadness and disgust (Lim, 2012; Schniter & Shields, 2013) may also be mobilized in the victim.

The victim's reaction is exacerbated as the difference between the WTR expected by the victim and the WTR implied in the offense increases (Sell, 2005). Inequality (1) governs offender behavior but it can also allow the victim to infer the offender's WTR via estimates of the costs and benefits involved.

$$\text{WTR}_{OV} \leq \frac{\text{Boff}_O}{\text{Coff}_V} \quad (2)$$