Guilt for Non-Humans

Guilt for Non-Humans

Luís Moniz Pereira [1], The Anh Han[2], Luis Martinez-Vaquero [3,4] and Tom Lenaerts[3,4]

[1]NOVA LINCS, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal. Email:lmp@fct.unl.pt

[2]School of Computing, Teesside University, Borough Road, Middlesbrough, TS1 3BA, UK. Email: T.Han@tees.ac.uk

[3]AI lab, Vrije Universiteit Brussel, Boulevard du Triomphe CP212, 1050 Brussels, Belgium. Email: l.martinez.vaquero@gmail.com

[4]MLG, Université Libre de Bruxelles, Pleinlaan 2, 1050 Brussels, Belgium. Email: tom.lenaerts@ulb.ac.be

Abstract

We argue that the emotion of guilt, in the sense of actual harm done to others from inappropriate action or inaction, is worthwhile to incorporate in evolutionary game models, as it can lead to increased cooperation, whether by promoting apology or by inhibiting defection. The study thereof can then transpire to abstract and concrete populations of non-human agents.

Psychological Background

Theorists conceive of shame and guilt as belonging to the family of self-conscious emotions (Lewis 1990) (Fischer and Tangney 1995) (Tangney and Dearing 2002), invoked through self-reflection and self-evaluation. Though both have evolved to promote cooperation, guilt and shame can be treated separately. Guilt is an inward private phenomenon, though it can promote apology, and even spontaneous public confession. Shame is inherently public, though it may too lead to apology and to the request for forgiveness (Smith 2008, pp. 96-107). Shame, however, hinges on being caught, failing to deceive, and the existence of a reputation mechanism.

The philosopher Martin Buber (Buber 1957) underlined the difference between the Freudian notion of guilt, based on internal conflicts, and *existential guilt*, based on actual harm done to others, which is the sense we are considering here. On transgression or error, the self renders judgment on itself. This self-evaluation may be explicit or implicit, conscious or unconscious. Shame and guilt typically arise from the blaming self-judgment about one's negative personal attributes (shame) or about negative harmful personal behaviour or failure to act to prevent harm (guilt). Shame is often conducive to hiding and anger, and guilt is often conducive to admission and reparative action. Guilt is considered empathic whereas shame not (Tangney et al. 2013, pp. 485-502). We leave out shame for now, because it involves reputation, and concentrate on guilt instead.

To avoid or attempt to prevent blame assignment that might result from inappropriate action or inaction, there exists a guilt mechanism concerned not just with a posteriori guilt for a harm actually intended, but functions as well a priori, preventing harm by wishing to avoid guilt. A posteriori outward admission of guilt may serve to pre-empt punishment, whenever harm detection and blame become foreseeable.

We know too that guilt may be alleviated by private confession (namely to a priest or a psychotherapist) plus the renouncing of past failings in future. Because of their private character, such confessions and atonements, given their cost (prayers or fees), render temptation defecting less probable. Public or open confession of guilt can be coordinated with apology for better effect, and the cost appertained to some common good (like charity), or as individual compensation to injured parties.

More generally, Frank has suggested humans have been endowed during evolution with the means to solve problems of commitment by means of "moral sentiments", to wit, those of anger, contempt, disgust, envy, greed, shame, and guilt (Frank 1988, pp. 46, 53).
Moral sentiments help solve such problems because honest manifestation of certain emotions make commitments more credible.

In particular, the promises of an agent known to be prone to guilt are therefore trustworthier. Heightened anger towards non-confessed guilt might be triggered by intention recognition, thereby putting pressure on guilt admission. Exhibition of guilt proneness by an agent may assuage other agents that defection by the agent was not intended,

even when it might be clear it was not, since intention ascription by others is not perfect.

If intentions provide such a role in determining due apology, how can one read an offender's mental states? We regularly judge the mental states of others, and the notion of *mens rea*[1] in criminal law depends on this ability. An offender's emotions, namely feelings of guilt, provide a measure of his mental states (Smith 2008, p. 96). Fake emotions better be discerned, of course. (Smith 2008, pp. 96-107) elaborates on detecting and distinguishing guilt, shame, embarrassment, remorse, and regret.

From (Smith 2008, pp. 101-103) we quote, abridging:

> 'According to (Rawls 1971), "shame is the emotion evoked by shocks to our self-respect" but we feel guilty when we "act [...] contrary to [our] sense of right and justice." Both involve our sense of morality, but in guilt "we focus on the infringement of just claims of others and the injury we have done to them, and on their probable resentment and indignation should they discover our deed."
>
> In shame "we feel struck by the loss to our self-esteem and our inability to carry out our aims: we sense the diminishment of self from our anxiety about the lesser respect that others may have for us and from our disappointment with ourself for failing to live up to our ideals." A single wrongdoing might provoke feelings of both shame and guilt, but the primary distinction involves the emphasis on either my disappointment with myself (shame) or my concern for the victims and norms I have transgressed (guilt).
>
> [p. 102] Understood in accordance with the earlier description, guilt would seem like an appropriate emotional component of apology because it accompanies the recognition of wrongdoings as such. When we identify and share a commitment to the value underlying a transgression, guilt would appear to designate the corresponding emotion. As an undesirable emotion, guilt also spurs us to undertake the reform and redress likely to free us from its clutches.
>
> [p. 103] Negative emotions can have a deterrent value in that potential offenders may resist urges to commit offenses if they wish to avoid the unpleasant feelings of guilt or shame that may accompany their deed. Negative emotions may also serve rehabilitative objectives because an experience of guilt may move an offender to reform her behaviour.'

[1] *Mens rea*, the intention or knowledge of wrongdoing that constitutes part of a crime, as opposed to the action or conduct of the accused. Compare with *actus reus*, action or conduct that is a constituent element of a crime, as opposed to the mental state of the accused. ORIGIN: mid 19th cent.: Latin, literally *'guilty mind.'*

Furthermore, according to (Tangney et al. 2013, pp. 494-496), guilt-prone individuals are inclined to manage anger constructively, and disinclined toward aggression. And they are less prone to defection and noise.

Guilt is the quintessential moral emotion: it promotes the acknowledgment of wrongdoing, the acceptance of responsibility, and the taking of reparative action. Expressions of guilt restore equanimity, reaffirm fairness, and compensate transgressions. Guilt leads to positive intra- and interpersonal processes, especially in contexts requiring cooperation.

Evolutionary Background on Guilt and Cooperation

If foreseen guilt prevents harm and absence of harm prevents possible retaliation and/or loss of reputation, then it would seem that a priori guilt would be evolutionarily advantageous. A posteriori guilt, on the other hand, would be evolutionarily advantageous because conducive to increased amount/possibility of apology, and we've seen apology is advantageous. Also apology reduces the pain of guilt.

Evolutionarily, guilt is envisaged as an in-built mechanism that tends to prevent wrong doing because of the internal self suffering it creates, and, should the wrong doing have taken place, it puts internal pressure on confession (admitting the wrong) and follow-up costly apology and penance, plus an expectation of forgiveness, so as to alleviate and dispel the internal suffering produced by guilt (Fessler and Haley 2012, pp. 7-36) (Tangney et al. 2013, pp. 485-502). Using the iterated Prisoner's Dilemma and Ultimatum games (Ketelaar and Tung Au 2003, pp. 429-453) found that guilt is a key factor in increasing cooperation among players.

It would pay for guilt to have spread once it appears, on two counts: as an inhibition mechanism and as a forgiveness stimulus. A third count, in our view, is that counterfactually thinking about guilt or shame is useful to prevent their future arising, a process of self-cleansing or self-debugging (Niedenthal, Tangney, and Gavanski, 1994). Counterfactual thinking, in turn, arises to explain causality in general, being subsequently used to handle guilt and self-improvement. Hence, guilt-dealing mechanisms seem to be evolutionarily advantageous for cooperation.

It seems clear that an evolutionary anthropological case about guilt has been made and accepted in the literature for intention recognition. Intentionality matters crucially to

distinguish intended actions from noise, from accidental actions, and from side effects, since non-intended harms should be discounted because they are unavoidable, and a revenge arms race would be pernicious to all.

The notion of intentionality is ascribed to agents, and in particular to gods and nature spirits. The latter ones have the power to exercise justice with regard to acts of killing animals and slashing cultivars (which are living beings). When it doesn't rain, or some tragedy happens, it is ascribed to wrath of the gods or animal spirits symbolised by totems. Performing human or other sacrifices is meant to atone and apologise for the harm done to other living beings.

That's how guilt has arisen evolutionarily, as humans know that the spirits know what you did (and even about thoughts and memories inside your head about doing it on purpose) though no one might have witnessed your deed. Guilt and apology are primed by their moral disapproval and by revenge.

Population morality in turn arises for a great number of reasons (namely starting with mutualism and following on to contractualism), plus in particular making sure that one's intentions can be explained and do follow the accepted rules and exceptions. The literature on morality already makes the case in detail, albeit according to distinct schools of thought. Some of these are closer than others to the mechanisms that we know in AI.

The Baldwin effect describes an inclination for general learning mechanisms to open the way to domain-specific adaptations. Agent-based simulation models have shown that a moderate bias toward prosocial behaviour is favoured in evolving populations where punishment for antisocial behaviour becomes dominant (Cushman and Macendoe 2009). Guilt may be a primer for a Baldwin effect by which harm is partly avoided, and where punishment for harm is diminished by forgiveness of costly confession and apology. Moreover, the pain of guilt, even a non-confessed one, acts as internal punishment, and so may serve as an internal evaluation mechanism to undermine one's defection (Damasio 1994) (Mameli 2004). Guilt signalling, having a cost in terms of accrued apology, comes under the "handicap principle" (Zahavi 1975) (Zahavi 1997), and that makes it "honest" or reliable, rather than deceitful.

Hence, on these counts, guilt will have been selected for prosociality.

Guilt Treating by Evolutionary Game Theory

The evolutionary issue about guilt is whether it is more worthwhile than the absence of guilt, with respect to emergence of cooperation. One would speak of guilty explicitly, and show that it's worthwhile, thus explaining its appearance on the evolutionary scene.

Guilt is widespread, for some reason, and we should show that it naturally connects with apology and forgiveness mechanisms, because of its emergent evolutionary advantage. Moreover, it does not seem too difficult to incorporate guilt into present frameworks involving apology and forgiveness (Martinez-Vaquero et al. 2015). It would mean duplicating each possibly defecting strategy into one experiencing guilt and a corresponding guiltless one.

This would open the way to the treatment of emotions as evolutionary mechanisms scaffolding cooperation, guilt being a widely acknowledged one. Furthermore, it would show that one does not need a specific kind of body (namely an anthropomorphic one) for guilt to be a functional useful emotion in population settings where cooperation is good value.

One may focus on emotions as being strategies in abstract evolutionary population games, sans specific embodiment (Pereira 2016). One adds guilt (and, for that matter, possibly guilt promoted counterfactual reasoning) as an evolved means to trigger costly apology, to expect a better chance of forgiveness, and to assuage the inner guilt that prompts the triggering.

The hypothesis, then, is that the emergence of guilt in a population is evolutionarily advantageous.

We can test this hypothesis via our already existing model comprising apology, revenge, and forgiveness, by piggybacking guilt onto them (Martinez-Vaquero et al. 2015). We might introduce a zero/one guilt parameter, which, on defaulting, not only increases the probability of apology (confession), but also, spontaneously pays a costlier apology, as the means to atone for internal guilt (through the redressing towards the co-player), rather than simply apologising.

On the other hand, the co-player will more readily accept a guilty apology and forgive. Indeed, this co-player attitude will favour in the population his own forgiveness by others, in case of his confession of guilt, instead of simple apology in the absence of guilt.

The prediction is that guilt will facilitate and speed-up the emergence of cooperation. In spite of its initial heavier

cost, in time the cost will be recuperated within the guilt-ridden population. One reason being that it is compensated by the costlier guilt apology of others, another reason being that it is more conducive to forgiveness.

To emphasise our point: the experiments would need to explore different initial situations from the ones already considered. Instead of albeit different but homogenous values of costly apology, the population would start with heterogeneous values of overall costly apology: a base cost (the apology) plus the added guilt cost.

Testing would mean that in the social imitation step the guilt and the forgiveness thresholds would also be copied, not just the strategy. One would start with a good mixture in the population of these two threshold factors, including zero guilt, within each of the strategies that defect and also within those that forgive, to see which factors pervade, for some apology compensation.

So, for each defaulting strategy, there would exist in the population both individuals with guilt and those without (say 50%-50% at the start). For the moment we would fix the forgiveness threshold. Guilt or its absence would be transmitted by social imitation too.

The base hypothesis is that when there is guilt in the starting population then the most frequent stationary distribution includes guilt and better cooperation. For which level of guilt this would happen would have to be experimentally found. Next, for that best level of guilt, we want to find a best level of probability of forgiveness, still starting with a mixed guilty/non-guilty population.

An additional possibility is to investigate guilt as a mechanism that diminishes defection. This would probably be tied to intention recognition, since guilt will have evolved as a fear about the detection of harm done (see above).

Acknowledgments

LMP acknowledges support from FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013. TAH acknowledges support from Teesside University, project URF-11200174. LMV and TL acknowledge the support of from the F.R.S.-F.N.R.S. (grant FRFC nr. 2.4614.12) and the F.W.O (grant G.0391.13N).

References

Buber. M 1957. Guilt and guilt feelings. *Psychiatry* **20** (2): 114–29.

Cushman, F. A., and Macendoe, O. 2009. The coevolution of punishment and prosociality among learning agents". In Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Damasio, A. 1994. *Descartes's Error*. New York: Avon.

Fessler, D. M. T. and Haley, K. J. 2012. The Strategy of Affect: Emotions in Human Cooperation". In *Genetic and Cultural Evolution of Cooperation*, P. Hammerstein (ed.), Dahlem Workshop Report. Cambridge, MA: The MIT Press.

Fischer, K. W. and Tangney, J. P. 1995. Self-conscious emotions and the affect revolution: Framework and introduction". In Fischer, K. W., Tangney, J. P. (eds.) *The Self-conscious Emotions: Shame, guilt, embarrassment, and pride* (pp. 3-22). New York: Guilford Press.

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

Ketelaar, T. and Tung Au, W. 2003. The effects of feeling guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect as information interpretation of the role of emotion in social interaction. *Cognition and Emotion*, 17(3).

Lewis, M. 1990. Thinking and feeling – the elephant's tail. In Maher, C. A.; Schwebel, M.; and Fagley, N. S. (eds.), *Thinking and Problem Solving in the Developmental Process: International Perspectives* (pp. 89-110). Hillsdale, NJ: Erlbaum.

Mameli, M. 2004. The role of emotions in ecological and practical rationality. In Evans, D. and Cruse, P. eds., *Emotion Evolution and Rationality* (pp. 158-178). Oxford: Oxford University Press.

Martinez-Vaquero, L. A.; Han, T. A.; Pereira, L. M.; and Lenaerts, T. 2015. Apology and Forgiveness Evolve to Resolve Failures in Cooperative Agreements, *Scientific Reports*, 5:10639.

Niedenthal, P.M.; Tangney, J.P.; and Gavanski. I. 1994. "If only I weren't" versus "If only I hadn't": Distinguishing shame and guilt in counterfactual thinking. *Journal of Personality and Social Psychology*, 67, 585-595.

Pereira, L. M. 2016. Software sans Emotions but with Ethical Discernment. In Silva, S. (ed.), *Morality and Emotion: (Un)conscious Journey to Being*. Frontiers of Cognitive Psychology series. Oxford: Routledge.

Rawls, J. 1971. *A Theory of Justice*.. Cambridge, MA: Harvard University Press.

Smith, N. 2008. *I Was Wrong: The Meaning of Apologies*. Cambridge: Cambridge University Press.

Tangney, J. P. and Dearing, R. 2002. *Shame and Guilt*. New York: Guilford Press.

Tangney, J. P.; Stuewig, J.; Malouf, E. T.; and Youman, K. 2013. Communicative Functions of Shame and Guilt. In *Cooperation and Evolution*", Sterelny, K.; Joyce, R.; Calcott, B.; and Fraser, B. (eds.), Cambridge, MA: The MIT Press.

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

Zahavi, A. 1977, A. *The handicap principle: a missing piece of Darwin's puzzle*. Oxford: Oxford University Press.