A cognitive model of punishment

Francesca Giardini (francesca.giardini@istc.cnr.it)
Institute of Cognitive Sciences and Technologies, Via San Martino della Battaglia, 44
00185 Rome Italy

Giulia Andrighetto (giulia.andrighetto@istc.cnr.it)
Institute of Cognitive Sciences and Technologies, Via San Martino della Battaglia, 44
00185 Rome Italy

Rosaria Conte (rosaria.conte@istc.cnr.it)
Institute of Cognitive Sciences and Technologies, Via San Martino della Battaglia, 44
00185 Rome Italy

Abstract

People use sanctioning behaviours differently according to what they believe and want to achieve, according to the context and to the situation. We need to understand the motivations for different forms of punishment in order to explain why sanctions and incentives have different effects on human behaviour. Aim of this work is to propose a cognitive model of three distinct kinds of punishing behaviours, differentiated in terms of the defining cognitive patterns.

Keywords: Cognitive modeling; Punishment; Cooperation.

Introduction

Punishment is a core mechanism to enforce and support social order, to promote cooperation and to prompt group beneficial behaviours. Social scientists have long debated on the nature and the effects of this mechanism, but there are many questions still open, as for instance the relationship between counter-punishment and cooperation. There is a growing body of evidence that altruistic punishment plays a crucial role in enforcing cooperation and in promoting group welfare (Fehr & Gachter, 2000, 2002), but some recent experimental results raised the problem of antisocial punishment, that is sanctioning people who behave socially. Herrmann, Thoni, and Gachter (2008) compared results on punishment and cooperation collected in sixteen different participant pools around the world. They showed the emergence of antisocial punishment in repeated public goods experiments, and proposed that differences can be explained in terms of different societal background. Nikiforakis and Engelman (2008) used a public good game with multiple punishment stages aiming at investigating whether retaliatory behaviours would escalate into a feud. Interestingly, cooperation rates declined but feuds were avoided by participants.

Although a number of accounts (for some representative work see (Bowles & Gintis, 2004; Henrich & Boyd, 2001; Henrich et al., 2006) have stressed the relevance of punishment in human societies, they suffer the flaw that they consider punishment as a unique behaviour. In our view, punishing actually consists in a complex behavioral repertoire in which it is useful to disentangle at least revenge (social-status punishment), retaliation (strategic punishment) and sanction (normative punishment).

Treating punishment as a single behaviour without caring for its cognitive foundations could be misleading especially if one is interested in explaining cooperation and its maintenance in evolutionary terms. There is neither a single form of punishment nor a single motive to punish other people, and the question is: How can we distinguish between punishment aimed at making the individual internalize the norm and pure revenge? How do people choose between punishment and revenge?

Cognitive modelling allows us to disentangle apparently indistinguishable acts and to understand the motives and objectives that pave the way to distinct ways of punishing. Taking revenge is not the same as punishing a wrongodoer or sanctioning a deviant behaviour, and explaining these differences and the related motivations could effectively advance research on cooperation and prosocial behaviours under several respects.

The rest of this article is organized as follows. Firstly, we will introduce a general theory of cognitive social action, in order to provide some basic concepts. Secondly, revenge will be analyzed, focusing on the explicit mental representations behind it. Therefore we will turn our attention to punishment, showing what is inside the punisher's mind. Finally, sanction will be described. Future work and conclusions will follow.

The cognitive roots of social behaviour

In general, this work aims at unveiling the proximate mechanisms of enforcement behaviours, in order to understand the mental mechanisms underlying revenge, punishment and sanction.

Before these arguments are developed, some terminological issues need clarification. As stated elsewhere (Conte & Castelfranchi, 1995), an agent is a goal-governed system.

> By this, we mean an entity, not necessarily autonomous, that has the capacity to act upon the external world in order to reduce the discrepancy

between the world itself and some regulatory state that is somehow represented within the entity (p.1).

A cognitive agent is endowed with cognitive representations of the external world and of its internal states as well. Agents have *beliefs* about themselves and the world and they act on the basis of their *goals* to reduce the discrepancy between the world and what they want.

There are several ways to influence agents, but here we refer to *cognitive influencing* (Cialdini & Goldstein, 2004; Conte & Castelfranchi, 1995), a process by which a given entity, say *Ii*, acts on another entity, *mj*, in such a way that a given goal of *mj*' s be strengthened or generated anew. Notice that, since *mj* is an autonomous intelligent system, *Ii* must act on her beliefs in order to strengthen or generate new goals and modify her behaviours. We will address here the *goal-generation process*, as strengthening an existent goal is only a weaker case of cognitive influencing. To strengthen or generate a new goal, *mj* must acquire a new belief, say *Bjp* (*Ii* will harm *mj*, if she does not apply his will). This belief will activate a previous goal of *mj* 's, *Gjp* (avoid harm), and the interaction between *Bjp* and *Gjp* generates a new instrumental goal in *mj*'s, *Gjq* (adopt *Ii*'s will)[1] .

This is a social plan of action, which is based on a complex variant of the *theory of mind.* In the classic theory of mind (Leslie, 1991; Baron-Cohen, 1991; Dennett, 1987; Premack & Woodruff, 1978), others' mental states are harboured in one's mind, giving rise to *social beliefs*, namely beliefs about others' mental states (e.g. beliefs, intentions, desires, emotions, etc). In cognitive influencing, instead, the influencing entity has *social goals* as well, i.e. goals about others' mental states.

As we will see in the following sections, the presence and the type of cognitive influencing permits to discriminate between apparently similar enforcing mechanisms that are actually very different.

The three punishing strategies can be arranged on two axes: cognitive complexity and intentionality of deterrence. In this way, revenge easily appears to be the lowest in cognitive complexity and to pursue deterrence as an emergent and unintended self-reinforcing effect. The opposite is true for sanction (high cognitive complexity and intentional deterrence), whereas punishment occupies an intermediate position.

Revenge

Revenge appears to be a common human trait, widespread in human history and societies. According to the Merriam-Webster dictionary, vengeance is *punishment inflicted in retaliation for an injury or offense.* The retaliatory aspect is the main feature of revenge and is what makes this form of reaction differing from the two forms of punishment described below. Vengeance is also strongly characterized by the presence of emotional aspects that contribute to the common view of revenge as a not fully rational behaviour.

This "flavour of irrationality" could have contributed to the paucity of interest in revenge, compared to punishment, among scholars. While justifications for punishing and individual motives to punish have been widely investigated, research on retaliatory actions has considered them either as tribal and archaic forms of norm enforcement (Boehm, 1986) or as genetic predispositions evolutionary evolved to react to aggressions (Elster, 1990).

Amegashie and Runkel (2008) present a differential game model of revenge in conflicts. In their model, revenge has a positive value in economic terms; this means that, however destruction is costly, given what has been suffered in the past, the victim derives satisfaction and then utility from exacting revenge in the present. Similarly, deQuervain et al. (2004) used neuroscientific methodology to investigate how brain regions reacted to defection in an interaction game. According to Knutson (2004) their results show that punishing a defector activates brain regions related to the anticipation of a reward, even when punishment was costly, thus explaining human preference for punishing violators. Interestingly, Nikiforakis and Engelman (2008) reported data on revenge causing collaboration to decline in the lab but without boosting a chain of reciprocal vengeance.

Broadly speaking, the term 'revenge' refers to two diverse but connected phenomena. On one side, revenge is a social ritual that requires and prescribes specific behaviors to group members to repair an offense. The Kanun, a customary set of laws used mostly in northern Albania and Kosovo, disciplined people's reactions to murder (blood revenge or *gjakmarrje*) and other offenses (*hakmarrje*) according to the roles and degree of kinship of all the people involved. Shirking revenge or taking it without respecting what is stated in the Kanun lead to the same result: honour can not be restored and the whole family or clan is to blame. Shackleford (2005) considers "cultures of honor", in which revenge is the primary form of reaction to aggressions, likely to emerge and be maintained where the state is weak and can not prevent or punish theft.

It is worth noticing that in general retributive concepts of law and the creation of institutions are considered as advancements to replace vengeance and avoide blood feuds[2], but the Kanun itself was a social institutions aimed at preserving social order (*KLD*, 1989).

On the other side, revenge is an individual behaviour

[1] By means of the so called adoption rule (Conte & Castelfranchi, 1995), according to which an autonomous agent (adopter) will have another agent's (adoptee) goal as her own, if she, the adopter, comes to believe that the adoptee's achievement of this goal will increase the chances that the adopter will in turn achieve one of her previous goals.

[2] In this work we are not interested in analyzing the emergence and function of blood feud and we consider revenge in isolation

found both in human (Zaibert, 2006) and non-human primates (Jensen, Call, & Tomasello, 2007), reacting to personally harmful actions.

As Elster observes, revenge is "the attempt at some cost or risk to oneself, to impose suffering upon those who made one suffer, *because they have made one suffer* (emphasis added)".

In our view, revenge serves a terminal goal, that of making the aggressor suffer, and this excludes any other concerns. Usually, vengeance occurs in groups of equals, in which the offense is perceived also as an attempt to reduce an individual's prestige, to declass him or her family. Repaying the offense becomes a way to reaffirm one's status in front of both the aggressor and the social group and this behaviour is far from being extincted in present societies.

It is worth noticing that revenge may act as a deterrent from further aggressions, but this is an emergent function that can not be even represented in the avenger's mind. Revenge is not pursued to affect the likelihood that the wrongdoer will repeat the aggression in the future, inducing her to cooperate next time or deterring her from further aggressions. The avenger wants to repay the damage she suffered with an equal or greater offense, no matter how much risky or dangerous this retaliation is. In a sense, we can say that the avenger is a "backward looker" that revolves around the past and acts in the present to rebalance what happened, without any concerns for his future.

Into the avenger's mind

We claim that vengeance entails a specific configuration of goals and beliefs and that this configuration differs from those implied by terminal and instrumental punishment. This means that, although the punisher and the avenger could perform the same action, their aims and intentions were deeply different as well as the resulting state of the world.

In order to describe revenge, we need first to introduce its actors. There are at least three roles agents play in revenge. There is the avenger (A), the Target (T), and the Onlookers (O). The avenger's beliefs and goals involve both the target (T) and the onlookers (O), whose presence, as we shall see, is crucial.

Looking into the avenger's mind, we find a set of beliefs that are necessary to trigger the desire to take revenge[3]. The offended agent should, at least, believe that (1) the offense he received was intentional, (2) T was the main or the unique responsible and then liable for punishment, (3) there is a material and/or symbolic dimension to be restored in front of T and O.

The above set of belief should be paired with a set of goals, also necessary to trigger the retaliatory response. We identify at least three distinct goals: one referred to the material action, and the other two related to the influence the avenger wants to exert on the victim's and audience's representations. In fact, revenge is not motivated only by the desire of making the target suffering, but achieving this goal is pivotal to the objective of changing the target's and audience's beliefs about the avenger. What matters is what the others believe about the avenger and not what they are expected to do next time they are required to cooperate, as it is in punishment. In this case, cognitive influencing is aimed at modifying only the beliefs of the target and the onlookers, as depicted in Figure 2.

$$\boxed{(\mathrm{Gx}) \longrightarrow (\mathrm{By})}$$

Figure 1: Cognitive Influencing in Revenge

The avenger's action is driven by the following goals: first, the goal of imposing a suffering on the target; second, the goal of changing the target's beliefs, making her aware that the avenger does not passively accept the aggression and is able and willing to strike back at the aggressor (influencing the target). Finally, there is the goal of changing the beliefs of the onloookers (influencing the onlookers). In revenge the audience plays a crucial role because the damage suffered is not only material, but it usually has a strong symbolic component. Honour, for instance, is an intangible asset that can be threatened by the aggressor and that can be restored only if there is an audience in front of which the retaliatory action is performed and that recognizes that action as an attempt of restoring the initial situation.

This picture needs to be enriched by some additional considerations. First, the avenger can strike back at the aggressor's family or closer relatives, because they share some common traits. Posner (1980) views this issue the other way around: family obligation to retaliate is needed to make the threat of revenge work as a deterrent.

Another relevant issue is the cost-benefit analysis the avenger could carry out in order to choose the best conduct. According to Elster (1990), the retaliator does not calculate pros and cons of her action, but simply react to the offense. In our view, the avenger considers benefits and costs, but in her utility function there is an element that overrule any other consideration, that is the symbolic gain in terms of respect, honour, power, etc. the revenge allows to take.

Kant, I. (1952). The science of right (W. Hastie, Trans.). In R. Hutchins (Ed.), Great books of the Western world: Vol. 42. Kant (pp. 397 446).

[3]Here we are not concerned with the actual punishing behaviour chosen by the actor, but we are interested in investigating which behaviours he considers the most appropriate

Punishment

Punishment is a more controversial phenomenon, as showe by the two following definitions explaining the competing views on it:

> Punishment is the practice of imposing something unpleasant or aversive on a person or animal, usually in response to disobedient or morally wrong behavior (Stanford Encyclopedia of Philosophy, Punishment).

> [...] individuals (or groups) commonly respond to action likely to lower their fitness with behaviour that reduces the fitness of the instigator and discourages or prevents him or her from repeating the same action (Clutton-Brock & Parker, 1995).

According to the first view, punishment is meant to *righting* a wrong, while the second one stresses the influencing aim of punishment, that of discouraging or *preventing* an agent from repeating the same action.

The first one is a *retributive approach* to punishment: a person deserves a punishment that is *proportionate* to the moral wrong committed. Unlike revenge, punishment is proportionate to the offence. Immanuel Kant (Kant, 1952) argued that punishment can never be administered merely as a means for promoting another good and should be pronounced over all criminals proportionate to their internal wickedness (p. 397). Its justification lies in righting a wrong, not in achieving some future benefits. The punisher wants the victim to perceive punishment as a *natural consequence* of offence: the greater the offence, the greater the punishment. We can find such a view either in the lex talionis of early Roman law and in Old Testament and Koran.

In the second view, punishment is assigned a *deterrent effect*: it reduces the frequency and likelihood of future offences. This approach is referred to as utilitarian and is most often attributed to Jeremy Bentham (Bentham, 1962). Based on the rational choice model, deterrence theory works by modifying the *costs* and *benefits* allowed within the circumstances so that the criminal activity becomes an *unattractive* option[4].

According to these two views on punishment, we can say that the punisher is either a *backward-looker* and a *forward-looker*. The punisher aims to *repay* the damage she or someone else suffered with an offence *proportionate* to the one suffered, and to minimize the chance that the attacker will repeat the aggression in the future, thus *deterring* him from further hostility.

This enforcing mechanism, controlling modern societies, is not at all easy to distinguish from revenge, but we suggest that the punisher and the avenger are aimed at modifying the target and the audience's minds in different ways: unlike the latter, the punisher has the *explicit* goal to interrupt the chain of aggressions, with the further effect of preventing blood feuds and giving more stability to the social order.

Into the punisher' mind

Here follows a description of the punisher mental configuration - in terms of beliefs and goals. In order to trigger the punishing response, the offended agent should display the following beliefs (it is not necessary that he has all of them): (1) the damage/offense had a locus of responsibility then liable for punishment, (2) there is a material and/or symbolic damage to be refund and finally (3) the offense/damage will repeated in the future, so that punishment might be useful to avoid such a reiteration.

The above set of beliefs should be paired with a set of goals in order to trigger the punishing response. We identify the following set of distinct goals. More precisely, P aims at imposing an offence proportionate to the one suffered (*retributive goal*), and/or at establishing or mantaining a dominance hierarchy, and at deterring T (and possibly O) from further hostility (*deterrence goal*).

In order for the latter goal to be satisfied, P can employ different means, here we will focus on cognitive influencing. In order to achieve it, P has to act in such a way that the following belief is generated in T's an O's minds "P will harm me/will impose a cost to me, if I do not apply his will that the aggression will not be repeated in the future". This belief, *By*, will possibly activate a *previous* goal of T and O, *Gz* (avoid harm/avoid the costs of punishment), and the interaction between *By* and *Gz* will generate a new *instrumental* goal in T and O's minds, *Gy* (abstaing from repeating the aggression in the future). Social emotions - such as feeling of guilt - play a crucial role in achieving deterrence.

$$\boxed{\text{Gx } ((\text{By}) \longrightarrow (\text{Gy}))}$$

Figure 2: Cognitive Influencing in Punishment

Sanction

A particular case of punishment is that intended to deter future offences in observance no more of the punisher's will, but of a specific (social) *norm*. We refer to this case as (informal) sanction. In our view, a sanction is a particular case of cognitive influencing in which the sanctioner wants to modify the future action of T, making him form *two* beliefs at once: (i) a *normative belief* about the existence of a certain norm, and (ii) and the belief that T did violate that norm. Such a plan, which is incorporated to the act of sanctioning, is aimed at inducing the target to abstain from further offences not only

[4]It has to be said that deterrence can also be achieved through reinforcement learning, as suggested by behaviorism.

in order to avoid the sanction, but in order to *respect* the norm.

In our view, a norm - be it social, legal or moral - is a two-sided, internal (mental) and external (social), object, coming into existence only when it emerges, not only through the minds of the agents involved, but also *within* their minds (see (Conte & Castelfranchi, 2006; "On the Immergence of Norms: a Normative Agent Architecture", 2007). In other words, norms work as such only when agents recognize them and take decisions upon them as norms. Only when the normative, i.e. prescriptive, character of an input is recognized by the agent, that input gives rise to a normative behaviour of that agent. In order for the norm to be satisfied, it is not sufficient that the prescribed action is performed, but it is necessary to comply with the norm because of the *normative goal*, that is, the goal deriving from the recognition and subsequent adoption of the norm. Thus, for a norm-based behaviour to take place, a normative belief has to be generated into the minds of the norm addressees, and the corresponding normative goal has to be formed and pursued.

Unlike the punisher, the sanctioner aims at drawing the target's and the audience's attention on the existence and violation of the norm and on the fact that there is an high rate of surveillance. Our hypothesis is that sanctioning is characterized by a *signalling* function that has the aim of making explicit the casual link between violation and sanction: "you are being sanctioned because you violated that specific norm". Focusing T attention on the fact that the sanction is a consequence of a norm violation, possibly has the effect of encouraging the sanctionee to accept it as an *entitled* act, thus avoiding reiterated aggression (like in revenge) (see also (Bandura, 1991; Xiao & Hauser, 2009).

We also claim that sanction has the further effect, possibly aimed at by the sanctioner, to encourage the target to ground future decisions on *internal* evaluative criteria, established by the norm. This argument needs further elaboration, of course, and in order to test our hypothesis, we plan to conduct a series of laboratory experiments adopting a game-theoretical framework.

While imposing sanctions to them, we often request our children, pupils, etc. to observe the norm for its own sake. Isn't this behaviour irremediably paradoxical? However, it is far from an exception: it appears to be a *pedagogic* strategy rather frequent at least in Westernized societies. In sanction, the penalty is inflicted with the aim to favour a full autonomous compliance with the norm. How is this possible? A plausible explanation calls into question mechanisms of *norm internalization* (Durkheim, 1951; Scott, 1971; Gintis, 2004; Bicchieri, 2006; Bowles & Gintis, 2003). In particular, under conditions and by mechanisms that require specification (see also, ("On norm internalization", 2009), agents internalize external enforcement, converting it into self-enforcement, based on self-esteem and moral emotions, like the feeling of guilt.

Into the sanctioner's mind

In order to trigger the sanctioning response, the agent (S) should believe that (1) a norm has been violated. Regarding the motivations, there are at least two distinct goals that S aims to achieve. The first one is that of generating or reinforcing into the T's and O's minds a normative belief (NB) about the *existence* of a certain norm, and the belief that T did *violate* that norm. We will call this goal, a *pedagogic* goal. The second goal of S is that of making the norm be respected thus avoiding that the violation would happen again (*deterrence goal*) (Gx). In order for the latter goal to be satisfied, S has to act in such a way that the normative goal (I want to comply with the norm) will be activated. Once the normative goal has been activated, the agent will decide whether to adopt it or not. He can decide to obey a norm for instrumenental and terminal reasons. In the former case, the agent comply with the norm only to avoid punishment. In terminal norm adoption, agents decide to comply with the norm because "noms must be obeyed".

Such a plan, which is incorporated to the act of sanctioning, is aimed at inducing the target to abstain from further offences *not* only in order to avoid the sanction, *but* ideally in order to respect the norm. This kind of cognitive influence is the most complex, since it entails not only goals and beliefs but also the Normative Goal.

$$\boxed{\text{Gx } ((\text{NBy}) \longrightarrow (\text{NGy}))}$$

Figure 3: Cognitive Influencing in Sanctioning

To some extent the advantages of sanctions are easily identifiable: norm compliance is expected to be more robust than is the case when conducts are ruled only by external punishment: under ideal conditions agents abstain from violating because they want to respect the norm and not only in order to avoid punishment. Hence, sanctioned agents are expected to be more consistent and compliant than punished and endogenously motivated agents. A further consequence is that agents come to be much better at *defending* the norms: a consequence of the latter prediction is that sanction is decisive, if not indispensable, for *distributed* social control. A positive desired effect of sanction is an overall lowering of the costs associated to the social enforcement.

Conclusions and Future work

In this work we applied cognitive modelling to investigate the mental underpinnings of three different systems of norm enforcement: revenge, punishment and sanction. We argue that these are distinct behaviours people

choose in accordance with what they believe and want, thus entailing specific mental configurations. We also argue that without unraveling these cognitive bases, we can not fully explain complex phenomena like cooperation and altruistic punishment. Moreover, we claim that the transition from one to the other has been allowed by specific cognitive patterns, and suggesting that these mental mechanisms selected among given social structures, at the same time reinforcing and being reinforced by them. This preliminary model will be enriched by a simulation-based study of the different forms of enforcement.

References

(1989). New York: Gjonlekaj Publishing Company.

Amegashie, J. A., & Runkel, M. (2008). The desire for revenge and the dynamics of conflicts. 1–18.

Bandura, A. (1991). Social cognitive theory of moral thought and action. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development.* Hillsdale, NJ: Lawrence Erlbauml.

Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading.* Oxford: Basil Blackwell.

Bentham, J. (1962). Principles of penal law. In J. Bowring (Ed.), *The works of jeremy bentham.* New York: Russell and Russell.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms.* New York: Cambridge University Press.

Boehm, C. (1986). *Blood revenge: The enactment and management of conflict in montenegro and other tribal societies.* University of Pennsylvania Press.

Bowles, S., & Gintis, H. (2003). Origins of human cooperation. In P. Hammerstein (Ed.), *Genetic and cultural origins of cooperation.* Cambridge: MIT Press.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology*, *61*, 17–28.

Cialdini, R., & Goldstein, N. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591– 621.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*, 209–216.

Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action.* London: University College of London Press.

Conte, R., & Castelfranchi, C. (2006). The mental path of norms. *Ratio Juris*, *19(4)*.

Dennett, D. C. (1987). Reprint of intentional systems in cognitive ethology: The panglossian paradigm defended. *Brain and Behavioral Sciences*, *6*, 343–390.

deQuervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004, August). The neural basis of altruistic punishment. *Science*, *305*, 1254–1258.

Durkheim, E. (1951). *Suicide.* New York: The Free Press.

Elster, J. (1990, July). Norms of revenge. *Ethics*, *100*(4), 862–885.

Fehr, E., & Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90(4)*, 980–994.

Fehr, E., & Gachter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

Gintis, H. (2004). The genetic side of gene-culture coevolution: internalization of norms and prosocial emotions. *Journal of Economic Behavior and Organization*, *53*, 57–67.

Henrich, J., & Boyd, R. (2001). Why people punish defectors. weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208)*, 79–89.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767-1770.

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science*, *319 (5868)*, 1362–1367.

Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences USA*, *104*, 13046–13050.

Kant, I. (1952). The science of right (w. hastie, trans.). In R. Hutchins (Ed.), *Great books of the western world: Vol. 42. kant* (p. 397-446).

Knutson, B. (2004, August). Sweet revenge? *Science*, *305*, 1246–1247.

Leslie, A. M. (1991). Theory of mind impairment in autism. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading.* Oxford: Basil Blackwell.

Nikiforakis, N., & Engelman, D. (2008). Feuds in the laboratory? a social dilemma experiment. *Research Paper University of Melbourne*, *1058*, 1–31.

On norm internalization. (2009). In *Proceedings of the 6th european social simulation association conference.*

On the immergence of norms: a normative agent architecture. (2007). In *Emergent agents and socialities: Social and organizational aspects of intelligence. papers from the aaai fall symposium.*

Posner, R. A. (1980, January). Retribution and related concepts of punishment. *The Journal of Legal Studies*, *9*(1), 71–92.

Premack, D. G., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Brain and Behavioral Sciences*, *1*, 515–526.

Scott, J. (1971). *Internalization of norms: A sociological*

theory of moral commitment. Englewoods Cliffs, N.J.: Prentice-Hall.

Shackleford, T. (2005). An evolutionary psychological perspective on cultures of honor. *Evolutionary psychology*, *3*, 381–391.

Xiao, E., & Hauser, D. (2009). Avoiding the sharp tongue: Anticipated written messages promote fair economic exchange. *Journal of Economic Psychology*, *30(3)*.

Zaibert, L. (2006). Punishment and revengel. *Law and Philosophy*, *25*, 81–118.