Norms Through Minds

Giulia Andrighetto Francesca Giardini Rosaria Conte

Laboratory of Agent Based Social Simulation (LABSS)

Institute of Cognitive Science and Technologies (ISTC)

CNR

Rome, Italy

Abstract

The aim of this work is to enlighten the role of cognitive influencing in norm emergence and compliance. The paper unfolds as follows: in the first part, norm immergence will be described as a necessary mechanism for norm emergence; in the second part, a cognitive analysis of punishment will be provided and the role of this enforcement mechanisms in norm compliance will be shown. Some remarks and ideas for future work will conclude the paper.

1 Norms not just as constraints

Very often, the social scientific study of norms goes back to the philosophical tradition that defines norms as behavioural regularities emerging from reciprocal expectations (Lewis 1969; Bicchieri 2006; Epstein 2006). Indeed, interesting sociological works (Oliver 1993) point to norms as public goods, the provision of which is promoted by 2nd-order cooperation (Heckathorn 1988; Horne 2007). This view inspired the most recent work of evolutionary game-theorists (Gintis et al. 2003), who explored the effect of punishers or strong reciprocators on the group's fitness. In such a view, norms are aimed at creating and imposing constraints on agents in order to obtain a given coordinated collective

behaviour.

The other side of norms is often ignored: the function of norms of inducing new goals into the agents' minds in order to influence them (not) to do something. Norms do not only operate by reducing existing choices. Norms also work by adding new alternatives therefore generating new goals[1]: they influence us to do something that might have never entered in our mind otherwise. To understand how this process works, the mechanism of cognitive influencing has to be introduced.

2. Cognitive influencing

Cognitive influencing (see Conte and Castelfranchi 1995) is the process by which a given entity, say I_i, acts on another entity, m_j in such a way that a given goal of m_j's (e.g. to comply with the norm n_i) is strengthened or generated anew. Notice that, since m_j is an autonomous intelligent system, I_i must act on her beliefs in order to strengthen or generate new goals and modify her behaviours.

To strengthen or generate a new goal, m_j must acquire a new belief, say B_jp ("I_i will harm m_j if she does not apply its will"). This belief will activate a previous goal of m_j's, $G_{mj}p$ ("avoid harm"), and the interaction between B_jp and G_jp generates a new instrumental goal in m_j's, G_jq ("adopt I_i's will").

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 (Dennett, 1987; Premack and Woodruff, 1978), mental states of others are harboured in one's mind, giving rise to social beliefs, namely beliefs about others' beliefs and about others' goals. In other words, an agent reading the mind of another agent comes to have beliefs about some mental states (beliefs and goals) of him.

[1] Throughout the paper, we will speak of goals from the point of view of computer science and autonomous agent theory. In particular, a goal is a wanted world-state that triggers and guides action (see Conte, 2009).

In cognitive influencing, on the contrary the influencing entity has social *goals* as well, i.e. goals about (a) others' (m_j's) beliefs ($G_{Ii}B_{mj}q$) and (b) about others' goals ($G_{Ii}G_{mj}q$). In case (a), I_i wants m_j to have a new belief ($G_{Ii}B_{mj}p$) so that, for example, he will modify his opinion about something or someone (see § 4.1). In case (b), I_i wants m_j to have a new belief ($G_{Ii}B_{mj}p$) so that he will generate a new goal ($G_{mj}q$)[2] (see § 4.2; 4.3). Despite the mind reader, the influencing entity has not only the aim to read the minds of others, but also to change them.

In the following sections, we will point out the role of cognitive influencing in norm emergence (§3) and in norm compliance (§4).

3. Norm immergence

We consider a norm – be it social, legal or moral – as *"a prescribed guide for conduct which is generally complied with by the members of society*" **(**Ullman-Margalit 1977). It has to be pointed out that the prescription through which a norm is transmitted is a special one: a prescription that is *requested* to be adopted because it is a norm and it is *fully applied* only when it is complied with for its own sake (although this "felicity condition" rarely applies *de facto*, see § 4.3). Even normative commands are often adopted under the effect of reinforcement. Nonetheless, this type of adoption is not satisfactory, so to speak, from the norm's point of view, if any such a perspective can ever be hypothesized. The *happiness condition* is that the norm is accepted, to say it in Hart's terms (1961), or internalized, to state it in Durkheim's terms (1951), because it is recognized as a norm. In other words, 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.

[2] In fact, I_i must transmit more than one belief to m_j (at least, for m_j to form the wanted goal, I_i must convince her that his harming power is credible).

Formatted: Font: 10 pt

Formatted: Font: 10 pt

Thus, for a norm-based behavior 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. Drawing upon Kelsen (1979), von Wright (1963) and a long tradition of deontic philosophy and logic-based theory of action, we define a normative belief as a belief that a given behaviour, in a given context, for a given set of agents, is either forbidden, obligatory, or permitted (Conte and Castelfranchi 1999; 2006). Our claim is that a norm emerges as a norm only when it is incorporated into the minds of the agents involved (Conte and Castelfranchi 1999; 2006); in other words, when agents recognize it as such.

In previous works (Castelfranchi, 1998, Conte et al. 2007), we described the process of norm emergence as a gradual and complex dynamics by which the macro-social effect, in our case a specific norm, emerges in the society *while* immerging in the minds of the agents producing it, generating a number of intermediate loops. The generation/emergence of social norms is a major circuit made of local loops, in which:

- partial or initial observable macroscopic effects of local behaviours occur
- retroact on (a subset of) the observers' minds, modifying them (producing new internal states, emotions, normative goals, normative beliefs, etc.)
- agents communicate internal states to one another, thus activating a process of normative influencing (see Conte and Dignum, 2001)
- these normative beliefs spread through agents' minds
- behaviours progressively conform to spreading states
- initial macroscopic effects get reinforced/weakened depending on the type of mental states spreading.

Thus, before any global effect emerges, specific local events affect the generating systems, their beliefs and goals, in such a way that agents influence one another into converging on one global macroscopic effect[3].

Once norms are installed in their minds, social members are enabled to behave accordingly. But this does not mean they will do so. The problem is that autonomous cognitive agents need to have internal reasons, i.e. goals, for doing something; their actions are intentional, motivated, and follow some decision. Hence, the question is, why should agents decide to conform to a recognized norm? In the following section, we will provide a very preliminary analysis of three different mechanisms of enforcement, revenge, punishment and sanction, aiming to show how they differ in influencing agents to behave in a certain way. Finally, we will focus on the sanctioning system in order to show its role in enforcing or producing the will to comply with a norm.

4 Enforcing Mechanisms

Theoretical and empirical studies about punishment in animal and human societies have demonstrated that this behaviour promotes and sustains cooperation in large groups of unrelated individuals (Fehr and Gatcher 2002; Gintis 2000, Sober and Wilson 1998). Although very interesting, these accounts present a common feature: they refer to punishment as a single action, a broad behavioural response under which different kinds of enforcment mechanisms are concealed.

We claim that punishment makes reference to a full behavioural repertoire rather than to a single behaviour; in particular we consider decisive to distinguish at least among revenge, punishment and sanction (even if other intermediate steps can be identified, see Andrighetto and Giardini in preparation). These are different forms of punishment that

[3] For an agent based model and simulation results describing the complex loop of norms' emergence see Andrighetto et al. 2007; 2008, Campennì et al. 2008).

Formatted: Font: 10 pt

stem from different beliefs and are aimed at achieving specific goals. From a pure behavioural point of view, an agent punishing another agent or taking revenge against him could apparently act in the same way. When we see a mother reproaching her child, we probably think that she is sanctioning him for doing something wrong, rather than taking revenge against him. But why? Which kinds of beliefs and goals are we attributing to the mother in order to consider her a sanctioner rather than an avenger? In the following analysis, still very preliminary and calling for further elaboration (see Andrighetto and Giardini, in preparation), we divide social punishing acts into three categories: revenge, punishment and sanction, and we provide a description of the mental configurations (in terms of beliefs and goals) characterizing them. We arrange these social reactions on an evolutionary trajectory that moves from revenge, considered as cognitively less complex, to punishment and sanction. More specifically, they differ in the kind of cognitive influencing they are aimed to obtain: from a simpler one aimed only to modify agents' beliefs to a more complex one aimed to change beliefs in order to also modify agents' goals.

Another dimension of change consists in the fact that the same motivation, as for instance "deterrence", is intentionally pursued in some actions (e.g. punishment and sanction), while in other cases is an emergent and unintended self-reinforcing effect (as it is in revenge)

4.1 Revenge

A preliminary distinction we need to draw regards social and non-social reaction to aggression or social damage. An action is social when it is meant to achieve a goal that mentions another agent's mental states (see § 2) and this allows us to rule out actions that are not meant to modify other agents' mental states. For instance, when retaliation is only aimed at restoring material power e.g., taking back the pen that someone has stolen, it is

not a social action. On the contrary, inducing some negative emotions in the initial aggressor, or making her learn that stealing pens is wrong, is a social action in the fullest sense.

Revenge, according to the Merriam-Webster dictionary is “punishment inflicted in retaliation for an injury or offence” or, in Elster’s terms (1990) it 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” (p. 862).

Vengeance 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. Long term, strategic planning does not seem to characterize it. The avenger wants to repay the damage she suffered with an *equal* or *greater* offence, no matter how risky or dangerous this retaliation is. Ideally, we can say that the avenger is a "backward-looker" that revolves around the past and acts in the present to rebalance what happened, with no concern for the future.

Revenge is motivated not only by the desire of making the target suffer, but also by the goal of changing the target's and audience's beliefs about the avenger, in order to restore the image that has been damaged by the aggression suffered. The avenger aims at

- repaying the damage he suffered with an equal or greater offence in order to
- change the target's and the audience's *beliefs* about himself: he wants to restore his image, damaged by the aggression suffered.

Presumably, the greater the offence, the more effective the image restoration. This argument needs further elaboration and in particular an analysis of how someone’s image can be restored has to be developed (see Andrighetto and Giardini, in preparation).

We claim that revenge is aimed to modify what others believe about the avenger. Cognitive influencing in revenge is directed towards the agents’ beliefs: a_i, the avenger,

has the goal to act on a_j's beliefs ($G_i B_j p$), in order to re-establish a symmetry in dominance unbalanced by the aggression suffered.

Ethnographic studies highlighted the transition from tribal to modern societies, in which retributive concepts of law and the creation of institutions replaced vengeance and avoided blood feuds (Boehm, 1986)[4]. Posner (1980) suggests that revenge and retribution may be partially determined by historical and economic circumstances, such as the private enforcement of law and high probabilities of detecting and punishing offences. When these conditions are met, a pure vengeance system may appear, although it is unlikely to be optimal.

We claim that to understand and estimate the effectiveness of a vengeance system in achieving and maintaining social order, it is essential to make explicit how this enforcement mechanism acts on the agents' minds, and in particular which is the kind of mind changing it is aimed to produce.

4.2 Revenge Vs punishment

Punishment is a more controversial phenomenon, as shown by the two following definitions:

> *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 and Parker, 1995).

According to the former view, punishment is meant to right a wrong, while the second one stresses the influencing aim of punishment, the one of discouraging or preventing an

[4] These systems are not completely extinguished, as the culture of honour in the southern United States demonstrates (Cohen and Nisbett, 1994; 1997).

Formatted: Font: Arial, 10 pt

agent from repeating the same action. Two competing justifications for the use of punishment follow from these two definitions (Carlsmith et al. 2002; Posner, 1980).

The first one is the *just desert* rationale (or, using another terminology, the retributivist approach to punishment): a person deserves punishment proportionate to the moral wrong committed. 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 central precept of this view is that punishment has to be proportionate to the harm. The punishment is an end in itself and needs no further justification. We can find such a view either in the *lex talionis* of early Roman law and in Old Testament and Koran.

A rival approach is the deterrence rationale: punishing an offender reduces the frequency and likelihood of future offences. This approach is referred as utilitarian and is most often associated with Jeremy Bentham. He argued that "general prevention ought to be the chief end of punishment, as it is its real justification" (Bentham, 1962). Deterrence theory works by changing the costs and benefits of the situation so that criminal activity becomes an unattractive option. It is based on a rational choice model. The economic rationale of punishment can be easily included in the deterrence approach (for a discussion of this point, see Posner, 1980).

From these two approaches to punishment, it follows that this practice must be justified by reference either to forward-looking or to backward-looking considerations. The former prevails in the utilitarian approach, its aim being to increase overall net social welfare by reducing (ideally, preventing) crime, the latter prevails in the just deserts view, in which punishment is seen either as a good in itself or as a practice required by justice, thus making a direct claim on our allegiance.

We consider the dichotomy between the just desert and deterrence rationale misleading, because they address two different phenomena. While the former approach refers to a behaviour exhibiting strong commonalities with revenge, only the latter deals with real punishment. This difference can hardly be realized if only behaviour is taken into account. Conversely, if we model and compare the mind of the avenger and the mind of the punisher, the difference becomes quite evident.

Our claim is that each type of agent wants to realize different world states: unlike the avenger, the punisher wants to act on the target's beliefs (G_iB_jp) in order to modify her goals.

	Gx (By)	Gx((By) -> (Gy))
Revenge	X	
Punishment	X	X

Formatted: Centered, Line spacing: Double

Formatted: Centered, Line spacing: Double

Formatted: Centered, Line spacing: Double

The avenger has the goal to change the target's beliefs in order to re-establish a symmetry altered by the aggression suffered. The punisher has the goal to change (a) the target's beliefs (he wants to be believed to be efficacious so that his will is likely to be respected), and (b) the target's goals, thus *dissuading* her from future aggressions. Unlike avenger, punisher is a forward looker: he wants to change the target's mind in order to orient her future beheviour. Therefore, punishment implies a more sophisticated cognitive equipment than revenge: a different sort of mind changing is at place. The punisher intentionally deter the target from aggressing him again, generating in her mind the goal of respecting the dominant.

Punisher wants to be respected by the target, but what does it mean? Respect (see Bagnoli 2003, 2007; Darwall 1977) is more than a single belief, it is a hybrid mental pattern (other examples in social cognition are expectations and social values), describing a state of the world *believed and wanted* at the same time. A respected leader is one that is

believed to deserve compliance, and therefore often complied with. Hence, by acting on the social beliefs of the target, the punisher induces her to form the corresponding goal (G_iG_jp) of actively respecting the dominant, by complying with his commands and reinforcing the belief that he deserves respect. By this means, the cognitive pattern is self-reinforced while reinforcing the social structure: future aggressions by the target are averted, as respect is a more efficient deterrent than actual punishment.

Unlike revenge, punishment is *proportionate* to the offence. More than making the punisher feel satisfied, it is necessary to cognitive influencing. The punisher wants the victim to perceive punishment as a natural consequence of offence: the greater the offence, the greater the punishment. But still more important is the target learning not to defy the punisher's dominance, in order to respect it in the future. The abovementioned hypothesis is still very preliminary and a deeper analysis of what respect is is urgently required.

4.3 Punishment Vs sanction

A particular case of punishment is the one intended to deter future offences in observance no more of the punisher, 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 signal that a norm has been violated. As punisher, sanctioner wants to modify the future action of the target, but in a specific way: making the norm immerge into the target's mind, so as to induce her to abstain from further offences in order to respect the norm. Norm immergence leads the target to form a new belief, namely a normative belief ("there is a norm prohibiting, prescribing, permitting..."), and the corresponding *normative goal*, that is, the goal deriving from the recognition and subsequent adoption of the norm (Andrighetto et al. 2007).

	Gx (By)	Gx((By) -> (Gy))	*Gx((NBy) -> (NGy))*
Revenge	X		
Punishment		X	
Sanction			X

Formatted: Centered, Line spacing: Double
Formatted: Line spacing: Double
Formatted: Centered, Line spacing: Double
Formatted: Line spacing: Double
Formatted: Centered, Line spacing: Double
Formatted: Line spacing: Double
Formatted: Centered, Line spacing: Double
Formatted: Line spacing: Double

In sanction a form of mind changing different from revenge and punishment is involved. The avenger has the goal to generate a new belief in the target's mind in order to re-establish a symmetry altered by the aggression suffered; the punisher has the goal to change the target's belief (he wants to be believed to be potent so that his will is likely to be respected), and (b) the target's goals, thus deterring her from future aggressions; the sanctioner has the goal to generate a particular kind of belief, i.e. a normative belief in to the target's mind and the corresponding normative goal, in order to induce him to observe the norm. Despite punisher, sanctioner acts on the norm's interest: the kinds of beliefs and goals he wants to generate and enforce into the target's mind have not a personal content, (e.g. I want to be believed to be potent), but they refer to the norm itself. While in punishment deterrence is based on the belief that the punisher is powerful and it has to be respected, in sanction it is based on the belief that there is a norm that "must be respected".

Once the normative goal has been generated, the addressee will decide whether to adopt it or not. He can decide to do his duty, to adhere to and obey a norm for several higher motives:

- instrumental reasoning: prizes or sanctions enforcing the norm, including others' approval and reputation;

- terminal goal: she has the terminal goal or value that "Norms be respected" (Kantian morality).

Agents reason about the norm, evaluate the consequences of violating it or not especially in the first case. For example, he can evaluate whether or not to stop at the red light at a desert crossroad, where no vehicles or policemen are passing by, and in which chances of collision or sanction consequent to transgression are therefore low. However, in our view this is not how norms are ideally aimed to work. They are rather sub-ideal cases based on norm-enforcement. As said in § 2, the felicity condition of the norm is fully applied only when the norm is complied with for its own sake. The normative imperative is aimed at being adhered to as such, just because it is a norm. Norms are aimed at being adopted, because they are norms and “norms must be obeyed”.

Nonetheless, either in ideal than in sub-ideal norm adoption, agents behaviour results from norm immergence: they act in accordance with what they believe to be a norm, if only to avoid the sanction that they have learned to expect from its violation.

Unlike punishment, 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. Sanction gives some autonomy to the agent: he will abstain from future aggressions/violations not only because of the respect of the leader, but because under some ideal conditions he wants to satisfy an impersonal will, the one expressed by the norm.

While imposing sanctions to them, we often request our children, pupils, etc. to observe the norm for their own sake. Isn’t this behaviour dreadfully inconsistent and irremediably paradoxical? Indeed, it is. However, it is far from an exception: it appears to be a pedagogic strategy rather frequently at least in our, Westernized societies. In the sanction, the penalty is inflicted with the aim to favour a full autonomous compliance with the norm,

which should verify the conditions for its complete satisfaction. How is this possible? A plausible explanation calls into question mechanisms of norm internalization (Durkheim, 1951; Scott, 1971; Gintis, 2004; Bicchieri, 2006, etc.). In particular, under conditions and by mechanisms that require specification (see also, Conte et al. 2009), agents internalize external enforcement, converting it into self-enforcement, based on self-esteem and moral emotions, like the feeling of guilt.

5 Concluding remarks

In this paper, we have discussed the role of cognitive influencing in norm emergence and compliance. After an initial analysis norms as social phenomena working not just by limiting some possible agents' choices but also by adding new alternatives and generating new goals, we described the process of immergence as necessary for norms to emerge in society. Then, we examined three different systems of enforcement, retaliation, punishment and sanction, arguing that the transition from one to the other has been allowed by specific cognitive patterns. We finally focused on sanction showing how this mechanism has to act on the agents' minds in order to induce them respect the norm.

In future works, we aim to carry out a simulation-based study of the different forms of enforcement, from retaliation to sanction, outlined above, and check whether and under what internal and external conditions they evolve, and what are their further effects both at the mental and social level.

6 Acknowledgments

This work was supported by

- the EMIL project (IST-033841), funded by the Future and Emerging Technologies program of the European Commission, in the framework of the

initiative Simulating Emergent Properties in Complex Systems.

- the European Science Foundation EUROCORES Programme TECT, funded by the Italian National Research Council (CNR) and the EC Sixth Framework Programme.

7 References

Andrighetto, G., Campennì, M, Conte, R., Paolucci, M. (2007). On the Immergence of Norms: a Normative Agent Architecture. In Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence, Washington DC.

Andrighetto, G., Campennì, M., Conte, R., Cecconi, F. (2008). How Agents Find out Norms: A Simulation Based Model of Norm Innovation. In *3rd International Workshop on Normative Multiagent Systems (NorMAS 2008)* 15-16 July, 2008, Luxembourg.

Andrighetto, G., Giardini, F. (in preparation). A Cognitive Model of Punishment.

Bentham, J. (1962). Principles of penal law. In John Bowring (Ed.), The Works of Jeremy Bentham (p. 396). New York: Russell and Russell.

Bicchieri, C. (2006). The Grammar of Society: the Nature and Dynamics of Social Norms, Cambridge University Press.

Boehm, C. (1986). Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies. Philadelphia: University of Pennsylvania Press.

Campennì, M., Andrighetto, G., Cecconi, F., Conte, R. (2008). Normal = Normative? The Role of Intelligent Agents in Norm Innovation. In The Fifth Conference of the European Social Brescia, September 1-5, 2008.

Carlsmith, K., Darley, J., & Robinson, P. H. (2002). Why do we punish? deterrence and just deserts as motives for punishment. Journal of Personality and Social Psychology, 83(284-299).

Castelfranchi, C. (1998). Simulating with cognitive agents: The importance of cognitive emergence. Multi-Agent Systems and Agent-Based Simulation, Heidelberg.

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

Cohen, D. and Nisbett, R. E. (1994). Self-protection and the culture of honor: Explaining southern homicide. Personality and Social Psychology Bulletin, 20, 551-567.

Cohen, D. and Nisbett, R. E. (1997). Field experiments examining the culture of honor: The role of institutions in perpetuating norms about violence. Personality and Social Psychology Bulletin, 23, 1188-1199.

Conte, R., G. Andrighetto, M. Campenni, and M. Paolucci. (2007). Emergent and immergent effects in complex social systems. In Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence.

Conte, R., Andrighetto, G., Campennì, M. (2009). On Norm Internalization. A Position Paper. The Sixth Conference of The European Social Simulation Association (ESSA 09), University of Surrey, Guildford, September, 2009 (submitted).

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

Conte, R. and Castelfranchi, C. (1999). From conventions to prescriptions. towards a unified theory of norms. AI and Law, 7:323–340.

Conte, R. and Castelfranchi, C. (2006). The mental path of norms. Ratio Juris, 19(4):501-517.

Conte, R and Dignum, F. (2001). From Social Monitoring to Normative Influence. *Journal of Artificial Societies and Social Simulation* vol. 4, no. 2, http://www.soc.surrey.ac.uk/JASSS/4/2/7.html

Dennett, D. (1987). The Intentional Stance. Cambridge, Mass: MIT Press/Bradford Books.

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

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

Epstein, J. (2006). Generative Social Science. Studies in Agent-Based Computational Modeling. Princeton-New York: Princeton University Press.

Fehr, Ernst & Simon Gatcher. 2000. “Cooperation and Punishment in Public Goods Experiments.” The American Economic Review 90(4):980–994.

Gintis, H. (2000.) Strong Reciprocity and Human Sociality. Working papers 2000-02, University of Massachusetts Amherst, Department of Economics

Gintis, H. (2004). The genetic side of gene-culture coevolution: internalization of norms and prosocial emotions. Journal of Economic Behavior & Organization, 53(1):57-67.

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

Lewis. D. K. (1969). Convention: A Philosophical Study. Cambridge Mass.: Harvard University Press.

Hart, H. L. A. (1968). *Prolegomenon to the Principles of Punishment.* (1959), reprinted in Hart, Punishment and Responsibility, Oxford University Press, pp. 1-27.

Heckathorn, D. (1988). Collective sanctions and the compliance norms - a formal theory of group-mediated social-control. American Journal of Sociology, (94):535– 562.

Horne, C. (2007). Explaining norm enforcement. Rationality and Society, (19(2)):139– 170.

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). Chicago: Encyclopedia Britannica.

Kelsen, H. (1979). General Theory of Norms. Hardcover.

Oliver, P. E. (1993). Formal models of collective action. Annual Review of Sociology, (19):271–300.

Posner, R. (1980). Retribution and related concepts of punishment. Journal of Legal Studies, 9, 71–

92.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a `theory of mind'? Behaviour and Brain Sciences, 4, 515-526.

Scott, J. F. (1971). Internalization of Norms: A Sociological Theory of Moral Commitment, Englewoods Cliffs, N.J.: Prentice-Hall.

Sober E. and Wilson D.S. 1998. Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.

von Wright, G. H. (1963). Norm and Action. A Logical Inquiry. Routledge and Kegan Paul, London.