Internalizing Norms:  
A Cognitive Model of (Social)  
Norms’ Internalization

Rosaria Conte, ISTC, Institute of Cognitive Sciences and Technologies, Italy  
Giulia Andrighetto, ISTC, Institute of Cognitive Sciences and Technologies, Italy  
Marco Campennì, LABSS - ISTC, Institute of Cognitive Sciences and Technologies and  
Università degli Studi di Roma “Tor Vergata”, Italy

ABSTRACT

Internalization is at study in social-behavioural sciences and moral philosophy; lately, the debate was revamped within the rationality approach to the study of cooperation and compliance because internalization is a less costly and more reliable enforcement system than social control. But how does it work? So far, poor attention was paid to mental underpinnings of internalization. This article advocates a rich cognitive model of different types, degrees and factors of internalization. In future work, it will be implemented on a normative agent platform to simulate the individual and social effect of internalization.

Keywords:  Agent Architecture, Autonomous Agents, Internalization, Norms, Rich Cognitive Modeling

INTRODUCTION

The Issue

The problem social scientists still revolve around is how autonomous systems, like living beings, perform positive behaviors toward one another and comply with existing norms, especially since self-regarding agents are much better-off than other-regarding agents at within-group competition. Since Durkheim, the key to solving the puzzle is found in the theory of internalization of norms (Mead, 1963; Parsons, 1967; Grusec & Kuczynski, 1997; Gintis, 2003).

Norm internalization is one of the red herrings running across all of the social-behavioral disciplines, and there are not many. Not only sociologists but also developmental, social and cognitive psychologists have perceived its crucial role in socialization. Drawing on the early work by Vygotsky (published in the US as late as 1978) and Piaget (1978), psychologists showed that a parental attitude oriented to elicit norm internalization predicts children’s later wellbeing and even their inclination to other-regarding behavior (Ryan & Deci, 2000).

Nonetheless, our scientific definition and understanding of the process of norm inter-
nalization is still fragmentary and insufficient. The main purpose of this article is to argue for the necessity of a rich cognitive modelling of norm internalization in order to (a) provide a unifying view of the phenomenon, accounting for the features it shares with related phenomena (e.g., robust conformity as in automatic behavior) and the specific properties that keep it distinct from them (autonomy); (b) model the process of internalization, i.e. its proximate causes (as compared to the distal, evolutionary, ones; see Gintis, 2003, 2004); (c) characterize it as a progressive process, occurring at various levels of depth and giving rise to more or less robust compliance; and finally (d) allow for flexible conformity, enabling agents to retrieve full control (Barugh et al., 2001) over norms which have been converted into automatic behavioral responses (Epstein, 2007). Thanks to such a model, it will be possible to adapt existing agent architectures (such as EMIL-A, cf. Andrighetto et al., 2007) and simulation platforms (EMIL-S, see Troitzsch, 2008) to test hypotheses concerning (a) individual and social effects of internalization, (b) factors favoring or hindering internalization, and (c) the evolution of internalization in future societies.

Throughout the article, the process of norm internalization is meant as a mental process that takes (social) norms as inputs and gives new goals of the internalizing agent (from now on, the internalizer) as outputs. Emotions, playing a significant but not necessary role in this process, will not be investigated at this stage.

Related Work

Contributions to explain internalization are sometimes based on reinforcement learning theory. Scott (1971), for example, theorized that norm internalization leads to robust compliance, provided the external sanctioning system is never completely abandoned. Unfortunately, this explanation is insufficient: it is incompatible not only with the view that “...social norms can get internalized to the extent that they do not need social enforcement” (Basu, 1998, p. 2), but also with experimental evidence. For example, subjects playing ultimatum games are found to follow fairness considerations even when unobserved (Bicchieri, 2006).

In the last few years, a strong renewal of interest around the notion of norm internalization appeared in the evolutionary game theoretic study of cooperation and prosocial behaviour. Gintis (2003) argued that increase in social complexity of early human society produced a rapidly changing environment, which in turn posed an adaptation problem to genetic mechanisms for altering goals. Internalization of norms is adaptive because it “facilitates the transformation of drives, needs, desires, and pleasures into forms that are more closely aligned with fitness maximization.” But how did it evolve?

Some authors (Bicchieri, 2006; Epstein, 2007) conceive of norm internalization as a process leading to a sort of automatic, or thoughtless conformity. People, observed Epstein (2007), blindly conform to the norm: the more they have done so in the past, the more they will redo it in the future. Agents learn not only which norms to conform to, but also how much they should think about them. In the author’s view, internalization is learning not to think about norms. Bicchieri (2006) has a sophisticated explanation, which leads to an equivalent conclusion. Agents learn what the norms are through shared expectations, which by definition they prefer to correspond to. Once found out what the norms are, they organize their beliefs into script-like structures, including the contexts, relationships and conditions in which they found out the norms. When later such belief structures are activated by ongoing activities or contexts, the corresponding norms will be activated as well, and conform to thoughtlessly. Bicchieri’s model was intended to be a model of norm compliance. The question is, does it also account for norm internalization, i.e. the phenomenon pointed out by Durkheim, Hart, etc. and defined by Scott as compliance “at a temporal or spatial remove from sanctions” (1971, p. xiii; italics added)?
Distributed Agency

[www.igi-global.com/article/distributed-agency/77664?camid=4v1a](www.igi-global.com/article/distributed-agency/77664?camid=4v1a)