Martin Neumann (2010)

Norm Internalisation in Human and Artificial Intelligence

Journal of Artificial Societies and Social Simulation **13** (1) 12
<http://jasss.soc.surrey.ac.uk/13/1/12.html>

Received: 28-Oct-2008 Accepted: 31-Dec-2009 Published: 31-Jan-2010

Abstract

In this article, principles of architectures relating to normative agents are evaluated with regard to the question whether and to what extend results of empirical research are incorporated in the architecture. In the human sciences, internalisation is a crucial element within the concept of norms. Internalisation distinguishes normative behaviour regulation from mere coercion. The aim of this article is to begin answering the question of to what extent normative agent architectures represent the theoretical construct of norm internalisation. The relevant research in this area may be found in socialisation research in psychology and sociology. Evaluation of conclusions from the empirical sciences allows to identify drawbacks and opportunities in existing architectures, as well as to develop suggestions for future development.

Keywords: Normative Agent Architectures, Norm Internalisation, Socialisation Theories, Theoretical Validity

Introduction

1.1 The development of normative agents constitutes one of the most active research fields in the domain of agent-based simulation (Meyer et al. 2009). Normative agent systems are of vital importance for theoretical research into the foundation of social order (Conte and Dellarocas 2001), as well as for the practical purposes such as e-commerce (Garcia-Camino et al. 2006; Vazquez-Salceda et al. 2005). Despite this theoretical and practical importance, the concept of norms is commonly introduced in the AI literature without much discussion. It is presumed that our intuition with regard to norms is clear and straightforward (Neumann 2008a). The evaluation of software architectures against empirical research provides a criterion for the assessment of simulation models that to date remains largely unemployed, namely, an assessment comparable to what is known as construct validity in survey research: to what extent do models of normative agents represent the theoretical construct of norms in target systems? In fact, the development of normative architectures to date has remained essentially uninfluenced by the conclusions of psychology, pedagogy or neurophysiology (Guerin 2008; Nullmeier 2004). The aim of the paper is thus to evaluate the state of the art in relation to architectures of normative agents, as measured against the state of the art with regard to the target system.

1.2 Research into normative software architectures should benefit from this approach in two respects: first, due to a critical appraisal of current architectures from the perspective of related social and psychological theories. Where does contemporary simulation research of normative agents stand in comparison to the cutting edge in sciences that are associated with the target system? Such a comparison will enable us to identify drawbacks and likewise opportunities in existing architectures. How realistic is the architecture, i.e. in so far as it is possible to justify modelling assumptions through findings in the empirical sciences? Secondly, what follows from a critical assessment of current approaches is the ability to generate suggestions for the development of future architectures. This is of particular relevance for models with the theoretical purpose to comprehend the wheels of social order. A theoretically well-founded design of individual normative agents provides a cross-validation of macro-level simulation results. For practical and engineering purposes, questions of utility rather than validity represent the decisive factor.

However, as the example of biologically inspired computing demonstrates, in the long run theoretical research might even be useful for engineering purposes.

1.3 However, a great variety of approaches to norms exist in the literature. This holds for the simulation literature as well as for the empirical sciences. One reason for the variety of concepts of norms is that their investigation is scattered over a vast range of different disciplines. For this reason, the examination has to concentrate on specific issues: an individual may become accepted into wider society through a variety of means. Norm internalisation, it is argued, represents one of the stronger mechanisms by which this process occurs. Since norm internalisation suggest social regulation of individual behaviour, as distinguished from mere coercion, the process of norm internalisation is of especial importance for an investigation of normative architectures undertaken from an empirical perspective. Indeed, the development of moral sentiments with subjectively binding force is a unique feature of the human species. The problem of internalisation concerns the mechanisms how individual humans develop this characteristic trait[1]. Although in the past decades consideration has been given to understanding processes of internalisation over an individual's entire life span, the central processes arguably take place during childhood. Indeed, the transmission of cultural values has even been denoted as a 'second birth' (Claessens 1972). The mechanisms by which this occurs are the focus of socialisation research. Socialisation is the bridge between the individual and society (Krappmann 2006).

1.4 However, it is evident that such a confinement also implies that an investigation of important findings in other fields has to be left for other research. For instance, since the literature on norms found in the political sciences or theories of law is more concerned with external behaviour regulation (i.e. laws), an examination of this research is excluded here. The reader may refer to Conte and Castelfranchi (1999) for a broad overview. An investigation of the mechanisms of norm internalisation has to be distinguished also from the macro-sociological question of how norms, once established, contribute to human behaviour regulation and correspondingly to the establishment of social order. The question how this can be modelled is investigated e.g. by Neumann (2008b). Moreover, a number of accounts that apply deontic logic to the development of normative architectures already exist. An overview of this research field is provided e.g. by Boella et al. (2007). A final selection criterion is motivated by the fact that, typically, Artificial Societies investigate stylised facts. This suggests further focusing the examination by excluding historical narratives that can be found in anthropological studies or in historical sociology (such as the work of Michel Foucault or Norbert Elias). Without doubt, this work is relevant. However, the notion of stylised fact suggests to outline a general theoretical perspective, rather than a comprehension of historical contingency.

1.5 The paper is structured in the following manner: Section 2 specifies the research question of theoretical validity in more detail. In Section 3 principles of one prominent example of a normative architecture, the BOID architecture, are introduced. Section 4 provides a critical examination of theoretical validity of these principles. Section 5 examines the particular perspective of the BOID architecture and possible directions for the design of further architectures. Section 6 concludes with a summary of what can be learned from this examination.

Modelling a theoretical construct

2.1 The question to what extend simulation models represent theoretical constructs belongs to the problem field of validation. Validation is an expanding research field. First and foremost, it has to be checked if simulation runs replicate findings from the social domain. However, a number of surveys of validating simulation studies emphasise that validation should be regarded as a multi-stage process and that the dichotomy between the context of discovery and that of justification cannot be maintained (e.g. Barlas 1996; Balci 1997; Moss 2008). On the contrary, validation must be conducted throughout the entire Life Cycle of a Simulation Study (Balci 1997). Hence a feedback loop between the model and the target system is established by which the model is calibrated against empirical data. Dependent on the objective of the model, this process can assume various forms. Some researchers concentrate on the calibration of parameters of a given theoretical structure (Windrum et al. 2007). Formal techniques present themselves as a means of calibrating the model to attain as accurate a representation of numerical data as possible. Such techniques might also include qualitative tests, for instance extreme value behaviour or a boundary adequacy test (Barlas 1996). Such methods focus primarily on an analysis of behaviour patterns of the simulation results. Researchers engaged in participatory modelling (Barreteau et al. 2003; Geller and Moss 2007) aim to incorporate the qualitative richness of narrative scenarios (Moss 2008). In this case, the validation process is less attuned to a formal analysis of simulation results (even though this is possible: cf. Moss and Edmonds 2005), than to a realistically descriptive specification of individual behaviour and social interaction in the design of the agents and their interactions (Moss 2008).

2.2 However, surprisingly little attention is paid to theory. Either a particular theoretical framework is simply taken for granted (this is especially so with economic models). In other cases, models are intended as a test of the consistency of a certain theory (e.g. Jacobsen and Bronson 1997). In such situations, it is the theory itself that is under scrutiny. Less attention is paid to the question of whether the model in fact is a correct representation of the theory. Other approaches deliberately start with the absence of a theoretical framework, aiming at a direct representation of empirical evidence without an intervening theory (Moss 2008). However, the questions of a) the choice of a certain theory, and b) how far the theory is then represented in the model, is to a far lesser extent examined in the literature on validating simulation models. Nevertheless, it will turn out that these questions are of particular relevance for modelling norm internalisation: on the one hand a number of different theories exist in this field. Evidence from the empirical science is ambiguous. For this reason it impossible to simply identify the 'best of our knowledge' to put it into the agents. On the other hand, it is hard to obtain direct empirical evidence of the mental processes at work. For this reason modelling norm internalisation cannot rely purely on observable data.

2.3 Indeed, the choice of specific theoretical assumptions is not innocent. Hare and Pahl-Wostl (2001), for instance, compared two models of rational agents and bounded rational agents, and concluded that the choice of agents' models has a significant impact on the simulation results. Gilbert (2002) investigated a number of different versions of the Schelling model, and discovered that it is impossible to differentiate between the different models using the simulation results. In this example different assumptions about underlying mechanisms yield very similar results. A purely empirical assessment of simulation models, conducted with reference only to simulation results, might remain ambiguous. Gilbert suggests amplifying the issue of validation through an evaluation of how theoretical constructs are represented. This calls for evaluation along the lines of theoretical insights from the empirical sciences. Within classical design of experiments, this is known as construct validity. The term construct validity covers the questions of whether the independent and dependent variables represent the theoretical construct they are intended to measure. A number of statistical methods have been proposed (cf. e.g. Schnell et al. 1995) as a means of arriving at a numerical operationalisation of this question. Most prominent are convergent validity and discriminant validity. Convergent validity is defined as a high correlation of results of different measurement procedures. Hare and Pahl-Wostl's examination can be regarded as an example where this element is not fulfilled: rational and bounded rational agents aim to approximate one and the same real world decision makers. However, simulation results are not convergent. Discriminant validity is defined as low correlation with criteria that aim to represent different aspects. Gilbert's examination can be regarded as an example where this element is not fulfilled. Assumptions that aim to represent rather different theories yield similar results. However, it is not the aim of the current investigation to arrive at any numerical representation. Rather, the goal is more in line with what Balci (1997) denoted, in addition to the well-known type I (rejecting a valid result) and type II (accepting an invalid result) errors in statistical research, as type III error: namely, solving the wrong problem. However, in contrast to Balci's suggestion, it is equally not the present investigation's aim to develop a scale (that eventually might even be binary) running from 'correct' to 'false'. By investigating numerical (including simulation) models in the earth sciences, Oreskes et al. (1994) have already warned that in open systems the notion of truth might be misleading. Rather, a major finding of the sociology of knowledge is that a theoretical construct is always a decision for a certain perspective on a problem. The purpose of this examination, then, is to enfold the perspective on the target system that is more or less implicit in the model assumption. As far as the problem of norm internalisation is concerned, it is of particular relevance to emphasise perspective rather than truth, since it is investigated in a number of different disciplines with correspondingly different perspectives on the problem. These different perspectives lead to different assumptions and conclusions. The general idea of this examination with the example of the BOID architecture is illustrated in the figure below. Note that it does not constitute a substitute for traditional techniques. Rather it suggests an additional procedure.

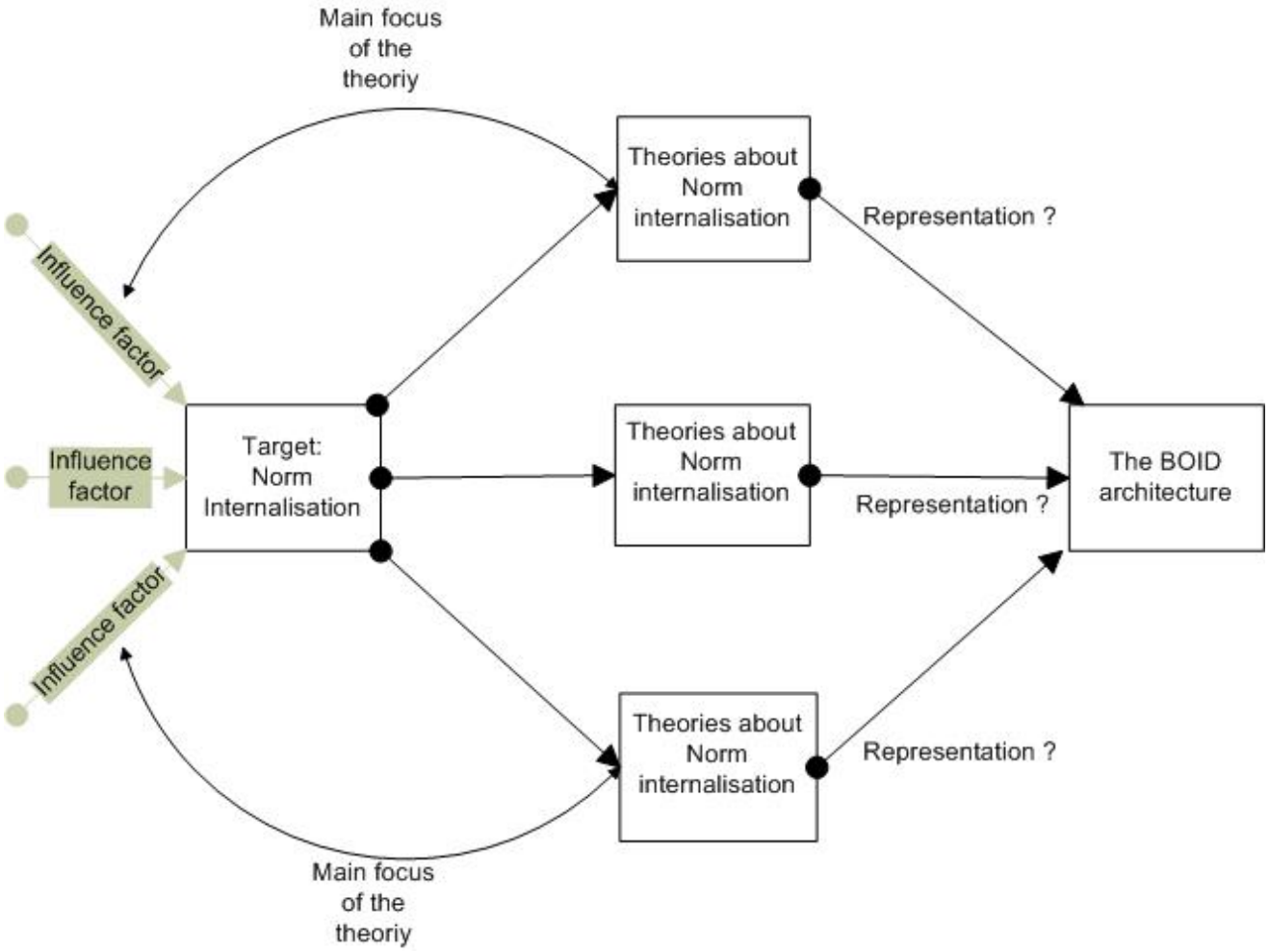

Figure 1. Representation of theoretical constructs by the BOID architecture

2.4 This approach does not establish a direct feedback loop between the target system and the model. Rather than assuming that the model offers a direct representation of the target system, an intermediate level of different theories about the target is added. The figure reads as follows: Three stages are differentiated. First, the empirical process of norm internalisation is clearly influenced by a number of different factors. Secondly, various theories exist to explain this process. These theories foster different conclusions. In part, the differences between the theories are due to the fact that their main focus is on different influence factors. These theories provide the theoretical construct to represent the empirical target system. Thirdly, it should be asked whether —and to what extent—the agent architecture represents any of these theoretical constructs. This should make explicit what intuitions about norms may be found in the agent architecture.[2]

Normative agent architectures: the BOID example

3.1 There is no unequivocal concept for the design of normative agents. Broadly speaking, normative simulation models follow two traditions: game theory and artificial intelligence (Neumann 2008b). However, game theoretic models typically investigate the process of norm spreading, rather than the process of norm internalisation. For this reason, the present survey concentrates on the latter approach. Moreover, it focuses on architectures rather than on implemented models. The development of architectures specifies the essential components of normative agents, which study how these processes could be modelled in principle. It has been argued that focusing on implementation before having a proper understanding of formal architectures risks loosing key norm-related intuitions (Dignum et al. 2002). Thus the specification of the theoretical construct should be found in particular within normative architectures. The development of normative architectures is a burgeoning research field. Not surprisingly, a number of different approaches can be identified. However, architectures of normative agents are predominantly informed in some way by BDI (Belief-Desire-Intention) architectures. It can be regarded as the point of departure for further developments. The BDI framework is intended to model human decision-making. One of its key insights is directed at amplifying logical models with cognitive components. While BDI architectures still apply logical operations, the notion of beliefs, desires and intentions refers to intuitive cognitive meaning in everyday language. Developed by the philosopher Michael Bratman as a model of rational decision making—in particular, as a means of clarifying the role of intentions in practical reasoning according to the norms of rationality (Bratman 1987)—the BDI framework

has been adopted by Rao and Georgeff for the development of software technology (Rao and Georgeff 1991). BDI agents constitute data structures that represent beliefs about the environment and desires of the agents. Desires enable the goal-directed behaviour of an agent. Furthermore, the decision-making process has to be managed, which is accomplished in two stages. To achieve a goal, a certain plan has to be selected and denoted as the intention of the agent. Secondly, the agent undertakes means-ends calculations to estimate those actions necessary to achieve its goals. Intentions are crucial in mediating between the agent's beliefs and desires and the environment. Hence, the decision-making process possesses the following structure: sensory input leads to a revision of beliefs. On the basis of desires, intentions are generated. One goal is selected, leading to a corresponding action.

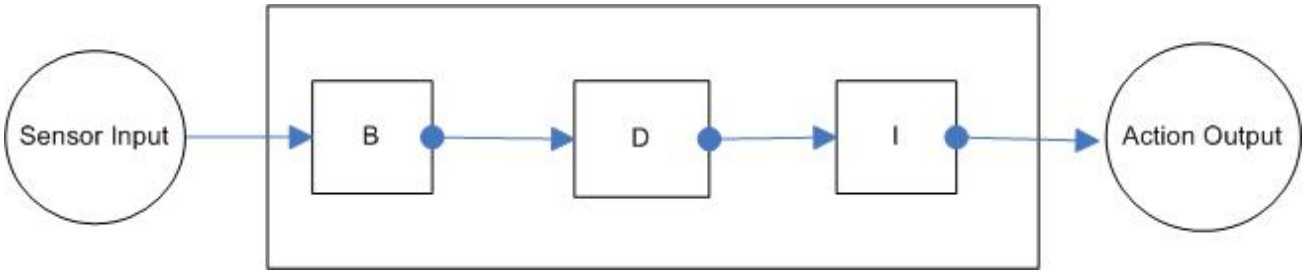

Figure 2. BDI architecture

3.2 However, BDI agents are desolate egoists. In groups, behaviour is more effective if agents orient their actions around other agents. To implement social behaviour in agents it has been suggested that another component should be added to the agent's architecture: obligations. Obligations have been described as 'desires of the society' (Dignum et al. 2002). In this component, social norms are implemented to include social rationality in the agent's design. The beginning of the investigation of normative agent systems can be traced back to an article by Shoham and Tennenholtz (1992). This article was groundbreaking insofar as it suggested implementing social constraints on individual behaviour that was not effected through a central controller but by behaviour restriction within the individual agents. They illustrate this by the example of two robots moving towards each other: the decision how to pass is governed by a social law (i.e. to move on the left hand or the right hand side). However, Shoham and Tennenholtz began this expanding research field with a concept of norms as simple constraints acting on individual behaviour (e.g. right hand driving). Agents are unable to recognise or deliberate about norms. For the example of the robots (or in more general terms, engineering purposes) this might be sufficient. For the analysis of Artificial Societies, however, a more sophisticated concept is desirable. More sophisticated accounts treat norms as mental objects (Castelfranchi et al. 2000). A number of diverging accounts exist that explore this general idea (cf. Neumann 2008a). While these differ in many respects, they all agree that the agent's cognitive design includes the notion that in certain situations a particular mode of behaviour is prescribed by norms and that the agent is able to deliberate about it. This intuition has been elaborated in the concept of obligations as a separate component in the agents' cognitive design. This principle conception can be found in a number of normative architectures.[3] Specifically, it allows for the conscious violation of norms. Norms intervene in the process of goal generation, which might—or might not—lead to the revision of existing personal goals in favour of normative goals. To cope with contradicting prescriptions, even more sophisticated accounts distinguish the concept of obligations from the notion of those norms that are regarded as a more abstract concept (Conte and Dignum 2001; Dignum et al. 2002). Obligations are explicit prescriptions that are always conditional to specific circumstances. They can be simply executed nor not. This is what agents are able to deliberate about. In the case of norms as abstract concepts, further deliberation is necessary. An instance of an abstract norm would be the idea of 'altruism'. In this case, further inference processes are needed for the formation of concrete goals. The majority of current approaches (Neumann 2008a), however, concentrate on the analysis of obligations.

3.3 A particular striking example of this approach is a straightforward extension of the BDI architecture, denoted as BOID (Belief-Obligations-Intentions-Desires) agent architecture. This can be regarded as the current state of the art. In the following it will be briefly characterised. A first attribute to be noted is that the idea itself of developing a separate obligations component is controversial (Broersen et al. 2001). In principle, norms could be implemented in a straightforward manner in the desires component of the agent. The rationale for including a separate component is geared towards ensuring an agent's autonomy: by explicitly separating individual and social desires, it is possible that the agent can deliberate over which component has priority. Conflicts may arise between different components. For instance, I may want to smoke a cigarette after dinner but am obliged to refrain from smoking in restaurants. If everything was stored in a single component, conflicts could not be modelled since logically everything could be deduced from a contradiction. However, if the desire to smoke and the obligation (i.e. society's desire) not to smoke is stored in different components, the agent can decide which desires (i.e. social or

individual) to fulfil. Thus, the agent is able to violate obligations, guaranteeing its autonomy.

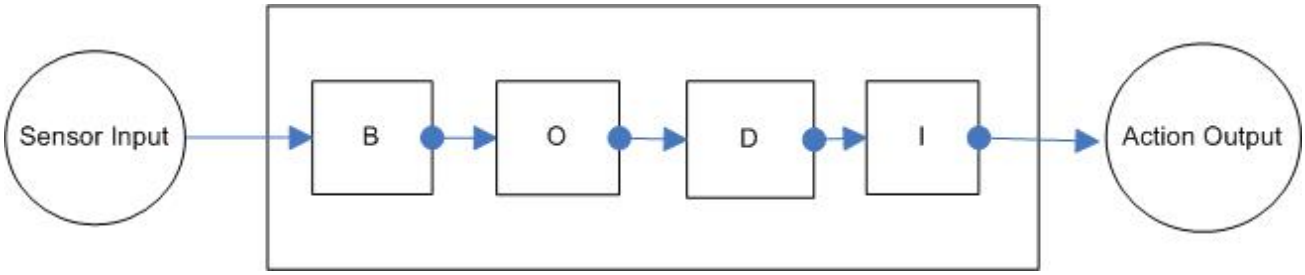

Figure 3. BOID architecture

3.4 Since there might be conflicts between the component's 'desires' and 'obligations', a new level of complexity is introduced in the agent architecture. These possible conflicts have to be resolved in some way. Different types of agents can be distinguished, dependent on which component is accorded priority.[4] For instance, this allows to differentiate between egoistic and socially responsible agents.

3.5 The BOID architecture is a very intuitive extension of the BDI approach. It can serve as a prime example to illustrate typical features of an agent's cognitive design, even though this particular architecture was not designed with the intention to model norm internalisation. However, to large degree these features can also be found in other approaches that are based on the notion of obligations. For this reason, the examination of the theoretical validity will concentrate mainly on this example.

3.6 With regard to the social environment, it is worth noting that the source of obligations, namely the society of other agents, remains rather abstract (This is different in some other architectures. Compare Boella and van der Torre 2003 or Sadri et al. 2006). For instance, the influence of social structure on the process of normative development is not considered. No feedback loop exists between the interactive structure and the agent's cognitive components. BOID agents are not socially embedded. With respect to cognitive design, the most obvious shortcoming appears to be that obligations cannot be generated endogenously,[5] and hence the component remains static. No developmental processes can be discovered in the agents. Moreover, norms are not related to a sense of identity or emotions, i.e. to body and soul. However, in contrast to mere coercion, these relations are regarded as a distinctive feature of normative human behaviour regulation. The BOID architecture does not rely on attempts of embodied intelligence. In short, the agents' architecture preserves the Cartesian dualism of mind and body. The decision-making process of BOID agents is the pure reasoning of an arm-chair philosopher. This reflects the fact that, originally, the BDI approach was developed as a philosophical theory of rational decision making. Rationality represents normative advice, not psychological theory. Kant, for instance, argued that pure reasoning would be the same for humans as well as for angels or Martians. Following Kant, pure reasoning is not inherent to our biological constitution. The same holds for BOID agents. They are neither socially embedded nor is the reasoning process shaped by any kind of psychological embodiment. However, even within robotics, the embedded intelligence approach is becoming increasingly influential (Anderson 2003). In robotics, it is increasingly accepted that minds differ from brains. Embedded intelligence is shaped by the history of interactions with the world. Physical interaction is perceived as necessary to solve problems such as symbol grounding or situational framing. It seems plausible to assume that this situation holds true all the more in the case of normative action. In particular, when modelling norm internalisation as distinct from mere coercion (i.e. for theoretical questions), it might be fruitful to take a closer look at the human example. How do humans become morally responsible agents?

Dissecting the BOID perspective

4.1 In his cloister, Ghandi dispensed advice to the untouchables in his commune: they were not to react to obscenities offered them by members of other castes. If the untouchables followed this advice, their tormentors would feel ashamed by their own obscenities. The Christian exhortation to turn the other check is similarly framed. Emotions such as shame, it is argued, have an important role in normative self-control. However, by what mechanism does an affect at the individual level become associated with a collective value? How can a shared idea of what is desirable or undesirable come to effect constraints on individual behaviour? How are mental concepts of socially accepted behaviour formed in the individual's mind? These questions may be usefully pursued with a focus on what is commonly referred to as norm internalisation in socialisation research.[6]

4.2 However, preceding scientific research, a long philosophical tradition exist concerning this issue.

This still guides the implicit assumptions of scientific theories. Broadly speaking, a dichotomy of two main philosophical approaches can be identified (Geulen 1991): one position assumes a harmony between, or identity of, individual and society. Representatives of this approach are, for instance, Aristotle, Leibniz and Hegel. The second position stands in contrast and postulates an antagonism between individual and society. Within this position, two further standpoints can be distinguished. Hobbes, for example, is representative of the argument that society should tame the individual. By contrast, Rousseau's perspective is paradigmatic of an approach that advocates the need for releasing the individual from society. Both philosophers share the assumption that an antagonism exists between the individual and society, although they disagree about the implications. This differentiation provides a useful guide to classify different perspectives.

4.3 The BOID architecture shows remarkable congruence with the perspective of the older scientific literature on internalisation that culminated in the work of Emil Durkheim and Sigmund Freud. This is reflected in the implicit assumptions of these theories as well as the BOID architecture. From a sociological perspective Emil Durkheim investigated the question of social integration (Durkheim 1893). He asserted that the individual consists of two parts: first, a private domain that is egoistic and guided purely by basic drives. The egoistic domain corresponds to the new-born child. The original human is a 'tabula rasa' in which social norms have to be implemented. Only through the process of socialisation do humans become socially and morally responsible—the second 'part' of the individual. Durkheim claimed that our best part has a social nature. Society, however, is coercive (Durkheim 1895), and can even compel individuals to commit suicide (Durkheim 1897). Norms are finally internalised once the individual no longer perceives this coercion (Durkheim 1907). This corresponds to Hobbes' perspective on the relation between society and the individual.

4.4 Starting from a clinical and psychological perspective, Sigmund Freud developed a theory of socialisation that in many aspects is surprisingly consonant with Durkheim's approach. As with Durkheim, Freud assumed the existence of an antagonism between individuals and society. This assumption can be discerned in his distinction between Ego, Id and Super-Ego. The Id represents the drives of the child-like portion of the person. It is highly impulsive and takes into account only what it wants, exclusively following the pleasure principle (Freud 1932). The Super-Ego enables control of the primary drives: it represents the moral code of a society and involves feelings of shame and guilt (Freud 1955). It is the site where social norms can be found. Finally, the Ego is the controlling instance: it coordinates the demands of Id, Super-Ego and outer world.

4.5 Freud's theory of Ego, Super-Ego and Id, then, parallels Durkheim's assumption that the internalisation of norms involves social coercion (Geulen 1991). From the perspective of both, society is in radical conflict with human nature. Society needs to tame the egoistic and amoral desires of the subject. It has to be controlled by internalised norms. However, norms are given as an external fact. Both Durkheim and Freud regard the individual as passive and internalisation as a unidirectional process. To divide the individual into a private and a social part implies that the social part is merely implemented into the individual. This is the model of a Nuremberg Funnel. Within this conception of the relationship, the social element remains a coercive force.

4.6 It is striking that the main idea of the BOID architecture, namely implementing norms in a separate component, shows remarkable congruence with these accounts. In particular Freud's architecture of the human psyche has some parallels to BOID architecture: the Id, guided by egoistic drives, taking into account only what it wants, belongs to the 'desires' component. Moreover, there is an obvious temptation to identify Freud's Super-Ego with the 'obligations' component. In fact, 'obligations' have been explicitly described as the desires of a society (Dignum et al. 2002)[7]. Thus, by differentiating cognitive modules, the BOID architecture is well in line with this aspect of Freud's theory. However, the theoretical congruence with the accounts of Freud and Durkheim is intrinsically tied to the view of the relation between individual and society that is implicit in the architecture: the dichotomy of obligations and desires refers to an antagonism between the individual and society. Initially, software agents have no need of other agents. In fact, the 'obligations' component is effective only if it is in conflict with the 'desires' component. If it were otherwise, an additional component would not be needed. This circumstance reveals the philosophical perspective of an antagonism between the individual and society that is precisely that also found in Freud's and Durkheim's accounts. It has to noted, however, that Durkheim and Freud do not represent the actual state of the art.

4.7 That this implicit antagonism is a certain perspective (rather than 'truth') becomes even more evident when it is compared to different accounts. At first sight, it seems inherent in the concept of autonomy that agents are able to deliberately violate norms. Otherwise the agents would simply have to execute a set of given norms. This was the concept of Shoham and Tennenholtz. The BOID architecture explicitly aims to be a more sophisticated and realistic representation of human normative reasoning. The concept of autonomy seems to imply an antagonistic perspective.

4.8 Yet diverging accounts can be found in the more recent psychological literature. In particular the theory of self-determination (Deci and Ryan 2000; Ryan and Deci 2000) draws a rather different picture. This theory investigates the question of how goals and desires are related to action. For artificial intelligence research, this is the problem that originally motivated the development of BDI agents. In human psychology, the issue belongs to the field of motivation theory (Trommsdorff 2005). The starting point of self-determination theory is to distinguish between intrinsic and extrinsic motivation. People simply do what they find interesting and important. This behaviour is intrinsically motivated without external reinforcement. This is the paradigm of autonomous action. There is emphatic empirical evidence to suggest that sanctions and even incentives work to undermine intrinsic motivation. Norms, however, constitute a socially determined pattern of behaviour. Thus norm obedience is always extrinsically motivated and not autonomously motivated in the first instance. Insofar this theory agrees with Durkheim, Freud and the BOID approach. However, it differs with respect to the question of how norms are internalised. Extrinsic motivation, it is claimed, can be internalised in different degrees (Deci and Ryan 2000). Deci and Ryan develop a fine grained scale from external control to integration into the self.

- *external regulation* such as sanctions and incentives represent purely external behavioural control. Behaviour patterns are only stable under external coercion. This mode of pursuing objectives is denoted as control orientation.
- *introjection* is the first degree of internalisation. Behaviour regulation remains partially dependent on external instances, but there is also partial internal behaviour regulation. However, this management of behaviour is closely related to an individual's public self-portrayal, typically bound up with emotions such as pride or shame. The salience of behaviour patterns remains low. Introjected behaviour remains dependent on control orientation.
- *Identification* is the next step, when people accept the underlying value of external behavioural regulation. For example, to accept that sports is good for health makes exercising more likely to be volitional. However, the associated behaviour remains instrumental (e.g. to being healthier), rather than a goal in its own right.
- *Integration* is finally attained when external guidelines have become part of the personal identity. A person is in line with him or herself if he or she orients behaviour around social norms. Integrated behaviour regulation is highly salient. This is the highest degree in transformations of external regulation into self-determination, and may be denoted as self-determined extrinsic motivation. Since integrated norms are in full accordance with personal values, action is regarded as autonomously motivated.

4.9 Hence, the scale from external regulation to integration is regarded as the scale from external control to autonomy. This enfolds a rather different concept of autonomy than the BOID architecture. The research here emphasises that full internalisation of norms is only realised when they have become part of one's identity. According to this theory, internalised norms thus form part of the person's own goals. Norm integration enables individual autonomy. From this perspective, BOID agents have not fully internalised norms. In terms of self-determination theory, the 'obligations' component merely describes norm introjection,[8] rather than full integration. Norms implemented in an 'obligations' component do not represent complete external regulation; by the same token, neither are they part of the agent's own desires. This is explicitly desired for the 'obligations' component: it is added to the architecture as a means of enabling norm compliance as well as violation. It is claimed that this process will preserve the agent's autonomy. However, in terms of self-determination theory, the transition from external regulation to integration is described as a transition from external control to autonomy. This is a crucial difference in the concept of autonomy. People act autonomously if norms are part of the personal identity. They are autonomous if they are morally responsible persons. In contrast, software architectures regard agents as autonomous only if norms precisely do not form part of their own desires. The difference in the concept of autonomy, however, shed light on the philosophical orientation with regard to the relation between the individual and society: if norm integration is regarded as the process to become an autonomous person, the antagonism disappears.

Beyond BOID

5.1 So far, it has been shown that the separation of different cognitive modules in the BOID architecture is in accordance with Durkheim's and Freud's account to regard the individual as being composed out of two parts: an egoistic, subjective one and a social component which represents the moral code of the society. Norms are merely implemented in the individual. This is the picture of the Nuremberg Funnel. This reveals an antagonism between the individual and society which allows only for norm introjection rather than integration. To identify the specific characteristics of the BOID architecture that are responsible for this perspective, it will now be contrasted with alternative theories from the empirical sciences. Moreover, this allows to develop suggestions for the design of future architectures of normative agents. First, the concept autonomy as it has been developed by the self-determination theory calls for a closer examination

of moral autonomy. The next question, how moral autonomy is acquired, leads to psychological and socio-psychological theories (focused on the concept of identity) that are out of the scope of the BOID account. According to these theories, moral autonomy is not a product of pure reasoning but calls for embodied and embedded agents. Indeed, such elements cannot be found in the BOID account. However, it can be shown that a disembodied account is bound to an antagonistic perspective. Elements of a socially embedded approach can be revealed in other normative architectures.

Moral autonomy

5.2 A characteristic feature of norm integration in terms of the self-determination theory is that the moral code becomes part of the personal identity. This calls for an examination of the notion of moral autonomy. The self-determination theory allows to infer a suggestion for a framework to represent how humans become morally responsible persons. Indeed, this is a characteristic feature of the human species. To describe the process of norm internalisation as described by the self-determination theory, a dynamic relation between the components 'obligations' and 'desires' would be required. Contingent on the salience of a norm, elements of the 'obligations' component should be imported to the 'desires' component. This shift would represent the assumption that internalised norms become part of the individual's sense of identity. In short, while BOID agents are normative agents, BDI agents may be regarded as moral agents. This process can be illustrated by the following figure:

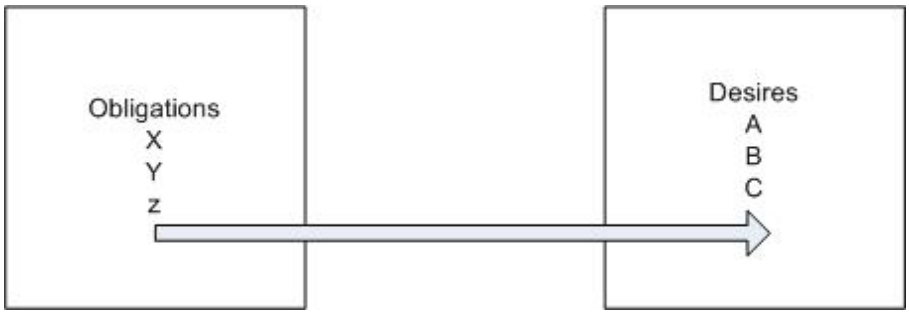

Figure 4. Moral transformation

5.3 However, this is only a broad framework. How are elements of the obligation component imported into the desires component? What are the mechanisms driving this process? It needs to be emphasised that in humans the degree of salience with regard to normative orientation is not equivalent to statistical frequency. It is a cognitive process that a moral code becomes part of the personal identity. It is an instance of the development of the power of judgement (*Urteilskraft*). This is the domain where Kohlberg's work of the development of moral judgements during the childhood is best-known. Lawrence Kohlberg investigated the formation of moral judgements using exhaustive longitudinal studies (Kohlberg 1996). He is well-known for his developmental theory, which postulates the existence of six degrees of development of moral judgements. These degrees of morality are ordered in three levels (Fittkau 1983).

5.4 Kohlberg's classification of moral judgements ranges from a pre-conventional, via a conventional, to a post-conventional level. The *pre-conventional level* (A) is typical for children younger than 9 years old. It consists of a stage of blind obedience to authorities (degree 1) and a means-ends morality (degree 2), characterised by fair exchange. The *conventional level* (B) is typical for the adolescence phase and for most adults. Kohlberg differentiates between a degree at which it is regarded as morally good to play a kind role (degree 3), and a degree in which the individual adolescents conceive themselves as a part of the social system (degree 4). Here they regard it as morally good to fulfil social obligations. The *post-conventional* level (C) is achieved only by a minority of adults. The degree of the social contract (degree 5) is characterised by a respect for, and support of, the principle values of society, even in cases where concrete laws are in conflict with them. The final degree 6 is achieved when moral judgements are based on universal ethical principles valid for the whole of humankind.

5.5 An example extensively studied by Kohlberg (1996) involved a burglary at a pharmacy: Mr Smith's wife is seriously ill, and her life can only be saved with a particular medicine. However, Mr. Smith doesn't have enough money to buy it and the pharmacist won't give the medicine without payment. Should Mr Smith commit burglary, i.e. commit a crime? His decision is not based on chance but on reasoning. For this reason, the situation might not lead to a binomial distribution of people committing, or not committing, the crime of burglary. Kohlberg poses this question to children of 7, 14 and 20 years of age. The focus of his research is the change in the reasons the children give for their judgement about how to act in this moral dilemma: while younger children typically orient their judgements around external authorities (act according to obligations: do not steal, because the police will catch you and you will be sent to prison), older children (sometimes)

appeal to universal moral principles (act according to morality: the absolute value of life allows for breaking the norm not to steal). This is an instance of autonomous moral reasoning.

5.6 This example reveals a number of insights: first, it shows that the process by which elements of the obligations component are imported into the desires component is a non-trivial process. It is a long-lasting cognitive development in human childhood. Secondly, the example shows that human actors can deliberately select between possibly competing obligations, and decide which to import into the personal identity. Autonomous moral reasoning is one way by which (morally responsible) humans can resolve conflicts between different obligations (saving life versus not stealing). Here is a crucial difference to the BOID strategy to separate different components (i.e. desires and obligations) as a means of conflict resolution. Thirdly, we learn that the character of the elements has been transformed. The absolute value of life is a rather more abstract deontic than avoiding being caught by the police. To resolve conflicts between competing obligations, a meta-level of abstract values has to be achieved. This refers to the discussion of different normative software architectures, namely the suggestion to distinguish mere obligations from a notion of norms as a more abstract concept (Conte and Dignum 2001). It calls for an abstraction process from concrete obligations to universal moral principles. This is what Kohlberg described for the case of human psychology.[9] The question now arises how agents can develop such abstract concepts. Crucial for this process is the development of new semantic categories. Hence, to represent this problem in agent architectures leads to the question of how agents learn.

Embodied morality

5.7 To represent learning processes, in particular concepts of reinforcement and social learning, are well-known in agent-based models. The main process of social learning is imitation. The process of reinforcement learning is learning by experience, i.e. the evaluation of positive or negative consequences of past actions. These concepts can be traced back to behaviouristic theories of learning. However, Kohlberg's degree 6 of moral judgements, for instance, implies to identify abstract principles behind concrete obligations. This is a creative process, namely the development of new semantic categories to interpret the perception of the world. Such a cognitive process is beyond the scope of theories of learning based on behavioristic principles. As a prime example, learning processes based on behaviouristic theories are incapable of representing the negative correlation between sanctions and internalisation as emphasised by Deci's and Ryan's theory of self-determination. To represent how humans become morally responsible persons calls for theories of cognitive development. Behaviourism is not capable of capturing processes of mental development; indeed, it specifically avoids commenting on the mind. In the psychological research this deficit led to a cognitive turn in the 1960s which so far remains largely unemployed in agent architectures.

5.8 The main locus of the process of cognitive development is the childhood. Compared to our evolutionary relatives, humans are characterised by an extended developmental process (Locke and Bogin 2006). For instance, bonobos raised in a human environment develop traits that can only rarely be found in the wild. However, they do not reach a cognitive complexity that goes beyond a ca. three year old child (Savage-Rumbaugh et al. 1998). Childhood, indeed, seems to be of particular relevance for the cognitive complexity of humans. Different streams of theories of cognitive development, which can be traced back to the influence of G.H. Mead, Piaget and Kohlberg, have one thing in common: their emphasis on developmental processes of growing children. During childhood cognitive processes take place whereby the brain develops into an adult. Such a reorganisation of the brain structure seem to be possible only in this period. This is indicated by evidence from Spitz' studies of children in orphanages (Spitz 1951) or Davis' reference to a woman known as Anna who faced the first years of life without social embedding and emotional care (Davis 1949). In later years humans seem to be unable to catch up early deficits. This indicates a process of biological self-organisation (triggered by social interaction) in which the human mind adapts to the environment (Maturana and Varela 1980; Goldspink 2000). While the neurophysiology of this development is out of the scope of this investigation, for the purpose of simulation it is sufficient to highlight some of the essential results of this process.

5.9 A prime example that describes the cognitive reorganisation of the human brain has already been identified by G.H. Mead (1934): the capability of role-taking. That is, to regard oneself from the perspective of the other. This ability enables individuals to anticipate the perspectives and expectations of others. In the process of role-taking, the individual develops a consciousness whereby the individual is itself a stimulus for the reaction of the other in situations of social interaction. This seem to be a characteristic feature oft the biology of the human species. Where understanding the internalisation of norms is concerned, it is crucial to recognise that the child is unable to regard itself from the point of view of another before it develops the capability of role-taking. Therefore, the child is unable to recognise itself as the norm addressee. Piaget (1932) denoted this as childish egoism. He distinguished two important stages in moral development: a heteronomous stage, in which norms are regarded as fixed and given by a normative authority

(mostly the parents), and the stage of autonomous morality, in which norms are determined by a free agreement. Obviously, this implies the capacity of role taking. According to Piaget this cognitive transition happens typically at the age of around 7 years. While the development of heteronomous morality might be modelled with reinforcement learning, autonomous morality and—in particular—the *transition* in the cognitive development of the child is out of the scope of reinforcement learning.

5.10 A central mechanism of this cognitive and moral development of children has been described by Piaget as assimilation and accommodation (Piaget 1947). While assimilation represents the process of subsuming experiences in an existing cognitive scheme, accommodation takes place if assimilation fails. The development of a new cognitive scheme, one dependent on the demands of the object and the situation, is denoted as accommodation; it can be considered as the development of new thinking tools. This constitutes a creative process. Hence, this is an implicit learning theory. To develop increasingly abstract concepts of moral values, as described by Kohlberg, processes of accommodation are required.

5.11 In conclusion, theories of cognitive development reveal crucial conceptual differences between the perspective of the software approach to norms and psychological accounts. Childhood is a place where principle processes of the self-organisation of the human brain take place. This developmental trajectory (including the development of morality) is a biological feature of the human species. Indeed, other species adapt to their environment in different ways. Insofar as this process is species specific, theories of cognitive development can be regarded as painting a picture of a biologically embodied mind. In contrast to behaviouristic psychology that has been tested with cats and dogs, these theories do not describe the cognitive development of, say, bonobos. From the perspective of cognitive theories, the lack of cognitive development—rather than simple conditioning or knowledge transfer (the Nuremberg Funnel)—in the agents' cognitive design becomes apparent (Guerin 2008). BOID agents do not have a cognitive development that can be compared the processes that take place during the human childhood. This fact highlights one of the main deficits of agent-based simulation models in so far as they attempt to represent the process of norm internalisation: namely, that agents do not possess a childhood.

5.12 Closely related to this deficit is the fact that software agents have no human needs, e.g. for appreciation or sexuality. Already Freud formulated a theory that took account of the development of the architecture of the human psyche, producing a psychological theory of norm internalisation. Freud famously centred his investigations around the sexuality of the early child (Freud 1905). In particular, he cast identification as a central mechanism and the focal point for the development of the Super-Ego. According to Freud, identification constitutes a reaction to an unrealisable erotic binding to the mother. Once the father attracts the notice of the child as an erotic rival, the child reacts to this situation by identifying with the father. The same mechanism is at work when the child is punished by the mother: an identification with the aggressor, namely, the punishing mother. According to Freud, this model explains the mechanism for the transmission of social norms in the child. Likewise, emotions typically related with norms such as pride or shame are not found either in the architecture (Wilks 2008).

5.13 Thus, a prime deficit in current models can be considered the lack of any representation of embodiment. The perspective of agent architectures is that of a ready-made actor. This preserves the Cartesian dualism of mind and body. This perspective correlates with that of an antagonism between individual and society. In representations of learning processes that are primarily inspired by the model of the Nuremberg Funnel the cognitive structure of the mind is unaffected by the implemented content. The content is simply added without any processes of accomodation. If social norms are merely implemented in the individual, they remain separate from individual desires. Based on this intuition, it is obvious that they should be implemented as a separate component in agents' cognitive design.[10] However, the development of morality implies, as highlighted particularly by Kohlberg, autonomous moral reasoning that cannot be reduced to the execution of conditioned behaviour or imitation. Cognitive development of biological humans differs from machine learning.

Embedded socialisation

5.14 However, the development of morality in biological humans is not determined by their biological constitution. This can be illustrated at the example Piaget's autonomous morality, based on the cognitive development of role-taking: playing with peers is particular important to reach this stage (Piaget 1932). Social interaction is an essential trigger for the cognitive development. Embodied intelligence is thus intimately tied to social embedding.

5.15 As we have seen already in the case of the self-determination theory, integration of moral values into the personal identity is a central element of moral autonomy. This calls for a closer examination of the concept of identity. The result of the picture of the developmental process

sketched by identity theories is that individual personality cannot be divided into private egoistic drives and an internalised social coercion. The constitution of individual identity, which directs *individual* goals and desires, is itself a socially embedded process. The investigation of the development of autonomous morality implies thus a philosophical perspective that stands in contrast to that of the BOID architecture of an antagonism between individual and society: identity theories imply a rather harmonic view on the relation between individual and society.[11] In fact, neither micro- nor macro-social structure is included in the BOID architecture. However, such processes are not out of the scope of software development. At least some key elements of the influence of social embedding on moral development can be integrated into normative architectures. In fact, steps in this direction are already to be discerned in the AI literature. Some suggestions will be made in the following.

Micro social interactions

5.16 Identity theories focus on micro-social relations. In contrast to an orientation based solely either on the individual subject or the society, identity theories emphasise the interaction of culture and individuals in the development of a morally responsible person (Bosma and Kunner 2001; Fuhrer and Trautner 1999; Keupp 1999). Identity theories can be regarded as a bridge between sociological and psychological accounts of norm internalisation. Within the context of identity theories, the term 'internalisation' should be carefully considered. It suggests a simple transmission of norms into the individual, and hence implies that the individual plays only a passive role in this process. In contrast to this picture, identity theories regard individual identity as the key link between persons and culture. Identity development is described as a process whereby culture and the individual mutually enfold each other.

5.17 With reference to William James (1890), criteria for identity are formulated as consistency, continuity and effectiveness. Moreover, identity consists of an inner and outer perspective. While the inner perspective is grounded on individual decisions, the outer is based on ascription of others (examples here are ethnic or gender identity). However, the individual might decide to identify with these ascription, for instance by participating in a woman's liberation group. This refers to a second distinction between personal and social identities (Tajfel 1970; Turner 1982; Turner and Onorato 1999). Personal identity is the self-construction of a personal biography. Hence, by participating in a woman's liberation group, ascription of others (the outer perspective) might trigger the individual biography, i.e. the personal identity. Social identity is determined by peer groups and reference groups, and refers to social networks. While peers are the group to which the individual factually belongs, the individual need not belong to the reference group. It is sufficient to identify with the values of this group. For instance, this identification might constitute sympathy for a political party. Social identity is decisively responsible for the process by which social norms and values become part of individual goals. This process is notably dependent on the salience of group membership.

5.18 Hence, social identity theory is essentially build upon micro social networks of peer and reference groups. In principle, this can be simulated. Network models belong to the standard repertoire of simulation technologies that simply have to be exploited for the purpose of developing normative agents. Even though a full representation of all aspects of identity would quickly become over-ambitious, networks of peer and reference groups could be implemented without a great deal of effort. For instance, the propensity for taking over group norms could be simulated, dependent on the salience of group membership. To model social identity, agents thus need to be embedded in micro-social structures.

5.19 Also steps in the direction to model role-taking do already exist. In Boella and van der Torre's architecture of a 'norm governed system' (Boella and van der Torre 2003), the agent's decision making process is governed by the belief that they are observed by other agents, and by the belief that the other agents have expectancies with regard to how they ought to behave. This can be considered as a first step in simulating role-taking. These rules of behaviour are determined by their social role. However, from a theoretical perspective of comprehending socio-psychological processes two shortcomings can be discerned: first, the process of developing this capacity is not regarded in this account. Secondly, from the perspective of identity theories it is a shortcoming that the agents regard themselves solely in terms of the question whether or not they fulfil their—externally given—social role. Recall that identity consists of an inner and an outer perspective. The architecture of Boella's and van der Torre's 'norm governed system' signals the outer perspective of identity, i.e. ascription of others. However, the inner perspective is dependent on one's personal decisions. This is not the case in this architecture.

Macro social structures

5.20 The notion of social roles in Boella and van der Torre's architecture refers to macro-social structure. The impact of macro social theories to a comprehension of norm internalisation will briefly be examined now. Norms receive particular emphasis within the sociological tradition of role theory. According to role theory, role behaviour is guided by a complex of norms (Parsons 1961; Popitz 1980). For the design of normative argent architectures it is a relevant conclusion of this macro social research that a process of interpretation is needed in order to recognise normative principles behind arbitrary behaviour. To recognise a norm, a certain behaviour has to be regarded as a sign for this norm. Norms thus have an inherent symbolic component (Claessens 1972). The concept of a symbol implies that something that is observed stands for something else that cannot be observed. This finding is of particular interest when identifying sanctions. On a behavioural (observable) level sanctioning merely represents aggression. An interpretation process is necessary before aggression can be regarded as a sanction. A ticket and a robbery might have the same observable consequences: my purse will be empty afterwards. However, the former is interpreted as a norm invocation, i.e. as a sign for something which is not observable (i.e. the norm). This insight has been elaborated in the domain of criminalistics (Popitz 2003). First, macro-social role theory emphasises that social roles exist—for example, policemen or traffic wardens—that are legitimised to sanction and that social roles exist, characterised as socialising instances, that are responsible for the transmission of social norms. Examples here would be parents or teachers. Moreover, theories of social stratification emphasise that the probability of an interpretation of aggression as a sanction increases with a concomitant increase in the social status of the aggressor.[12]

5.21 These macro-social aspects are not a feature of the BOID perspective on norms. However, in principle, simulation models are capable of representing elements of sociological role theory. It would be a simple extension of the general account to equip agents with social roles. Examples are again Boella and van der Torre (2003) or Sadri et al. (2006). In these accounts, agents are equipped with specific social roles. However, the roles require programming-in. The models lack a dynamic perspective on social roles. Since social roles are simply given (and public) in these models, agents do not need a process of interpretation to distinguish aggression from sanctioning. However, the cognitive processes involved in such an interpretation can primarily be discerned in the recognition of status differentials. Since related concepts of reputation and prestige are already subjects of agent-based models (e.g. Hahn et al. 2007; Sabater-Mir et al. 2006) this seem to be a feasible problem in principle.

Conclusion

6.1 In this article it has been investigated inasmuch the theoretical construct of norm internalisation as it can be found in socialisation research is represented in normative architectures. This has been done at the example of the BOID architecture. What to take away from this?

6.2 First, it has to be emphasised that theories of norm internalisation are of great importance for the theoretical validity of normative agents. Internalisation is one of the stronger mechanisms by which an individual may become integrated into wider society. At the same time society penetrates into the individual mind by the process of internalisation. Thus the process of norm internalisation is essential for a comprehension of the relation between the individual and society. Since the very beginning of sociological research this has been a central question of sociology. If a model of normative agents is developed for the theoretical purpose to understand human societies, an assessment of how the theoretical construct of norm internalisation is represented by the model is an essential stage in a validation process.[13] However, empirical sciences also reveal that no unequivocal 'true' theory exist. For this reason theoretical validity cannot be condensed into a single number. Conversely, it cannot be expected that the one and only 'true' model will ever be developed. What can be done is to make the theoretical perspective explicit. At first sight, a claim for explicitness might appear as a rather carping critique. However, it has quite substantial consequences. It can make explicit, what is implicitly missing. Indeed, it unfolds a chasm between certain concepts of internalisation and the BOID account. Even though the BOID architecture aims to provide a more realistic representation of norms in human societies than earlier attempts, with regard to a representation of the mechanisms of internalisation it is still far from the cutting edge in the empirical sciences. This can be traced back to some very fundamental shortcomings of traditional software agents.

6.3 Dissecting the implicit assumptions in the BOID architecture reveals a certain perspective on the problem: namely, the separation of different components is in broad accordance with older accounts of Durkheim and Freud. Obligations can be regarded as a Freudian Super-Ego. The philosophical orientation that implicitly underlies the BOID architecture is inspired by these classical accounts: by opposing obligations and desires, an antagonism between individual and society is assumed. From the perspective of more recent accounts obligations have to be classified as not fully internalised norms. They remain merely introjected. In fact, the BOID architecture is contradictory to theories of norm internalisation that are based on the concept of

identity. This shifting perspective has far-reaching consequences for the conception of both norms and agent architectures.

6.4 The concept of identity implies that norms become part of the goals of the individual agent. To model a full integration of norms into the personality (as described by the self-determination theory), a dynamic relation between obligations and desires would be required that enables a transfer of elements from the obligation component into the desires component. In fact, this is a technical description of moral autonomy. This is a distinctive feature of the human species. The construct validity of normative agents would benefit substantially if a normative architecture could account for a representation of autonomous moral reasoning. However, the notion of moral autonomy implies a conception of norms that is rather different than the classical accounts of Freud's Super-Ego or Durkheim's double layered agency. If norms become part of the personal identity, the antagonism between individual and society disappears. This reveals a different philosophical perspective than accounts to separate various cognitive components. The different perspective on norms calls for a different design of the agents.

6.5 In particular, the human capacity of becoming a morally responsible person is not the result of pure reasoning. On the contrary, a recourse to empirical sciences reveals that the development of autonomous moral reasoning is only enabled by biological embodiment and social embedding. This could inform the development of software architectures. Models of norms are good at simulating the spreading of norms or the effect of norms on a global level (Neumann 2008b). However, so far the feedback loop from the aggregate level back to the individual agent's mind is represented to a far less extend. Indeed, this can be traced back to a major shortcoming of software agents, namely that they need to be regarded as disembedded and disembodied. This reveals a Cartesian dualism. Such an account is bound to an antagonistic perspective on the relation between individual and society. The lack of embodiment becomes particularly apparent in the fact that no childhood exist. So far ontogenetic developmental processes cannot be represented by agents. Likewise, agents lack emotions and sexuality. Moreover, the development of the human brain is shaped by engagement with the social environment. Playing with peers is essential for the development of morality. Thus the further development of architectures that aim to represent the cognitive processes at work in moral reasoning could profit substantially, if the personal history of interactions and emotional engagement with the world would be integrated in the cognitive design of the agents. So far, such feedback processes between inter- and intra-agent processes are not much explored in the design of normative agents. The BOID architecture retains its roots in the field of pure reasoning. Humans, however, are different. It appears debatable whether societies of Martians would exhibit the same kind of normative structures than humans.

Acknowledgements

This work has been undertaken as part of the Project 'Emergence in the Loop' (EmiL: IST-033841), funded by the Future and Emerging Technologies programme of the European Commission, as part of the framework initiative 'Simulating Emergent Properties in Complex Systems'. This article evolved in a long process of intense discussions. Contributions of Birk Hagemeyer, Joachim Giese-Mandler, Chris Goldspink, Annette Duffner, Edmund Chattoe-Brown and two critical as well as constructive anonymous referees are gratefully acknowledged.

Notes

[1] However, not the evolutionary history of the human species but the mechanisms of individual development are scrutinised.

[2]It should be emphasised that this evaluation does not scrutinise the validity of the theoretical results reached by the traditional methods of the empirical sciences. Here, simulation might provide a tool for evaluating different theoretical assumptions. This would call for different simulation models, based on different theoretical assumptions and a comparison of their results with empirical data (cf. Hare and Pahl-Wostl 2001).

[3]A non-exhaustive list of examples includes, for example, the architecture described by Boella and van der Torre (2003), and obviously the BOID architecture itself (e.g. Broersen et al. 2001), the 'architecture of a normative System' (Boella and van der Torre 2006), Castelfranchi's et al. (2000) conception of 'deliberative normative agents', the account described in the paper: from desires, obligations and norms to goals (Dignum et al. 2002), and the 'architecture for autonomous normative agents' (Lopez and Marquez 2004).

[4]For instance, given the assumption of realism (i.e. beliefs override other components), egoistic agents may have the following structure of priorities (Broersen et al. 2001):

- BDIO - beliefs override desires, which override intentions, which override obligations
- BDOI - beliefs override desires, which override obligations, which override intentions
- Social agents can have the following structure of priorities:
- BODI - beliefs override obligations, which override desires, which override intentions
- BIOD - beliefs override intentions, which override obligations, which override desires
- BOID - beliefs override obligations, which override intentions, which override desires

[5]In game theoretical models, this even represents part of the definition of the situation: to cooperate or defect in prisoner's dilemma or trust games, to accept or reject a proposal in ultimatum games, etc. The proposed norm is framed in the description of the situation.

[6]Socialisation research is placed in sociology as well as in psychology. These are different disciplines with correspondingly different traditions and backgrounds. No comprehensive discussion of this field in its historical development can be provided here. Hence, as it holds for the selection of socialisation theory, also the selection of the literature on socialisation has to be restrictive. First, only those aspects are considered that allow for an assessment of the BOID approach and to generate suggestions for further development. Moreover, within the literature that fulfils this condition, the criterion for further selection is to rely on established findings. Only those theories are considered that are incorporated into standard handbooks and textbooks. This implies that approaches which might be interesting and might direct future research, but are not as mature at the moment, are not considered.

[7]That said, an equivalent to the functions performed by the Ego cannot be found in software agents. The Ego resolves conflicts between the Id, the Super-Ego and reality. At first sight, one might assume that this function is performed by the intentions component. Intentions mediate between agents and the environment. However, the resolution of conflicts is realised in software architectures by the ordering of priorities between different components (cf. footnote 4). All the same, Broersen et al. (2001) have proposed implementing a separate component that is explicitly responsible for conflict resolution. They suggest that this component manages the control loop in the decision-making process until conflicts are resolved. Since the only variable components are intentions, this is accomplished by the process of selecting adequate intentions (Broersen et al. 2001). For instance, I might have the plan (intention) to visit my mother-in-law. However, my car is broken, which means that I abandon this plan (intention revision). This conflict resolving component would be equivalent to Freud's Ego.

[8]The degree of norm internalisation in game theoretic models would have to be classified as external regulation. Norm obedience is only guaranteed by sanctions. There is no norm internalisation in these models. Hence, in this respect, introjected norms in the BOID architecture go a step further.

[9]At this point it would be interesting to examine whether the design of agents could benefit from findings from neuroscience. However, this would be the task for another article.

[10]It should be noted, though, that game theoretic models of norm spreading manage to get along without a separate obligations component. That said, in these models the desire of agents is restricted to utility maximisation, which has to be expressed numerically in some way. Agents are faced with a strategic (typically binary) decision situation. An active element of normative orientation in the choice relating to the ends of action cannot be found in a game theoretic approach. The ends of individual action remain underspecified. Agents do not 'know' norms but react to environmental conditions, which might include actions that can be described as sanctions. Agents' behaviour can only be interpreted as normative by an external observer.

[11]However, this does not exclude crisis (cf. e.g. Erikson 1968).

[12]In game theoretic the behaviour modification of agents is due to the anticipation of sanctions. Insofar as the anticipation of sanctions is regarded as an instance of norm internalisation (Ziegler 2000), this is an expression of the perspective of an antagonism between the individual and the society, represented by the sanctioning agent. However, it is not necessary to distinguish sanctions from aggression in this account. It is sufficient that the agents are in fear of the damage caused by aggression. Only a certain behaviour regularity of the aggressive agent is sufficient for an anticipation of the aggression. While game theoretic models are able to simulate the spreading of a behaviour regularity (such as driving on the left-hand side of the road), these agents are not able to recognise it as a norm, since the cognitive component is missing. Aggression can only be interpreted by an outside observer (the modeller) as sanctioning.

[13]Obviously, other research questions lead to other theories as relevant for construct validation. For instance, economics investigates interactions between adults and often between strangers. This leads to other theoretical concepts than those that can be found in socialisation research. Norm internalisation is typically not about adults. However also in economics the role of norms and values is increasingly realised (cf. e.g. Hofstede et al. 2009). Even though in this case the problem is how norms, once established in an adult population, influence the behaviour, the cognitive conditions are grounded in childhood. The main processes that enable a normative orientation take place during childhood. Childhood is important for the possibility to adopt to a - normatively framed - social environment as an adult. For this reason, the evaluation of normative architectures as measured against the standard of socialisation research belongs to the central components of theoretical validity of normative agents.

References

ANDERSON, M L (2003) Embodied cognition. *Artificial Intelligence* 149 (1). pp. 91 - 130.

BALCI, O (1997) Principles of Simulation Model Validation, Verification and Testing. *Transactions of the Society for Computer Simulation International* 14 (1). pp. 3—12.

BARLAS, Y (1996) Formal aspects of model validity and validation in system dynamics. *System Dynamics Review* 12 (3). pp. 183—210.

BARRETEAU, O, Le Page, C and D'Aquino, P (2003) Role-playing games, models and negotiation processes. *Journal of Artificial Societies and Social Simulation* 6 (2)10 http://jasss.soc.surrey.ac.uk/6/2/10.html.

BOELLA, G and van der Torre, L (2003) Norm Governed Multiagent Systems: the delegation of control to autonomous agents. In: *Proceedings of the IEEE/WIC IAT Conference.* IEEE Press.

BOELLA, G and van der Torre, L (2006) An architecture of a normative System: counts-as conditionals, obligations, and permissions. In: *AAMAS.* ACM Press.

BOELLA, G, van der Torre, L and Verhagen, H (2007) (Eds.) *Dagstuhl Seminar Proceedings 07122: Normative Multi-agent Systems* http://drops.dagstuhl.de/portals/index.php?semnr=07122.

BOSMA, H and Kunner, E (2001) Determinants and Mechanisms in ego development. A review and Synthesis. *Developmental Review* 21 (1). pp. 39—66.

BRATMAN, M (1987) *Intentions, Plans and Practical Reasoning.* Stanford: CSLI Publications.

BROERSEN, J, Dastani, M, Huang, Z and van der Torre, L. (2001) The BOID Architecture: Conflicts between Beliefs, Obligations, Intentions, and Desires. In *Proceedings of the 5th International Conference on autonomous agents.*

CASTELFRANCHI, C, Dignum, F and Treur, J (2000) Deliberative normative agents: Principles and Architecture. In: Jennings, N R and Lesperance, Y (Eds.) *Intelligent Agents VI. Agent Theories, Architectures, and Languages. LNCS Vol. 1757*. Berlin: Springer.

CLAESSENS, D (1972) *Familie und Wertsystem: eine Studie zur zweiten sozio-kulturellen Geburt des Menschen.* Berlin: Duncker & Humblot.

CONTE, R and Castelfranchi, C (1999) From conventions to prescriptions. Towards an integrated view of norms. *Artificial Intelligence and Law* 7 (4). pp. 323—340.

CONTE, R and Dellarocas, C (Eds.) (2001) *Social Order in Multiagent Systems.* Kluwer, Norwell.

CONTE, R and Dignum, F (2001) From Social Monitoring to Normative Influence. *Journal of Artificial Societies and Social Simulation* 4(2)7 http://jasss.soc.surrey.ac.uk/4/2/7.html.

DECI, E and Ryan, R (2000) The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry* 11 (4). pp. 227—268.

DIGNUM, F, Kinny, D and Sonenberg, L (2002) From desires, obligations and norms to goals. *Cognitive Science Quarterly* 2 (3/4) http://people.cs.uu.nl/dignum/papers/CSQ.pdf.

DURKHEIM, E ([1893] 2004) *Über soziale Arbeitsteilung. Studie über die Organisation höherer Gesellschaften.* Frankfurt a.M.: Surkamp.

DURKHEIM, E ([1895] 2002) *Die Regeln der soziologischen Methode.* Frankfurt a.M.: Surkamp.

DURKHEIM, E ([1897] 2006) *Der Selbstmord.* Frankfurt a.M.: Surkamp.

DURKHEIM, E (1972/1907) *Erziehung und Soziologie.* Düsseldorf: Schwann.

ERIKSON, E.H. (1968) *Identity, Youth and Crisis.* New York: Norton.

FITTKAU, B (Ed.) (1983) *Pädagogisch-psychologische Hilfen für Unterricht, Erziehung und Beratung.* Aachen: Hahner Verlagsgesellschaft.

FREUD, S (1905) Drei Abhandlungen zur Sexualtheorie. In *Gesammelte Werke* Vol. 5. London: Imago.

FREUD, S (1932) Neue Vorlesungen zur Einführung in die Psychoanalyse. In *Gesammelte Werke* Vol. 15. London: Imago.

FREUD, S (1955) *Abriss der Psychoanalyse.* Frankfurt a.M.: Fischer.

FUHRER, U and Trautner, N (2005) Entwicklung und Identität. In Asendorpf, J (Ed.) *Enzyklopädie der Pschologie—Entwicklungspsychologie* Vol. 3. Hofgrebe: Göttingen.

GARCIA-CAMINO, A, Rodriguez-Aguilar, J A, Sierra, C and Vasconcelos, W (2006) *Norm-Oriented Programming of Electronic Institutions: A Rule-Based Approach.* In ACM Press.

GELLER, A, Moss, S (2007) The Afghan nexus: Anomie, neo-patrimonialism and the emergence of small-world networks. *CPM Report 07-179,* Centre for Policy Modelling, Manchester Metropolitan University Business School http://cfpm.org/cpmrep179.html.

GEULEN, D (1991) Die historische Entwicklung sozialisationstheoretischer Ansätze. In Hurrelmann, K and Ulich, D (Eds.) *Neues Handbuch der Sozialisationsforschung.* Weinheim: Beltz.

GILBERT, N (2002). *Varieties of emergence.* Paper presented at the Agent 2002 Conference: Social agents: ecology, exchange, and evolution, Chicago.

GOLDSPINK, C (2000) Modelling social systems as complex: Towards a social simulation meta-model. *Journal of Artificial Societies and Social Simulation* 3 (2)1. http://jasss.soc.surrey.ac.uk/3/2/1.html.

GUERIN, F (2008) Constructivism in AI: Prospects, Progress and Challenges. In *AISB 2008 Convention Proceedings* Vol 12: Computing and Philosophy.

<1-- gueth, w and kliemt, h (1998) the indirect evolutionary approach: bridging the gap between rationality and adaption. *Rationality and Society* 10 (3). pp. 377—399. -->

HAHN, Ch, Fley, B, Florian, M, Spresny, D, and Fischer, K (2007) Social Reputation: a Mechanism for Flexible Self-Regulation of Multiagent Systems. *Journal of Artificial Societies and Social Simulation* 10(1)2 http://jasss.soc.surrey.ac.uk/10/1/2.html

HARE, M and Pahl-Wostl, C (2001) Model uncertainty derived from choice of agent rationality - a lesson for policy assessment modelling. In Giambiasi, N and Frydman, C (Eds.) *Simulation in Industry: 13th European Simulation Symposium.* Gent: SCS Europe Bvba.

HOFSTEDE, G, Jonker, C and Verwaart, T (2009) Modeling power distance in trade. In David, N and Sichman, J (Eds.) *Multi-Agent-Based Simulation IX. Lecture Notes in Artificial Intelligence 5269.* Berlin: Springer.

JACOBSEN, C and Bronson, R (1997) Computer Simulated Empirical Tests of Social Theory: Lessons from 15 Years' experience. In Conte, R, Hegselmann, R and Terna, P (Eds.) *Simulating Social Phenomena.* Berlin: Springer.

JAMES, W (1890) *The principles of psychology.* New York: Holt.

KEUPP, H (1999) *Identitätskontruktionen.* Hamburg: Rowohlt.

KOHLBERG, L (1996) *Die Psychologie der Moralentwicklung.* Frankfurt a.M.: Surkamp.

KRAPPMANN, L (2006) Sozialisationsforschung im Spannungsfeld zwischen gesellschaftlicher Reproduktion und entstehender Handlungsfähigkeit. In Schneider, W and Wilkening, F (Eds.)

Enzyklopädie der Psychologie—Entwicklungspsychologie Vol. 1. Hofgrebe: Göttingen.

LOCKE, J. and Bogin, B (2006) Language and life history: A new perspective on the development and evolution of human language. *Behavioral and Brain Sciences* 29 (3). pp. 259-279.

LOPEZ, F and Marquez, A (2004) An architecture for autonomous normative agents. In: *5th Mexican international conference in Computer Science, ENC 04 Los Alamitos, USA,* IEEE Computer Society.

MATURANA, H R and Varela, F (1980) *Autopoiesis and Cognition*, Dordrecht: Reidel.

MEAD, G H ([1934] 1968) *Geist, Identität und Gesellschaft.* Franfurt a.M.: Surkamp.

MEYER, M, Lorscheid, I and Troitzsch, K G (2009) The Development of Social Simulation as Reflected in the First Ten Years of JASSS: a Citation and Co-Citation Analysis. *Journal of Artificial Societies and Social Simulation* 12(4)12 http://jasss.soc.surrey.ac.uk/12/4/12.html.

MOSS, S (2008) Alternative Approaches to the Empirical Validation of Agent-Based Models. *Journal of Artificial Societies and Social Simulation* 11(1)5 http://jasss.soc.surrey.ac.uk/11/1/5.html.

MOSS, S and Edmonds, B (2005) Sociology and Simulation: Statistical and Qualitative Cross-Validation. *American Journal of Sociology,* 110 (4). pp. 1095—1131.

NEUMANN, M (2008a) A classification of normative architectures. In *Proceedings of the WCSS08.* Washington.

NEUMANN, M (2008b) Homo Socionicus: A case study of simulation models of Norms. *Journal of Artificial Societies and Social Simulation* 11(4)6 http://jasss.soc.surrey.ac.uk/11/4/6.html.

NULLMEIER, E (2004) Wissensbasierte Systeme - was Sie schon immer wissen wollten. *Beitrag für die Tagung Wissensmanagement in der Wissenschaft" am 26./27. März 2004.*

ORESKES, N, Shrader-Frechette, K and Belitz, K (1994) Verification Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263. pp. 641—646.

PARSONS, T (1961) An Outline of the Social System. In Parson, T, Shils, E A, Naegele, D and Pitts, J (Eds.) *Theories of Society.* Glencoe: Free Press.

PIAGET, J ([1932] 1983) *Das moralische Urteil beim Kinde.* Stuttgart: Klett-Cotta.

PIAGET, J ([1947] 1955) *Psychologie der Intelligenz.* Zürich: Rascher (Ass & Akk).

POPITZ, H (1980) *Die normative Konstruktion der Gesellschaft.* Tübingen: Mohr.

POPITZ, H (2003) *Über die Präventivwirkung des Nichtwissens. Dunkelziffer, Norm und Strafe.* Berlin: BWV.

RAO, A M and Georgeff, M (1991) Modeling rational agents within a BDI architecture. In *Proceedings of the KR91.*

RYAN, R and Deci, E (2000) Self Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well Being. *American Psychologist* 55 (1). pp. 68—78.

SABATER-MIR, J, Paolucci, M, and Conte, R (2006) Repage: REPutation and ImAGE Among Limited Autonomous Partners. *Journal of Artificial Societies and Social Simulation* 9(2)3 http://jasss.soc.surrey.ac.uk/9/2/3.html

SADRI, F, Stathis, K and Toni, F (2006) Normative KGP Agents. *Computational and Mathematical Organization Theory* 12 (2). pp. 101—126.

SAVAGE-RUMBAUGH, S E, Shanker S G and Taylor T J (1998) *Apes, Language and the Human Mind.* Oxford: Oxford University Press.

SCHNELL, R, Hill, P and Esser, E (1995) *Methoden der empirischen Sozialforschung.* München: Oldenbourg.

SHOHAM, Y and Tennenholtz, M (1992) On the synthesis of useful social laws for artificial agent societies. In: *Proceedings of the 10th AAAI Conference.*

SPITZ, R (1951) The Psychogenic Diseases of Infancy: an attempt at their Etiological Classification. *Psychoanalytic Study of the Child* 6. pp. 255-275.

TAJFEL, H (1970) Experiments in intergroup discrimination. *Scientific American* 223 (5). pp. 96—102.

TROMMSDORFF, G (2005) Entwicklung sozialer Motive: pro—und antisoziales Handeln. In Asendorpf, J (Ed.) *Enzyklopädie der Pschologie—Entwicklungspsychologie* Vol. 3. Göttingen: Hofgrebe.

TURNER, J C (1982) Towards a cognitive redefinition of the social group. In Tajfel, H (Ed.) *Social Identity and intergroup relations.* Cambridge: Cambridge University Press.

TURNER, J C and Onorato, R S (1999) Social Identity, Personality, and the Self-Concept: A self categorising perspective. In Tylor, T et al. (Eds.) *The psychology of the social group*. Mahwah: Erlbaum.

VAZQUEZ-SALCEDA, J, Aldewereld, H and Dignum, F (2005) Norms in Multiagent Systems; From Theory to Practice. *International Journal of Computer Systems and Engineering* 20 (4). pp. 225—236.

WILKS, Y (2008) What would a Wittgensteinian computational linguistics be like? In *AISB 2008 Convention Proceedings* Vol 12: Computing and Philosophy.

WINDRUM, P, Fagiolo, G and Moneta, A (2007) Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation* 10(2)8 http://jasss.soc.surrey.ac.uk/10/2/8.html.

ZIEGLER, R (2000) Hat der Homo Oeconomicus ein Gewissen? In Metze, R et al. (Eds.) *Normen und Institutionen: Entstehung und Wirkung.* Leipzig: Leipziger Universitätsverlag.