Social Norms accentuate agent trust and attenuate agent autonomy

Henry Hexmoor and Prapulla Poli
University of Arkansas
Computer Science & Computer Engineering Department
Engineering Hall 313, Fayetteville AR 72701
{hexmoor, ppoli} @uark.edu
Phone: 479 575 2420 Fax: 479 575 5339

Abstract

This paper presents the influence of norms on agent trust and autonomy. Different attributes of norms were considered that accentuate trust and attenuate autonomy. Experimental results illustrate changes in trust and autonomy with different patterns of norm adherence. We also define stability as a result of predictable agent behavior and present related experiments.

1. Introduction

Many social concepts such as benevolence, social networks, social laws, conventions, and power are constructively used in modeling the dynamics of a multi-agent system. Study of social norms is essential for creating predictable behavior among agents in an E-society. The principles of social norms have various impacts in modeling agents' trust and autonomy in a multi-agent system. This paper explores variations in trust and autonomy where agents have equal social power in a social network and adhere to a set of pre-defined norms while performing a set of tasks. This extends our earlier work on building a model of trust and autonomy [Hexmoor and Poli, 2003a and 2003b].

Agents' interactions in a society are often guided by social actions such as norms [Hexmoor, Battula 2003]. A norm can be understood as a behavioral constraint on agents that interact with each other in a multi-agent society [Nicole, et. al, 1999]. Norms are common in human societies. To mention a few, there are norms of traffic, norms for business transactions, norms in schools, and so on. Norms can be represented as goals or obligations that guide, control, or regulate proper and acceptable behavior among agents [Henricus 2000] and even can serve as filters to generate and select goals [Boman 1999]. Much of the work on norms was carried by researchers who have implemented norms in multi-agent system. Castelfranchi argues for a deliberative normative agent with norm impacts on goal, plan generation and selection [Castelfranchi 1999]. We explore applications of norm attributes to multi-agent environments and observe how trust and autonomy among the agents change.

The primary role of norms in the formalization of multi-agent systems is to provide predictable behavior and coordination among agents in a multi-agent system. This can be considered to be a type of behavior stabilization, which plays the same role for such systems as intentions do for single agent systems Boella 2003]. This can be illustrated considering the traffic rules in a city. An example is when people follow the traffic rules carefully without violation then there would be fewer accidents and the entire system remains more stable. However, when people violate the traffic signals it may cause difficulties leading to unstable situations. An experiment in section 4 illustrates this point.

Norms guide agents in performing the tasks by imposing certain conditions or sanctions. These norms have been developed to enhance the reliability of communication, and to make actions of agents more predictable and verifiable. They also facilitate coordination of agent actions [Conte, et. al, 1999]. Norms take different forms in different areas such as *social theory* versus *legal theory*, etc., [Henricus 2000]. In social theory, norms are defined as the rules to be obeyed since they are agreed upon, and are obeyed because of one's conscience etc., Norms in legal theory are accepted out of fear for the authority issuing the norm. They are accepted from a sense of duty. They may be accepted since they solve the problems of coordination and cooperation etc [Henricus 2000]. We consider norm affinity, norm commonality, and other aspects of norms for observing effects on trust and autonomy of agents.

Agents that do not follow norms might face sanctions that have major impact on their autonomy. Not all norms have sanctions. We limit our focus on norms that are in force only when they are followed by at least half the number of agents. The association of violation with sanctions is explained by the notion of *social control*, "an incessant local (micro) activity of its units" [Castelfranchi 2000], aimed at restoring the regularities prescribed by norms. Boyd argued the importance of punishment (i.e., sanctions) in societies [Boyd, et. al, 2003].

The main objective of this paper is to show that various attributes of norms contribute to altering trust and autonomy values, which are in turn used as part of an agent decision to perform a task or to delegate it.

Trust between agents depends on many parameters including competency, histories, reciprocity, benevolence, culture, and reputation [Castelfranchi 1998]. Interpersonal trust is also a function of familiarity and social ties. Interpersonal trust, which is the focus of this paper, is similar to institutional trust. Institutional trust is the trust that exists among individuals due to their participation in social norms and values of various institutions they are members of. Usually, trust levels accumulate and diminish gradually unless there are radical changes in agent attitude toward one another, such as major changes in benevolence [Abdul-Rahman 2000]. Another conceptualization is that, trust is not a precursor to delegation but one between collaborating individuals who communicate. Trust is in the degree of belief in validity of messages. In this notion of trust, capability, benevolence and exchanges of trustee is not in question but the agents' interaction with different power levels is considered. In summary, we suggest that agent X's trust in agent Y about task T (we will denote that by Trust (X, Y, T)), is partly a function of agent's X's perception of agent Y's benevolence towards it, partly a function of agent X's perception of agent Y's capability toward task T, partly due to balance of tit-for-tat [Hexmoor and Poli 2003b], partly due to the power variations with others [Poli and Hexmoor 2003] and partly due to different concepts of norms. This approach to conceptualizing trust lends itself to formulating delegation between two individuals, which requires trust between delegator and delegee [Castelfranchi and Falcone 1998], [Sichman et. al 1993].

Autonomy is a broad topic and touches on many disciplines. For instance, in philosophy, there are many theories about autonomy that discuss freedom, self-control, and individualism [Christman, Anderson 2003], [Schneewind 1997]. What is clear is that an agent must have nontrivial cognitive abilities to reason about its action and decisions to consider autonomy as a property an agent owns. An agent's autonomy towards a task is affected by its capability and the sense of freedom it receives from other agents [Hexmoor and Vaughn 2002]. This sense of freedom can be approximated by a combination of factors such as power, norms and trust. Autonomy is subject to constraints and context of the agent's environment and as such is not strictly determined by interpersonal dependencies.

In the remainder of this paper we will begin by elaborating our model of trust, autonomy, and delegation. In section three, we discuss our implementation we have used for our experimental results. In section four, we describe a series of experiments that illustrate our models. In section five, we draw some conclusions.

2. A model of trust, autonomy and delegation

Our model of trust is aimed at capturing a precondition to the formation of intentions to delegate a task, i.e., asking for a task to be done by another agent. An agent's assessment prior to delegation may include an analysis of risk and utilities, creating an intermediate notion of trusting value, prior to adoption of an intention. In most applications, trust has the consequence of reducing the need for the trusting agent to supervise or monitor the trusted agent.

The variety of definitions has added to the confusion about, and misconceptions of trust. In multi-agent systems, trust has been related to models of other social notions such as autonomy, delegation, dependence, control, power, and norms, which influence interactions between agents. In this paper, we treat trust as a dyadic relation, i.e., the amount of trust each agent has on other agents. We define Trusting value to be the amount of trust an agent has on other agents with respect to a particular task [Hexmoor and Poli 2003a]. This value among the agents is calculated by the following expression:

$$\text{Trusting value (A, B, t)} = (1/4)^* [\text{capability (B, t)} + \text{Benevolence (B, A, t)} + 10^* \text{Delegation of Harmony (A, B)} + 10^* \text{Norm Affinity}] \quad (1)$$

Here A, B are agents and t is the task to be performed by agent B. capability (B, t) is the agent B's ability to perform a task t and we assume both A and B perceive the same value. Benevolence (B, A, t) is agent B's (i.e. trustee's) level of well wishing towards agent A (i.e. trusted) in performing a task t. Delegation of Harmony (A, B) is the number of times agents A and B have agreed to the delegation request from one another after internally weighting all other considerations by the sum of number of times agents A and have made a delegation request. The Delegation of Harmony value range from 0.0 to 1.0. Delegation of Harmony is only computed when agents interact at the interpersonal level [Hexmoor and Poli 2003b]. Norm Affinity is the number of norms two agents share in common which is the notion of *norm commonality*. Restated, norm affinity is the closeness between two agents over time as they share norms in common and is calculated using the following expression

Norm Affinity = cn / ((a - cn) + (b - cn) +cn)

cn – Total number of norms that agents A and B follow in common
a – Total number of norms that agent A follow
b – Total number of norms that agent B follow.

This norm value is between 0 and 1.Norm Affinity ranges from 0.0 to 1.0.

Trust, the way it is defined in this paper is a dyadic trust where trust between two agents is calculated. Social ties are considered to be established among agents who are assumed to relate at interpersonal level [Hexmoor and Poli 2003b]. These ties become stronger with time. When these ties are combined with norms, interactions are controlled by norms that are collective rather than dyadic [Hardin 2002]. This is because all agents are guided by following norms. If any agent does not follow norms, then those agents will experience sanctions that might affect their autonomy level.

Cook offered an alternative explanation for combined synergy between dyadic trust and social ties [Cook and Hardin 2000]. In this alternative, communal norms substitute the need for dyadic trust. Instead the majority generally follows common norms and this controls social order just as interpersonal relationships world. This is evident in the contrast between banking systems in the developed world where communal norms function instead of interpersonal trust and ties in the financial institutions of the newly independent Russian Republic. Another equivalence is the nature of evolution of communal laws. Communal laws are developed due to strong ties and norms of cooperativeness. This offers equivalence between a society governed by a network of dyadic trust and one that relies on a combination of social ties and norms.

Norms and sanctions play major roles in human societies. Autonomy is predominantly affected by sanctions. The following equation presets autonomy computation in our model. Autonomy value for a particular agent is the amount of trust others have on that particular agent to perform a task, and is computed by the following expression.

Autonomy value (A, t) = (1/4)* [capability (A, t) + Average(T) + 1/(n-1)* Balance of reciprocity + fear of norm (A)] (2)

Balance of reciprocity for an agent A is counting two values and subtracting two values:

- Add the number of times delegated tasks by agent A has been agreed upon divided by the number of such agents
- Add the number of times agent A has made a delegation request regardless of accepting that request divided by the number of such agents
- Subtract the number of times agent A has agreed to a delegation request by another agent divided by the number of such agents making the request
- Subtract the number of times agent A has been asked for delegation regardless of whether A has agreed to work on the task divided by the number of such agents.

capability (A, t) is the agent A's ability to perform a task t. Average(T) is the average trust of all the agents on agent A and is measured by

$$\text{Average (T)} = 1/(n-1) \sum_{i=1}^{n} T_i$$

where $T_1, T_2 \ldots T_n$, are the trusting values of the agents on agent A on a particular task t. The amount of trust an agent has on itself determines its competence for performing a task. Fear of norm is calculated from the following equation

$$\text{fear of norm (A)} = (m/n)*(1/n)*\sum_{i=1}^{n} (S)$$

m/n is the ratio of number of norms that an agent has to the total number of norms. S is sanction and is calculated for all norms that an agent follow, which is summation of fear values of all the norms it follow. i is number of iterations which range from 1 to n (n - number of norms). The coefficients are used to normalize the equations to a fixed range. We call this *autonomy* value of the agent as trusting value of an agent on itself. Autonomy (equation 2) of an agent is same as the trusting value of the self-agent. Obviously, Equation (1) affects Equation (2). Agents use trusting values to evaluate their own autonomy in deciding and performing the tasks.

Autonomy is compared with the trusting values of all the agents to determine which agent should perform a task. An agent selects a task considering the norms that it follows. When an agent has more than one task to opt, the agent checks how many norms permit it to perform a particular task by adding the value each norm assign to the task (which is pre-defined). Agent performs the task for which the permit value is higher. Every agent has an individual task assigned to perform. The autonomy of an agent to perform the pre-defined task is compared with the value of autonomy of the agent for whom it selected to perform the task. If the agent's self-autonomy is higher than the autonomy of the selected agent then it performs its own task for itself. When multiple agents determine to perform a unique task, an agent with higher power is performs the task. For agents with equal powers their autonomies with respect to the task are compared. For agents with equal autonomy their capabilities with respect to the task are compared and the agent with the higher capability performs the task. If the agent's capabilities are equal, the task is performed by one of the agents selected randomly.

The success or failure of an agent can be determined by comparing the capability values of an agent with a randomly generated number ranging between 0 to10. If the random number has a value greater than the capability value of an agent, it is considered as a failure and if the number is smaller, then it is considered as a success. An agent may perform one task each time and no two agents can do the same task. The capability, trust and relations among the agents are updated with the success or failure in performing the tasks. The average autonomy and trusting values of the agents are calculated to observe a relation between the two with respect to the norms before updation.

As discussed in section 1, norms have a great deal of impact on agents in establishing stable behavior. Stability of an agent is defined, as ratio of the number of norms that an agent follows for a task to the total number of norms that govern the task. Stability value for an individual directly contributes to stability of a group of agents. There is much work that is carried out by researchers on stability in multi-agent systems [Balakrishnan and Srinivasan 1997], [Chli et. al 2003]. We observed stability of different groups of agents that followed different number of norms in performing their tasks. Agents in each group perform tasks while under influence of assigned norms. When an agent selects a task, a test is performed to check if the agent is following the prescribed number of norms for the assigned task. For each agent in the group stability is calculated and later the average stability of the group is calculated.

3. Simulated Test bed

In our implementation simulation, N agents considered N tasks repeatedly, i.e. each agent has its own task, which is same in each time period. This does not mean that each agent has to perform the assigned task. Agents may perform tasks assigned to other agents. The tasks are performed with certain norms. The agents perform tasks based on the norms they follow. The ranges of capability, benevolence, autonomy and trust are between 0.0 and 10.0.

In our simulation we assume in general agents perform certain tasks and develop trust, capability, and benevolence among them. In the algorithm shown in Figure 1 the aim is to focus on the performance of agents. The following is pseudo code for our simulation.

```
1.Initialize the values of capability matrix (C[][] ) to
   random values between 0 to 10.
2.Initialize the values of Benevolence (B[][][] ) to 0.0
3.Initialize the values of tasknorms[][] between 0 to 2
3.while (tasks remain) {/* main body of the algorithm*/
4.   for all agents and tasks { /* trusting values */
5.     if (a = b)   /* a, b – variables stand for agents*/
          TV [t][a][b]  = C[a][t] + average(T) +
 (Balance of reciprocity)/(n-1) + fear of norm (A)
6.    else TV [t][a][b]  = C[a][t] + B[t][a][b] +
     10*Delegation of Harmony(a, b) + 10* Norm Affinity
7.   A[a][t] = C[a][t] + average(T) + (Balance of
   reciprocity)/(n-1) + fear of norm (A) /* autonomy */
8. compare A[][] with TV[][][] to find the suitable
  agents performing task t
9. compute the number of tasks being executed per
   iteration and unsuccessful attempts.
10. C[][] = C[][] + i   /*Update C[][]  with success */
    C[][] = C[][] – i  /*Update C[][] with failure */
11. B[][][] = B[][][] + i  /*Update B[][][] with success */
    B[][][] =  B[][][] – i   /*Update B[][][] with failure */
12. AA[a][t]  = A[a][t]/(n*n) /*average autonomy*/
13. ATV[t][a][b] = TV[t][a][b]/((n*n)*(n-1))  /*where a!=b
; average Trusting values */
       }  /* for loop*/
     }  /*while loop */
```

Figure 1. Algorithm to compute average trusting values and average autonomy

- Average(T) is the average of trusting values of all agents with respect to self agent on particular task.
- B[][][] is the benevolence matrix of the agents.
- B[][][] initialized to 0.0 represents the lowest benevolence.

- B[][][] initialized to 10.0 represents the highest benevolence.
- TV[][][] is the matrix that holds the trusting values of agents with respect to tasks.
- AA[][] is an average autonomy of all the agents with respect to tasks and is used in plotting the graph.
- ATV[][][] is the average trusting values of all the agents except the self-trusting values with respect to tasks and is used in plotting the graph.
- fear of norm defined in section 2.
- n is the number of agents.
- Tasknorms[][] is the matrix that shows which norms permits a task

4. Experiments and Discussions

This section presents results of using our abstract simulation of agents and tasks. Experiments were performed to observe variations in average trust and average autonomy of groups of agents with various norm components. Number of norms that an agent follows is randomly selected. Each norm is assigned to at least half the agents. We are not modeling norms with low adoption rates. Each norm is modeled to carry a sanction with a value that is randomly selected between 0 and -1. In the first experiment, the results were observed for 25 units of time where average autonomy of the agents is calculated. In the second experiment the results were observed for 25 units of time where average trusting values of the agents are noted. As said earlier, each norm is followed by at least half the agents. Agents share norms in common when they are in a multi-agent system. Sharing of norms is termed *norm commonality*. Agents develop certain trust relation with other agent considering the norms that agents share in common, which is termed *norm affinity*. Average trust of agents was observed with varying norm commonalities and norm affinities among group of agents. Benevolence values started at zero but changed over time with the delegation of tasks by the agents. Power among the agents is fixed and is a randomly generated.

Figure 2 shows variations of average trusting value among a group of agents with varying norm affinity. The average trust among a group of agents increases with increasing norm affinity among the agents. This is because, the more the agents' interact while following same norms, the more the agents are aware of one another by which agents trust levels increase. When the norm affinity among the group of agents is zero (i.e., Norm Affinity = 0) the average trusting value among the agents was 1.0. When the norm affinities among the group of agents were 3 and 5 the average trusting values among the agents were 2.86 and 2.91 respectively. When the norm affinities were 10 and 15 the average trusting values among the group of agents were 3.23 and 3.83 respectively. As the norm affinity among the agents increased the average trusting value among the group of agents increased.

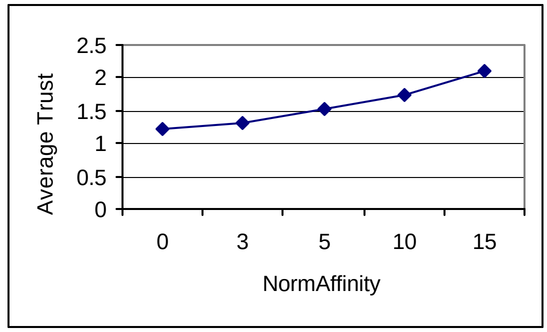

Figure 2. Average trusting value of different group of agents that had different norm affinities were recorded at 25th cycle of simulation

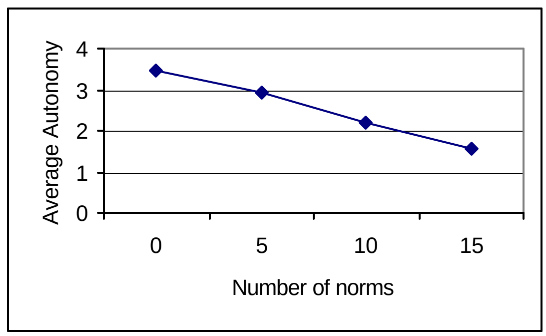

Figure 3. Average Autonomy of different group of agents at 25th cycle unit

The more the number of norms an agent follows the more the sanctions will be. Each norm has its own sanction that when violated apply to the agent that did not follow the norm. The fear of norm has impact on an agent's autonomy. As the number of norms the agents follows increase the average autonomy of the agents decreases. Figure 3 shows the average autonomy of a group of agents with varying norms. We performed experiments where norms were equally distributed among agents in each experiment. When agents have no norms to follow (i.e., number of norms = 0) average autonomy of the agents were higher (i.e., average autonomy = 3.44). Average autonomy of the agents was 1.58 when agents followed the most number of norms (i.e., number of norms = 15). As the number of norms that agents follow increase the average autonomy declines.

Further experiments were performed to show how the commonalities in norms affect the trust levels among the group of agents. The average trust among the group of agents with high norm commonality, i.e., the number of norms that agents share are high, have higher trust levels than the group of agents that have low number of norms in common. As the number of adopted norms increase the average trust level of the group of agents increase. In this experiment, we setup 5 different groups each with different norm commonalities. Average trust of each group was observed. The group of agents that had no norms in common (i.e., norm commonality = 0), had an average trust of 1.19. Average trust of the groups that have 3 and 5 norms in common were 1.28 and 1.32 respectively. The group of agents that have all norms in common have the highest average trusting value of 1.72.

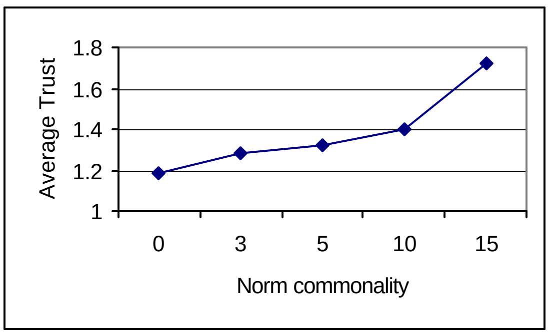

Figure 4. Average trusting value of different group of agents with different norm commonalities were recorded at 25th cycle of simulation

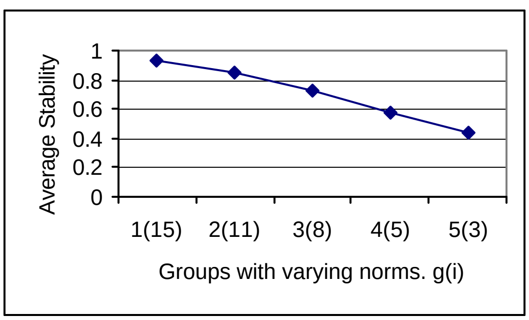

Figure 5. Stability of five groups. Each group has the same number of agents following different norms. Results were observe at 25th cycle unit for each group

Another set of experiments were performed to observe variations in stability of agents' as agents perform the tasks under various number of norms. Each task is assigned certain number of norms. In the experiment agents select a task depending on their capability for performing a task, i.e., their autonomy. The experiment shows stability curve when agents follow norms differently. The experiment was performed considering 5 different groups. The values of each group in the experiment were noted at 25th cycle unit. Figure 5 shows the variations in stability. This graph clearly illustrates that stability declines as the number of norms the group of agents follows in performing a task decrease. The total number of norms prescribed in our experiment to perform a task was 15. The group of agents that followed all the norms attained an average stability of 0.93 and the group of agents that followed only 3 norms out of 15 norms exhibited an average stability of 0.44. 'g(i)' in the graph represents the number of norms (i) the agents in group g are following.

5. Conclusion

A simple model of autonomy and trust was presented that relied on norms that agents followed. Presenting many parameters will have obscured our observations. This model is deliberately kept simple to illustrate the role of norms that agents obeyed in the relationship between autonomy and trust separately. An agent experiences autonomy with respect to a task if it is capable of performing or when it follows the norms without violations. The notion of norm is shown to affect trust and autonomy. Norms largely accentuate average trust among the group of agents and attenuate average autonomy of the group of agents. Also, predictability of agent behavior, which we called stability, is shown to declines as the number of norms that agents should follow to perform a task decrease.

Acknowledgements

This work is supported by AFOSR grant F49620-00-1-0302.

References

A. Abdul-Rahman, S. Hailes 2000. Supporting Trust in Virtual Communities. In: Proceedings *Hawaii International Conference on System* Sciences *33, Maui, Hawaii, 4-7 January-2000*.

H. Balakrishnan and S. Srinivasan 1997. Analyzing stability in wide-area network performance. *ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, 1997.

G. Boella and L. van der Torre 2003. Attributing Mental Attitudes to the Normative Systems, In *AAMAS'03*, July 14–18, 2003, Melbourne, Australia.

M. Boman 1999. Norms in artificial decision-making. Artificial Intelligence and Law, 7(1): 17-35, March 1999

R. Boyd, H. Gintis, S. Bowles, and P. J. Richerson 2003. The evolution of altruistic punishment. In Proceedings *of the National Academy of Sciences*, 100:3531–3535, 2003.

C. Castelfranchi 2000. Engineering social order. In *Proceedings of ESAW00*, Berlin, 2000.

C. Castelfranchi, F. Dignum, M. Catholijn, Jan Treur. Jonker 1999. Deliberate Normative Agents: Principles and Architecture (1999) pp. 364-378, ATAL-99.

C. Castelfranchi, & R. Falcone 1998. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In: Proceedings of *the Third International Conference on Multi-Agent Systems,* Paris, France, pp 72-79, 1998.

M. Chli, P. De Wildez, J. Goossenaerts, V. Abramov, N. Szirbik, L. Correia, P. Mariano, R. Ribeiro 2003. Stability of Multi-Agent Systems. In the IEEE Systems, Man and Cybernetics 2003.

J. Christman, J. Anderson 2003. (Eds) Forthcoming 2003. *Autonomy and the Challenges to Liberalism: New Essays*, http://faculty.la.psu.edu/jchristman/autonomy/autpapers.html

R. Conte, R. Falcone, and G. Sartor 1999. Agents and Norms: How to fill the gap. *AI and Law***,** *special issue on Agents and Norms.*

K. S. Cook and R. Hardin 2000. "Networks, Norms, and Trust-worthiness." In Social Norms, edited by Karl-dieter Opp and Michael Hechter. New York: Russell Sage Foundation.

R. Hardin 2002. In *Trust and Trustworthiness* Vol IV in the Russell sage foundation series on trust. pp 183-184.

V. Henricus 2000. Norm Autonomous Agents, PhD thesis, Department of Computer and Systems Sciences, The Royal Institute of Technology and Stockholm University, Sweden, 2000.

H. Hexmoor 2001. Stages of Autonomy Determination. In: *IEEE Transactions on Man, Machine & cybernetics-Part C* (SMC-C), Vol. 31, No. 4, pp 509-517, November- 2001.

H. Hexmoor and J. Vaughn 2002. Computational Adjustable Autonomy for NASA Personal Satellite Assistants. In: *ACM SAC-02*, Madrid, 2002.

H. Hexmoor, S. Battula 2003. Discovering Human Desired Norms of Interaction Improves Social Adeptness in Agents, In AAAI Spring Symposium on Human Interactions with Autonomous Systems in Complex Environments, Stanford, CA, AAAI press.

H. Hexmoor and P. Poli 2003a. Benevolence, Trust and Autonomy. In: Proceedings of the International Conference on Artificial Intelligence (IC-AI'03) Las Vegas, Nevada, 2003.

H. Hexmoor and P. Poli 2003b. Effects of Reciprocal Social Exchanges on Trust and Autonomy. In: Proceedings of Third International Conference on Intelligent Systems Design and Applications (ISDA-03) Tulsa, Oklahoma, 2003.

J. Nicole, Saam and H. Andreas 1999. Simulating Norms, Social Inequality, and Functional Change in Artificial Societies, JASSS, 1999.

P. Poli and H. Hexmoor 2003. Effects of Power on Trust and Autonomy, In AAMAS-03 workshop titled: Autonomy, Delegation, and Control, H. Hexmoor (Ed), Melbourne, Australia.

J. B. Schneewind 1997. *The Invention of Autonomy: A History of Modern Moral Philosophy*, Cambridge Univ Press.

J. S. Sichman, Y. Demazeau, R. Conte, C. Castelfranchi 1993. A Social Reasoning Mechanism On Dependence Networks. In: Proceedings of *the Eleventh European Conference on Artificial Intelligence*, 1993.

J. D. Thomas and K. Sycara 1998. Heterogeneity, stability, and efficiency in distributed Systems. *Proceedings of the 1998 International Conference on Multi Agent Systems*, 1998.