

Benevolence, Trust and 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. *In this paper we will discuss tradeoffs between trust and autonomy. We show that benevolence play a central role in connecting these two social notions. We will show empirical results that illustrate the affinity between trust and autonomy.*

Keywords: Agent, Trust, Autonomy, and Benevolence

1. Introduction

Whereas the sense of being trusted contributes to an agent's autonomy, the sense of trusting contributes to an agent's motivation to delegate. In this paper we will compare agent communities with differing benevolence among them. Differing benevolence levels contribute to different trends of trusting and trusted values, which lead to decision to self-act versus to delegate. When agents know one another, they develop a balance of good will and help toward one another. We model this in general in terms of an attitude of benevolence among agents [9]. Agents might have general attitude of benevolence towards other agents, but in this paper we treat benevolence as a dyadic relation. If an agent X has high benevolence towards agent Y, when agent Y asks agent X for a task, agent X is very willing to accept the request. Furthermore, agents usually reciprocate with benevolence, i.e., benevolence usually begets more benevolence [11 and 12]. However, benevolence is not necessarily symmetric. Agent X's benevolence towards agent Y is not

necessarily the same as agent Y's benevolence towards agent X. Benevolence does not subtract an agent's autonomy. On the contrary, it might increase it. Since being benevolent contributes to being trusted by others, which might lead to being a delegee, i.e., being given many tasks. If such an agent is highly competent as well, the agent will accumulate successes. Combination of increased competence and being trusted in will contribute to agent's sense of autonomy. On the contrary, lack of benevolence toward other agents will lead to lack of being trusted and that leads to lack of autonomy. Benevolence might be thought of as a good social norm to adopt. Although in general, norm abiding does not directly affect trust or autonomy. Norm adoption affects an agent's reputation, which may have effects on trust.

There have been many studies of trust, which have introduced many definitions of trust, e.g., [5, 3, 1, and 6]. We suggest that agent X's trust in agent Y about task T (we will denote by Trust (X, Y, T)), is partly a function of agent's X's perception of agent Y's benevolence towards it and partly a function of agent X's perception of agent Y's capability toward task T. There are many other accounts of interpersonal trust [8]. Our conceptualization is closest to Castelfranchi's where they base trust on competence and willingness of trustee [2]. This approach to conceptualizing trust lends itself to formulating delegation between two individuals, which requires trust between delegator and delegee. Although we agree on the competence component of trust, we feel that Castelfranchi's definition could apply to complete strangers and

does not take the interpersonal sense of trust we are interested in. The “willingness” component does not capture trustee’s attitude towards trusted. Interpersonal trust is also a function of familiarity and social ties. Social ties affect trust temporally. Usually, trust levels accumulate and diminish gradually unless there are radical changes in benevolence. 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 of trustee is not in question but truth telling is.

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 [10 and 4]. 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. In this paper, we consider personal autonomy, with endogenous and exogenous sources, which an agent uses to determine the nature of interaction with other agents regarding a specific task [7]. The agent’s perceived degree of autonomy is used in deciding to consider the task for self or others. In this paper we will limit endogenous sources to be the agent’s capability regarding the task. The exogenous sources might be powers, the agent perceives from other agents or permissions it gets from others. We model the permissions in terms of trust it perceives from others agents.

In the remainder of this paper we will begin by describing our model of trust and delegation. In section three we present implemented simulation we have used for our empirical results. In section four, we will describe a series of experiments that show effectiveness of approach. In section five, we draw conclusions about relationships between trust and autonomy.

2. A Model of Trust and Autonomy

Our model of trust is aimed at capturing a precondition to the formation of intentions to delegate. 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, and power, 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 the other agents. We define **trusting value** to be the amount of trust an agent has on the other agents with respect to a particular task. This value among the agents is calculated by the following expression:

$$\text{Trusting value (A, B, t)} = \text{capability(B, t)} + \text{benevolence(B, A, t)} \quad (1)$$

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 **benevolence(B, A, t)** is how an agent B (i.e. trustee) well wishing towards agent A (i.e. trusted) in performing a task t.

The Autonomy of an agent is the ability in performing a task by itself and is computed by the following expression.

$$\text{Autonomy (A, t)} = \text{capability (A, t)} + \text{Average(T)} \quad (2)$$

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

$$\frac{1}{n-1} \sum_{i=1}^n T_i$$

where $T_1, T_2 \dots T_n$, are the trusting values of the remaining 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. We call this *autonomy* value of the agent as trusting value of an agent on itself. Said differently, the autonomy (equation 2) of an agent is same as the trusting value of the self-agent. Obviously, equation 1 affects equation 2. Let’s take an extreme example of two agents A and B with B benevolent towards A with all tasks but A is not benevolent towards B with any task. Equation 1 will promote trust of agent

A in B. If B accepts delegated tasks and successfully executes them, B increases its autonomy using equation 2. The benevolent agent, B, gains in autonomy. A's autonomy has little chance of increase since it is not trusted by B.

Autonomy is compared with the trusting values of all the agents to determine which agent should perform a task. Every agent has an individual task assigned to perform. This autonomy of an agent to perform the pre-defined task is compared with the autonomy of the overall tasks determined. The agent performs a task for which the autonomy is highest. When multiple agents determine to perform a unique task, the task is performed by an agent whose autonomy is higher with respect to the task. 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.

From the derived expressions it can be observed that benevolence directly affects the trusting value and indirectly affects the autonomy of the agents.

3. A Simulated Testbed

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 either benevolence fixed or benevolence changing during a run. The initial value of fixed benevolence is set to either 0.0 or 10.0. At the beginning, an initial value of 0.0 is considered as no benevolence among the agents and an initial value of 10.0 is considered when there is high benevolence among agents. The ranges of capability 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[])
between 0.0 to 10.0.
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)
6. else TV [t][a][b] = C[a][t] + B[t][a][b]
7. A[a][t] = C[a][t] + average(T) /* 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[] + 2 /*Update C[] with success */
C[] = C[] - 2 /*Update C[] with failure */
11. B[][] = B[][] + 2 /*Update B[][] with success */
B[][] = B[][] - 2 /*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.
- n is the number of agents.

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 to 10. 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 lesser it is considered as a success. An agent may perform many number of tasks and the same tasks may be repeated. The success of

a task is dependent only on the capability of the agent (as compared to the random number). Benevolence is important when the agent may or may not cooperate (and perform the task delegated to it) and benevolence does not play a part in the performance. The capability and benevolence among the agents are updated, from which the autonomy and trusting values are updated. The update is performed by adding an increment value of 2 to capability and trust of the successful agents. The values are updated by 2 for simplicity. The average autonomy and trusting values of the agents are calculated to observe a relation between the two with respect to benevolence. The rate of successful tasks and unsuccessful attempts is measured as a factor of time with respect to benevolence.

4. Experiments and Discussions

This section presents results of our abstract simulation of agents and tasks. Four experiments were performed considering three agents and three tasks. The fifth experiment was conducted with six agents and six tasks. The results were observed for 25 units of time. In each time unit the average autonomy, average-trusting values, number of successful tasks and the number of unsuccessful attempts were noted. Two cases of benevolence among agents were considered for simulation results.

Benevolence = 0.0: This is the lowest benevolence considered. Two runs were performed, one with constant benevolence and the other changing.

Benevolence = 10.0: This is the highest benevolence. As in the previous case, two runs were performed, one with constant benevolence and the other with changing.

In the remainder of this section we will state our observations about the relationships between average autonomy, average trust and benevolence of all the agents.

Initially, the lowest and highest benevolence values are 0.0 and 10.0 respectively. Figures 2 and 3 show the results of the agents, with the lowest benevolence. Figure 2 shows the average trust and average autonomy with benevolence at zero. In one run the benevolence is fixed at 0.0 for the entire run and in the next run benevolence is 0.0 initially, but its values change with time. As seen in Figure 2, average trust and

average autonomy values stayed low with time when the benevolence was fixed, whereas the trust and autonomy of the agents with changing benevolence increased gradually. Both the values of average trust and average autonomy of constant and changing benevolence remained at a near constant level after a certain time period. As the trust level increased very slightly (i.e. 1.38) at time $t = 1$, autonomy increased to a higher level of 1.85 at time $t = 1$ when the benevolence is fixed. When the benevolence is changing the trust level increased to a higher level (i.e. 1.81) at time $t = 3$, autonomy increased to a much higher level of 3.62 at time $t = 3$. Autonomy values are at a higher level than the trust values: This is because, when agents are less benevolent towards each other they do not trust each other influencing the agent to work independently there by increasing their autonomy level. As seen in Figure 2 the rate of change in average autonomy and average trust are higher when benevolence is changing. In Figure 2, T0F and A0F are the average trust and average autonomy curves of the agents when the benevolence is fixed. T0C, A0C represent the curves of the average trust and average autonomy of the agents when benevolence is changing.

Figure 3 shows the results of cumulative tasks considered or completed over time. When benevolence is low and fixed, the number of tasks being performed (S0F in Figure 3) was less whereas the range of success (S0C in Figure 3) in completing the tasks was high for changing benevolence with time. When benevolence is fixed at 0.0 the number of unsuccessful attempts (U0F in Figure 3) is either the same or less than the number of unsuccessful attempts for changing benevolence (U0C in Figure 3). With low benevolence, there might not be much trust among the agents in performing the task and the chance of increase in success rate is high as benevolence changes with time. Comparing the slopes of successful tasks and unsuccessful attempts of both fixed and changing benevolence can explain this discussion more clearly. From Figure 3 it can be noted that the slope of all the curves is almost zero at time $t = 0$. At time $t = 10$, the slopes of successful tasks and unsuccessful attempts were respectively 1.1 and 2.2 for fixed benevolence, for changing benevolence the values were respectively 2.2 and 1.1. The slope of successful tasks and

unsuccessful attempts with respect to fixed benevolence at time $t = 25$ (i.e. 0.0) were 1 and 2 and that of changing benevolence were 2 and 1 respectively. Therefore, the change in benevolence increases the slope of the successful tasks and decreases the slope of unsuccessful attempts with time.

Figures 4 and 5 show the results of the agents with the highest benevolence (i.e. 10.0). Figure 4 shows how the average autonomy and average trust are related with high benevolence. In one run, benevolence is kept constant at 10.0 and in the next run the value of benevolence changes with 10.0 as the initial value for all the agents.

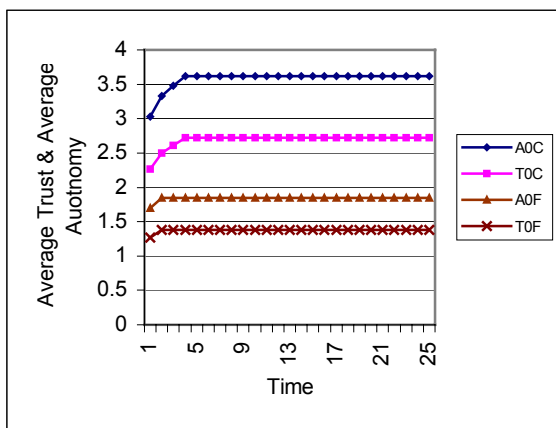


Figure 2. Average trust and average autonomy with respect to time at Benevolence = 0.0

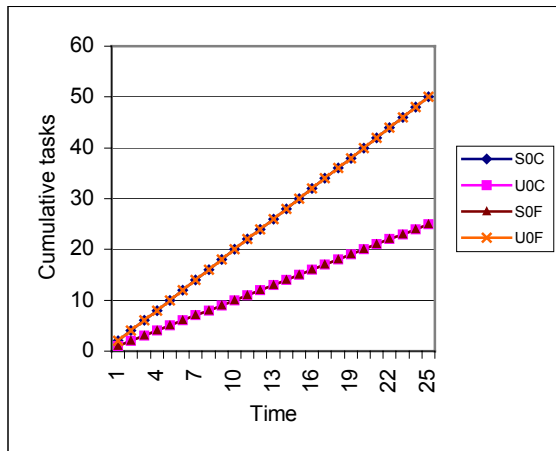


Figure 3. Cumulative tasks with respect to time at Benevolence = 0.0

Since benevolence is directly proportional to the trusting values and indirectly proportional to the autonomy, when the agents are more benevolent towards each other the trust among the agents will be higher and also the average trust among the agents is higher than their average autonomy. As the average trust

increases gradually, the average autonomy of agents also increase. Both average trust and average autonomy remain stable from a certain level. The average trust and average autonomy of the changing benevolence were higher than that of the average trust and average autonomy when the benevolence is fixed. From this it can be said that the autonomy and trust of the agents are higher when benevolence among the agents is changing over time rather fixed. In Figure 4, T10F and A10F are the average trust and average autonomy curves of the agents when the benevolence is fixed. T10C, A10C represent the curves of the average trust and average autonomy of the agents when benevolence is changing. The variations in average trust and average autonomy of the agents are affected by the capability of the agents, as the capability is also a factor.

Figure 5 shows the results of cumulative tasks considered or completed over time. With benevolence high (10.0) and fixed, the rate of unsuccessful attempts (U10F in Figure 5) was higher and the number of tasks (S10F in Figure 5) being performed was few. The range of success in completing the tasks (S10C in Figure 5) was higher when benevolence was changing with time and the unsuccessful attempts (U10C in Figure 5) were fewer (almost zero). The chances of increase in success rate are high as benevolence changes with time. Considering the slopes of the curves with time, the successful and unsuccessful attempts of both fixed and changing were almost 0 for time $t = 0$. At time $t = 10$ the slopes of the curves for fixed benevolence were 2.1 and 1.2 and that for changing were 3.2 and 0.1. At $t = 25$ the changes in the slopes were 1.96 and 1.04 for fixed benevolence whereas for changing benevolence the slopes were 2.96 and 0.04. Therefore, with a changing benevolence, the number of unsuccessful attempts can be reduced and the rate of successful tasks can be increased with time.

When comparing the trust and autonomy with low and high benevolence, it is clear that trust among the agents and autonomy of the agents will be high with high benevolence. Hence, being benevolent will improve the trust and autonomy levels of the agents. Also, the number of unsuccessful attempts was very few almost down to zero and the number of completed tasks was higher. The trust and autonomy along with

the number of completed tasks will increase when benevolence is kept changing over time. The higher the benevolence, the higher are the agents trust and autonomy. They remain constant after certain period of time, though.

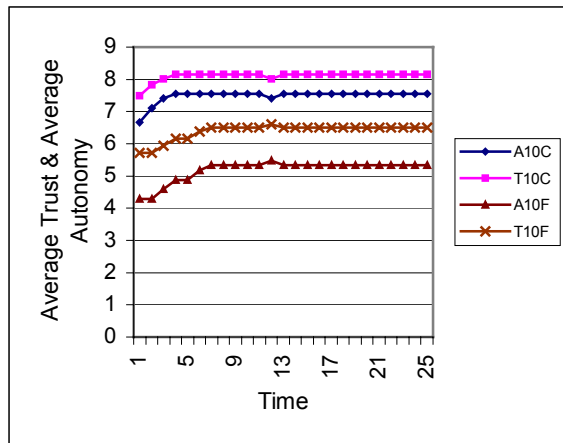


Figure 4. Average trust and average autonomy with respect to time at Benevolence = 10.0

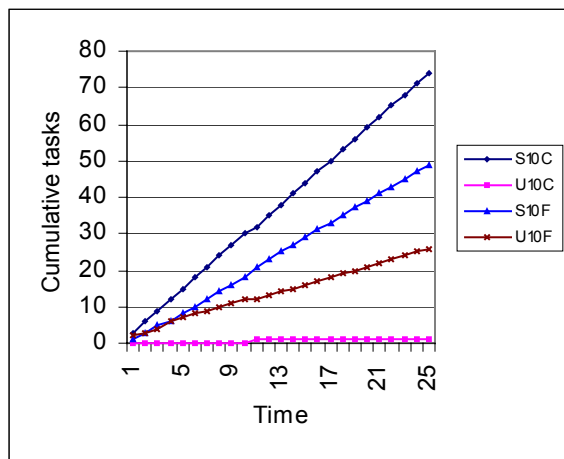


Figure 5. Cumulative tasks with respect to time at Benevolence = 10.0

The fifth experiment was performed considering six agents and six tasks at a time. The results were observed for 25 units of time. For each time unit the average autonomy and trusting values were noted. In this experiment two groups were considered each consisting of three agents. The agents in the first group have high benevolence (i.e. 10.0) towards the second group and among themselves. The second group of agents has low benevolence (i.e. 0.0) towards the first group of agents and among themselves. The average trusting values of the first group on the second and vice versa were noted for each time unit. The average autonomy of the first and second group of agents for each time unit were also noted and a graph of the average trusting

values and average autonomy with respect to time was plotted as shown in Figure 6. We discuss the average trusting values and average autonomy of the two groups by considering their slopes from the Figure 6. The slopes of the trusting values of the first and second group were 6.75 and 3.05 at $t = 1$ and at time $t = 3$ the slopes were 2.32 and 1.0 respectively. At time $t = 12$ the slopes were 0.57 and 0.24 respectively. From these slopes it is observed that the trusting values of the first group with higher benevolence have higher slopes than the second group with lower benevolence, since the first group of agents are more benevolent towards the second group. Considering the values of the average autonomy of the first and second groups, at time $t = 0$ the average values were 5.25 and 4.33 respectively. At time $t = 10$ the values were 5.29 and 4.55 respectively and at time $t = 15$ the values were 5.33 and 4.44 respectively. From these values it can be observed that the average autonomy of first group was higher than that of the second. The average autonomy and average trust of both the groups increased gradually and then continued with a constant slope over time.

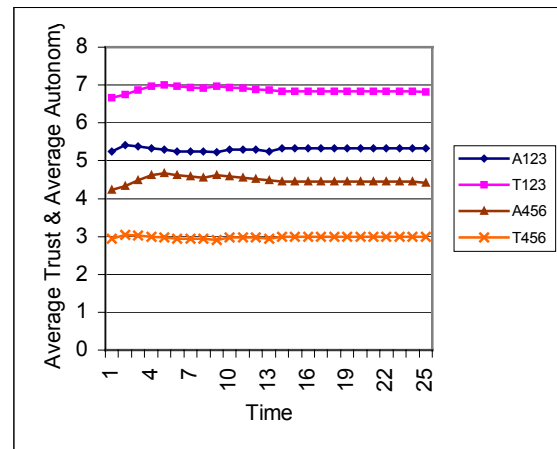


Figure 6. Average trust and average autonomy with respect to time

Since the benevolence of the first group towards the second is higher, their trusting values towards the second group were higher. Similarly, as the benevolence of the second group towards the first group is low, their trusting values towards the first group are low. When considering the autonomy of the two groups, though both the groups' autonomy increased and stayed constant over time, the autonomy values of the first group were higher than that of the second group. This is because of the high

benevolence among the agents in the first group. Hence, the higher the benevolence higher will be the trusting values and the autonomy of the agents.

5. Conclusion and Future work

We presented a simple model of autonomy and trust that relied on benevolence between agents. This model is deliberately kept simple to illustrate the role benevolence plays in the relationship between autonomy and trust. An agent trusts a second agent in consideration of delegating a task to it if the second agent is capable of performing the task and it is benevolent towards the first agent. An agent experiences autonomy with respect to a task if it is capable of performing it and it is trusted by other agents with regard to the task. There are many other parameters that affect trust and autonomy. These parameters in general have to do with the relationship among agents and their interactions. Our simple model can be easily extended to include other parameters. Presenting many parameters will have obscured our observations. We have seen from our experiments that when benevolence among agents is low, their autonomy is higher than their trust. When benevolence is high, trust is greater than autonomy. When benevolence is kept low, fewer tasks are completed and vice versa. The latter makes sense in terms of fewer agents considering task delegation.

Acknowledgement

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

References

- [1] M. Prietula, K. Carley, 2001. Boundedly Rational and Emotional Agents Cooperation, Trust and Rumor, In *Trust and Deception in Virtual Societies*. Edited by Cristiano Castelfranchi and Yao-Hua Tan.
- [2] C. Castelfranchi and R. Falcone. *Principles of trust for MAS: Cognitive anatomy*, social importance, and quantification. In *Proceedings of the Third International*

Conference on Multi-Agent Systems, pages 72--79, Paris, France, 1998.

- [3] C. Castelfranchi, Y. Tan (Eds) 2001. *Trust in Virtual Societies*, Dordrecht: Kluwer Academic Publishers.
- [4] J. Christman J. Anderson (Eds) Forthcoming 2003. *Autonomy and the Challenges to Liberalism: New Essays*, J. Christman J. Anderson (Eds) Forthcoming 2003. *Autonomy and the Challenges to Liberalism: New Essays*, <http://faculty.la.psu.edu/jchristman/autonomy/autpapers.html>
- [5] D. Gambetta, Can We Trust Trust?, in Gambetta, Diego (ed.) *Trust: Making and Breaking Cooperative Relations*, electronic edition, Department of Sociology, University of Oxford, chapter 13, pp. 213-237, 2000.
- [6] R. Hardin, 2002. *Trust and trustworthiness*, Sage publishing.
- [7] H. Hexmoor and J. Vaughn, 2002. Computational Adjustable Autonomy for NASA Personal Satellite Assistants, In *ACM SAC-02*, Madrid.
- [8] 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*.
- [9] A. M. Mohamed, 2000. *Benevolent Agents*, PhD thesis. Univ. of South Carolina Center for Information Technology.A
- [10] J. B., Schneewind, 1997. *The Invention of Autonomy :A History of Modern Moral Philosophy*, Cambridge Univ Press.
- [11] S. Sen, 1996. Reciprocity: a foundational principle for promoting cooperative behavior among self-interested agents," *Proceedings ICMAS-96*, pages 322-329, Kyoto, Japan.
- [12] R. L. Trivers, 1971. The Evolution of Reciprocal Altruism, In *Quarterly review of Biology*, volume 46, pages 35-56, 1971.

