Benevolent Agents: Are They Autonomous and Rational?

Abdulla M. Mohamed[1]
Center for Information Technology
Department of Electrical & Computer Engineering
University of South Carolina
Columbia, SC 29201, USA
(803) 777-9540
mohamed@engr.sc.edu

Michael N. Huhns
Center for Information Technology
Department of Electrical & Computer Engineering
University of South Carolina
Columbia, SC 29201, USA
(803) 777-5921
huhns@sc.edu

ABSTRACT

Philosophers, sociologists, psychologists, and biologists have studied the concept of benevolence for many years. Recently, researchers in AI have begun considering it, but they have chosen a definition based on the mathematical utility for an individual agent. This definition is incomplete. As a result, many researchers criticize benevolence, thinking it contradicts both autonomy and rational theory. In this paper, we argue that benevolence should also have a classical basis that recognizes the moral goodness of an agent and includes social awareness. We present a complete definition of benevolence and show that benevolent agents are indeed autonomous and rational.

Keywords

Benevolence, autonomy, rationality, cooperation.

1. INTRODUCTION

Nwana and Wooldridge state that agent technology is the most "rapidly" growing area in the world of computer science. But, there is no agreement among the researchers on what an agent is [13]. Similarly, there is no agreement on what is a benevolent agent. Definitions of benevolence for agents are split into two different strands. Researchers like Castelfranchi, Conte, Jennings, Wooldridge, d'Inverno, and Luck define benevolent agents as those that accept all other agents' requests for help. For example, d'Inverno and Luck describe a benevolent agent as "an agent for the requesting agents" [10]. Other researchers, such as Rosenschein and Genesereth, define benevolent agents in terms of the similarities of their goals. They believe that benevolent agents have common or nonconflicting goals, and they call this part of the paradigm the *benevolent agent assumption* [19].

Some researchers, such as Castelfranchi, criticize benevolent behavior. They believe that a benevolent agent is not autonomous, because it adopts other agent's goals upon request. In addition, they think that there is a contradiction between benevolence and rational behavior. They consider a benevolent agent irrational, because there are no benefits for it in achieving others' goals. In other words, a benevolent agent is not a rational agent, because it uses its own limited resources for the benefit of other agents without any reward. Based on Castelfranchi's concept of rationality, any action should have a cost, and an agent pays the cost to achieve its goals. Thus, an agent should choose its action rationally and save its resources to achieve its own goals [4,5].

2. BACKGROUND

Figure 1 highlights the major studies of benevolence in different scientific fields, such as philosophy, biology, game theory, social psychology, computational social psychology, and artificial intelligence.

2.1 Prominent Definitions of Benevolent Agents

Goal adoption is classified into three types, namely, terminal, instrumental, and cooperative adoption. Terminal adoption, also called benevolent, occurs when an agent adopts others' goals without any personal advantages in mind, and the goal will not help the agent to achieve any of its internal goals. Instrumental adoption occurs when an agent adopts others' goals with some personal advantage for itself. For example, feeding chickens helps them grow (satisfying their goal), and at the same time, it provides us with more food to eat (satisfying our goal). Finally, cooperative adoption happens when an agent adopts a goal because it is shared with another agent [7].

In Castelfranchi's later work, he modified his view of benevolent agents: this work emphasized the fact that a benevolent agent must adopt other agents' goals and interests without being asked by the recipient agents, and even without the recipients' expectations [7].

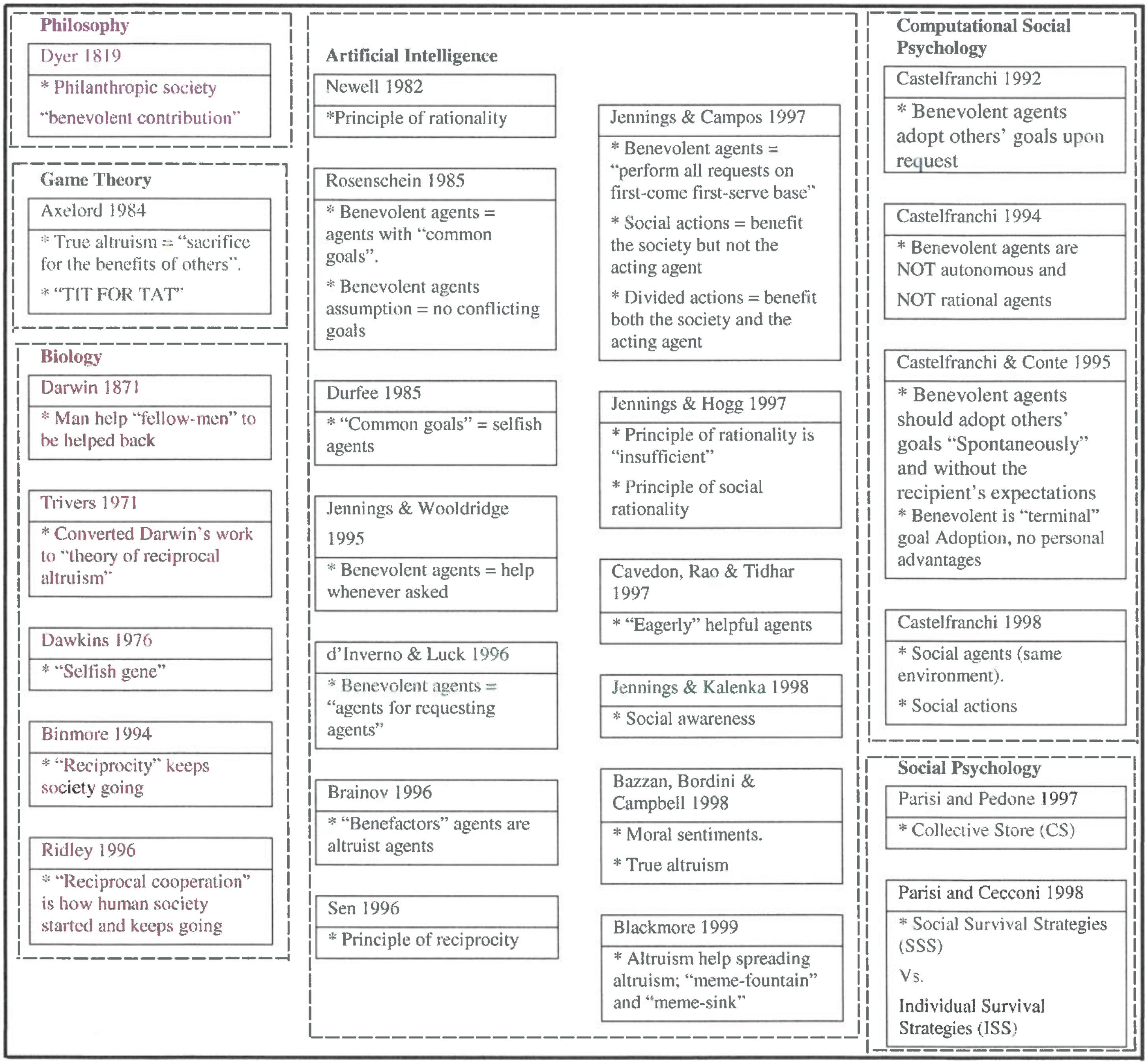
Jennings and Wooldridge define a benevolent agent as one that helps another agent whenever it is asked [12]. Similarly, Jennings and Campos termed benevolent agents as those that perform all goals that they are capable off on a first-come first-serve basis and accept all requests [14]. Moreover, Jennings and Kalenka, while describing a good decision-making function, selected benevolence. The function of benevolent decision is to "accept all requests made" [16].

Rosenschein defines in his Ph.D. dissertation benevolent agents as those that "hold common goals" [20]. In addition, Rosenschein and Genesereth state that all DAI studies assumed that all agents

[1] Mr. Mohamed is a Ph.D. candidate in computer engineering at the University of South Carolina.

have nonconflicting goals. Researchers had focused on how agents could help each other achieve their common goals or how they could use common resources without interfering with each other. In reality, not all agents are benevolent; they don't all have common goals or help each other benevolently. Each agent has its own goals and intentions that it would like to achieve [19].

Durfee, Lesser, and Corkill think that Rosenschein and Genesereth [19] miscalled the agent that shares some goals a benevolent agent. In contrast, they think that these agents are selfish, because they only take actions that will help them achieve their own "interpretation" of the goals [13].



Philosophy

Dyer 1819
* Philanthropic society
"benevolent contribution"

Game Theory

Axelord 1984
* True altruism = "sacrifice for the benefits of others".
* "TIT FOR TAT"

Biology

Darwin 1871
* Man help "fellow-men" to be helped back

Trivers 1971
* Converted Darwin's work to "theory of reciprocal altruism"

Dawkins 1976
* "Selfish gene"

Binmore 1994
* "Reciprocity" keeps society going

Ridley 1996
* "Reciprocal cooperation" is how human society started and keeps going

Artificial Intelligence

Newell 1982
*Principle of rationality

Rosenschein 1985
* Benevolent agents = agents with "common goals".
* Benevolent agents assumption = no conflicting goals

Durfee 1985
* "Common goals" = selfish agents

Jennings & Wooldridge 1995
* Benevolent agents = help whenever asked

d'Inverno & Luck 1996
* Benevolent agents = "agents for requesting agents"

Brainov 1996
* "Benefactors" agents are altruist agents

Sen 1996
* Principle of reciprocity

Jennings & Campos 1997
* Benevolent agents = "perform all requests on first-come first-serve base"
* Social actions = benefit the society but not the acting agent
* Divided actions = benefit both the society and the acting agent

Jennings & Hogg 1997
* Principle of rationality is "insufficient"
* Principle of social rationality

Cavedon, Rao & Tidhar 1997
* "Eagerly" helpful agents

Jennings & Kalenka 1998
* Social awareness

Bazzan, Bordini & Campbell 1998
* Moral sentiments.
* True altruism

Blackmore 1999
* Altruism help spreading altruism; "meme-fountain" and "meme-sink"

Computational Social Psychology

Castelfranchi 1992
* Benevolent agents adopt others' goals upon request

Castelfranchi 1994
* Benevolent agents are NOT autonomous and NOT rational agents

Castelfranchi & Conte 1995
* Benevolent agents should adopt others' goals "Spontaneously" and without the recipient's expectations
* Benevolent is "terminal" goal Adoption, no personal advantages

Castelfranchi 1998
* Social agents (same environment).
* Social actions

Social Psychology

Parisi and Pedone 1997
* Collective Store (CS)

Parisi and Cecconi 1998
* Social Survival Strategies (SSS)
Vs.
Individual Survival Strategies (ISS)

Figure 1: Highlights of benevolence research by different scientists

2.2 Benevolence Vs. Autonomy

d'Inverno, Luck, and Wooldridge, throughout their framework of social structure, assume autonomous agents are not benevolent. They suggest that, "Crudely, the benevolence assumption states that agents will always attempt to do what is requested of them: they are not autonomous". Also, they think "benevolence is reasonable for many distributed problem-solving systems, it is not an appropriate assumption in most multiagent scenarios" [11].

Castelfranchi criticizes benevolent behavior because benevolent agents are not autonomous. Castelfranchi says, "We don't want universally and genetically benevolent agents; they are neither autonomous nor rational". He wants agents who adopt other agents' goals only if there is a benefit for the helping agent itself, which is an instrumental and cooperative agent [5].

d'Inverno and Luck stated that an autonomous agent would behave in a strictly selfish manner. Thus, benevolent, altruistic, trusting, sympathetic, and cooperative agents are not truly autonomous agents [9].

Bazzan, Bordini and Campbell quote d'Inverno and Luck [9] and claim that, "Cooperation will occur between two parties only when it is considered advantageous for each party to do so. Thus, autonomous agents are selfish agents, and benevolence could exist for only selfish reasons" [1].

2.3 Benevolence Vs. Rationality

Castelfranchi, Miceli and Conte asked the famous question about benevolent and rational behavior: why should agents adopt each other's goals? They believe that benevolence contradicts rational theory. Looking from the goal adoption theory, they think that benevolence can exist but it is unnecessary [3].

Jennings and Hogg mapped the *Principle of Social Rationality* to a utility-based function in order to maintain a balance between individual and societal needs. Their utility-based function is the sum of the differences of the benefits and losses (costs) of both the acting agent and its society due to some action. In addition, they defined two other functions: the *Expected Individual Utility* (EIU) and the *Expected Social Utility* (ESU). A social agent will put more emphasis on ESU, whereas a selfish agent will put more on EIU [15]. Clearly, benevolent agents will concentrate on ESU since benevolence is a social concept, but they will not totally ignore EIU.

3. PROPOSED DEFINITION OF BENEVOLENT AGENTS

Benevolent agents have been defined, characterized, and analyzed by a number of researchers, primarily computational and social psychologists. But other fields of science, such as philosophy and biology, addressed the concept of benevolence a long time before AI got started. Some used the term benevolence; others used altruism to describe the same phenomenal behavior.

In 1871, Darwin suggested that a man helps other fellow men hoping to be helped back by others in the future [8]. One hundred years later in 1971, Trivers converted Darwin's idea to the *Theory of Altruism* [21]. Philosophers and biologists approach and describe benevolence as a pure concept of virtue, compassionate, and moral sentiments. They describe the benevolent action as the doing of a kind action to another, from mere good will and without any obligation; it is a moral duty only.

On the other hand, computational scientists analyzed and measured benevolence in terms of individual costs and benefits. Except for very few researchers, such as Bazzan, Bordini, and Campbell, they ignored the origin of benevolence, whose long history in philosophy and biology explores virtue and moral duty. They thought that benevolence should not be taken for granted, but should be considered as an important "phenomenon" that develops in societies of autonomous agents from exploration of agent emotions. Also, they think that in the present MAS theories, benevolence description is missing the emotional components [1].

So, what is the right approach to define and study benevolence? Should it be a pure moral or a pure individual benefits approach? A combination of both is what we are going to use. In other words, we will take the concept of benevolence from where it originated—philosophy and biology—and apply it to computational agents. We will study benevolence as a concept of goodness, social duty, and utility function.

An agent is benevolent if:

1. The agent *voluntarily* helps other agents without being commanded to do so.
2. The agent's benevolent actions are intended to benefit the society to which the agent belongs.
3. The agent should not expect an *immediate* reward or benefit for its benevolent actions. If it did, then the agent is instrumental, not benevolent [7].
4. The agent's benevolent action is taken while the agent is pursuing one of its own goals in such a way that it should neither prevent nor help the agent accomplish its goal.

Figure 2 shows a classic example of benevolence: *the mattress on the road*: a mattress in the middle of a road can cause a traffic jam, because all vehicles will have to slow down to maneuver around it. It results in a delay for everyone. A benevolent agent will stop and moves the mattress out of the way, so other agents can proceed on their way without any further delays. Such an action would cause the benevolent agent more delay than if it just tried to avoid the mattress like everyone else, and the agent receives no immediate reward or compensating "benefit" for its action.

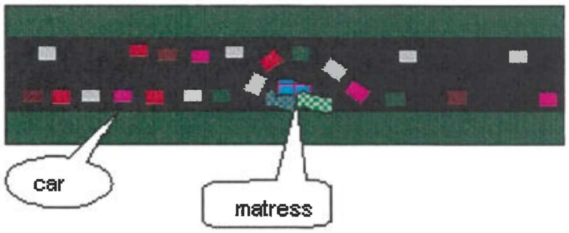


Figure 2: Mattress on the road

4. DISSCUSION

4.1 Benevolent Agents Are Autonomous Agents

We strongly believe that benevolent agents are autonomous agents. In our definition of benevolent agents, we emphasize the fact that benevolent action is *voluntary*, not performed upon request. The benevolent agents decide by themselves to take benevolent actions; their decisions are not controlled by other agents. The benevolent actions are taken out of the goodness, love, and friendship of the benevolent agents toward the other members of their society.

Castelfranchi, d'Inverno, Luck, and Wooldridge think that benevolence contradicts autonomy. Their definition of benevolent agents is based on the fact that benevolent actions are taken upon request from other agents. But our definition of benevolent agents states clearly that benevolent actions are taken *voluntary* without any requests or obligations. In the case of *the mattress on the road* example, an agent will stop to move the mattress off the road because it decides to do so, not because some other agent on the road (car) instructed or requested it to take such an action. And if it decides not to stop, it will not be punished by society. Thus, their criticism is not valid, and benevolent agents are indeed autonomous agents.

Castelfranchi and Conte made a very interesting point that proves benevolent agents are autonomous agents. They stated that benevolent actions should be taken without the other agents' expectations [7]. Thus, the benevolent actions are taken without any requests, so benevolent agents are autonomous agents. Moreover, unanticipated actions that benefit the society, more than the acting agents, will have more impact on the other individuals than anticipated actions. Such unanticipated actions are part of Bazzan, Bordini and Campbell's moral sentiments of agents [1]. We strongly believe that all moral actions are autonomous actions, since they are driven by the agents' goodness and loyalty to the group.

4.2 Benevolent Agents Are Rational Agents

We also believe that benevolent agents are rational agents. What is a rational agent? Simply, it is an agent that does the right thing. According to our definition of benevolent agents, benevolent actions should benefit the benevolent agents' society and will not stop them from reaching their goals. This will benefit the benevolent agent in the long run, i.e., it is an indirect benefit. In other words, if the society is doing well, then all its members, including the benevolent agent, must be doing well too. Another motivation is the belief that the agent's benevolent actions may encourage others to act benevolently in the future, thereby providing compensation in the longer term. This relates to Blackmore's work on memes where she states that altruism spreads altruism (meme-fountain) [2]. It is important to understand that a benevolent entity can exist only in an environment with other entities, never alone.

Benevolent agents do not take a benevolent action if they will be harmed. As we stated in our definition, benevolent actions are taken while the agents are pursuing their goals in such way that they should not prevent the agents from reaching their goals. In *the mattress on the road* example, an agent will not stop to move the mattress off the road if one of its passengers is having a heart attack and needs to be rushed to the hospital. An agent will stop and pickup the mattress if this action will not stop the agent from reaching its goal. For example, an agent whose purpose for being on the road is to get familiar with the town will stop and remove the mattress from the road. This action will not harm the agent, but will help other agents on the road.

Castelfranchi, Miceli and Conte think that benevolent agents are irrational agents, because they waste their resources helping others without any benefits for themselves. Based on our definition of benevolent agents and the above example, we see that benevolent actions benefit the society without stopping benevolent agents from reaching their goals. They also encourage others to behave benevolently. In other words, benevolent actions benefit the society immediately and the benevolent agent in the long run. Thus, we strongly believe that benevolence does not contradict rationality, as claimed by Castelfranchi, Miceli, and Conte [3].

5. FUTURE APPLICATIONS OF BENEVOLENT AGENTS

Software agents are unlikely to encounter mattresses, so where might a benevolent agent in an information system have an opportunity to behave benevolently? The agent could clean up stalled or failed transactions, close sockets that were left open by a process that terminated early, or remove locks set by failed or former processes. When it does not have either the authority or ability to take action, it can simply provide notifications to agents or systems that do.

One of the main areas where benevolence will play a major role is in Internet computing. Many agents will soon be populating the Web, and if they behave benevolently by helping each other search for information, for example, it will greatly reduce the traffic on the Internet, thereby benefiting everyone.

5.1 Collective Store Database

Parisi, Pedone, and Cecconi [6,18] discuss the ideas of individual survival strategies (ISS) and social survival strategies (SSS). Social survival strategy employs a collective store (CS) to which all individuals in a group contribute some of their resources. The collective store in turn redistributes the resources to group members by some allocation criteria or converts the resources into something new. Resources may include essential provisions, money, or knowledge—or CPU time and data storage space. Through simulations, the researchers concluded that a group using a collective store could survive severe environmental conditions, while individuals without a collective store would perish. In addition, the raw resources that individuals contributed could be transformed into new resources that no single individual could produce.

The concept of a collective store strategy implies benevolent behavior of the agents. By examining our definition of benevolence, we clearly can see that all benevolence criteria are met. Each agent contributes its resources (or its surplus) to the collective store willingly (autonomy) and without any guarantee that it will receive the appropriate amount, or even anything at all, back from the store. In addition, the store will decide who needs the resource the most and provide them with it, which in turn will benefit the group (rationality).

The collective store could be implemented as a large database of query results and information, see Figure 3. And benevolent query agents contribute their search result to this collective store database. When heavy Internet traffic degrades the search environment, the collective store database could help those agents seeking information on the Web. This is the basis for Internet search services such as Excite, Lycos, and AltaVista, except that users do not have to contribute anything in exchange for using these services. However, agents making greater contributions to a collective store might be given higher priorities in the subsequent use of the store. The collective store could refine the data submitted by different agents and derive new results through data mining techniques. Moreover, a collective store can gather data from agents that have better Web access capabilities and redistribute them to those with poorer capabilities, such as low-bandwidth PDAs.

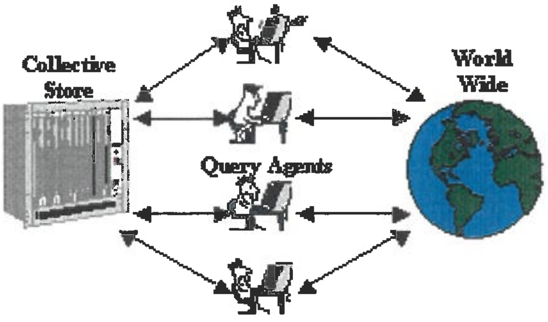


Figure 3: Collective Store Database

5.2 Benevolent Query Agents

As we begin spending more time on the Web, the demand will rise for agents that can perform our daily Web chores for us. Each agent will represent its owner, serving as the owner's surrogate for Web tasks and transactions. To be an effective surrogate, agents will have to be imbued with their owners' preferences and characteristics, such as cooperation, friendliness, sociability, and benevolence. Then the Web will be a friendlier and more productive environment for work, learning, and recreation. Research prototypes of such benevolent agents are operating now, and will soon be making their way onto the Web.

One of the most common Web agents is a query agent. A query agent searches the Web to find an answer to a user's request, and in so doing it may visit many sites and databases. When asked, a benevolent query agent would freely share its query results with other agents on the Web, even though it may have consumed substantial resources to get this knowledge and might have to consume more to share it. Through one agent's benevolence, other agents charged with similar queries would not have to explore all the sites or databases the first explored: they can simply use its results. Thus, benevolent agents can help reduce Internet traffic, leading to faster Web processing for all

6. CONCLUSION

There has been a lot of research on cooperation among agents, but benevolence has not been addressed comprehensively. There are many different and incomplete definitions of benevolent agents, which cause some researchers to criticize benevolence. They think that benevolent agents are neither autonomous nor rational, because they are asked to take actions that benefit the other agents, but not themselves. Based on our model of benevolent agents, we can clearly see that benevolent agents are both autonomous and rational agents. They decide on their own to take benevolent actions, not by force or obligation to other agents. In addition, the society of the agents will benefit from the benevolent actions, which in turn benefit the benevolent agents. Benevolent agents will be compensated for their benevolent actions in the longer term when other agents conduct similar behavior, which is why benevolence only makes sense within a society of agents that is driving toward the same global goal. On the other hand, benevolence might not be suitable for other multiagent systems, especially when there is competition for the same goal, such as money, power, etc. Internet computing is one of the fastest growing areas that utilize agents. Social behavior such as benevolence will find its way very soon into Internet applications as we move toward a more sociable web.

7. REFERENCES

[1] Ana L. Bazzan, Rafael H. Bordini and John A. Campbell, "Moral Sentiments in Multi-Agent Systems," *Pre-proceeding of the 5th International Workshop on Agents Theories, Architectures, and Languages (ATAL-98),* Research Note UCL-CS [RN/98/29], 1988.

[2] Susan Blackmore, "*The Meme Machine*", Oxford University Press Inc., New York, 1999.

[3] Cristiano Castelfranchi, Maria Miceli, and Rosaria Conte, "Limits and Levels of Cooperation: Disentangling Various Types of Prosocial Interaction," *Decentralized AI – 2,* Yves Demazeau and Jean-Pierre Muller (Eds.), Elsevier Science Publisher B.V., Holland, 1991.

[4] Cristiano Castelfranchi, "No More Cooperation, Please! In Search of the Social Structure of Verbal Interaction," *Communication from an Artificial Intelligence Perspective*, Ortony, A., Slack, J., and Stock, O. (Eds.) Springer-Verlag, Berlin, Germany, 1992.

[5] Cristiano Castelfranchi, "Guarantee for Autonomy in Cognitive Agent Architecture," *proceeding of ECAI-94 Workshop on Agent Theories, Architecture, and Languages*, Springer-Verlag, Berlin, Germany, 1995.

[6] Federico Cecconi and Domenico Parisi, "Individual versus social survival strategies," *Journal of Artificial Societies an Social Simulation*, vol. 1, no.2, The SimSoc Consortium, Department of Sociology, University of Surrey, Guildford, United Kingdom, 1998.

[7] Rosaria Conte, and Cristiano Castelfranchi, "*Cognitive and Social Action,*" UCL press, London, UK, 1995.

[8] C. Darwin, "*The Descent of Man and Selection in Relation to Sex*", John Murray, London, 1871.

[9] M. d'Inverno and M. Luck, "Understanding Autonomous Interaction", ECAI *'96: Proceedings of the 12th European Conference on Artificial Intelligence*, W. Wahlster (ed.), 529-533, John Wiley and Sons, 1996.

[10] M. Luck and M. d'Inverno, "Engagement and Cooperation in Motivated Agent Modelling",

Distributed Artificial Intelligence Architecture and Modelling: Proceedings of the First Australian Workshop on Distributed Artificial Intelligence, Zhang and Lukose (eds.), Lecture Notes in Artificial Intelligence, 1087, 70-84, Springer-Verlag, 1996.

[11] M. d'Inverno, M. Luck and M. Wooldridge, "Cooperation Structures", *Fifteenth International Joint Conference on Artificial Intelligence*, pages 600-605, Nagoya, Japan, 1997.

[12] Nicholas Jennings and Michael Wooldridge, "Agent Theories, Architectures, and Languages: A Survey", *Intelligent Agent: ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, Nicholas Jennings and Michael Wooldridge (Eds.), Springer-Verlag, Berlin, 1995.

[13] E. H. Durfee, V. R. Lesser, and D. D. Corkill, "Cooperation Through Communication in Distributed Problem Solving Network", *Distributed Artificial Intelligence,* M. N. Huhns (ed.), 29-58, Kaufmann, San Mateo, California, 1987.

[14] N. R. Jennings and J. R. Campos: "Towards a Social Level Characterization of Socially Responsible Agents", *IEEE Proceedings on Software Engineering*, pages 11-25, 1997.

[15] N. R. Jennings and L. M. Hogg "Social Rational Agents-Some Preliminary Thoughts", *Proceeding of AISB Workshop on Practical Reasoning and Rationality*, Manchester, UK, 1997.

[16] N. R. Jennings S. Kalenka, "Socially Responsible Decision Making by Autonomous Agents", *Proceeding of Fifth Int. Colloq. on Cognitive Science*, San Sebastian, Spain, 1998. (Invited paper)(To appear)

[17] H S Nwana and M Wooldridge, "Software Agent Technology", *Software Agents and Soft Computing*, H S Nwana and N Azarmi (Eds.), Springer-Verlag, Berlin, Germany, 1997.

[18] Roberto Pedone and Domenico Parisi, "In What Kinds of Social Groups can Altruistic Behavior Evolve?" *Simulating Social Phenomena*, Conte, R., Hegselmann, R., and Terno, P. (eds.), Springer-Verlag, Berlin, 1997.

[19] J. S. Posenschein and M. R. Genesereth, "Deals Among Rational Agents," *IJCAI-85*, pp. 91-95, Aravind Joshi (eds.), Morgan Kaufmann Publishers Inc., California, 1985.

[20] J. S. Rosenschein, "*Rational Interaction: Cooperation Among Intelligent Agents*", Ph.D. Dissertation, Computer Science Department, Stanford University, Stanford, CA 94305, 1985.

[21] R. L. Trivers, "The Evolution of Reciprocal Altruism", *Quarterly review of Biology*, Vol.46, pages 35-56, 1971.