Trust and Norms for Interaction

Michael Luck, Steve Munroe, Ronald Ashri
University of Southampton
United Kingdom
mml, sjm01r, ra@ecs.soton.ac.uk

Fabiola López y López
Benémerita Universidad Autónoma de Puebla
México
fabiola@cs.buap.mx

Abstract – *Cooperation is the fundamental underpinning of multi-agent systems, allowing agents to interact to achieve their goals. However, agents must manage the risk associated with interacting with others who have different objectives, or who may fail to fulfill their commitments. In this paper, we consider the role of trust and norms in a motivation-based view of agency. Motivations provide a means for representing and reasoning about overall objectives, trust offers a mechanism for modeling and reasoning about reliability, honesty, etc, and norms provide a framework within which to apply them.*[1]

Keywords: norms, trust, interaction, multi-agent systems

1 Introduction

After over a decade of research into agent-based computing, it is clear that the concepts underlying the paradigm are now firmly embedded in the general computational infrastructure. Regardless of whether the labels used correspond to those underpinning notions of agents, modern computing is concerned with distribution rather than centralisation, with flexibility rather than rigidity, and with independence rather than user-intervention. Agent-based computing seeks to address these concerns through the development of systems comprising multiple interacting entities that combine to offer better results than might be achieved if these entities were to act alone.

In support of such aims, much progress has been made. There are now a variety of agent toolkits, standards, methodologies, and development environments to facilitate practical application development (an extensive review of which is available here [13]), as well as coordination mechanisms (e.g. [12, 8]), negotiation frameworks (e.g. [15]), regulatory frameworks [6], and models for representing and reasoning about trust [17].

The introduction of large-scale open systems of this kind is likely to lead to a new set of problems, however, relating to the *effects* of interactions between agents. Indeed, what we are beginning to witness is the emergence of computational societies, of electronic organisations, and of all the variety of good and bad consequences that they bring with them. Just as in human societies, we need to consider the impact of regulations and their absence, of opportunistic and malicious behaviour, and we need to find ways to organise and manage systems in order to mitigate their potential deleterious effect on a system as a whole. While some work has been done on each of these concerns, their combination in large-scale open systems has not been addressed, yet they are fundamental requirements if the visions of Grid computing [9] and ambient intelligence [17], for example, are to be realised.

In this paper, we examine how the notions of motivation, norms and trust can facilitate reasoning about complex behaviour in large-scale open systems and inform the development of mechanisms to control behaviour, both when each notion is viewed in isolation and as part of a more unified approach. If our aim is to support the development of effective mechanisms to deal with such systems, then we must provide an understanding of how motivations, norms and trust *interact*, so that we can eventually provide guidance in the task of developing mechanisms to manage open systems.

We begin by examining motivations, and explain how they can act as *control mechanisms* for individual agent behaviour. Subsequently, we discuss how norms can be represented and their impact on agent behaviour and the society in general be understood. Next, we examine trust and discuss how it is related to both motivations and norms. Finally, we outline a conceptual framework that provides a view of the interplay between trust, motivations and norms.

2 Motivation

Much of computing, especially AI, is conceptualised as taking place at the *knowledge level*, with computational activity being defined in terms of *what* to do, or *goals*. Computation can then be undertaken to achieve those goals, as is typical in planning, for example. How-

[1] 0-7803-8566-7/04/$20.00 © 2004 IEEE

ever, the reasons for the goals arising are typically not considered, yet they may have important and substantial influence over their manner of achievement. If goals determine *what* to do, these reasons, or *motivations*, determine *why* and consequently how.

The best illustration of the role of motivation in computing is perhaps in relation to autonomous agents which, in essence, possess goals that are *generated* within, rather than *adopted* from, other agents [14]. These goals are generated from motivations, which are higher-level non-derivative components that characterise the nature of the agent. Motivations can be considered to be the desires or preferences that affect the outcome of a given reasoning or behavioural task. For example, *greed* is not a goal in the classical artificial intelligence sense since it does not specify a state of affairs to be achieved, nor is it describable in terms of the environment. However, it may give rise to the generation of a goal to rob a bank. The motivation of greed and the goal of robbing a bank are clearly distinct, with the former providing a reason to do the latter, and the latter specifying how to achieve the former.

Goals specify what must be achieved without specifying how, and in that sense, enable individual agents to choose the best means available to them in deciding how to achieve them. Although this gives a large degree of freedom in the dynamic construction of multi-agent systems, virtual organisations, etc., it provides little by way of direction, guidance, or meta-level control that may be valuable in determining how best to achieve overarching aims. Motivations address this both by providing the reasons for the goal, and by offering constraints on how the goal might best be achieved when faced with alternative courses of action.

Motivated agents are guided in their choice of activities by examining how different activities affect their motivations. Those activities that best serve the agent's motivational interests are those that are pursued. This approach is similar to the more traditional approach that uses the notion of *utility* to guide agent activity. Utility-based agents act under the principle of utility-maximisation, in which activities with higher utility are chosen over those with lower utility. In this way motivation and utility perform the same function for agents — they both guide the agent's choice of activities — but, whereas utility is an economic abstraction of value or benefit that is overlaid on an agent's choices by the agent designer, motivation is an internally derived measure of value determined both by a set of internal state variables (such as hunger or thirst, for example) and the external environment. For example, in the presence of food, an agent may or may not choose to eat depending on the state of its internal environment (specifically its hunger motivation). In this sense, motivation grounds the generation of measures of value such as utility in the agent's internal state, and thus is in a sense prior to and generative of such notions. This enables an agent to calculate the utility of a given course of action on-the-fly, based on its effects on the agent's current motivational state.

Motivation thus potentially offers a substantially higher level of control than is available at present, and which will become increasingly important for agents that need to function in an autonomous yet persistent manner while responding to changing circumstances. Agent technology is a powerful computational tool but, without the constraints that might be provided by motivations, agents may lack the required behavioural control. A motivational approach may also permit agents to better understand and reason about another agent's choices and behaviour, particularly in application to interactions in large-scale open systems where fine-grained control cannot be exerted, even if desired.

The scope here is great, with motivations providing a means by which the reasons underlying such interactions can contribute to, and constrain, the ongoing formation, operation and dissolution of multi-agent systems and virtual organisations, for example. Indeed, the interplay between an agent's motivations and the rules and regulations, or norms, of the group to which the agent belongs may have many implications on both the effectiveness of the individual agent and the group.

3 Norms

It has been argued by many [2, 3, 4] that agents interacting in a common society need to be constrained in order to avoid and solve conflicts, make agreements, reduce complexity, and in general to achieve a desirable *social order*. This is the role of norms, which represent what *ought* to be done by a set of agents, and whose fulfillment can be generally seen as a public good when their benefits can be enjoyed by the overall society, organisation or group [1]. Research on norms and agents has ranged from fundamental work on the importance of norms in agent behaviour [4, 21] to proposing internal representations of norms [3, 22], considering their emergence in groups of agents [23], and proposing logics for their formalisation [19, 24]. Despite such efforts to understand how and why norms can be incorporated into agents and multi-agent systems, there is still much work to do.

The easiest way to represent and reason about norms is by seeing them as built-in constraints where all the restrictions and obligations of agents are obeyed absolutely without deliberation. In this view, the effort is left to the system designer to ensure that all agents respond in the required way and, consequently, that the overall system behaves coherently. However, this may result in inflexible systems that must be changed off-line when either the agents or the environment change. By contrast, if a dynamic view of norms is taken, the flexibility of the overall system can be assured [26]. Towards this end,

agents must be endowed with abilities first, to adopt new norms and then to comply with them. By introducing agents able to *adopt norms*, we allow the representation of multi-agent systems composed of heterogeneous agents, independently designed, which can dynamically belong to different societies (or multiple societies) with the ability to adopt different roles [11]. This is a useful property for agents working virtual organizations, coalitions and human society simulations. Moreover, if this process is autonomous, agents may also have the possibility of selecting the society to which they want to belong, based on their own motivations and preferences.

3.1 Norm Compliance

Turning to the process of *norm compliance*, agents can be represented as either entities that always comply with norms, or entities that autonomously choose whether to do so. Both possibilities may cause conflicts between a society and the individuals within it. On the one hand, if norm compliance is assumed, social goals (achieved through norm obedience) are guaranteed. However, personal goals may be frustrated by obeying all the imposed norms because agents may lose opportunities that new situations offer to their individual interests. On the other hand, if the decision of whether to comply with a norm is left to the agent, although personal interest may be satisfied, the system becomes unpredictable when not all norms are obeyed, and consequently society performance may be degraded. In this situation, enforcement mechanisms can be introduced as a means of persuading agents to obey the norms. That is, agents might comply with norms in order to either avoid a punishment or obtain a reward. As a result, we need to consider agents able to deal with norm adoption and compliance, as well as with the sanctions and rewards associated with them. Moreover, both adoption and fulfillment of norms are important decision processes where agent autonomy, as defined by the motivations of the agent, plays a significant role.

Norms are thus the mechanisms of a society to influence the behaviour of the agents within it. They can be created from different sources, varying from built-in norms to simple agreements between agents, or more complex legal systems. They may persist over different periods of time, for example until an agent dies, as long as an agent remains in the society for which the norms were issued, or just for a short period of time until a normative goal is satisfied.

It is important to mention that *adoption of* and *compliance with* norms are two different, but related, processes. The first involves the agent's acknowledgment of three facts: it is part of the society, it is an addressee of the norms, and the issuer of the norm is entitled to do so. By contrast, compliance with norms involves an agent's commitment to obey the norm and therefore to achieve the associated normative goals. During the norm adoption process, norms are recognised as duties by the agent. It knows the norm and, most of the time, it is willing to obey it. However, at run-time the situation of an agent may change, making it difficult to maintain its compromise of obeying the norm, especially if that norm is causing conflict with its individual goals. Therefore, before complying with a norm, an agent must evaluate whether its fulfillment will satisfy its personal current motivations and preferences. In other words, an autonomous agent must not only decide which goals to pursue, how these goals can be achieved and which external goals can be adopted [14], it also must decide which norms to fulfill, based on its motivations. Sometimes norms are obeyed as an end just because agents have intrinsic motivations to be social. Other times, agents only obey norms if a punishment is applied for not doing so, if they are rewarded, and others still are guided by their internal motivation to be trusted. However, norms are sometimes violated, and to understand why, we must also analyse the motivations agents have to do so.

3.2 Enforcement Mechanisms

Some *enforcement mechanisms* are needed as a means of ensuring that personal interests do not overcome social rules. Usually enforcement mechanisms are associated with punishments and rewards so that agents are obliged to obey norms because of either the fear of being punished or the desire to gain something. However, as some sociologists point out [10], punishments and rewards will only affect an agent's decision to comply with norms if they either hinder or benefit one of the agent's goals. That is, punishments cannot be taken into account if none of the agent's interests (individual goals) is hindered. For example, the norm of wearing fashionable clothes may have an associated punishment of not being socially accepted. However, this applies just to a specific group of agents, and there may be others less interested in being accepted, who therefore consider the fulfillment of that norm as unworthy. Rewards are similarly a means to *motivate* agents only if one of the agent's goals receives benefits from such fulfillment. Thus, we can say that punishments and rewards do not have any effect on an agent's decision if they are not associated with some of the agent's individual goals.

Now, since punishments and rewards are defined as goals, in order to determine their effects on an agent's overarching goals, we need to understand when a goal can either hinder or benefit another goal. In general, a goal can hinder another when they are in conflict. Sometimes this is easy to observe because the state of one goal is simply the negation of the other, such as being outside a room and inside it at the same time. However, more generally conflicting situations are more difficult to observe. For example, cleaning a room and

watching a TV programme can be in conflict if they are intended at the same time and in different places. Goals receiving benefit are similar in that the easiest way to observe situations where a goal benefits from another is when both goals represent the same state but are achieved by different agents.

3.3 Motivations for Norm Compliance

In general, norms are broken when their fulfillment may hinder personal goals that agents consider as worthy for their personal interest, or when agents have internal motivations to reject external orders. Whatever the causes to violate a norm, both society and individuals may be affected. On the one hand, societies issue norms to be obeyed as mechanisms to achieve social goals, and it is expected that all society members comply with them. On the other hand, agents have individual goals that may be frustrated in order to comply with their duties. For example, in the case of the obligation to pay taxes, the society as a whole expects the norm to be fulfilled as means of achieving social welfare, but such an obligation may hinder the personal goals of taking holidays abroad or buying something. In this case, the decision concerns only the agent which, based on its motivations and current situation, must decide what is more important. Some careless agents may take this decision just by considering both the normative and their personal goals, but others may also take into consideration the consequences of being either punished or rewarded. For example, if an agent decides not to pay tax and continues with its goal towards holidays, it must accept the consequences of being punished. Conversely, if agents are cautious they must consider both the possibility of being punished and how much the punishment may affect their other personal goals.

4 Trust

The discussion so far indicates that autonomous agents decide whether to comply or not with norms issued by society following an analysis of the trade-offs between the consequences of complying or not with a norm. Furthermore, this decision-making process is influenced by how individual agents are motivated to act within a society. This raises two important issues with respect to how an agent society should be regulated and how individual agents should take decisions about whether there is a risk involved in attempting to cooperate with other agents. Firstly, an agent society as a whole should be concerned with the extent to which individual agents will be willing to comply with norms, and the effort that should be expended to ensure norm compliance through enforcement and the severity of sanctions. Secondly, individual agents should be concerned with the extent to which other agents that they interact with, either through commitment or contract, will be willing to perform any task resulting from the interaction (either because agents are more willing to defect in pursuit of more utility somewhere else or because there is uncertainty about whether agents can achieve the task).

In both cases, computational models of *trust* (defined as the positive expectation that an interaction partner will act benignly and cooperatively in situations in which defecting would prove more profitable to itself [5]) have an important role to play. They can provide a conceptual framework for determining the appropriate level of legislation that a society should impose through the definition and enforcement of norms, and enable individual agents to decide how to manage the risk of interacting with others. Current research in trust mechanisms for agents has focused on how individual agents may derive trust valuations for other agents, either through an analysis of the *history of interactions* with an agent [18], or by requesting *reputation information* from an existing social network [20], in which reputation is understood as a third party's estimate of trustworthiness. There has been little work that considers the *interplay* between norms, motivation and trust, even though a framework relating all three can provide a useful tool for aiding in the design of agent societies and individual agents.

5 Trust, motivations and norms

The types of norms and motivations prevalent within a society can significantly impact on the trust that an agent can place both on the society, as an issuer and enforcer of norms, and as an entity that provides redress in the face of malicious behaviour, and on individual agents to act benignly and to fulfill their contracts or commitments. Thus, it is important to provide a clear understanding of the interplay between norms, motivations and trust and in this section we provide an outline of these links.

In order to provide some means of reasoning about the interplay between these three aspects we use a three-dimensional space to identify different types of societies, based on the nature of norms, motivations and trust prevalent within the society, and use some examples to draw links between them. We illustrate this three-dimensional space in Figure 1. In Figure 1, the y-axis represents norms and their enforcement, with an increase in the value of y indicating the prevalence of more strict norms and enforcement. This can constrain the motivations of agents and prevent them from acting maliciously if they intend to do so. The x-axis represents motivations, with an increase in the value of x representing a prevalence of malicious motivations, indicating that agents are more likely to defect if they see more utility in alternative interactions. Finally, the z-axis represents trust, with an increase in the value of z indicating an increase in the trust that agents place in other agents and, therefore, an increase in willingness to cooperate with others. Finally, the squares labeled

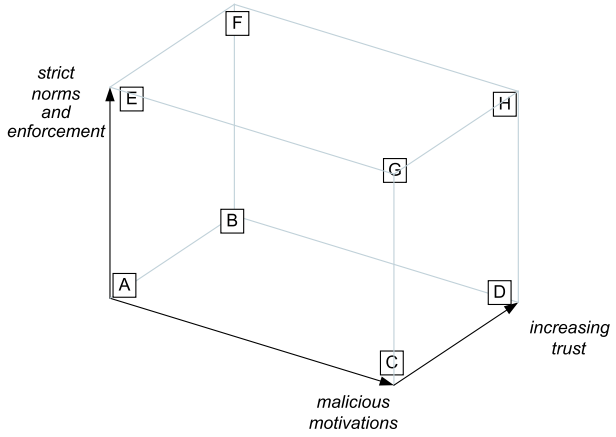

Figure 1: The interplay between norms, motivations and trust

by a letter indicate different types of societies within this space. There is, of course, another dimension (not shown) which concerns the trust agents place in the society, and we will discuss this throughout.

The most extreme case is that of society D, in which we have lax norms and enforcement, a prevalence of malicious agents, but agents still placing trust in each other. In such a situation, agents should carefully consider which agents they choose as interaction partners since there is a high possibility that those partners will defect and there is little protection through societal regulations. In fact, agents should display as little trust as possible in such situations and move towards the C type of society, where in the face of malicious behaviour and no norms, trust in other agents is very low.

In the cases of societies G and H the level of trust that agents show in each other is less important since the society itself is heavily regulated through strict norms and enforcement. Such a situation is similar to electronic institutions, in which each action is prescribed at design-time and compliance is ensured [6, 7].

Now, in the cases where agents are more inclined to behave benignly (A, B, E, and F), there is less risk in placing trust in individual agents and the amount of control imposed through norms is less important. In fact, agents that are not willing to trust others, as in the case of society A and E, may miss out in opportunities for cooperation that could lead to positive results because of their lack of trust in agents that are not intending to act maliciously. In addition, for societies with a prevalence of benign agents excessive control, such as in the cases of societies E and F, will hinder interactions and individual agents without providing real benefits for the society, since agents are, in any case, not likely to act maliciously. The most efficient situation is that of society B, where agents are allowed freedom in the types of interactions they can have, are not likely to act maliciously, and no effort is expended to control them.

This approach space provides an effective means of characterising and contrasting different types of societies so as to understand the interplay between norms, motivation and trust. A designer can then make use of such knowledge to decide the level of control that should be enforced, and individual agents can use it to decide the level of trust they can have in other agents.

6 Conclusions

Research in agent-based computing is now moving to tackle the problems posed by open agent societies where agent behaviour cannot be predicted and regulation and control mechanisms are necessary to mitigate potentially deleterious effects. In this respect, motivations, norms and trust all have an important role to play as part of the conceptual models that designers can access to develop solutions. Furthermore, these three issues are interrelated, and understanding how they are related plays an important part in devising mechanisms for different situations.

In this paper we discuss the notions of motivations and norms, drawing from our existing work on these issues [25, 16], and relate them to trust through a consideration of the different types of societies that result from the prevalence of stricter norms and enforcement, and agents inclined to behave maliciously. Through a characterisation of the different types of societies we provide the foundation required to then support the choice of appropriate control mechanisms for the society and trust evaluation mechanisms for individual agents. In the future, we aim to identify through experimentation whether clear guidelines about the best course of action can be defined for a wider range of society types.

References

[1] C. Castelfranchi, R. Conte, and M. Paolucci. Normative reputation and the cost of compliance. *Journal of Artificial Societies and Social Simulation*, 1(3), 1998.

[2] R. Conte. Emergent (info)institutions. *Journal of Cognitive Systems Research*, 2:97–110, 2001.

[3] R. Conte and C. Castelfranchi. Norms as mental objects. From normative beliefs to normative goals. In C. Castelfranchi and J. P. Muller, editors, *From Reaction To Cognition*, volume 957 of *LNCS*, pages 186–196. Springer, 1995.

[4] R. Conte, R. Falcone, and G. Sartor. Agents and norms: How to fill the gap? *Artificial Intelligence and Law*, 7(1):1–15, 1999.

[5] P. Dasgupta. Trust as a commodity. In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 49–72. Blackwell, 1998.

[6] M. Esteva, D. de la Cruz, and C. Sierra. ISLANDER: an electronic institutions editor. In C. Castelfranchi and W. Johnson, editors, *Proceedings of The First International Joint Conference on Autonomous Agents and Multi Agent Systems AAMAS'02*, pages 1045–1052. ACM Press, 2002.

[7] M. Esteva, J. Rodriguez-Aguilar, J. Arcos, C. Sierra, and P. Garcia. Formalising agent mediated electronic institutions. In F. Dignum and C. Sierra, editors, *Agent Mediated Electronic Commerce*, LNAI 1991, pages 126–147. Springer-Verlag, 2001.

[8] C. Excelente-Toledo and N. Jennings. The dynamic selection of coordination mechanisms. *Journal of Autonomous Agents and Multi-Agent Systems*, 9(1–2):55–85, 2004.

[9] I. Foster and C. Kesselman, editors. *The Grid: A Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, 1998.

[10] S. Kerr. On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, 18(4):769–782, 1975.

[11] S. Kirn and L. Gasser. Organizational approaches to coordination in multi-agent systems. Technical report, IImenau Technical University, Germany, 1998.

[12] V. Lesser, K. Decker, T. Wagner, N. Carver, A. Garvey, B. Horling, D. Neiman, R. Podorozhny, M. NagendraPrasad, A. Raja, R. Vincent, P. Xuan, and X. Zhang. Evolution of the GPGP/TAEMS Domain-Independent Coordination Framework. *Autonomous Agents and Multi-Agent Systems*, 9(1):87–143, 2004.

[13] M. Luck, R. Ashri, and M. d'Inverno. *Agent-Based Software Development*. Artech House, 2004.

[14] M. Luck and M. d'Inverno. Motivated Behaviour for Goal Adoption. In C. Zhang and D. Lukose, editors, *Multi-Agent Systems: Theories, Languages, and Applications, 4th Australian Workshop on Distributed Artificial Intelligence*, volume 1544 of *LNCS*, pages 58–73. Springer, 1998.

[15] P. McBurney and S. Parsons. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language and Information*, 11(3):315–334, 2002.

[16] S. Munroe, M. Luck, and M. d'Inverno. Towards a motivation-based approach for evaluating goals. In *The Second International Joint Conference on Autonomous Agents & Multiagent Systems*, pages 1074–1075. ACM Press, 2003.

[17] S. D. Ramchurn, D. Huynh, and N. R. Jennings. Trust in multiagent systems. *The Knowledge Engineering Review*, 2004 (to appear).

[18] S. D. Ramchurn, C. Sierra, L. Godo, and N. R. Jennings. A Computational Trust model for Multi-Agent Interactions based on Confidence and Reputation. In R. Falcone, S. Barber, L. Korba, and M. Singh, editors, *Workshop on Deception, Trust, and Fraud in the Second International Joing Conference in Autonomous Agents and Multi-Agent Systems*, pages 69–75, 2003.

[19] A. Ross. *Directives and Norms*. Routledge and Kegan Paul Ltd., 1968.

[20] J. Sabater and C. Sierra. REGRET: a reputation model for gregarious societies. In C. Castelfranchi and L. Johnson, editors, *Proceedings of the 1st International Joint Conference onAutonomous Agents and Multi-Agent Systems*, pages 475–482, 2002.

[21] R. Tuomela and M. Bonnevier-Toumela. Social norms, task, and roles. Technical report HL-97948, University of Helsinki, Helsinki, 1992.

[22] R. Tuomela and M. Bonnevier-Toumela. Norms and agreements. *European Journal of Law, Philosophy and Computer Sience*, 5:41–46, 1995.

[23] A. Walker and M. Wooldridge. Understanding the emergence of conventions in multi-agent systems. In V. Lesser and L. Gasser, editors, *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95)*, pages 384–389. AAAI Press/MIT Press, 1995.

[24] R. Wieringa, F. Dignum, J. Meyer, and R. Kuiper. A modal approach to intentions, commitments and obligations: Intention plus commitment yields obligation. In M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems*, pages 80–97. Springer-Verlag, 1996.

[25] F. L. y Lopez, M. Luck, and M. d'Inverno. Constraining autonomy through norms. In *The First International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 674–681. ACM Press, 2002.

[26] F. Zambonelli, N. Jennings, and M. Wooldridge. Organisational abstractions for the analysis and design of multi-agent systems. In *Proceedings of the First International Workshop on Agent-Oriented Software Engineering*, 2000.