Chapter 2

Autonomy: Variable and Generative

MICHAEL LUCK*, MARK D'INVERNO** and STEVE MUNROE*
* *Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, United Kingdom*

*** Cavendish School of Computer Science, University of Westminster, 115 New Cavendish Street, London W1W 6UW, United Kingdom*

Key words: motivation, adjustable autonomy, generative autonomy

Abstract: *In the paper we discuss variable and generative forms of autonomy. Variable autonomy is discussed in terms of the practicalities in designing autonomous agents, dealing as it does with the notion of degrees of autonomy and hence issues of agent control. The major part of the paper discusses an absolute, theoretically grounded notion of autonomy: the ability to generate one's own goals. This theoretical account of autonomy is embedded in the larger SMART framework and is intimately linked with the issue of motivation. Autonomous agents are motivated agents in that for the generation of goals an agent needs a set of higher order, non-derivative sources of action, or in our terminology, motivations. Autonomous agents in the SMART framework form the basis and source of action in multi-agent systems, which can thus propagate through the other entities in the system, such as non-autonomous agents and objects. We conclude with a discussion regarding the situations an autonomous agent would be willing to relinquish its autonomy thus linking the generative and variable notions of autonomy.*

1. INTRODUCTION

Autonomy is one of the most used but least *operationalised* words in the fields of intelligent agents and multi-agent systems. This is strange in computer science where typically concepts must be clearly and precisely defined in order for them to be incorporated into theories, models and implementations. In reality, this is seldom a problem as it is often assumed that the autonomy of an agent is something that arises from the overall flexibility of the agent's behaviour as it goes about its business in its environment away from human direction. Thus an agent engaging with others in its environment in a seemingly intelligent and flexible way, making decisions that reflect its goals, overcoming obstacles etc., would most probably be stamped with the descriptive label of being autonomous. However, this simply indicates that autonomy is merely a description of a variety of flexible and perhaps adaptive behaviours. Nothing in the agent itself, no process or architecture, can be identified as the controller or source of the agent’s *autonomy* and, as a consequence, autonomy begins to acquire the nature of an *emergent* property.

Autonomy is undeniably a critical issue in the field (Castelfranchi, 1995; Barber and Martin, 1999), yet many, it seems, are content to assume it as an emergent property without giving it any real concrete definition. Agents are often taken to be autonomous by default, and as such there seems little need to add the tautologous prefix in explicitly considering *autonomous* agents. Some in the field, however (Balkenius, 1993; Castelfranchi, 1995), view autonomy in agents as an important yet problematic issue that demands attention.

Seeing this problem with the lack of focus in the use of the term autonomy, some researchers are beginning to think more carefully about what autonomy really means. And, in the literature, two very different conceptions of what autonomy *should* mean are beginning to emerge. On the one hand, some researchers operationalize autonomy as the level or degree to which an agent can achieve its goals without intervention and thus strongly relate the notion of autonomy to an agent’s *dependence* upon others. High dependence under this definition equates to low autonomy. However, there is another emphasis that considers autonomy as an absolute enabler for *generating* an agent’s own goals in response to different situations. While the achievement of some of these goals may depend on other agents, this dependence does not affect the autonomy of the goal-generating agent.

In this paper, we consider these two views of autonomy in more detail, focusing in particular on the latter view, which we have adopted as the basis for an extensive theory of agenthood over recent years. We begin with a short consideration of autonomy as independence between agents before moving on to consider autonomy as goal generation to greater depth. We introduce the concept of motivation, review relevant work in this area, and then describe how motivation in our model underlies generative autonomy in our Structured and Modular Agents and Relationship Types (SMART) framework. We end our description of SMART by describing the implications of this model for interacting agents requiring assistance from others. Finally, we review the two views and try to draw some conclusions.

2. TWO VIEWS ON AUTONOMY

2.1 Adjustable Autonomy

One view of agents holds that once we've solved all the technical problems, it will be possible to have agents that are able to explore the (virtual) world and perform all sorts of tasks for their users, all with complete autonomy and integrity with regards to user likes and dislikes. These agents can be likened to surrogate selves that embody the user’s desires and aspirations, traveling about the cyber-sphere exploiting opportunities in the user's stead, making money, closing deals, securing contracts, and so on. While this picture is perhaps a little rosy, it is certainly a desirable one. However, the likelihood of any of the above coming true depends (amongst other things) on the key issue of *trust*. If organisations and individuals are to use software agents to look out for their interests in electronic worlds as suggested above, then the agents carrying out such interactions must be trustworthy; users will want guarantees that the software will not incur losses through faulty or inept operation. This is perfectly natural and proper, and occurs in the real world when human agents are engaged to act on behalf of others. In such situations, the way to avoid problems with new tasks is simply to employ *training*; usually involving the supervision of some or all of the task at hand with information flowing backwards and forwards between the two parties. In this way, an employer can be sure that an employee understands the nature of the task and is aware of the potential problems, and can require the employee to request assistance if problems arise. Now, this relates to autonomy in that the more an employer trusts an employee to achieve a task correctly, the more autonomy in relation to the task will be bestowed upon the employee. If this is the case with human agents, then perhaps we can use the same system with electronic agents.

Increasingly, some researchers are aiming to design agents that have an *adjustable autonomy* such that it can be reeled in or out depending on the circumstances. Barber and Martin (1999), for example, link an agent's autonomy to its ability to influence the decision-making process for a given problem. In their view, an agent acting alone has complete autonomy in that it holds all the decision-making power. Similarly, an agent making all decisions for other agents, as well as itself, has complete autonomy and power over itself and its subjects. An agent that shares decision-making with others is in a consensus relation with them, and thus its autonomy is limited in proportion to the number of agents involved in making those decisions. Finally, an agent that has no involvement in the decision-making process consequently has no autonomy and is command driven.

What these models offer is a way of representing autonomy in an explicit way that then enables it to be measured and manipulated. Different dimensions of autonomy are suggested by Brainov and Hexmoor (2001), such as simple autonomy from the user; autonomy from the environment (which changes as a response to the predictability of that environment) and group autonomy (how free the agent is from interference by others). Once identified, these aspects of autonomy can be measured and adjusted according to the experimenter's whim. Perhaps the major immediate difficulty with adjustable autonomy is the problem inherent in recognizing when an adjustment in autonomy is required. Should there be some way to oversee and measure the performance of an agent in order to make necessary changes in its autonomy, such as increasing its reliability on another's (perhaps the user's) judgment? If so, then how often should the agent's performance be checked? Or should the agent itself decide when it should give up its independence and seek the aid of others? Some solutions have been offered (for example, see (Barber and Martin, 1999)) but there remain many open issues.

2.2 Autonomy and Goal Generation

Our own notion of autonomy focuses on an agent's ability to generate its own goals. A dictionary definition will tell us, among other things, that autonomy amounts to freedom of will (and we will add that it includes the ability to exercise that will). In short, this means that it provides the ability to exercise choice, which is particularly relevant in the context of goals and goal-directed behaviour, as in Castelfranchi's notions of goal (or motivational) autonomy (Castelfranchi, 1995). Delving further, we can see that the literal translation of autonomy from the Greek *auto-nomy* is *self law* or *self government*, and presupposes the ability to generate one's own rules (or in our terms, goals) for living. The self-generation of goals therefore becomes the defining characteristic of autonomy. In this view, autonomous agents are able to generate their own goals, to select between multiple alternative goals to pursue, and to decide to adopt goals from others (to further their own ends).

Thus from a purely *conceptual* or theoretical point of view removed from practical considerations, autonomy can naturally be regarded as absolute, without dimension or measure of degree – one can either generate one's own goals or one cannot (ignoring possibilities of degrees of goal generation, of course). Yet, this *strong view* of autonomy contrasts with much of the practical work with agents described earlier in which autonomy is taken to be the same as *independence*, a very distinctly *relative* notion. In what might be called this *weak view*, a non-autonomous agent either depends on others or is an automaton, while an autonomous agent can either be independent or depend on others. It is this last point that seems to suggest that autonomy is not the same as independence – an agent does not simply lose its autonomy by virtue of depending on another for a particular goal; situations of dependence occur also for autonomous agents.

Practically then, the notion of independence can be used as an approximation for autonomy with the added benefit that it admits the dimensions and measures of degree that are missing from the strong view. In this sense it might be considered as a valuable practical operationalisation of autonomy, and provides a way to characterise different dependence situations.

3. AUTONOMY THROUGH MOTIVATION

For all the difficulty in pinning down autonomy, it is key in our view to understanding the nature and behaviour both of individual agents, and of interactions between them. In a series of papers over a number of years, we have described and formally specified an extended theory of agent interaction, based on *goals* and *motivations*, which takes exactly this standpoint. The theory describes the SMART framework for categorizing different agents (Luck and d'Inverno, 2001), and has been used as a basis for investigating aspects of the relationships between agents (d'Inverno and Luck 2000), providing an operational account of their invocation and destruction (d'Inverno and Luck 1997), as well as for reformulating existing systems and theories, including those relating to dependence situations (d'Inverno and Luck 1996).

In essence, autonomous agents possess goals that are *generated* within rather than *adopted* from other agents. These goals are generated from *motivations*, which are higher-level non-derivative components that characterise the nature of the agent. As we will discuss in more detail shortly, they can be considered to be the desires or preferences that affect the outcome of a given reasoning or behavioural task. For example, as we have pointed out elsewhere, *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.

This view of autonomous agents is based on the generation and transfer of goals between agents. More specifically, something is an agent if it can be viewed as satisfying a goal that is first created and then, if necessary and appropriate, transferred to another. It is the adoption of goals that gives rise to agenthood, and it is the *self-generation* of goals that is responsible for autonomy. Thus an *agent* is just something that is either useful to another agent in terms of satisfying that agent's goals, or independently purposeful. Importantly, agents rely on the existence of others to provide the goals that they adopt for instantiation as agents. In order to escape an infinite regress of goal adoption, however, we define *autonomous agents* to be just agents that generate their own goals from motivations.

Social behaviour arises as a result of individual agents interacting with each other (through cooperation, competition, and other such forms of interaction) so as to exploit the resources available in a rich and dynamic multi-agent domain. If agents are to make use of others to help them in their tasks, such social behaviour is critical. Underlying this cooperation is the transfer or adoption of goals from one agent to another, a subtle and complex process that depends on the nature of the agents involved.

3.1 What is Motivation?

According to Halliday, the word *motivation* does not refer to a specific set of readily identified processes (Halliday, 1983). Though for practical purposes motivation can be discussed in terms of *drives* and *incentives*, the push and pull of behaviour. Drives are internally generated signals that tell the organism that it has violated a homeostatic balance such as hunger, thirst etc. There are also the circadian drives such as sleep and wakefulness. Incentives originate outside of the organism and can vary in their attractiveness to the organism arousing more or less motivation. Incentives can be both positive and negative for example a positive incentive usually causes approach behaviours such as a person deciding to buy a car due to the attractiveness of its specifications. A negative incentive causes avoidance behaviours such as a shy person avoiding social interaction. Motivation has long been seen as a key concept in the organisation of behaviour within the psychological and ethological sciences. Our focus, however, is on providing an effective control mechanism for governing the behaviour and reasoning of autonomous agents through the use of motivations. Though we focus on a computational approach, in this section we will discuss related work.

In Cognitive psychology researchers come close to the intended meaning of motivation that we propose here. Kunda (1990) informally defines motivation to be, “any wish, desire, or preference that concerns the outcome of a given reasoning task” and suggests that motivation affects reasoning in a variety of ways including the accessing, constructing and evaluating of beliefs and evidence, and decision making. Much work has been done experimentally to explicate these thoughts but work is just beginning to put them in a computational context.

One early example is Simon (1979), who takes motivation to be “that which controls attention at any given time,” and explores the relation of motivation to information-processing behaviour, but from a cognitive perspective. Sloman and Croucher (1981), and Sloman (1987) alone have elaborated on Simon's work, showing how motivations are relevant to emotions and the development of a computational theory of mind.

Problem solving can be considered to be the task of finding actions that achieve the current goals. In this way goals provide the reason and context for behaviour. But how are the goals to be chosen? Typically the agent chooses a goal if the environmental conditions support the pre-conditions necessary for the goal; that is the *external context* determines goal selection. However, in real biological agents often the same environmental cues elicit different behaviour. This can be attributed to the current motivations of the agent. This *internal context* is often missing in computational agent based systems. Clearly, this is inadequate for research concentrating on modeling autonomous agents and creatures, which requires an understanding of how such goals are generated and selected. Additionally, it is inadequate for research that aims to provide flexibility of reasoning in a variety of contexts, regardless of concerns with modeling artificial agents. Such flexibility can be achieved through the use of motivations which can lead to different results even when goals remain the same (Luck, 1993).

In Sloman’s development of Simon’s Ideas (Simon, 1979), Sloman argues explicitly for the need for a “store of ‘springs of action’ (motives)” (Sloman and Croucher, 1981). For Sloman, motives represent to the agent what to do in a given situation and include desires, wishes, tastes, preferences and ideals. Key to Sloman’s conception of motives is their role in processing. Importantly, Sloman distinguishes between two types of motives. First-order motives directly specify goals, whereas second order motives generate new motives or resolve conflicts between competing motives – these are termed *motive generators* and *motive comparators*. According to Sloman, a motive produced by a motive generator may have the status of a desire. This relatively early work presents a broad picture of a two-tiered control of behaviour: motives occupy the top level, providing the *drive* or *urge* to produce the lower level goals that specify the behaviour itself. In subsequent work, the terminology changes to distinguish between *nonderivative motivators* or goals and *derivative motivators* or goals, rather than between motivators and goals themselves. Nevertheless, the notion of derivative and nonderivative mental attitudes makes one point clear: that there are two levels of attitude, one which is in some sense innate, and which gives rise to the other which is produced as a result of the first.

In a different context, the second of Waltz's `Eight Principles for Building an Intelligent Robot' requires the inclusion of “innate *drive* and evaluation systems to provide the robot with moment-to-moment guidance for its actions” (Waltz, 1991). In elaborating this principle, Waltz explains that the action of a robot at a particular time should not just be determined by the current sensory inputs, but also the “desires” of the robot, such as minimizing energy expenditure (laziness), and maintaining battery power levels (hunger).

Moffat & Frijda (1995) use a similar concept which they term ‘*concerns*’ which are “dispositions to prefer certain states and/or dislike others”. In their model the agent selects the most relevant information coming in through its sensors. The relevance of an event comes from the agents concerns. Thus for example if the agent detects food in its environment and this event is relevant to its hunger concern a goal may be generated to move towards the food and eat it. The most relevant event causes a signal to be emitted which in turn causes the relevant goal to be instantiated.

All this varied research into robotics, artificial life, and autonomous agents and creatures has provided the impetus for a growth of interest in modeling motivations computationally, and a number of different representations for motivations and mechanisms for manipulating them have been developed at both subsymbolic and symbolic levels (e.g. (Balkenius, 1993; Halperin, 1991)).

3.2 Motivated Behaviour in Autonomous Agents

Responses made to a given stimulus can vary depending both on the internal state of the agent and/or the external situation (i.e. the environment). If the external situation remains constant, differences in response must be ascribed to changes in the internal state of the responding agent. These differences are due to the motivations of the agent.

An agent can be thought of as having a fixed range of identifiable motivations of varying strength. These motivations can be regarded as being innate, and certain behaviours may be associated with one or more motivations. For example, sexual courtship behaviour might be associated with the motivation for reproduction. Executing the courtship behaviour may enable the agent to procreate with the partner, which typically will mitigate the motive to reproduce. These behaviours are known as *consumatory behaviours*; other behaviours such as courtship displays make the conditions of the consumatory behaviour come true, and are known as *appetitive behaviours*.

This view of motivation is somewhat simplified, and although much behaviour occurs in functional sequences with appetitive behaviours leading to consumatory ones, complex interactions between motivations and behaviours are possible (Hinde, 1982). For example, a single situational cue could relate to many motivations which in turn could release many activities, or cause an action which in turn leads to other behaviours, or even cause some motivations to decrease so that others would increase in turn. In addition there are inhibitory relationships between behaviours in animals and also relationships that increase the strength of other behaviours. Moreover, the combination of motivations may lead to different or variable behaviours.

These are all difficult issues which must be addressed in attempting to construct accurate behavioural models of real and artificial agents. Our concern, however, is not with providing such accuracy, but in constructing *simple* yet *adequate* models which will allow effective control of behaviour.

3.3 A Simple Example of Motivation In Autonomous Agents

We can define autonomous agents to be agents with a higher-level control provided internally by motivations. Thus we can specify motivations of *curiosity*, *safety*, *fear*, *hunger*, and so on. In a simple agent design, we might then associate the motivation of *safety* with the goal of *avoiding obstacles* which, in turn, is associated with the actions required to achieve such results. Motivations will also vary over time according to the internal state of the agent. For example, if the agent spends a long time without food, then the hunger motivation will increase. When the agent feeds, the hunger motivation will decrease.

Each motivation thus has a strength associated with it, either variable depending on external and internal factors, or fixed at some constant value. A motivation can thus be represented by a triple, <*m*, *v*, *b*> known as an *m-trip*le where *m* is the kind of motivation, *v* is a real number, the strength (or intensity, (Sloman, 1987)) value associated with that motivation, and b is a boolean variable taking the value *True* when the strength value, *v*, is fixed, and *False* when it is variable. An autonomous agent can be regarded as embodying a set of *n* motivations, M, which comprises the *m-triples*, <m_1, *v*, *b*> …<m_n, *v, b*>. Thus the set of motivations, M, is a function of the kind of agent being considered, while each motivation in this set at a particular point in time is a function of an instance of a particular kind of agent and its environment together. In order to act on motivations, a threshold value for strength may be necessary, which must be exceeded to force action. Alternatively, the highest strength value may be used to determine the motivation currently in control.

More sophisticated mechanisms are possible such as those described by Norman and Long (1995; 1996), Sloman (Sloman 1987; Beaudoin and Sloman, 1993), and Moffat and Frijda (Moffat and Frijda, 1995; Moffat et al., 1993). In addition, other representations for motivations and mechanisms for manipulating them have been developed at both subsymbolic and symbolic levels (e.g. by Schnepf (1991), Maes (1989a, 1989b, 1991) and Halperin (1991)). All are possible instantiations of the model described in the remainder of this paper, but the details are unimportant at present. It is enough to note that the abstract model provides the framework within which such mechanisms can be incorporated according to the particular need.

4. THE SMART FRAMEWORK

As has been described elsewhere in more detail (Luck and d'Inverno 1995), we propose a four-tiered hierarchy comprising *entities*, *objects*, *agents* and *autonomous agents*. Underlying the SMART view of the world is the fundamental assumption that all components are entities. Some of these entities are objects, and some the objects are agents. In turn, some of the agents can be further specialised to autonomous agents. In this section, we briefly outline the agent hierarchy (Shown as a Venn diagram in Figure 1.). Many details are omitted – a more complete treatment can be found in (Luck and d'Inverno 1995).

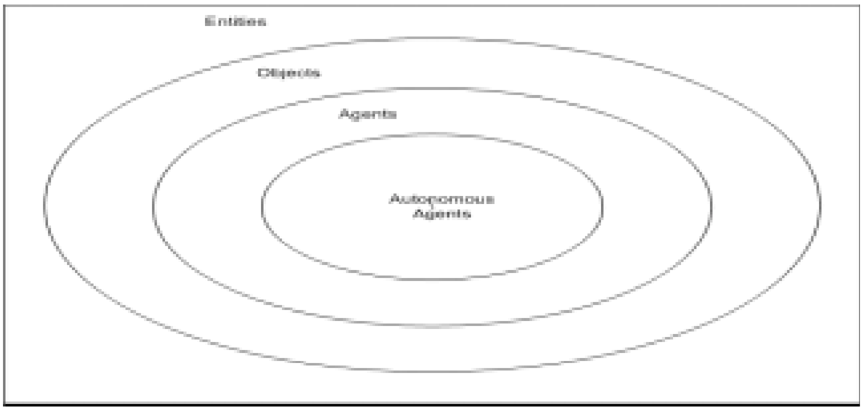

Figure 1. Entity Hierarchy overview

Entities simply provide a way to denote components in the world before we have any recognisable structure for them, or before we can classify them as objects, agents or autonomous agents. Although we will not provide a mathematical treatment in this paper, the use of entities also enables a simple and elegant formal description to be provided. (Elsewhere, we provide extensive mathematical descriptions of the SMART framework in the Z specification language, based on the notion of entities.) Objects can then be defined to be just things that have abilities and attributes and with no further defining characteristics. Similarly, agents are just objects that are useful, typically to other agents, where this usefulness is defined in terms of satisfying some goal of these other agents. In other words, an agent is an object with an associated set of goals. Now, a particular object may give rise to different instantiations of agents that are created to satisfy some need of another agent. If we define agenthood in this way, then we also rely on the existence of these *other* agents to provide goals that are adopted in order to give some initial reason for creating or instantiating an agent in the first place. Carried to its logical end, however, we arrive at a situation where agents are only defined in relation to already existing agents, and a continuing chain of agent instantiation results. In order to escape an infinite regress of goal adoption, therefore, we can define autonomous agents, which are just agents that can generate their own goals *from motivations*. Thus autonomous agents, which set their own agendas, are distinguished from *server* agents by virtue of their ability not simply to satisfy or achieve goals, but to create them.

For example, a knife can be an object. It has attributes specifying that it is solid, made of steel, is silver in colour and has a sharp edge. Its capabilities specify that it can cut things. If I cut a steak with a knife, then the knife is my agent for cutting the steak. The knife may not actually possess the goal, but it is certainly satisfying, or can be *ascribed*, my goal of cutting the steak. A robot that rivets a panel onto an aeroplane fuselage is also an agent, and if it has motivations such as hunger and achievement, then it is an *autonomous* agent.

As explained above, this paper will not offer a mathematical treatment, but to illustrate the simplicity and elegance of the key notions underlying the SMART framework, we provide some very simple formal definitions.

$$\begin{aligned} \mathit{Entity} = = [&\mathit{attributes}\colon \mathrm{P}\ \mathit{Attribute};\ \mathit{capableof}\colon \mathrm{P}\ \mathit{Action};\\ &\mathit{goals}\colon \mathrm{P}\ \mathit{Goal};\ \mathit{motivations}\colon \mathrm{P}\ \mathit{Motivations}] \end{aligned}$$

$$\mathit{Object} = = [\ \mathit{Entity} \mid \mathit{capableof} \neq \{\ \}\ ]$$

$$\mathit{Agent} = = [\ \mathit{Object} \mid \mathit{goals} \neq \{\ \}\ ]$$

$$\mathit{AutonomousAgent} = = [\ \mathit{Agent} \mid \mathit{motivations} \neq \{\ \}]$$

In summary, if there are attributes and capabilities, but no goals, then the entity is an *object*. If there are goals but no motivations, then the entity is an *agent*. Finally, if neither the motivation nor goal sets are empty, then the entity is an *autonomous agent*. Thus, we have a simple but precise framework that identifies and characterises agents and autonomous agents, and distinguishes them clearly.

4.1 Goal Generation

Now, given that the key to our notion of autonomy is the ability of an agent to generate its own goals and set its own agenda, we turn our attention to that particular aspect. As stated above, the SMART framework involves the generation of *goals* from *motivations* in an autonomous agent, and the adoption of goals by, and in order to create, other agents. In previous work we have given a complete formal description and specification of how autonomous agents, *defined* in terms of their high-level and somewhat abstract *motivations*, can construct goals.

Autonomous agents will always try to find ways to mitigate motivations, either by selecting an action to achieve an existing goal as above for simple agents, or by retrieving a goal from a repository of known goals. Thus, SMART requires a repository of known *goals* that capture knowledge of limited and well-defined aspects of the world. These goals describe particular *states* or *sub-states* of the world with each autonomous agent having its own such repository.

Now, in order to retrieve goals to *mitigate* motivations, autonomous agents must have some way of assessing the effects of competing or alternative goals. Clearly, the goals which make the greatest positive contribution to the motivations of the agent should be selected unless a greater motivational effect can be achieved by *destroying* some subset of its goals. The motivational effect of generating or destroying goals not only depends on the motivations, but also on the goals of the agent. For example, an autonomous agent should not generate a goal that it already possesses or that is incompatible with the achievement or satisfaction of its existing goals.

In general, agents may wish, or need, to use the capabilities of other entities. They can make use of the capabilities of these others by *adopting* their goals. For example, if Michael needs to write a paper that he cannot complete alone, he must get assistance from others. More specifically, he must persuade someone else to adopt his goal before the paper can be completed. Similarly, if he needs to work at home on the paper, he may need to use a floppy disk to take the current version of the paper home, and then use a laptop computer to continue writing and editing the paper. Indeed, in the course of writing the paper, both inanimate objects such as the computer and the disk may be needed, as well as other people to collaborate with. Each of these objects (including the people) can be ascribed, or viewed, as adopting Michael's goals in order that his desire for success can be satisfied. This notion of goal adoption underlies social behaviour, and an understanding of the ways in which it can be achieved is fundamental for effective modeling and simulation of agent systems.

Thus, entities may serve the purposes of others by adopting their goals, but the ways in which they adopt goals depends on the kind of entity they are. In the description given above, goals may be generated only by autonomous agents, but both non-autonomous (server) and autonomous agents can adopt goals. With autonomous agents, goal adoption amounts to a problem of *negotiation* or *persuasion*, requiring an analysis of the *target* autonomous agent. With non-autonomous agents, goal adoption requires an analysis of both the agent intended to adopt the goal, and any other agent *engaging* that agent. With objects, no analysis is required, since agents are *created* from objects with the relevant associated goals.

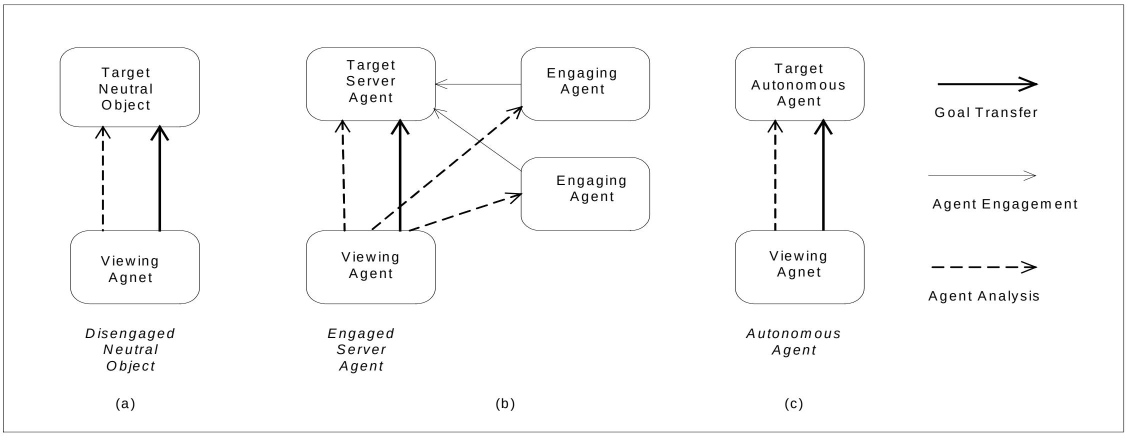

Figure 2. Goal Adoption in Neutral-Objects, Server Agents and Autonomous Agents.

Figure 2 shows three fundamental cases of goal adoption, which we consider in detail below. In the figure, there are three kinds of agent. A *target* agent or object is one that is intended to adopt goals. An *engaging* agent is one whose goals are currently (already) adopted by the target agent. Finally, a *viewing* agent is an agent that seeks to engage a target agent or object by having it adopt goals. It is a viewing agent because the way in which goal adoption is attempted is determined by its view of the situation. We consider the three cases of goal adoption below.

In the simplest case, goal adoption by non-autonomous agents occurs by instantiating an agent from a non-agent object or a *neutral-object* with the goals to be adopted. In this case, no *agent* exists before the goals are adopted, but the act of goal transfer causes an agent to be created from a neutral object using those particular goals. Thus, for example, a cup in Steve and Mark's office, which is just a neutral-object, becomes an agent when it is used for storing Steve's tea. In this case it *adopts* or is *ascribed* his goal of storing liquid. It is possible to create the agent from the object because the cup is not being used by anyone else; it is not *engaged* by another agent. An entity can only be a neutral object if it is not engaged.

If the target object is *engaged* by other agents then it is itself an agent, so the protocol for goal adoption changes. In this case, there are several ways to *engage* the target object. The first involves supplying the target object with more goals that do not affect the existing agency obligations. (Obligations here simply refer to the existing relationships between entities by which one is engaged by another.) In this case the agent is *shared* between the viewing agent and the existing engaging agents. The second involves trying to persuade any engaging agents to *release* the engaged object so that it becomes a *neutral-object* and can therefore subsequently be engaged by the viewing agent as required. The third possibility involves *displacing* the engaging agent so that the engaged object becomes a neutral-object and can then subsequently be ascribed other goals. This possibility is dangerous since it may cause conflict with the previous engaging agents.

As an example, suppose that a cup is currently in use as a paper-weight for Steve, so that the cup is *Steve's* agent with his goal of securing loose papers. Suppose also, that Mark wishes to use the cup to have some tea. The first way for Mark to engage the cup is for him to attempt to use the cup without destroying the existing agency relationship between Steve and the cup. Since this would involve an awkward attempt at making tea in, and subsequently drinking from, a stationary cup, he may decide instead to try other alternatives. The second alternative is to negotiate with Steve to release the cup so that it can be used for storing tea while the third alternative is for Mark to displace the goal ascribed to the cup by removing the cup from the desk and pouring tea into it. The cup is no longer an agent for Steve and is now ascribed the goal of storing tea for Mark. It has switched from being engaged by Steve to being engaged by Mark, and this is equivalent to the agent reverting to an object and then being re-instantiated as a new agent. This method may not be an appropriate strategy, however, because in destroying the agent obligation of the cup as a paper-weight, there is a risk of conflict between Steve and Mark.

In the example above, the second possibility for goal adoption by server-agents involves Mark persuading Steve to first release the cup from its existing agenthood. The cup would then become a neutral-object and could be instantiated as required by Mark. In general, such persuasion or negotiation may be more difficult than the direct physical action required for goal adoption in non-autonomous entities. Autonomous agents are motivated and as such, only participate in an activity and assist others if it is to their motivational advantage to do so (that is, if there is some motivational benefit). They create their own agendas and for them, goal adoption is a *voluntary* process as opposed to an *obligatory* one for non-autonomous agents. In a similar example, Michael might ask Steve to assist in writing a paper, but Steve may refuse. This notion of volunteering to do something refers to the *choice* that distinguishes an autonomous agent - autonomous agents have the ability to decide whether to cooperate or not, in line with their own agendas.

In general, goals must be adopted through explicit autonomous agent initiative, as opposed to an ascription of goals for non-autonomous agents. However, in some contexts the ascription of goals to autonomous agents may be meaningful. Suppose, as a dramatic yet unlikely example, that Steve incapacitates Mark in some way but manipulates the unconscious victim so that he functions as a hat stand. In this situation, the autonomous agent, Mark, could be *ascribed* the goal of holding hats even though he has not explicitly adopted this goal. Such cases can be described by considering the autonomous agent as an agent in an obligatory relationship.

5. DISCUSSION

As we have explained elsewhere (Luck and d'Inverno, 1995), Franklin and Graesser's definition of an *autonomous agent* as a system that pursues "its own agenda" (Franklin and Graesser, 1997) reinforces the perspective in this paper. However, it is important to recognise that an autonomous agent in this view is still free to generate the goal to relinquish its autonomy by entering into a supervisory or group relationship *if it is in its own interest to do so*. As humans, we do this all the time. Indeed, a major benefit of societal living is the access to members in the group who specialize in one activity or another; it is only sensible to defer to an expert's judgment in those areas we know little about. Similarly, an agent in a multi-agent system should be able to draw upon expert opinion or instruction in those cases where it has little domain knowledge. And a prerequisite for this is the willingness to allow others either to influence decisions or to make decisions on one's own behalf. As stated earlier, however, all this exists within the context of enlightened self-interest. The answer to whether we can control autonomy depends on the viewpoint adopted. In the strong view, it is by definition impossible to control autonomy externally. At the same time, however, we can design agents with appropriate motivations and motivational mechanisms that constrain and guide agent behaviour as a result of internal imposition. In this way, control is *on-board*, and more and better processing of environmental information is required.

The SMART agent hierarchy distinguishes clearly between objects, agents and autonomous agents in terms of goals and motivations. Such an analysis of the entities in the world not only provides appropriate structures so that different levels of functionality may be established, but also information as to how multiple entities or agents can cooperate to solve problems which could not be solved alone. By basing the distinctions on function and purpose, we do not arbitrarily differentiate between knives and robots, for example, especially when it is not useful to do so. Instead, our motivation and goal-based analysis allows us to concentrate precisely on important aspects of multi-agent interaction and problem-solving. In that context, we have considered the roles of goal generation and adoption. We have specified how and why goals must be generated in some autonomous agents in response to motivations, grounding chains of goal adoption, and further, how goals are adopted by objects, agents and autonomous agents in this agent model.

This paper has looked at the issues surrounding the notion of autonomy in agent systems. We have discussed how there is a growing need for an explicit operationalization of the term in order that issues surrounding autonomy can be addressed. In the literature to date there appear two distinct yet related notions of autonomy. The first of these refers to the level to which an agent is free from dependence on other agents in the decision-making process. As such, autonomy here is a relative and continuous concept admitting many levels ranging from complete autonomy through consensus levels of autonomy to a complete lack of autonomy (as in command-driven agents). Difficult issues arise when considering how to determine when, and by how much, an autonomy level should be changed. Should the impetus come from some supervisory power external to the agent or should the agent itself decide when to relinquish its autonomy? This last possibility links in with the second notion of autonomy as the ability to generate one's own goals. Here, autonomy cannot be erased by dependence. Agents are free to generate the goal to submit to another's authority or to share authority in the generation of a goal or in a decision-making process if, by doing so, the agent's interests are best served. It is this latter notion that has been the main focus of this paper.

One last consideration is needed with regard to autonomy: when is it needed? Autonomy's main advantages of flexibility and robustness in the face of dynamic, open worlds can be distinctly undesirable in certain agent domains. Indeed, the strong view of autonomy can be very dangerous if used for example in military applications for tank or missile control. Indeed introducing autonomy into any form of safety critical domain demands extreme caution and extensive testing and may well be best served by other forms of agent control. Thus, we also need to consider the kinds of situations to which autonomy is suited. Whilst we have offered an absolute theoretical viewpoint of autonomy in the form of goal generation as well as the weaker alternative of dependence, which provides a practical realisation of autonomy that is useful for many, it is important to understand the difference in purpose and context of these notions, and not to be dogmatic in practical situations. Clearly, there is value in studying the general concept of autonomy, regardless of practical concerns, but we must also address ourselves to the practical issues. Ultimately, it matters little what we call autonomy (just as it matters little whether we call a program an agent) as long as it gives us the required robustness and flexibility we desire.

In that sense, there is likely to be a convergence of the two views. The strong view offers better-defined mechanisms for directly controlling autonomy, but with less obvious means of manipulation. Future work on autonomy should seek to provide such means of manipulation to enable better user-control, and to allow application in the kinds of domains where user-intervention and control may be critical. This could be through non-invasive ways of coercing agents into certain decisions and behaviour, or possibly through some analogue of invasive courses of action like drugging, hypnosis, etc. Either way, the issues involved in agent autonomy are important, and some early results in this area are likely to provide a rich vein of future work.

6. REFERENCES

Balkenius, C., 1993. The roots of motivation. In J. Meyer, H. L. Roitblat, and S. W. Wilson, editors, From animals to animats 2, *Proceedings of the Second International Conference on Simulation of Adaptive Behavio*r. 513, MIT Press/Bradford Books.

Barber, S. and Martin, C., 1999. Agent Autonomy: Specification, Measurement, and Dynamic Adjustment*. In Proceedings of the Autonomy Control Software Workshop, Autonomous Agents 1999 (Agents'99),* 8-15, Seattle WA.

Beaudoin, L. P. and Sloman, A., 1993. A study of motive processing and attention. In *Prospects for Artificial Intelligence: Proceedings of AISB*93, 229–238, Birmingham.

S. Brainov and H. Hexmoor, Quantifying Relative Autonomy in Multi-agent Interaction, *In Proceedings of the IJCAI'01 Workshop Autonomy, Delegation, and Control: Interacting with Autonomous Agents,* 27-35, Seattle, 2001.

Calstelfranchi, C., 1995. Guarantees for Autonomy in Cognitive Agent Architecture, In *Agent Theories, Languages and Languages (ATAL'94), Lecture Notes in Artificial Intelligence 890,* 56-70, Springer.

d'Inverno, M. and Luck, M., 1996a. A formal view of social dependence networks. In Zhang, C. and Lukose, D., editors, *Distributed Artificial Intelligence Architecture and Modelling: Proceedings of the First Australian Workshop on Distributed AI, Lecture Notes in Artificial Intelligence 108*7, 115–129, Springer-Verlag.

d'Inverno, M. and Luck, M., 1996b. Formalising the contract net as a goal directed system. In Van de Velde, W. and Perram, J.W., editors, *Agents Breaking Away: Proceedings of the Seventh European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Lecture Notes in Artificial Intelligence 103*8, 72–85, Springer-Verlag.

d'Inverno, M. and Luck, M., 1997. Development and application of a formal agent framework. In Hinchey, M. G. and Shaoying, L., editors, *Proceedings of the First IEEE International Conference on Formal Engineering Method*s, 222–231. IEEE Press.

d'Inverno, M. and Luck, M., 2000. Sociological Agents for Effective Social Action, in Proceedings of the Fourth International Conference on Multi-Agent Systems, 379-380, IEEE Computer Society.

Franklin, S. and Graesser, A., 1997. Is it an agent or just a program?: A taxonomy for autonomous agents. In Muller, J.P., Wooldridge, M.J. and Jennings, N.R., editors, *Intelligent Agents III- Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages, Lecture Notes in Artificial Intelligence, 1193,* 21-35. Springer.

Halliday, T., 1983. Motivation. In Halliday, T. R. and Slater, P.J.B., editors, *Causes and Effect*s. Blackwell Scientific.

Halperin, J.R.P., 1991. Machine motivation. In Meyer, J.A. and Wilson, S.W., editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animat*s, 238–246. MIT Press/Bradford Books.

Hinde, R.A., 1982. *Ethology: Its nature and relations with other science*s. Fontana Press.

Kunda, Z., 1990. The case for motivated reasoning. *Psychological Bulleti*n, 108(3):480–498.

Luck, M., 1993. *Motivated Inductive Discover*y. PhD thesis, University of London, London.

Luck, M. and d'Inverno, M., 1995. A formal framework for agency and autonomy. In *Proceedings of the First International Conference on Multi-Agent System*s, 254–260. AAAI Press / MIT Press.

Luck, M. and d'Inverno, M., 1996. Engagement and cooperation in motivated agent modeling. In *Distributed Artificial Intelligence Architecture and Modeling: Proceedings of the First Australian Workshop on Distributed AI, Lecture Notes in Artificial Intelligence 108*7, 70–84. Springer.

Luck, M. and d'Inverno, M., 2001. Autonomy: A Nice Idea in Theory. In *Intelligent Agents VII, Lecture Notes in Artificial Intelligence* 1986, 351-353, Springer.

Maes, P., 1989a. The dynamics of action selection. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligenc*e, 991–997, Detroit.

Maes, P., 1989b. How to do the right thing. *Connection Scienc*e, 1(3):291–323.

Maes, P., 1991, A bottom-up mechanism for behaviour selection in an artificial creature. In J. A. Meyer and S.W. Wilson, editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animat*s, 238–246. MIT Press/Bradford Books.

Moffat, D. and Frijda, N.H., 1995. Where there's a will there's an agent. In M. Wooldridge and N. R. Jennings, editors, *Intelligent Agents: Theories, Architectures, and Languages, Lecture Notes in Artificial Intelligence 89*0, 245–260. Springer.

Moffat, D., Frijda, N.H. and Phaf, R.H., 1993. Analysis of a model of emotions. In *Prospects for Artificial Intelligence: Proceedings of AISB9*3, 219–228, Birmingham.

Norman, T.J. and Long, D., 1995. Goal creation in motivated agents. In Wooldridge, M.J. and Jennings, N.R., editors, *Intelligent Agents: Theories, Architectures, and Languages, Lecture Notes in Artificial Intelligence 89*0, 277–290. Springer.

Norman, T.J. and Long, D., 1996. Alarms: An implementation of motivated agency. In M. Wooldridge, J.P. Muller, and M. Tambe, editors, *Intelligent Agents: Theories, Architectures, and Languages, Lecture Notes in Artificial Intelligence 103*7, 219–234. Springer.

Schnepf, U., 1991. Robot ethology: A proposal for the research into intelligent autonomous systems. In J. A. Meyer and S.W. Wilson, editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animat*s, 465–474. MIT Press/Bradford Books.

Simon, H.A., 1979. Motivational and emotional controls of cognition. In *Models of Though*t, 29–38. Yale University Press.

Sloman, A., 1987. Motives, mechanisms, and emotions. *Cognition and Emotio*n, 1(3):217–233.

Sloman, A. and Croucher, M., 1981. Why robots will have emotions. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligenc*e, 197–202, Vancouver, B.C.

Waltz, D.L., 1991. Eight principles for building an intelligent robot. In J. A. Meyer and S.W. Wilson, editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animat*s. 462-464, MIT Press/Bradford Books.