# A Dialectic Architecture for Computational Autonomy

Mark Witkowski[1] and Kostas Stathis[2]

[1]Intelligent and Interactive Systems Group, Department of Electrical and Electronic Engineering, Imperial College, Exhibition Road, London SW7 2BT, U.K.
m.witkowski@imperial.ac.uk
[2]Intelligent Computing Environments, Department of Computing,
School of Informatics, City University, London EC1V 0HB, U.K.
kostas@soi.city.ac.uk

**Abstract.** This paper takes the view that to be considered autonomous, a software agent must possess the means by which to manage its own motivations and so define new goals. Using the motivational theories of Abraham Maslow as a starting point, we investigate the role that argumentation processes might play in balancing the many competing aspects of a whole agent's motivational agenda. This is developed into an Agent Argumentation Architecture (AAA) in which multiple "faculties" argue for different aspects of the total behavior of the Agent. The overall effect of these internal arguments then defines the overt "personality" of the agent.

## 1 Introduction

In this discussion paper we consider the nature of autonomy, what it means to be autonomous and how a greater degree of autonomy might be achieved in independent machines. We shall use the notion of a software agent as the exemplar type of system to discuss the nature of autonomy, but also draw on ideas from psychological theory.

Autonomy should be considered as separate from automatic or independent operation [11], [12], [15], [25]. Luck and d'Inverno [11] have proposed the view that an object can be regarded as an agent once it has goals and the means to effect them, and an autonomous agent one that has motivations, and so the ability to create goals according to some internal, hidden and changeable agenda. By this definition the overwhelming majority of "autonomous agents" of the type characterized by [6] and [13] would be more properly defined as independent. In this view autonomous systems cover a wide spectrum of degrees of autonomy. The earliest obvious precursors, Cybernetic or homeostatic [27] machines use pre-defined (but possibly self-adjusting) control strategies to maintain a pre-defined set point. As with these automatic systems, the reactive or behaviorist model [2] is certainly independent of the programmer, but tied to the strategies laid down by its program. In the standard BDI agent model [21], for instance, immediate pre-programmed strategies are replaced by a goal driven approach in which decisions about which specific actions to be taken are deferred until the goals are activated.

This paper investigates the situation where a software agent has control of its goals and decision-making process, but also the ability to set and maintain its own agenda of goals and select actions for expression according to an individualized class of arguments and internal priorities. Full autonomy and thus complete freedom from the programmer requires that the agent can learn new activities, adopt new goals and, for complete autonomy, devise new learning strategies also.

The term "autonomy" is derived from the ancient Greek Αυτονομια, meaning self-regulation by having self-laws. Formal models of Deontics, the formalization of duties, obligations and prohibitions, seem counterproductive in this context. At their heart, the paradox of the notion of something that is obligatory, yet coupled to sanction to be applied if that obligation is not fulfilled ("contrary to duty"). Note that this paper is not a discussion of morality, of rights and wrongs, but rather of mechanism. A fully autonomous agent cannot be obliged (as in deontics) to conform to law, but must decide the consequences under its own prioritization and proceed or not (e.g. without sanction, law has no effect). An agent need not be isolated from other agents, it will need to interact and cooperate with peers. Furthermore, an agent does not need to act as a servant, but might elect to act as such.

We are interested in producing a generic architecture, building on notions proposed by Kakas and Moraïtis [8], for embodied and non-embodied entities to achieve full autonomy for its own sake. But we will then consider implications for practical technology, which may in turn tell us something about the human condition. Why, for instance, do we consider ourselves autonomous? Having made the proposal for such an architecture in the body of this position paper, we return in the discussion section to these issues. The issues under discussion here are at the edge of what current techniques in reasoning can be expected to achieve, and are perhaps at the very boundary of what logic can be expected to represent, at least in its current form ([9] for a discussion). Detailed considerations of the logic formalization fall outside the scope of this position paper.

Section two presents a view of the motivational theories [14] of Abraham Maslow (1908-1970), and asks whether they can offer any insight into how a software agent may be made more completely autonomous. A software agent might do so by taking more immediate control of its own behavioral agenda, setting its own goals and determining the overall outcomes and consequences to the agent in terms of those various and varying motivating factors. We shall use the approach laid out by Maslow as a starting point, but argue that its provisions do not entirely apply in the case of Software Agents.

Section three we will look at extensions to the last class of autonomous agents, loosely based on notions of the BDI architecture ([1], [21]), in which processes ("faculties") encapsulating a number of different top-level goals (which appear as "motivations" to an observer of the agent, or to the introspective agent) must both propose new actions or goals that support the area they are responsible for, and argue that these goals should be adopted by the agent as a whole. We shall take the view this overall process can and should be viewed and modeled as one of a dialectic argumentation game, in which individual faculties must both argue for the value of their individual contribution, of which they are a part.

Argumentation has found favor recently as a way of modeling legal arguments using logic ([10], [19], and [20] for review). We note that there are significant differences between legal and internal argumentation, and that the categories of argument must therefore be different. Section five will discuss a procedural layer, providing for a game-like protocol by which the argumentation might take place.

## 2 Personality and Motivations

This section takes as its starting point a discussion of the motivational theories of Maslow [14]. Maslow's work has been influential in the understanding of human drive and its relationship to the expression of personality (and thence to our notions of autonomy as individuals). It is important because it attempts to relate a model of underlying processes to observable traits, and as such stands apart from work which primarily serves to categorize and measure personality (e.g. [5], [7], and [22] for review).

Maslow describes five classes of motivation, forming a dynamic "needs hierarchy". At the base level are **Physiological needs**, the immediate requirements to maintain the function of the organism. These needs, in the living organism, will include homeostatic [27] functions, such as blood sugar or oxygen balance, or hydration levels. Physiological needs may also drive complex behaviors, such as food acquisition, that are not well modeled with a control theoretic approach. At the next level **Safety Needs** predominate. In this respect Maslow specifically refers to the search for stability and freedom from fear and anxiety in the individual's continuing context, rather than freedom from immediate danger. At the third level **Belongingness and Love Needs** emerge. These refer to the apparent human requirement to seek out and maintain immediate contact with other individuals in a caring and cared for context ("the giving and receiving of affection"). Maslow argues that failure to achieve or denial of these needs leads to a wide range of distressing psychological symptoms. At the fourth level **Esteem Needs** emerge. Maslow divides these needs into two primary categories, the need for self-esteem, "… the desire for strength, achievement … independence and freedom" and for the esteem of others, "… status, fame and glory, dominance, recognition, attention, importance, dignity, or appreciation". It is clear that these two categories (as with the other major needs) cover a broad spectrum of possible drivers for activity. At the final level **Self-actualization Needs**, the individual is driven to develop its capabilities to the highest degree possible, "what humans can be, they must be". This level will include invention and aesthetic (musical, artistic, and, presumably, scientific) achievement.

In describing this as a needs hierarchy, Maslow postulates that until a lower level need is satisfied, the next level will not emerge. We suggest that this does not represent a true hierarchy (in the sense that one level facilitates, or is facilitated by, the next), rather that the lower, say physiological needs, are just generally more urgent than the others, so the argument for satisfying them becomes correspondingly stronger and not easily overturned by less urgent topics. Such a mechanism still appears to the observer as a "hierarchy" in the manner indicated by Maslow. Yet it is clear that, in humans at least, the higher-level "needs" can completely subsume the lower. An artist might starve in his garret to produce works of great personal aestheticism, which nobody else appreciates. An ambitious person might seek public esteem at the expense of personal relationships and happiness – yet still be a glutton.

Kakas and Moraïtis [8] describe the power structure between the basic Maslow levels (motivations $M_i$ and $M_j$, for instance) using a priority relation (h_p($M_i$,$M_j$)) indicating that $M_i$ has a higher priority than $M_j$. Morignot and Hayes-Roth [17] suggest a weighting vector to arbitrate between levels. These schemes have a gross effect on "personality". The levels appear to have, and need to have, a more subtle interaction, sometimes winning out, sometimes losing.

Maslow elegantly captures this idea: “sound motivational theory should … assume that motivation is constant, never ending, fluctuating, and complex, and that this is an almost universal characteristic of practically every … state of affairs.” The model we propose divides the agent’s reasoning into many independent but interacting faculties. Each faculty is responsible for arguing the case for the activity or approach for which it is responsible and any that support it, and arguing against any that would contradict it. The strength of each faculty is determined primarily by the number and applicability of arguments it has available, but ultimately on notions of implicit preference which definitively answers the question “what is it that I want more?”

Maslow argues that there is no value in trying to enumerate a fixed list of drives (despite producing long lists of need descriptive adjectives to illustrate his levels), and that drives overlap and interact. Yet to produce an equivalent “motivation theory” for a software agent, we must exactly isolate the specific factors that will motivate the behavior of the individual agent and produce some form of explanation as to why they should be incorporated into the design in addition to proposing a mechanism by which they might appear as “never ending, fluctuating, and complex”.

In this part of the paper we consider a partial equivalence between the human interpretations of the Maslow motivations to that of a software agent. Agent faculties can be divided into several main grouping. (1) **Operational:** those relating to the immediate protection of the agent and its continuing operation. (2) **Self-benefit:** Those relating to aspects of the agent’s behavior that is directly related to its ongoing protection and individual benefit. (3) **Peer-interaction:** those relating to individually identified entities, human or artificial, with which it has specific relationships. (4) **Community-interaction:** those relating to the agent’s place in an electronic society that might contain both other software agents and humans, with which it must interact. This category might include both informal and institutionalized groups. (5) **Non-utilitarian:** longer-term activities, not directly related to tasks that offer an immediate or readily quantified benefit.

We assume that the key resource that a software agent must maintain is access to processor cycles, associated data storage and the communications medium reflecting the immediate protection criteria (1). Without immediate access to an active host the agent is completely ineffective and effectively dead. The situation is somewhat more apparent with an embodied agent, such as a robot, where access to uninterrupted power and avoidance of physical damage are clear criteria [17].

Category (2) above equates broadly to the safety needs. In this category an e-commerce agent might seek to accumulate financial strength to pay for a reliable infrastructure strategy, or construct a viable migration plan.

Category (3) addresses how to interact with immediate, individually identified humans and other software agents. Each will have its own personality, and the agent must tailor its interactions with them in specific ways to maintain appropriate relationships and be able to achieve its goals in the future. This broadly equates to the care and kinship needs, though it may be that an autonomous agent might take a hostile view towards a third party, perhaps arguing that “my friend’s enemy is my enemy”. That is, the relationship between an established ally is more important than the third party, and that it would be jeopardized by perceived collaboration with that third party. An alternative, more social, view would propose the agent has a duty of care towards them anyway. How an autonomous agent would represent or express a personal dislike remains to be explored.

In category (4) the agent accumulates arguments relating to the peer groups, identified, but not individualized and towards the greater society in which it, and any human partner with which it is associated must operate. In the case of an e-commerce assistant agent, this will almost certainly include aspects of the full legal system, and would certainly include the norms and standards of the particular trading circles in which the agent and partner choose to operate.

Level (5) remains problematic for software agents in the absence of a “feel-good” qualia for agents, but one might speculate that successful agents, those that have accumulated an excess of resources by careful management or good luck would have the opportunity to continue exploring their world for self-improvement alone. One could speculate even further that some might continue to accumulate excess resources for its own sake, or enter into philanthropic activities for less-fortunate agents or agent communities of their choosing.

## 3 A Game-based Architecture for Dialectic Argumentation

The Agent Argumentation Architecture (AAA), shown in Figure 1, consists of the following components, an Argumentation State, a Knowledge Base (KB), a number of Faculties (F), an “attender” module (managing the flow of incoming information) and a “Planner/Effector” module (responsible for making plans from goals and performing actions as required). The interaction between components is organized as a complex game consisting of argument sub-games for achieving goals. We draw from the representation already available in [23] to describe games of this kind in terms of the valid moves, their effects, and conditions for when these argument sub-games terminate and when goals may be reestablished.

The **Argumentation State (AS):** A communal structure for the current state of the game, including the arguments put forward, and not yet defeated or subsumed, but accessible by the faculties and the input and output modules.

The **Knowledge Base (KB):** Acts, conventionally, as a long-term repository of assertions within the system. For the purpose of the discussion that follows we shall assume that the elements held in KB take the form of conjectures, rather than expressions of fact. To assume this implies that the knowledge of the agent is non-monotonic, credulous (as opposed to skeptical), and allows inconsistency. On the other hand, a monotonic, skeptical and consistent knowledge base hardly allows for an argumentation process, as it is always obliged to agree with itself.

We will partition the KB according to the Maslow motivation types (KB = $\{K_{m1} \cup K_{m2} \cup K_{m3} \cup K_{m4} \cup K_{m5}\}$, as indicated by the solid radial line in figure 1) and according to the faculties (KB = $\{K_{f1} \cup \ldots \cup K_{fn}\}$, as indicated by the dotted radial lines in figure 1). We assume that the number of faculties (n) will be greater than 5, and that some faculties will impinge on more than one motivational category.

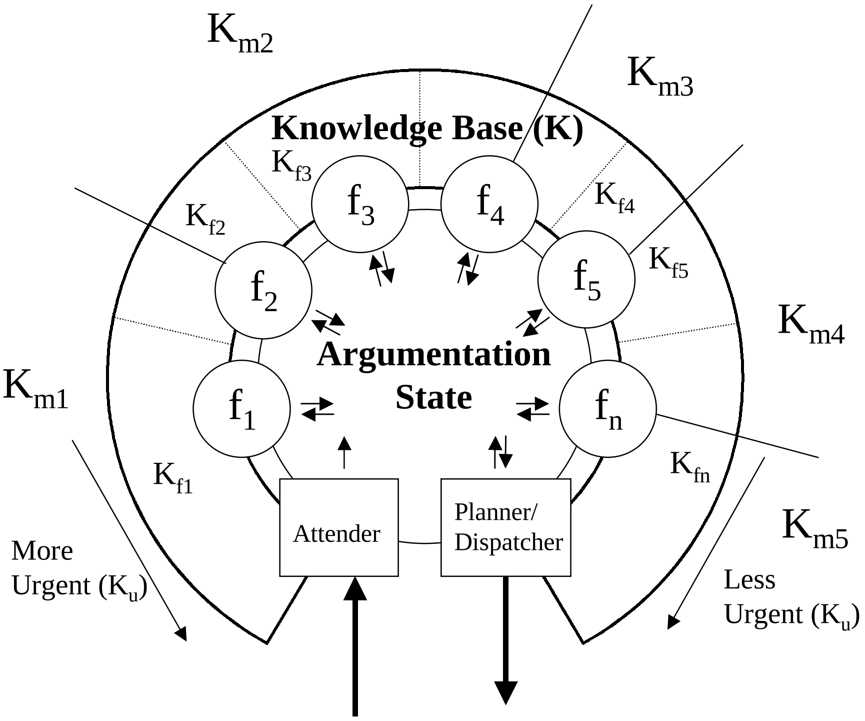

Figure 1: The Agent Argumentation Architecture (AAA)

The **Faculties:** Faculties ($F = \{f_1 \dots f_n\}$) are responsible for particular aspects of the whole agent's possible agenda, which taken together will comprise every aspect that the agent can address. Each faculty may therefore argue that a goal should be established to actively achieve some aspect of that agenda (or avoid some situation that would be detrimental to the agenda). Equally it must monitor the goals and actions proposed by other faculties to determine whether the consequences of those goals or actions would interfere with or contradict some aspect of its own agenda. If some proposal supports the faculty's agenda, it will argue in support of the proposal, or argue against it if the agenda is contradicted. A faculty could, of course be ambivalent towards a proposal, which can have both positive and negative consequences, the faculty being supportive, unsupportive or neutral according to the relative merits of the two positions. Each faculty is arguing the whole agent's best interests are served by pursuing (or not pursuing) certain courses of action, but from its specific viewpoint. There is no winner or loser; each faculty is (or at least, should be) working towards the overall advantage of the whole agent. Essentially each faculty has an "opinion" of what is best for the agent as a whole, but from its limited viewpoint, and must successfully argue its case against other possibly competing views for this to prevail and become incorporated into the agent's overt behavior.

The **"Planner/Effector" module:** Is responsible for creating prototype plans from agreed goals and effecting actions from agreed plans.

The **"Attender" module:** We assume that there is a continuous stream of incoming information, which will comprise of at least the following types of item:

requests to perform activities on behalf of other agents, broad suggestions of things the agent might like to do or adopt (for instance, adverts, suggestions by providers that the recipient would be advantaged by purchasing or doing something), “news” items, indicating the outcome of various previous actions and activities, and solicited or unsolicited assertions from other agents, which the current agent may or may not wish to adopt according to its current motivational strategy. These are delivered to the AS and discarded soon, each faculty having a brief time to inspect and adopt them into the KB if required.

## 4 The Role of Argumentation

In this section we consider how argumentation, [10], [19], [20] might play a role in the architecture for a fully autonomous agent described in the last section. The majority of work in argumentation has been conducted as an approach to the mechanization of the legal process, (often) seeking to model how lawyers conduct cases. In Prakken and Sartor’s view [20] a (legal) argumentation system will have four notional layers, a *logical layer* defining the structure of arguments and the underlying semantics, a *dialectical layer* defining how conflicting arguments will be resolved, a *procedural layer* defining the discourse rules for the argumentation and a *strategic or heuristic layer* providing rational ways to conduct the argumentation. Much attention falls onto the second of these, as it captures the form of the argumentation process. As legal argumentation is primarily combative it is often considered as an n-ply game in which one party presents an assertion, which the other party then attempts to overturn, leading in turn to the possibility of a counterargument, and so on. The procedure terminates when there is no further effective counterargument to be made and the argument is therefore won or lost.

Legal argumentation is, in the English and US tradition, both combative and largely retrospective, comprising accusations and rebuttals between prosecution and defense. In the legal process there is an overall and external framework to arbitrate between the parties (the statutes) and sanctions to be applied to those found guilty. In the proposed model for autonomy, the argumentation will be about current or future events, goals and actions, and the consequences that may follow if one route or another is taken. There will be no overall guiding external principle against which a definitive decision about what to adopt as the overt behavior of the agent might be judged. Existing models of legal argumentation [20] rely on four major types of argumentation strategy (though not necessarily all in the same system): **Conflicts**, arguments reaching differing conclusions may be *rebutted*, where one aspect of an established argument is shown not to hold in the current circumstances, *assumption attack*, the assumption of non-provability is presented with a proof, and *undercutting*, where one argument challenges the rule of inference used by another. **Argument comparison**, where lines of argument are decided by recourse to a higher principle. **Argument status**, where defeated arguments may be reinstated if the rebuttal is itself defeated. In **defeasible argumentation**, general rules may be defeated by more specific ones, where a defined priority relationship exists.

The agent argumentation system is driven by the search for, and resolution of conflicts. Conflicts arise where two faculties arrive at different conclusions at the end of individual chains of consequences, or one asserts that an action by the other will be detrimental to its agenda. We identify six classes of argument appropriate to the AAA:

**Goal Proposal Move:** Some faculty $f_n$ determines that prevailing circumstances imply that a goal must be asserted or will become asserted shortly.

**Conflict of Interests Moves:** For a goal or action proposed by faculty $f_n$ with intended consequence *x*, faculty $f_m$ asserts that some *y*, which it believes is also consequential leads to a conflict, either by asserting an action that $f_m$ determines as detrimental (interest conflict) or interfering with a current plan of $f_m$ (resource conflict). $f_n$ may retract, propose an alternative or a covering action.

**Alternative Argument Move:** Faculty $f_m$, having detected a conflict of interests, proposes an alternative solution to achieve $f_n$'s original goal in a manner that does not conflict with its agenda (cooperative).

**Retraction Move:** Faculty $f_n$ retracts a proposed action or goal, because it is no longer applicable to $f_n$ because of changed circumstances (including conflict with another faculty).

**Undercut Move:** Faculty $f_n$ challenges the assertion that there is a conflict, and attempts to undercut $f_m$'s consequence chain, by arguing $f_m$ has included an invalid step or overestimated its significance.

**Covering-action Move:** Some action *a* has a positive outcome for $f_n$, but gives rise to a potential liability according to $f_m$, $f_m$ may propose a prior action that would avoid the liability. This is a "duty of care" argument, complicating the behavior of the agent to ameliorate possible (but not certain) negative outcome. This argument is particularly valuable where the undesirable consequence is rare, but highly damaging.

In this model, the agent is required under normal circumstances to completely exhaust the argumentation process, i.e. to explore all the rational routes to a decision [1]. A hung decision would, in an isolated system, represent a considerable dilemma with the agent unable to act. In a connected system, new information is always arriving on the AS, and this new information may be sufficient to tip the balance one way or the other. This is, perhaps, the ideal, but likely the agent cannot wait.

Underlying much of human decision-making would appear to be a complex web of preferences, allowing us to choose between the otherwise rationally un-decidable alternatives. In the first instance the system should determine whether there are any preference relations specific to the two items in conflict (prefer(*x*,*y*)). The preference for a rule may be tied closely to the confidence an agent has in it [28]. This being so the situation is resolved. Failing that, more general classes of preference should be invoked.

The agent will have a current preference ordering relationship between, say, the urgency 'U' of rules in the knowledge ($k_n \subset K$) base. For example the preference ordering of U ($U_p$) indicates the agent's current rank ordering of K ($k_1$ .. $k_n$), expressed as: $U_p$: prefer(u($k_{34}$), u($k_{12}$), u($k_{100}$) …, u($k_m$)). Similarly the agent can place a rank ordering on the roles of different faculties ($F_p$: prefer($f_n$, $f_q$, …, $f_m$)), or on different motivational areas ($M_p$: prefer($m_2$, $m_1$, $m_3$, $m_4$, $m_5$), as in [8]). The agent may further make explicit the rank ordering of these classes of rankings (e.g. prefer($U_p$, $F_p$, $M_p$)), and so be in a position to change them. It is clear that, in humans at least, these preference orderings are quite variable, modified by mood, emotion and recent events.

The outcome of the argumentation process equates to Von Wright's [26] notion of an *extrinsic* preference, subject to reason and rationality, and the rank ordering to *intrinsic* preference, hidden and apparently outside rationality. The manner in which preferences might be dynamically updated, while interesting and perhaps central to notions of full autonomy, falls outside the scope of this paper.

## 5 The Procedure

This section outlines the activities (the *procedural layer*) by which an agent's faculties might interact to define a goal directed strategy. Overall, the activities are essentially asynchronous, with faculties responding to items appearing on the AS, and are bounded by the computational resource available to them and the whole agent.

**Activity 1)** Each faculty is responsible for proposing immediate actions or goals onto the communal AS. It is expected that a goal or action will be one that maintains or enhances the whole agent, but strictly from the viewpoint of the topic related to the proposing faculty (i.e. they are expected to be topic "selfish"). This is the primary creative component of the whole agent. Proposals can be immediate, short or long term. In general, the lower numbered "motivations" will give rise to immediate and short-term proposals, the higher numbered ones mid to long-term proposals, reflecting the scale of urgency. Proposals for immediate actions will often relate to safety or highly opportunistic situations (a purely reactive faculty could only propose actions). In an agent of any sophistication there will be many specific goals, of varying duration, active at any one time. In general, an agent will have many more suggestions for actions and goals than it could reasonably service. Proposals are therefore opportunistic and made with the likelihood that they will be rejected. Any action or goal proposed at this stage represents a "desire" on the part of the whole agent.

**Activity 2)** Every proposed action or goal must be vetted by each of the other faculties to determine whether it, or any of its known consequences, would violate (or augment) the agenda of that faculty in the current circumstances. Arguments against a proposal can include: the proposal, or one of its known consequences, would directly contradict an aspect of the vetter's agenda; that the proposal would, if enacted, drain resources from a previously agreed course of action, disrupting, or making that course of action untenable. To not respond to a proposal at this stage is surely to tacitly accept it. One might suppose that this, and the other, vetting stages would be subject to a policy-based [1] strategy to reduce computational load and delay.

**Activity 3)** Those goals that pass through stage 2, are passed to a conventional means-ends planner. The planner might, at this stage, find no reasonable instantiation of the goal in the current circumstances, and it would be (temporarily) abandoned. Otherwise the planner will deliver to the communal AS a proposed sequence of actions for all the faculties to consider. At this stage the planner is only required to produce a viable sequence of proposed actions, perhaps based on a least cost heuristic or some other (perhaps simplistic, perhaps sophisticated) "optimal strategy" metric. It may contain considerable consequential risks and liabilities to the agent.

**Activity 4)** With the extra detail of the instantiated plan, each faculty is required once again to review the plan, raising arguments or objections if any action proposed as part of it would cause immediate or consequential violation of any faculty's primary motivations. At this stage a plan might be rejected outright or be returned to the planner for modification, to avoid or amend the contended step or steps. Again, to not respond at this stage is to tacitly accept the plan and its consequences. The agreed sequence of actions become part the whole agent's intentions at this stage, and are passed to the plan effector module.

**Activity 5)** At the allotted time, or under the planned conditions, the effector module will briefly present the action as instantiated at the moment of execution to the AS, giving the assembled faculties one last chance to argue that it should be suppressed, primarily due to changes in circumstances since it was previously agreed. Cotterill [3] refers to this as "veto-on-the-fly", arguing that it brings a significant advantage to an agent; it is also reminiscent of Bratman's [1] rational reconsideration step. If the action is challenged this will appear as a hesitation, even if the action finally goes ahead.

## 6 Discussion

This model presents an autonomous agent as a peer group of faculties, each with a role within, and responsibility to, the whole agent. This is not a "democracy". There is no voting, but argumentation in which the "last man standing" [4] (the faculty with the most effective arguments, or whose arguments are preferred) becomes the winner, and whose suggestions become part of the overt behavior of the whole agent. Acceptance of this criterion effectively guarantees that each argumentation game will terminate in a finite time. Additional urgency constraints may, of course, be required.

The observable "personality" of the agent is therefore defined by the balance and effectiveness of the individual faculties. In a well-balanced agent system based on these principles, the range of faculties will cover all the aspects of the whole agent's individual and social requirements. Equally, each faculty will have an equitable share of the overall resources and only propose a reasonable number of goals, deferring if the arguments placed against it are stronger (i.e. not continue to assert weaker arguments, once a stronger one is presented, or to repeat defeated arguments unless a material change to circumstances has occurred).

It will be interesting to speculate as to the effects of various failings or imbalances in the relative strengths of each component faculty. If a faculty has few, or only weak, arguments, the whole agent will appear weak, or even oblivious, to that part of the normal social interaction. If "individual" completely dominates over "society", the agent shows symptoms of "pathological behavior", ignoring social convention and the needs of others. If the argumentation process that vets the initial plan (activity 4) or the release of actions the agent's behavior is inadequate (activity 5), the agent appears *akratic* (mentally incontinent) in the social context [15]. If a specific faculty were to persistently propose goals, particularly the same one, even if repeatedly rebuffed the whole agent would appear distracted. If these were not successfully rebuffed, then the agent would appear to be obsessive. This could, of course, take many forms, depending on the specific faculties involved.

The model and architecture presented here has much of the flavor of Minsky's view of a society of mind [16], with individual, focused, component parts each contributing some specialization or expertise in the service of the whole. There is further a clear analogy between this society of mind within an agent and a society of individual agents. A strongly autonomous agent makes a significant impact on that society by being able to express its internal arguments and preferences to influence other. Every action made contributes to the agent's perceived personality. It remains unclear whether “motivation” in the sense used by Maslow reflects a *post-priori* explanation of observable behavior arising from competitive individual “interest groups” (faculties) or a genuine drive mechanism. Maslow puts high emphasis on motivation as a sensation overtly available to the (human) agent; we place emphasis on the mechanism, giving rise to apparently complex motivated behavior within a deterministic framework.

Whether fully autonomous agents of this type will ever find a place in the world of e-commerce remains to be seen, but the outlook is not encouraging, Luck and d’Inverno [12] are certainly skeptical. Almost certainly not in the immediately foreseeable future, as most people want and expect their computers and software to behave in a (reasonably) consistent manner and perform the actions as requested. For instance work to date ([24], [29]) on agents in the connected community, presupposes that software agents have restricted flexibility and are primarily subservient. By definition a fully autonomous agent may refuse, and may even perform actions specifically contrary to the wishes or intentions of a “user”. It could easily be seen as a recipe for apparent erratic and unpredictable behavior. On the other hand it may be that the extra flexibility offered by full autonomy will play a part in generating a whole new class of interesting cooperative activities and the development of independent communities of such agents, literally trading on their own account and purchasing the computational and support resources they need from dedicated suppliers. After all the richness and diversity of human society is the product of the interactions between uncounted millions of apparently autonomous entities.

However, e-commerce is not the only application for software agents, and true autonomy may find unexpected value in a range of applications from entertainment to providing companionship for the lonely. In each case the unexpected and unpredictable may prove to be an added bonus, otherwise missing from the pre-programmed.

For the engineer or computer scientist with a philosophical leaning, discovering what would make a software agent autonomous would in turn have the most profound implications for our understanding of how humans come to view themselves as autonomous and in possession of apparent “freewill”. True autonomy in artificial agents is worth studying for this reason alone. Our suspicion is that we have hardly scratched the surface of this problem – or begun to perceive its full potential.

## References

[1] Bratman, M.E. (1987) *Intention, Plans, and Practical Reason*, Cambridge, MA: Harvard University Press

[2] Brooks, R. (1991) Intelligence without Representation, *Artificial Intelligence*, Vol. 47, pp. 139-159

[3] Cotterill, R. (1998) *Enchanted Looms Conscious Networks in Brains and Computers*, Cambridge University Press

[4] Dung, P.M. (1995) On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-person Games. *Artificial Intelligence*, Vol. 77, pp. 321-357

[5] Eysenck, H. (1991) Dimensions of Personality: 16: 5 or 3? Criteria for a Taxonomic Paradigm, *Personality and Individual Differences*, Vol. 12(8), pp. 773-790

[6] Jennings, N.R., Sycara, K.P. and Wooldridge, M. (1998) A Roadmap of Agent Research and Development, *J. of Autonomous Agents and Multi-Agent Systems*, Vol. 1(1), pp. 7-36

[7] John, O.P. (1990) The "Big Five" Factor Taxonomy: Dimensions of Personality in the Natural Language and in Questionnaires, in Pervin, L.A. (ed) *Handbook of Personality: Theory and Research*, New York: Guildford, pp. 66-100

[8] Kakas, A. and Moraïtis, P. (2003) Argumentation Based Decision Making for Autonomous Agents, AAMAS-03, to appear

[9] Kakas A.C., Kowalski, R.A. and Toni, F. (1998) The role of abduction in logic programming, Gabbay, D.M. et al (eds.) Handbook of Logic in Artificial Intelligence and Logic Programming 5, Oxford University Press, pp. 235-324

[10] Kowalski, R.A. and Toni, F. (1996) Abstract Argumentation, *Artificial Intelligence and Law Journal*, Vol. 4(3-4), pp. 275-296

[11] Luck, M. and d'Inverno, M. (1995) A Formal Framework for Agency and Autonomy, in Proc. 1st Int. Conf. On Multi-Agent Systems (ICMAS), pp. 254-260

[12] Luck, M. and d'Inverno, M. (2001) Autonomy: A Nice Idea in Theory, in Intelligent Agents VII, Proc. 7th Workshop on Agent Theories, Architectures and Languages (ATAL-2000), Springer-Verlag LNAI, Vol. 1986, 3pp

[13] Maes, P. (1994) Agents that Reduce Work and Information Overload, Communications of the ACM Vol. 37(7), pp. 31-40

[14] Maslow, A.H. (1987) *Motivation and Personality*, Third edition, Frager, R., *et al*. (eds.), New York: Harper and Row (first published 1954)

[15] Mele, A.R. (1995) *Autonomous Agents: From Self-Control to Autonomy*, New York: Oxford University Press

[16] Minsky, M (1985) *Society of Mind*, New York: Simon and Schuster

[17] Morignot, P. and Hayes-Roth, B. (1996) Motivated Agents, Knowledge Systems Laboratory, Dept. Computer Science, Stanford University, Report KSL 96-22

[18] Norman, T.J. and Long, D. (1995) Goal Creation in Motivated Agents, in: Wooldridge, M. and Jennings, N.R. (eds.) *Intelligent Agents: Theories, Architectures, and Languages*, Springer-Verlag LNAI Vol. 890, pp. 277-290

[19] Prakken, H. (1997) *Logical Tools for Modelling Legal Argument. A Study of Defeasible Reasoning in Law*. Dordrecht: Kluwer Academic Publishers

[20] Prakken, H. and Sartor, G. (2002) The Role of Logic in Computational Models of Legal Argument: A Critical Survey, in: Kakas, A. and Sadri, F. (eds), *Computational Logic: From Logic programming into the Future (In Honour of Bob Kowalski)*, Berlin: Springer-Verlag LNCS Vol. 2048, pp. 342-381

[21] Rao, A.S. and Georgeff, M.P. (1991) Modeling Rational Agents within a BDI-Architecture, Proc. Int. Conf. on Principles of Knowledge, Representation and Reasoning (KR-91), pp. 473-484

[22] Revelle, W. (1995) Personality Processes, *The 1995 Annual Review of Psychology*

[23] Stathis, K. (2000) A Game-based Architecture for Developing Interactive Components in Computational Logic, *Journal of Functional and Logic Programming*, 2000(1), MIT Press.

[24] Stathis, K., de Bruijn, O. and Macedo, S. (2002) Living Memory: Agent-based Information Management for Connected Local Communities, *Journal of Interacting with Computers*, Vol. 14(6), pp 665-690

[25] Steels, L. (1995) When are Robots Intelligent Autonomous Agents? *Journal of Robotics and Autonomous Systems*, Vol. 15, pp. 3-9

[26] Von Wright, G.H. (1963) *The Logic of Preference: An Essay*, Edinburgh: Edinburgh University Press

[27] Wiener, N. (1948) *Cybernetics: or Control and Communication in the Animal and the Machine*, Cambridge, MA: The MIT Press

[28] Witkowski, M. (1997) Schemes for Learning and Behaviour: A New Expectancy Model, Ph.D. thesis, University of London

[29] Witkowski, M., Neville, B. and Pitt, J. (2003) Agent Mediated Retailing in the Connected Local Community, *Journal of Interacting with Computers*, Vol. 15(1), pp 5-33