

Agent Autonomy: Social Integrity and Social Independence

Marcus J. Huber

Intelligent Reasoning Systems
4976 Lassen Drive
Oceanside, California, 92056
(760) 806-1497
marcush@home.com

Abstract

As interactions between agents become more common, it will become very important to be able to characterize and perhaps even guarantee an agent's level of autonomy. We will both want agents to perform tasks on their own while at the same time both remaining controllable by ourselves and secure from control and manipulation by others. Furthermore, as multiagent societies become more sophisticated and such artifacts as authority structures are introduced, it will become important to provide a means by which the agent can introspect and perhaps dynamically alter its level of autonomy contextually. Most intuitions of autonomy seem to involve the notion that it is related to dependence/independence. Our model of autonomy captures the notion that, in one sense, autonomy represents security from corruption and manipulation by external influences, i.e., its social integrity. Our model also captures that in another sense autonomy represents an agent's ability to perform its tasks without dependence upon others, i.e., its social independence. This paper presents a multidimensional conceptualization of autonomy and introduces a pragmatic interpretation of our scheme that is applicable to the characterization of the autonomy level of any software entity but which is especially amenable to agent-based systems.

1 Introduction

No man is free who is not a master of himself.
Epictetus

This paper has two goals. The first goal is to motivate that autonomy is a complex and multifaceted concept that cannot be determined looking solely at a single agent and that is best captured using a multidimensional formalization. The second goal is to introduce our multidimensional conceptualization of autonomy that captures many of the aspects of autonomy, primarily the key features of social integrity and social dependency, and then define a relatively simple and pragmatic implementation.

Autonomy is an amorphous notion discussed by philosophers throughout the ages and it is unlikely that there will ever be a single, unanimously embraced definition. Pretty much everyone would agree, however, that it is tied very closely to the notion of *independence*. Also contentious is whether an agent exhibits autonomy in shades of gray or whether it is an all or nothing attribute. Some have defined autonomy as the relative amount of work that the agent does on its own, without help from a human. A variation

on this definition is that autonomy is the relative amount of work an agent can do without *any* help (from a human or an agent). A third very different definition might be whether or not the agent always does that it is asked to do. An analogous definition to this last is whether an agent believes everything that it is told.

Are any of the above definitions of autonomy correct? The answer is both yes and no. They each do seem to capture some aspect of autonomy and are probably suitable for some application or domain. At the same time, they all seem to be missing one or more important aspects of the notion; they all seem too narrowly defined. Compared to the second definition, the first definition suffers from not considering an agent's getting help from other agents; an agent could get everything done without help from a human yet have been completely dependent upon other agents and still be considered autonomous. The first two definitions allow for autonomous agents that blindly do everything they are told, which seems to be contrary to the concept of autonomy. The latter two definitions cover an agent's "right of refusal" but miss the aspect of how much work an agent can do independently.

We can get further insight into autonomy by looking at the work by other researchers. Covrigaru and Lindsay (Covrigaru and Lindsay 1991) believe that an entity's level of autonomy can be determined by looking only at the characteristics of an individual agent. According to them, an entity is more likely to be autonomous the more of the following features it exhibit, including goal-directed behavior, movement, self-initiated activity, adaptability, flexibility, robustness, and self-sufficiency. Covrigaru and Lindsay's definition is somewhat appealing as it does capture a number of features that seem desirable for an autonomous agent to exhibit. Two of their key features, self-initiation and self-sufficiency, suggest that they believe autonomy is in some part a factor of an agent's isolation from other agents. They make an explicitly statement that it does not matter whether goals came from internal processing or received from external sources.

Luck and d'Inverno, in contrast, define an autonomous agent to be anything that has motivations (Luck and d'Inverno 1995), where a motivation is defined to be any desire or preference that can lead to the generation and adoption of goals. In their formalization therefore, they consider goals to be a derivative of motivations, which are themselves non-derivational. Luck and d'Inverno's definition also does not concern itself with where the agent's motivations originate; they imply internally, but this is not required.

Both the definitions by Covrigaru and Lindsay and by Luck and d'Inverno permit an agent's motivations or goals to be completely dominated, even manipulated, by an external force and such an agent could still be considered autonomous. We find this antithetical to the notion of autonomy, and by the fact that neither work considers the process by which the agents internalize the possibly external goals, nor that they consider how much an agent can be manipulated or otherwise "corrupted" from pursuing its own reasoning as *it* deems best, we find their definitions seriously lacking.

Castelfranchi [6] defines autonomy as the amount of separation between external influences and an agent's goals. Castelfranchi recommends a "double filter" upon goal autonomy. First, he requires that an agent perform some form of reasoning about externally derived information before internalizing it. Second, he requires external influences must be filtered through beliefs before an agent's goals can be modified. This definition of belief and goal autonomy is quite compatible with our own, more general

notion of autonomy. We believe, however, that it is too narrowly defined and needs to be reinterpreted in a broader context. For example, their particular definition cannot be applied to agents or software that has no concept of beliefs and goals while their general ideas (extended here) certainly can.

Huhns and Singh define a set of autonomy measures that are generally compatible with our own but capture autonomy at a more abstract level. In their recent summary of the field of intelligent agents (Huhns and Singh 1998), they distinguish between Absolute autonomy, Social Autonomy, Interface Autonomy, Execution Autonomy, and Design Autonomy. Absolute Autonomy describes an extreme agent that completely ignores other agents and does whatever it wants to do. Absolute Autonomy is an extreme of their own Social Autonomy, which touches upon the notion of level of isolation from other agents. Design Autonomy relates to how much freedom an agent's designer has in terms of designing the agent and therefore is not relevant here. Interface Autonomy is also a design-related feature and is also not relevant to this discussion. Their Execution Autonomy measure captures the notion of how much freedom an agent has in making decisions while executing its plans. We do not directly address this autonomy measure but we will see later that it falls out naturally as a consequence of our definition of autonomy.

Continuing with our examples, we will try to illustrate some of the key principle in our conceptualization of autonomy that we are presenting in this paper. Imagine first an agent that we know thoughtfully considers every request from another agent, perhaps to ensure that it is maximizing its own utility, and a second agent that performs only a cursory syntactic check before accepting or rejecting a goal from the other agent. What would we say about the autonomy level of these two agents relative to each other? Almost certainly, we would say that the first agent has a higher level of autonomy since it is less corruptible; it is less likely to be manipulated by other agents. The quote from Epictetus above nicely illustrates this notion. We believe that one of the most fundamental aspects of an agent's autonomy is its intrinsic capacity and capability to insulate itself from external influences, its relative inviolability to unacknowledged manipulation and corruption [Huber, 1999].

We recognize that the conceptualization of autonomy as the relative level of an agent's incorruptibility does not by itself capture all of the aspects of autonomy. Such a scheme has a flavor of being a measure of how secure an agent is from being controlled or manipulated by external, perhaps malevolent influences, but this does not completely cover all of how an agent might be influenced. In a multiagent environment, for example, most agents will not have all of the capabilities or resources it requires to accomplish its tasks and will have to rely upon other agents. So, while agents may have a very high autonomy level according to our theory of relative incorruptibility, they may have a dependence upon other agents and resources that certainly has *some* aspect of limiting the agent's autonomy; even though nothing is manipulating the agent's goals, beliefs, intentions (etc.) directly, they are still not completely in control of their fate. For example, an agent may contract another agent to perform a subtask and then have to wait until the other agent completes the subtask before going on about its own business. In this case, the contracting agent becomes dependent upon the subcontracting agent, signifying to us a loss of autonomy. This reduction in independence impacts the contracting agent in

a much less direct manner than direct manipulation of beliefs or goals but it may in fact have the exact same effect: the contracting agent's adoption of goals, selection and execution of behaviors, even inferences and adoption of beliefs, may be constrained or altered as a result.

We recognize that no single representation of autonomy will be completely correct for all purposes but, at the same time, we feel that there needs to be a broader, more encompassing, multidimensional characterization of autonomy that will result in wide and flexible applicability. We therefore present a theory of autonomy that unifies the social aspect of autonomy with the aforementioned security aspect introduced earlier.

As a side note, we feel as many do that autonomy is a significant feature of anything calling itself an intelligent agent. However, we allow for a weak definition of agency in this paper as the definition of autonomy presented within this paper is independent of any particular model of agency and is even applicable to non-agent software constructs.

2 Representing Autonomy

It is not the greatness of a man's means that makes him independent,
so much as the smallness of his wants.

William Cobbett

How might we characterize autonomy within intelligent agents and software constructs? It might seem reasonable that we could look at a single agent to ascertain values for attributes that characterize its level of autonomy (Covrigaru and Lindsay 1991). However, we believe that this is a critical misconception. Consider an agent that exhibits goal-directed behavior, a feature often claimed as being key to agency and by Covrigaru and Lindsay as indicative of autonomy. Is the agent autonomous? Now, let us say that all of that agent's goals are dictated by another agent. Is the agent autonomous now? We think that the response to the first question is typically a "maybe" or a hesitant "yes", and that the reaction to the second question is typically a firm "no". It seems that the agent in the second case, where it is known that the goals are controlled by another agent, is clearly not autonomous, while the autonomy level of the agent in the first case, where the source of the agent's goals is unknown, is highly uncertain but in some vague sense felt to be higher. What is the difference? It is not until we introduce the relationship of the agent's goals to another agent that we seem to be able to reach a firm decision. We believe that the one crucial aspect of autonomy that all definitions of autonomy need to include is that autonomy is always in terms of a relationship to external influences, in our case agents and humans.

Now let us now assume that we have a definition of autonomy, even a very simple definition. What can we do with it? What good is it? The primary use is that now we can enhance our agents with an explicit representation of this knowledge, giving them the ability to model and reason about the autonomy level of other agents and themselves. So enhanced, our agents can reason about the most appropriate autonomy level to exhibit

with respect to other agents in as finely grained a manner as it wants to. An agent that is part of a team of agents might have a particular autonomy level with respect to the team in general, different values with respect to individual teammates, and yet other values with respect to agents outside its team. Agents with a human master should almost certainly have an autonomy relationship with its master that is distinct from what it has with other agents.

Nothing that we are saying in this paper should come across as a claim that a high autonomy level is necessarily better than a low autonomy level. The significance and proper interpretation of an agent's level of autonomy will be domain and situation specific. For example, in some situations, a low autonomy agent that accepts all work might be considered "dependable" (a good thing) or it might be abused by an agent that dumps all its own work to it (a bad thing). An agent that has a high level of autonomy might be able to get a lot of work done without interference or guidance from a user (a good thing) but may be more difficult to regain control of when it becomes necessary (a bad thing).

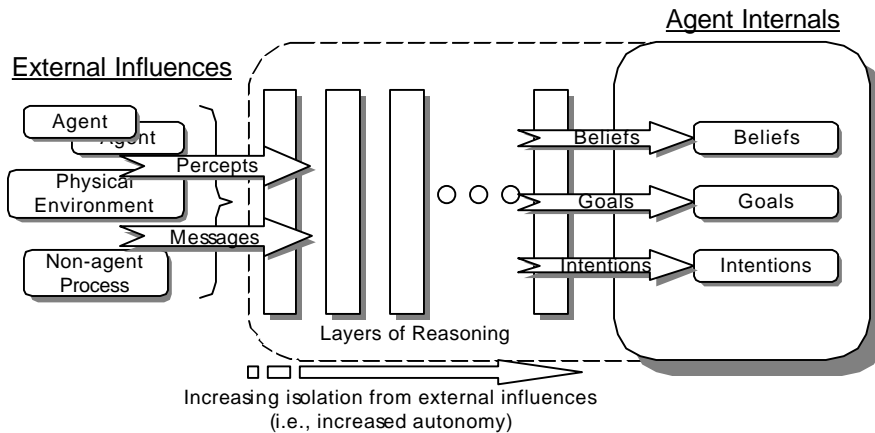


Figure 1. An agent's level of autonomy may be characterized by how much isolation, in terms of representational and reasoning layers, the internal constructs of an agent have from external influences. Illustrated is a BDI agent.

2.1 Autonomy as a Measure of an Agent's Social Integrity

More and more agents have been fielded both within research as well as industry and have naturally more frequently encountered other agents. Many are also striving to give more capabilities and responsibilities to agents. However, as we place more agents out in the computational world and give them more opportunities to operate under their own cognition, we grow more concerned that they are going to be all right. We become concerned about how secure our agents are from being manipulated or corrupted by malevolent humans and other agents and we become concerned about our ability to retain and/or regain control over the agents. Basically, we want our agents to be to some degree

autonomous while at the same time we want to be confident that our agents' integrity has not been compromised.

Within this paper, we define autonomy in part to be how much mastery an agent has over itself; i.e. how difficult it is for an agent to be corrupted or manipulated by outside influences without it being cognizant of it. We consider this aspect of autonomy its *social integrity*. Our concrete reification of this perspective on autonomy is that an agent's reasoning provides the agent its integrity, where each layer or step of reasoning represents an opportunity for the agent to examine the incoming information and to accept or reject internalizing the information (see Figure 1). These intervening reasoning layers typically (although not necessarily) transform information from one semantic form into another as it moves from external influences toward the agent's internal structures, becoming less and less "contaminated" by outside influences at each transformation. As such, each layer provides a certain amount of security in integrity by imposing a certain amount of "distance" between external influences and the agent. Of course, one layer of reasoning will tend to pose relatively more or less of a barrier to external influences and therefore be "thicker" or "thinner" than other layers. We can take this relative strength into account by weighting each layer appropriately.

Defining autonomy with respect to each internal construct seems reasonable to us. With this granularity of representation, we could distinguish between all of the following:

1. agents whose beliefs are easily manipulated by external influences (e.g., an agent that "believes everything that it reads" (via perception) or "believes everything that it's told" (via communication with other agents)) from those that are less "gullible" because they perform more reasoning to verify that the information is correct, accurate, and compatible with its own beliefs.
2. agents that blindly accept goals from other agents from those that thoroughly consider the goals to make sure they are not contrary with the agent's own goals.
3. agents with a high level of social integrity with respect to one internal structure (e.g., critical agents) while at the same time have a low level of social integrity with respect to another structure (e.g., gullible agents).

We are interested in defining autonomy with respect to all of an agent's key internal control and representational structures, so we greatly extend and revise Castelfranchi's idea of levels of isolation from other agents. Because we are interested in supporting the representation of an explicit measure of autonomy with respect to an agent's beliefs, we are forced to remove Castelfranchi's requirement that the agent must filter everything through its beliefs[6]. An example of our social integrity scheme applied to a BDI architecture is shown in Figure 1, where we illustrate autonomy measures with respect to beliefs, goals, intentions (some BDI implementations may also want to include plans/capabilities).

To calculate the autonomy level relative to an internal structure, we compute the weighted sum of the procedural layers' separation between external influences and the for all of the possible *influence paths* for the internal structure in question. There may be multiple paths through the agent's architecture for each internal structure. For example, beliefs may eventually be modified through perception interpretation routines that convert raw sensor information into beliefs and new beliefs

may also arise as part of an agent's inferencing capabilities. A value must be computed for each of these influence paths.

We then take the minimum of these influence path values. Using the minimum function results in a conservative measure as it provides a value that represents the *shortest influence path* into a particular internal structure of an agent. The shortest influence path for an internal structure is the measure of social integrity for that particular structure.

To compute the agent's overall autonomy value, we consider the minimum of all of the structural integrity values computed. Again, this is the most conservative interpretation; the agent is only as secure as its weakest point and the structure with the lowest integrity measure represents the most easily manipulable or corruptible structure.

Defined more formally (an illustration of the contributing factors involved below are shown in Figure 2):

- η is the number of internal agent structures
- $s(n)$ is the n^{th} internal structure, $1 \leq n \leq \eta$
- $u(s(n))$ is the number of influences paths for structure $s(n)$
- then $\sum_{n=1}^{\eta} u(n)$ is the total number of influence paths for the agent
- $i(m, s(n))$ is the m^{th} influence path for structure $s(n)$, $0 \leq m \leq u(s(n))$
- $\lambda(i(m, s(n)))$ is the number of reasoning layers for influence path $i(m, s(n))$
- $l(r, i(m, s(n)))$ is the r^{th} reasoning layer for influence path $i(m, s(n))$, $0 \leq r \leq \lambda(i(m, s(n)))$
- $\rho(l(r, i(m, s(n))))$ is the coefficient of weighting for reasoning layer $l(r, i(m, s(n)))$

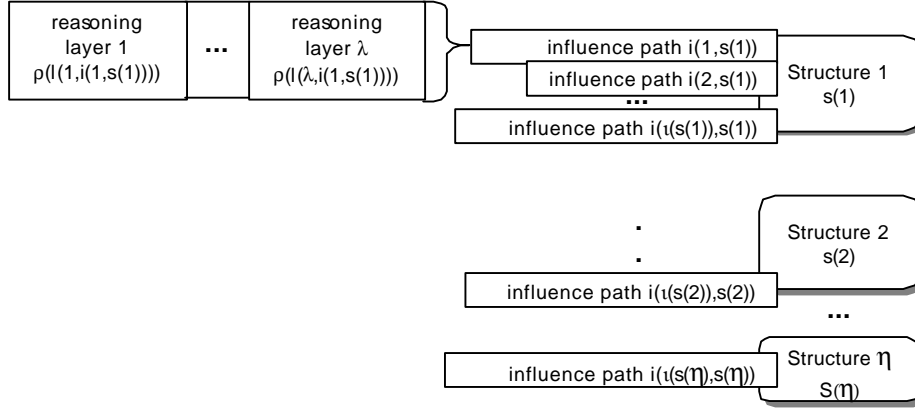


Figure 2. Illustration of factors involved in computing an agent's social integrity measure of autonomy.

Definition 1. Influence Path Measure of Autonomy

The autonomy measure for influence path m and internal structure n is calculated as the sum of the coefficients of weighting of the layers of reasoning.

$$\alpha_{\text{IPM}: m, n} = \sum_{r=1}^{\lambda(i(m, s(n)))} \rho(l(r, i(m, s(n))))$$

Definition 2. Structure Measure of Autonomy

The autonomy measure of internal structure n is the minimum of all of the possible influences paths for that structure:

$$\alpha_{SM: n} = \min(\alpha_{IPM: 1, n}, \alpha_{IPM: 2, n}, \dots, \alpha_{IPM: U(s(n)), n})$$

In other words, $\alpha_{SM: n}$ represents the value of the *shortest influence path* for structure n .

Definition 3. Social Integrity of Autonomy

The autonomy measure of an agent's social integrity is the minimum of all of the agent's structural autonomy values:

$$\alpha_{INTEGRITY} = \min(\alpha_{SM: 1}, \alpha_{SM: 2}, \dots, \alpha_{SM: \eta})$$

In other words, α_{SI} is the overall measure of autonomy, with respect to social integrity, for the agent and is a measure of the agent's softest spot, so to speak.

With this formulation, a value of zero would indicate that one or more external influences have a direct path into the agent's internal structure(s) and therefore possibly have complete control over the agent. A value of ∞ would indicate that external influences have no possible way of influencing the agent's internal structures and would probably indicate a completely sociopathic agent (equivalent to Huhns and Singh's Absolute autonomy [Huhns and Singh 1998]).

Computing a single value for the agent's autonomy as we have done permits a simple comparison between two possibly heterogeneous agents' relative vulnerability. A corporation or institution might then be able to specify a policy stating a minimum value of social integrity that all agents representing it might have to exhibit before being allowed to be fielded with other, non-institutional agents.

A high value for one agent and a low value for another agent indicates how relatively easy it might be to corrupt the second agent compared to the first, but it does not specify which particular structure represents the easier target's weakest structure or influence path, both of which would be necessary to target a particular agent for exploitation. We could consider using a vector of autonomy values, one entry per agent internal structure, i.e.,

$$\alpha_{INTEGRITY} = \langle \alpha_{SM: 1}, \dots, \alpha_{SM: \eta} \rangle$$

that would more comprehensively describe the agent's autonomy and identify the relative strengths of the various structures. However, heterogeneous agents will have potentially very different internal architectures and therefore very different key internal structures, making comparison of agents difficult or impossible with a vector scheme.

This theoretical and pragmatic interpretation of autonomy is applicable to intelligent agent architectures (such as BDI architectures) as well as non-agent software constructs. In fact, the general principles are applicable to all programs, not just those purportedly embodying mentalistic states as do intelligent agents. To apply the ideas presented here, an agent programmer must identify those internal structures that are critical to the decision making and execution behavior of the agent. Once this has been done, the agent programmer must then determine the number of levels of reasoning intervening between inputs to the agent and those critical internal structures. Agent architectures typically make this process simple because the critical representational structures are explicit and the influence paths to these structures are therefore more easily identifiable than non-agent frameworks.

In order for an agent to alter its level of autonomy within our scheme, the agent would need to add or remove levels of reasoning between its internals and external influences, or alter the level of checking that a particular level performs. As depicted in Figure 1, more or less autonomy could be realized by adding or removing layers or reasoning or by increasing or decreasing the amount of filtering and checking existing layers perform.

2.2 Autonomy as a Measure of Social Dependence

Dependencies arise between agents as a matter of necessity to accomplish complex tasks that are beyond the capabilities of a single agent, that require simultaneous activity, or that perhaps require resources not controlled by the agent. For example, an agent may become dependent upon another agent to accomplish part of one of its own tasks. In some cases, an agent need another agent to accomplish a task required as a prerequisite to accomplishment of one of its own tasks. An agent in situations such as these may have a high measure of autonomy with respect to social integrity yet still be highly constrained because of these dependencies on other agents. In other words, other agents cannot directly influence the agent's internal state (beliefs, intentions, etc.) but they still have some form of control over the agent. In the worst case, the dependent agent may be unable to do any work on its own until another agent completes its work or releases a resource. This lack of control over what and when an agent can do something implies to us a reduction in some aspect of its autonomy. Because this aspect of lack of control is due to dependencies upon other agents, we call this the *social dependency* measure of autonomy.

Social dependencies arise for a number of reasons, including: authority and organizational structures; commitments and obligations; capability and resource limitations that induce task dependencies (e.g., decomposition, necessary sequencing); and degree (or lack) of faith in competence of others agents. The dependency relationship holding between the agents can be labeled according to the type of the dependency, including (but not limited to) boss/worker (authority structure), source/sink (dataflow structure), teammate/teammate (team formed either as part of organization or voluntarily through negotiation), and contractor/contractee (negotiated task or resource dependency). These dependency relationships can be categorized by whether the dependency comes from *superior* agents (e.g., those higher in the authority structure or beforehand in a dataflow structure), *peer* agents (e.g., teammates or those on otherwise equal terms), or *inferior* agents (e.g., those lower in the authority structure, or a subcontractor, or afterward in a dataflow).

In other words, agents assume dependencies from all sorts of sources:

- Agents lose dependence when there exists an authority relationship over the agent. Also, agents that must rely upon agents "upstream" of it to provide it information are less independent.
- Agents lose independence with respect to agents that are otherwise peers with themselves when they enter into an agreement to either perform work for a peer agent (i.e., their goal is not their own) or when a peer agent is performing a task for it and it depends upon that agent to be done.
- Agents lose dependence when they offload tasks to inferiors upon which they depend for their own continued behavior.

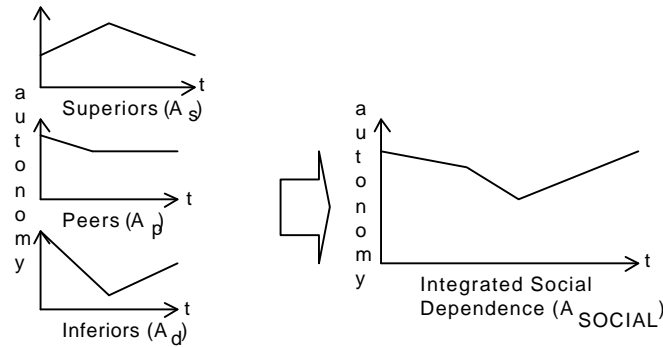


Figure 3. Social dependency autonomy measure.

The various social dependency values will likely change over time. This dynamicism will arise, for example, when agents acquire tasks on behalf of other agents which it then subsequently completes, requires resources that it cannot immediately obtain and then at a later time gains access to them, delegates tasks to other agents which are subsequently completed, etc. We illustrate this in Figure 3, above. Our formalization of social dependency autonomy captures the notion that autonomy drops as social dependencies increase and that autonomy rises as social dependencies decrease.

Slightly more formally, let:

α_s = % of tasks imposed upon the agent from a superior agent

α_{pa} = % of tasks accepted from peers

α_{pd} = % of tasks contracted (and dependent upon completion) to peers

α_i = % of tasks imposed upon inferior agents (and dependent upon completion)

then

$\alpha_{SOCIAL} = \alpha_s + \alpha_{pa} + \alpha_{pd} + \alpha_i$ = % of tasks dependent upon others (relative to all agent tasks)

Note that the summation function computing α_{SOCIAL} is exclusive in that a dependent task should not be considered in more than a single factor (i.e., α_{SOCIAL} ranges over 0.0 ... 1.0).

Given the above characterization of social (in)dependency, we end up with a single value with which to handle and compare with other agents' similar attribute. In contrast to the social integrity measure, which may involve very different constituent parts in its final computation, the social dependency measure will be similar between agents would be comparable at both the summary level (α_{SOCIAL} above) as well at finer levels (perhaps at the level of the list of $\langle \alpha_s, \alpha_{pa}, \alpha_{pd}, \alpha_i \rangle$). The characterization of autonomy above might be associated with particular tasks or goals as do Barber and Martin [1][2][3] to provide a finer level of detail if such is deemed necessary. While complementary in some ways as supporting computation of our α_{SOCIAL} measure, Barber and Martin's work however make claims such as "an agent's autonomy is not a function of ... dependence on other agents", which we find to be antithetical to the notion of autonomy and therefore is not completely congruous with our own theories.

We believe that our model of social (in)dependence as a factor in autonomy captures the intuition held by many current models of autonomy that relate the simple intuition of

how many of an agent's activities it can perform without aid relative to the total number of activities the agent must perform to achieve all of its tasks. Of course, since these models of autonomy say nothing about how manipulable the agent is by other agents, we believe their work to cover only a small portion of a complete model of an agent's autonomy.

2.3 Integrating the autonomy measures

To compute an overall autonomy value for an agent, we combine the social integrity autonomy value and the social dependency autonomy value. A simple yet effective scheme would be a simple weighted sum of the two measures, where the two weighting coefficients represent the relative importance of the two autonomy measures (integrity and social dependence) to the agent's overall autonomy measure (in its own mind or in its

Slightly more formally, let, $\pi_{\text{INTEGRITY}}$ and π_{SOCIAL} be the weighting coefficient for social integrity and social dependence, respectively, then the agent's overall autonomy value α would be:

$$\alpha = \pi_{\text{INTEGRITY}} * \alpha_{\text{INTEGRITY}} + \pi_{\text{SOCIAL}} * \alpha_{\text{SOCIAL}}$$

We leave how these coefficients are determined unspecified within this paper. However, we see them being set, for example, by agent developers at agent design or invocation time to suit the domain or perhaps determined dynamically by the agents themselves as they operate. This is shown in Figure 4, where we assume a relatively static integrity measure and a dynamic social dependency measure.

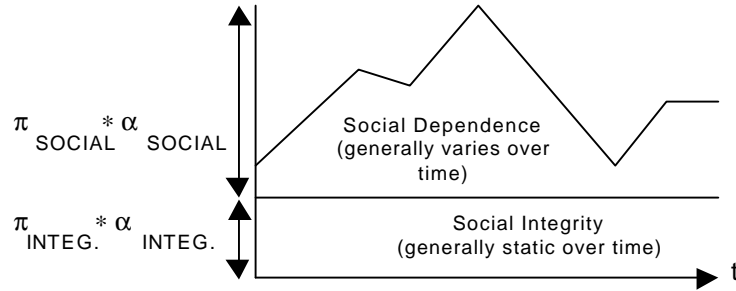


Figure 4. Overall autonomy level calculation over time.

3 Summary

Absolute liberty is absence of restraint; responsibility is restraint;
therefore, the ideally free individual is responsible to himself.

Henry Brooks Adams

According to our definitions above, autonomy is in part an integrity relationship between external influences and an agent's significant internal structures and in part a social dependency relationship that an agent has with other agents related to its tasks. To summarize, to determine the social integrity autonomy value for any agent or software construct requires looking at the design of the particular agent or piece of software and calculating how secure the agent is from manipulation and corruption by other agents and software constructs. For agent architectures, the designs tend to closely match their theoretical underpinnings and have significant structures and algorithms associated with the theories' significant constructs. For example, BDI (Belief-Desire-Intention) theoretic architectures [4][5] have highly formalized internal structures that would naturally have autonomy values computed for each of these particular constructs. In prior work [8], a BDI architecture called JAM [9] was analyzed for its integrity autonomy in terms of capabilities (i.e., a plan library) as well as the "standard" beliefs, goals (desires), and intentions. An analysis of the Soar agent architectures [11], an example of another concrete agent architecture, would most likely have an autonomy values computed for the relative inviolability of its important structures of workspaces and capabilities (in addition to goals, beliefs, and perhaps other structures as well). Our autonomy characterization will work for software constructs as well, but given non-agent software's likelihood of have more implicitly defined structures and reasoning mechanisms, it may be more difficult to characterize.

To determine the social dependency autonomy value for any agent or software construct requires looking at the dynamic runtime relationships that the particular agent or piece of software has with other agents/software. To do this requires identifying and quantifying the agent's relative dependence upon other agents with respect to accomplishing its own tasks and, ostensibly, the freedom or lack therefore that it has relative to those dependencies.

To determine an agent's overall autonomy value then, we provide a simple weighted sum of the social integrity and social dependency values, with weighting based upon domain and/or agent-specified priorities.

We believe that the characterization of autonomy presented within this paper has widespread applicability, both with respect to number of agent and software frameworks that may be so characterized and to the wide range of domains in which such a characterization should prove beneficial.

4 References

[1] Barber, K.S., and Martin, C.E., Specification, Measurement, and Adjustment of Agent Autonomy: Theory and Implementation, University of Texas Technical Report TR99-UT-LIPS-AGENTS-04, 1999.

- [2] Barber, K.S., and Martin, C.E., Autonomy as Decision-Making Control, University of Texas Technical Report TR00-UT-LIPS-AGENTS-08, 2000.
- [3] Barber, K.S., A. Goel, and Martin, C.E., Dynamic Adaptive Autonomy in Multi-Agent Systems, *Journal of Experimental and Theoretical Artificial Intelligence*. Also available as University of Texas Technical Report TR99-UT-LIPS-AGENTS-05, 1999.
- [4] Bratman, M. *Intentions, Plans, and Practical Reason*. Cambridge, Mass.: Harvard University Press. 1987.
- [5] Bratman, M., Israel, D., and Pollack, M. Plans and Resource-bounded Practical Reasoning. *Computational Intelligence* 4:349-355. 1988.
- [6] Castelfranchi, C. Guarantees for Autonomy in Cognitive Agent Architecture. In *Intelligent Agents – Theories, Architectures, and Languages*, 56-70, Michael Wooldridge and Nicholas Jennings editors, Springer-Verlag. 1995.
- [7] Huhns, M., and Singh, M. Agents and Multiagent Systems: Themes, Approaches, and Challenges. *Readings in Agents*, 1-23. Michael Huhns and Munindar Singh eds., San Francisco, California: Morgan Kaufmann. 1998.
- [8] Huber, M.J. Considerations for Flexible Autonomy Within BDI Intelligent Agent Architectures, Working Notes of the AAAI Spring Symposium on Agents with Adjustable Autonomy, Stanford, CA, pgs 65-72. 1999
- [9] Huber, M. J. JAM: A BDI-theoretic Mobile Agent Architecture. *Proceedings of the Third International Conference on Autonomous Agents*. 1999.
- [10] Konolige, K., and Pollack, M. A Representationalist Theory of Intention. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*. Chambery, France. 1993.
- [11] Laird, John E. and Allen Newell and Paul S. Rosenbloom, "SOAR: An Architecture for General Intelligence", *AI Journal*, 1-64, 1987.
- [12] H. J. Levesque, P. R. Cohen, and J. Nunes. On Acting Together. In *Proceedings of the National Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, Inc., California, 1990.
- [13] Michael Luck and Mark d'Inverno. A Formal Framework for Agency and Autonomy. In *Proceedings of the First International Conference on Multi-Agent Systems*, 254-268. MIT Press. 1995.
- [14] Pell B., Bernard D.E., Steven A. Chien, Erann Gat, Nicola Muscettola, P. Pandurang Nayak, Michael D. Wagner, and Brian C. Williams. An Autonomous Spacecraft Agent Prototype. In *Proceedings of the First International Conference on Autonomous Agents*, Marina del Rey, CA 1997.