

Types and Limits of Agent Autonomy

Gordon Beavers and Henry Hexmoor

Computer Science & Computer Engineering Department
Engineering Hall, Room 313, Fayetteville, AR 72701
{gordonb, hexmoor}@uark.edu

Abstract

This paper contends that to fully understand interactions between agents, one must understand the dependence and autonomy relations that hold between the interacting agents. Individual variations, interpersonal dependencies, and environmental factors are determinants of autonomy that will be discussed in this paper. The paper concludes with a discussion of situations when autonomy is harmful.

1. Introduction

Autonomy has been a primary concept for characterizing agents. This paper calls attention to several facets of understanding autonomy from the point of view of an agent. We are primarily interested in enabling agents to explicitly reason about autonomy while interacting with other agents. Autonomy is conceived as an internal and integral component of reasoning about interactions. This paper is not concerned with negotiation about autonomy among agents who may share a similar conception of autonomy, but rather we seek to expose the origins of this concept within a given individual as a result of the relationships in which that individual participates. This paper will investigate various ways in which autonomy can be interpreted and implemented in artificial agents, the utility of these implementations, and how these interpretations are linked to usage in a human context.

The simplest notion of autonomy is that of local determination. An agent that determines its actions for itself based only on its internal state is generally considered autonomous, that is, if the determination of the agent's behavior is local and without input from other agents, the agent is autonomous. Thus, a reactive agent running on a deterministic program would be considered autonomous, but since such an agent has no intentions, and is incapable of introspection, autonomy seems to be a concept of limited usefulness in the context of strictly reactive agents. In contrast, other investigators in the area of autonomy claim that the concept of autonomy is appropriate only for intentional agents that have introspective abilities. An intermediate position is that autonomy is essentially a social notion, which can be understood in terms of social dependencies. The continuum of positions with respect to autonomy will be surveyed in section 2.

One aspect of autonomy we will explore is individual variation among agents such as agent personality [6]. Autonomy is partly a psychological characteristic, and thus varies across agents according to personality. There are individuals who lead and others who prefer to follow. Another aspect of autonomy to be considered is interpersonal dependencies such as dependencies in the physical world and interpersonal social dependencies. Earlier papers, e.g. [2] have claimed that autonomy can be fully understood in terms of social independence. While this notion of autonomy is satisfactory for simple agents, it becomes inadequate as agents become more richly endowed with psychological characteristics, in particular, when agents can be considered to have a personality. A third area of exploration is social regulation and an agent's attitude toward laws, norms, and values in the agent's environment such as found in agent organizations, institutions, and general agent society.

2. Interpretations of Autonomy

It is common to refer to agents that determine their behavior without the influence of other agents as autonomous. This usage reflects the notion that an autonomous agent is independent and self-governing. These, however, are vague notions. What is generally meant is that an autonomous agent is empowered to choose to act contrary to the desires of other agents. Consider two simple agents each of which must make a choice between two options A and B. The first makes the decision based on input from a sensor monitoring some physical condition of the agent's environment. The second makes the decision based on input from another agent. The first agent acts independently of any other agent, but the second agent's action is dependent on the action of another agent. The first is autonomous with respect to the decision, but the second is not autonomous according to common notions. Is there any meaningful distinction between these agents? Both are reacting to the condition of an input. Neither has any knowledge of other agents, social constructs, or social structures. Is autonomy a fruitful concept in considering such simple agents? Probably not. We, therefore, confine our attention to intentional agents, and note the positive correlation of usefulness of the concept of autonomy with social abilities of the agent.

An intentional agent can exhibit various degrees of autonomy as determined by how much influence local and non-local there is in its decision making process. Local limitations on autonomy include constraints on intentional and physical abilities, while non-local limitations can be exerted through social values, e.g., avoid damage to the concerns of other agents, norms, e.g. keep to the right, legal restrictions, e.g., do not exceed the speed limit, and the need to cooperate.

An advocate of a more sophisticated position with respect to agent autonomy might claim that autonomy is a social notion, and as such, can only be a property of agents endowed with social properties interacting with other such agents in a social situation. Within this camp, there can be many variations with one of the simplest being that of [2]. Castelfranchi asserts that "all forms of social autonomy

should be defined in terms of different forms of social independence” and further that “each and any (sic) component of the architecture or necessary condition for a successful action can define a dimension/parameter of autonomy, since it can define an abstract ‘resource’ or ‘power’ necessary for the (sic) goal achievement, i.e. it can characterize a specific ‘lack of power’ and than (sic) a possible dependence and social non-autonomy.” The question of what agent characteristics are required for an agent to be considered social must be addressed. Must an agent be intentional to be social? The Castelfranchi quote above implies that the mere existence of a dependence relation in a social setting is enough to establish an autonomy value. Is more required? Must an agent be intentional? Must an agent be capable of introspection and self-evaluation to be considered social?

Castelfranchi implies that dependency and autonomy are inverses of each other, but it is possible that an agent could be dependent on another agent and yet still remain fully autonomous, if the agent has the capability of ending the dependency relation at will. This suggests a distinction between first and second order autonomy. An agent who is dependent on another agent, but who retains control over that relation has given up first order autonomy, but retained second order autonomy. Dependencies likewise can be either first order or second order. When agent A depends on input from agent B to establish A’s dependency relations, A has a second order dependency on B.

A stronger position on autonomy requires that an agent not only be social, but also be capable of reliably assessing its own capabilities. The claim is that an agent is not fully autonomous unless it has justification for believing that it is autonomous. Capability must be part of autonomy in the sense that an agent is autonomous with respect to a particular task or action only if that agent has the ability to complete the task or action. The greater the agent’s warranted belief in its own capability the greater the agent’s autonomy with respect to the relevant tasks.

In some environments, agent intentions and decisions are highly constrained by the environment. For example, an agent driving on an icy road has fewer options than an agent driving on a clear road. These constraints limit the agent’s autonomy in the sense that a prudent agent will refrain from activities that have a sufficiently high probability of resulting in undesirable consequences. On the other hand, if the environment is safe for the agent, then choices are unconstrained, and autonomy is expanded.

We now provide a critique of two claims found in [2]. First, Castelfranchi makes the assertion that “all forms of social autonomy should be defined in terms of different forms of social independence.” While it is the case that dependency tends to diminish autonomy, lack of dependency does not necessarily yield autonomy, and further the effect of a dependency on an individual agent depends upon that particular agent’s personality as discussed in the next section. The effect of a dependency is not uniform across all agents. A richer understanding of autonomy requires the consideration of agent mental properties that are not under the control of social relationships. Individual assertiveness and self-confidence are examples of such properties. That is, we are claiming that personality affects subjective

autonomy estimates, which are an aspect of social autonomy, but subjective autonomy estimates are not definable in terms of social independence.

Second, Castelfranchi also claims that every component of the agent's architecture and every condition necessary for a successful action define a dimension of autonomy. Castelfranchi's linking of autonomy to resources and powers necessary for goal achievement would seem to make autonomy assessment objective, since, in any given situation it is an objective matter of fact which resources and powers are necessary for goal achievement. We argue that consideration of objective resources and powers is inadequate to capture a full understanding of autonomy, that autonomy has an essential subjective aspect that is overlooked by objective measures. In addition, autonomy can be affected by non-necessary, contingent conditions. Castelfranchi's concept of autonomy is objectively determined and lacks aspects of intentionality and personality.

Castelfranchi is correct to assert that autonomy is a relational notion between agents, actions, and goals. However, he incorrectly assumes that autonomy is completely determined by these relationships and the agent's "powers". Castelfranchi also correctly asserts that autonomy considerations must extend to the agent's relationship to its environment, since the ability to perceive and react in a reasonable manner to the environment is necessary for reliable goal achievement.

With respect to the first claim, Castelfranchi provides the following unproblematic assertion : (1) "If an agent Y depends on an agent X for its internal or external power/resource p relative to its goal G then Y is not autonomous from X relative to its goal G and resource p." It is, however, an unjustified leap from this assertion to (2) "all forms of social autonomy should be defined in terms of different forms of social independence." (1) is an implication, while (2) is essentially a biconditional. The converse of (1) is not true, that is, there is an agent who is not autonomous, to some degree, with respect to some goal because of its internal mental state, e.g., diffidence or self-doubt, but who is not dependent on any other agent or the environment. Learning and personality characteristics such as self-confidence and aggressiveness increase independence. Castelfranchi claims that autonomy is independence, nothing more. We need to distinguish "being autonomous" from "acting autonomous". Both internal and external conditions affect autonomy.

With respect to the second claim, here is an example where a contingent matter prevents goal achievement. An agent may have adequate resources and power to help a friend agent but may reason about lack of recent reciprocity from the friend and refrain from helping. This is an example where contingency of balance of reciprocity between two agents contributes to determining autonomy. Although objectively resources and power were present for goal achievement, a subjective matter entered in autonomy decision-making. Dependence has both local and global aspects.

Consider how the gravity of a decision affects autonomy, that is, how consequences interact with autonomy. A casual treatment would say that autonomy is independent of the consideration of consequences. If an agent drives too fast on icy roads, unfortunate consequences are likely to ensue, but the agent's autonomy to select a speed is not affected by the possibility of disaster. A prudent agent will

exercise its autonomy by choosing a lower speed, but its autonomy in the choice of speed is not affected by the possible consequences. This analysis is superficial. The consideration of consequences has affected the choices open to the agent, and thus has a second order effect on the agent's first order autonomy. In other momentous decisions, the connection between autonomy and consequences is harder to place. Consider an agent who is disarming a bomb by removing a timer. Cut the wrong wire and the bomb detonates. The gravity of the situation affects the manner in which the agent exercises its autonomy, namely very carefully, that is, it has a second order effect on the agent's autonomy but does not affect the agent's first order autonomy with respect to disarming the bomb. The gravity has a second order effect on autonomy in this case, since the gravity affects how the agent acts, not on whether the agent acts. Autonomy is not just a matter of whether an agent chooses to act, but also how it chooses to act.

4. Individual Variations

Individuals vary in their conception of autonomy. These variations are due, in part, to differences in personality [6]. In the case of agents, we take personality to be a collection of persistent patterns of behavior such as trusting and agreeableness. The agent community has begun developing agents with synthetic personality, see [5] and [8]. deRosier and Castelfranchi mention laziness, helpfulness, dominance, and conflict-checking as agent personality traits [3]. Obviously, personality determines the individual's conception of autonomy. For a rough understanding of individual variations of autonomy we suggest two dimensions. The first dimension we term *rigidity*. Rigidity constrains an individual's interactions by demanding (1) certainty, (2) independence, (3) norms and values, (4) boundary precision, and (5) control. Highly rigid individuals demand precision and are less flexible. Disorders of perfectionism and obsessive compulsiveness commonly accompany high rigidity. One aspect of rigidity is the need for certainty of effects of actions (by self and others) and certainty of information available to them. With higher certainty, individuals are more confident in their actions and rely on their choices. When there is a lot of uncertainty, an individual agent may not feel in control of its environment and may feel less confident about its decisions. Another aspect of rigidity is the need for independence. A highly dependent personality requires the company of others. A dependent agent may require constant direction and reinforcement. In the opposite extreme, an independent individual will avoid others and be fiercely boundary conscious and not be a team player. A third aspect of rigidity is mindless adherence to laws, conventions, norms, and values. A legalistic individual will be unyielding in following rules. On the other hand, an individual may seek to avoid rules and regulations. The fourth aspect of rigidity is boundary maintenance. Individuals differ in their tolerance of limits and some tolerate flexible values while others demand precise boundaries and limits. A fifth aspect of rigidity is the need for control and dominance. This is related to the need

for a particular level of independence. Some individuals have higher need to be in control while others are content with less control.

Our second dimension we will call *capacities* which is the individual's capacity to objectively perceive an agent's attributes that constrain interaction. Individuals that do not accurately perceive uncertainties can believe that they are in deterministic environments and may conceive of their autonomy being circumscribed. Those that perceive more uncertainties are better able to form a more complex sense of autonomy. Individuals who see themselves to be highly dependent, will sense relatively lower levels of autonomy as opposed to those who are oblivious to these dependencies. Sensing the extent to which an individual's decisions are guided by norms and values can give an individual a sense of security and lack of freedom. The ability to perceive control is another aspect of an individual's capacity. Sensing higher levels of control is generally reassuring for an individual and will lead to a sense of freedom and experience of higher autonomy. The exact relationship between control and autonomy appears to be very complex. At one end of the capacities dimension are individuals with the highest capacities and in the other end there are individuals with the lowest capacities.

Individuals fall in a unique place in this two dimensional personality space shown in Figure 1. This space is useful in suggesting the agent type to construct. For a domain that requires stable functions with moderate timing coordination precision, we suggest agents with low capacity and low rigidity. Functions are stable if they are context independent. Coordination precision is moderate if coordination constraints among functions are not strict. Automation of routine tasks in a stable environment is an example of such an environment. In environments that require stable functions but with high coordination precision, we suggest agents with low capacity and high rigidity. Coordination precision is high if coordination constraints among functions are strict. Air traffic control is an example of such an environment. For a domain that requires context sensitive functions with moderate precision, we suggest agents with high capacity and low rigidity. An environment that requires negotiation and collaboration is an example of such an environment. For a domain that requires context sensitive functions with high precision, we suggest agents with high capacity and high rigidity. Unpredictable planetary surface exploration is an example of such an environment.

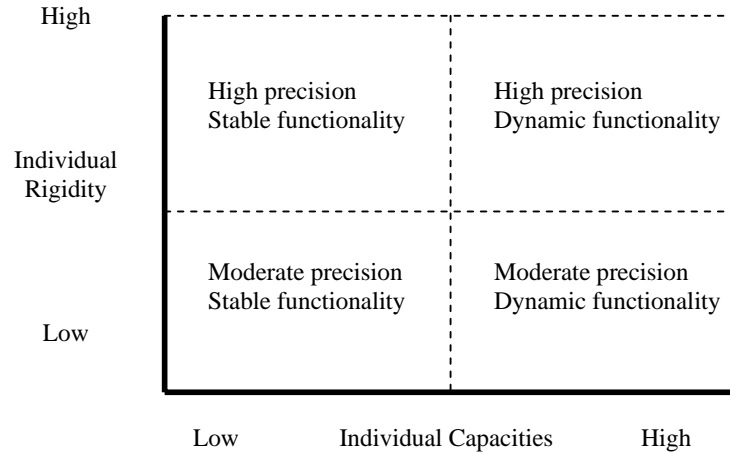


Fig. 1 Space of Individual Variations

5. Environmental Variations

Whereas freedom is implied from autonomy, freedom to act or decide does not imply autonomy. Freedom is the social deontic component of autonomy. Autonomy is also more than this exogenous sense of freedom. An individual must satisfy endogenous conditions, beyond socially extended autonomy, in order to be autonomous.

Autonomy can be interpreted as a combination of socially warranted freedoms and internally perceived freedoms. When individuals are in groups, organizations, and institutions, their individual autonomy is altered by the properties of the group. This effect may be caused by their membership, their representation of group to others, or their participation in collective behavior. In settings where social obligations are rigid and matter a great deal, autonomy tends to be how an individual relates to those social rules. In contrast, there are many environments where few or no taboos or strict codes of conduct constrain agent behavior. In such environments, individuals are free to be unaffiliated.

Individuals, as members of groups, are subject to the social climate within which they are embedded. Their memberships and allegiances constrain their decision making and provide them with a set of behaviors, rights, duties, and obligations governed by a set of norms and values. This may augment or detract from their individual autonomies. Autonomies are affected by the individual's degree of commitment and loyalty to the groups to which they belong. As we explored above in our consideration of individual variation, individuals do vary in their tendency toward commitment and loyalty. However, here we are suggesting that the existence of groups to which an individual may belong produces a set of attributes for an agent's autonomy deliberation. We point to a need to develop reasoning methods for autonomy that account for an agent's membership in groups.

Given that an individual is a member of a particular group, when it represents

the group in its interaction with non-group members, the individual assumes the group's persona and autonomy and augments it to its own. A representative must embody the essential components of the group it is representing and use it in reasoning about the group's autonomy.

In collective action or collective decision-making, the group as a whole owns the action or decision. An individual's autonomy toward the collective action or decision can only be conceived of as a contribution. This can be in terms of voting or vetoing power in the case of decision-making. In the case of physical or social action, an agent's autonomy is represented by the extent to which the individual facilitates the group's intent.

6. Limits of Autonomy: When is it too much?

Agents that interact with humans must be designed with particular attention to human safety and to not endanger human goals as well. Problems can arise when agents are fully autonomous and their actions are un-interruptible. Both agents that are excessively passive and agents that are excessively active can be harmful. Humans in the loop of agents need to be able to adjust the activity level of agents dynamically, that is, the autonomy level of agents needs to be adjustable. This seems to suggest a multi-tiered approach to agent autonomy. Agents may operate with certain nominal autonomies under normal circumstances but under heightened safety or concern conditions, agents should be switched to other modes when authorized human users manipulate their autonomy. Changing agent roles in interaction has been reported in mixed-initiative work [1]. Another promising area is empowering agents to reason about shared norms and values [7]. Since accounting for all contingencies is not realistic, agents need to have the ability to reason for themselves about whether their human supervisor would approve of their choices and assumed autonomies.

An agent's assumed level of autonomy may have unintended effects on other agents in a multi-agent setting. When tasks among agents are coupled, cooperating agents need to use a shared notion of autonomy to take into account one another's actions. Sharing autonomy is useful for harmonizing agent autonomies in order to account for one another's influence and to avoid negative influences [4]. Collective autonomy requires mutual trust among agents. Groups of agents may develop a notion of autonomy that belongs to the entire group. Individuals will conceive of autonomy of their group. However, the group's autonomy can only be altered by collective actions such as negotiation. There can be problems when an individual in such a group misunderstands the group's autonomy when it represents the group. This can cause harm to the group or others.

6. Conclusions

We have argued that autonomy is not uniformly conceptualized, and we must also account for variations between individuals and environments. This subjectivity is shown to counter arguments by Castelfranchi and others to the effect that autonomy is determined strictly by interpersonal dependencies. This work has promise in the selection and design of complex agents that must match the requirements of their environment for autonomous behavior. We briefly highlighted situations when autonomy is harmful.

Acknowledgements

This work is supported by AFOSR grant F49620-00-1-0302.

References

1. J. Bradshaw, G. Boy, E. Durfee, M. Gruninger, H. Hexmoor, N. Suri, M. Tambe (Eds), 2002. Software Agents for the Warfighter, ITAC Consortium Report, AAAI Press/The MIT Press.
2. C. Castelfranchi, 2000. Founding Agent's Autonomy on Dependence Theory, In proceedings of ECAI'01, pp. 353-357, Berlin.
3. C. Castelfranchi and F de Rosi, 1999. Which User Model do we need to relax the sincerity assertion in HCI? UM'99 Workshop on *Attitude, Personality and Emotions in User-Adapted Interaction*.
4. H. Hexmoor, 2001. Stages of Autonomy Determination, IEEE Transactions on Man, Machine, and cybernetics- Part C (SMC-C), Vol. 31, No. 4, Pages 509-517, November 2001.
5. D. Moffat. 1997. Personality parameters and programs. I. R. Trappl and P. Petta, editors. Creating personalities for Synthetic Actors, Pages 120-165, Springer.
6. L.A. Pervin and O.P. John, (Editors) 2001. Handbook of Personality Theory and Research, Guilford publications.
7. D. Shapiro, 2001. Value-driven agents, Ph.D. thesis, Stanford University, Department of Management Science and Engineering.
8. R. Trappl, and P. Petta, (eds.) 1999. Creating Personalities for *Synthetic Actors*. Springer-Verlag, Berlin.