Chapter 1

A Prospectus on Agent Autonomy

HENRY HEXMOOR*, CRISTIANO CASTELFRANCHI** and RINO FALCONE**
* *Computer Science & Computer Engineering Department, Engineering Hall, Room 313, Fayetteville, AR 72701*

** *Institute of Cognitive Sciences and Technologies - National Research Council, Rome, Italy*

Key words: interaction, dependence, mixed initiative, adjustable autonomy

Abstract: This paper summarizes the state of art in agent autonomy. It dispels myths and builds a foundation for study of autonomy. We point to a renewed interest in good old-fashioned AI that has emerged from consideration of agents and autonomy. This paper also serves as a reader's guide to the paper in this book. We end with sobering thoughts about the future of the human relationship with machines.

1. A GROWING DOMAIN

There have been many recent meetings that have explored the issues of autonomy in agents—in its meaning, in its operationalization, and its consequences. In the early 1990s a series of workshops focused on architectures and this gradually shifted to address autonomy more directly (AAAI 1991-1995, 2000, IJCAI 1999, AA 1999). Another strong driving factor was the proliferation of space missions that more explicitly focused on the problem of interaction with an autonomous agent and the "adjustability" of this autonomy (Dorais, et al., 1998; Bradshaw, et al., 2001). This clearly is a consistent trend in AI and human machine interaction.

This volume contains several papers on the state of art in the field and illustrates maturing trends ranging from analytical and formal approaches to concrete attempts and application problems. This book also challenges and provokes us to return to issues of good old-fashioned AI, as we explain next.

2. WELCOME BACK GOOD OLD AI: TOWARDS (INTELLIGENT) AUTONOMOUS AGENTS?

Some sort of "agentification" is sweeping the entire field of AI. In fact, the agent-based approach is becoming synonymous with AI. It is not only a matter of cultural fashion or of emergence of topics as it was the case with introduction of expert systems. A prime reason for this transformation is that the agent-based approach offers solutions for advanced computational problems that are robust, distributed, flexible, and scalable. But there are also cultural and theoretical underpinnings. Agents have resuscitated all the original challenges, provocations, and ambitions in the good old-fashioned AI of Turning. Crucial issues like emotions, mind-body relation, experience and coupling with an external world, learning, etc., that were central in the discussion between Turing and philosophers about the possibility of realizing artificial intelligence, are now resuscitated. In the last 15 years, each of these areas has become an important branch of AI. Among those issues are *autonomy* and *initiative*. Initiative is pro-actively reasoning about, the course of problem solving. This might include dialogues or tasks. Autonomy is reasoning about delegation and dependence, and is closely affiliated with initiative. This has been at the core of contention between Turing and Popper. Popper's final objection and obstruction to AI is the following:

> One day, Turing was talking on the radio and he issued the challenge, claiming something like "Describe me what - in your view- a computer cannot do, ... and I will build one on purpose". The famous philosopher Karl Popper was rather scandalised and wrote him a letter claiming that there is something in particular that computers do not have and cannot have: *initiative*! We cannot 'describe' initiative, although it is something that any little child and animal has.

In fact, we are building agents and robotic systems that exhibit nontrivial initiative and therefore agents can be considered to be autonomous. Sharing initiative with machines is also the reason we are concerned with "mixed initiative" in human-computer interaction, or about "adjustable autonomy". Here we will not obsess with such foundational issues. Instead, we will keep in mind the range and significance of our objectives-- that is, to build methods that harness the power of machines with initiative.

3. WHAT IS AUTONOMY?

From science fiction to affectionate references, human fantasies about naming machines and interacting with them socially serve a useful function for human psychology but do not advance our understanding of autonomy. It is possible to build useful social agents as in (Breazeal 1999) that affect human emotion. However, mere anthropomorphism or human-like emotion-based behavior does not affect autonomy. Autonomy consideration makes sense when the machine can nontrivially and purposively change its interaction.

Autonomy is a characterizing notion of agents, and intuitively it is rather unambiguous. We recognize the quality of autonomy when we perceive or experience it. Yet, it is difficult to limit it in a definition. The desire to build agents that exhibit a satisfactory quality of autonomy has included agents that have a long life, are highly independent, can harmonize their goals and actions with humans and other agents, and are generally socially adept. We will try to approximate our intuition, to dispel false attributions, and to point the way to scholarly thinking about autonomy.

Let's consider two types of interaction for study of autonomy. The first is *interaction between human and machine*. In this type of interaction, autonomy concerns are predominantly for the agent to acquire and to adapt to human preferences and guidance. The reference point in this style of interaction is always the human, and this gives us a relative sense of autonomy. In this relative

sense, we are concerned with relative deviations in the agent's attitudes and functioning with respect to the human user. The word "autonomous" connotes this relative sense of the agent's autonomy from the human. A device is autonomous when the device faithfully carries the human's preferences and performs actions accordingly. For instance, consider an agent in service of a human. The agent is said to be fully autonomous when it has access to the complete set of choices and preferences of its user. Here the user is a distinguished entity that might judge or change an agent's autonomy. The idea that an agent's autonomy can be adjusted to match the pace of the human is termed *adjustable autonomy* (Musliner and Pell, 1999).

Relative autonomy also makes sense in another type of interaction, which is *among a group of agents*. In these interactions an agent's autonomy can be considered relative to another agent or an environmental factor. There is no user but any other agent may be the reference point.

Autonomy is a social notion, and in fact research has been linked to many social theories. Delegation theory is one such social theory. In many cases the user (or the delegating agent) needs local and decentralized knowledge and decision from the delegated agent. This agent-- delegated to take care of a given task-- has to choose from among different possible recipes (plans), or to adapt abstract or previous plans to suit new situations; it has to find additional (local and updated) information; it has to solve a problem and not just to execute a function, an action, or implement a recipe; sometimes it has to exploit its "expertise". In all these cases this agent takes care of the interests or goals of the former "remotely", i.e., far from it and without its monitoring and intervention (control), and autonomously. This requires what is called an "open delegation": basically the delegation "to bring it about that ..." (Castelfranchi and Falcone, 1998). The agent is supposed to use its knowledge, its intelligence, and its ability, and to exert a degree of discretion.

Control and autonomy are related. When agents consent to a balance of control between them, their balance of autonomy is somewhat complementary to their control. In other words, control affects autonomy. The reverse does not hold. Between an agent that has larger autonomy and another with lower autonomy, there may not be a control relationship. Over a common set of choices, autonomy of the agent who has agreed to be controlled over the choices is lower than the agent who is controlling. Exerting control is the degree to which power is asserted. Experiencing control is the amount of power an agent feels power imposed. Accomplished control is the amount to which an agent concedes to power asserted. Naturally, the degree of autonomy is affected by control to the extent the agent allows control. To the extent to which this measure is binary, the relationship between control and power is close to being complementary.

So far we have described an operational sense of autonomy. Beyond this functional perspective, one can consider an agent's internal manipulation of its own capabilities, its own liberties and what it allows itself to experience about the outside world as a whole. Margaret Boden defines behavioral autonomy as the agent's capacity to be original and not guided by outside sources (Boden 1996). In this view, an agent formulates a liberty over how it functions in the world. It is not nuanced based on one thing or another. There have been few elaborations of this in the literature.

It is important to point out certain works that remain outside the pursuit of autonomy. When machines perform tasks that require cognitive reasoning, we judge the machine's ability to explicitly reason when attributing autonomy. However complex, machines that do not have an explicit reasoning module should not be considered in the same light. An automated chess-playing machine may not have much autonomy. Autonomy is considered if this machine is aware of its interactions and uses that in its action selection.

Mere adaptation or compliance is not enough either. A robot might suitably conform to a variety of circumstances and interactions. However, unless it is purposeful, it is passive. We would like to also exclude much of systems that use machine learning. Human subjective judgment of a machine's appropriate and timely behavior is not a litmus test of its autonomy. We can design a

machine that is, or learns to become, socially adept. In general, this is independent of the machine's autonomy. We would consider a change in its autonomy if the machine learns to reassess its relationship to the world outside itself and have the potential to change its interactions accordingly. In summary, awareness and purposefulness in interactions are required before autonomy can be considered.

4. AGENT CENTRIC CONCEPTION OF AUTONOMY

Autonomy is a fundamental component of agenthood. Work in diverse fields such as Alife and biology have provided useful perspectives. In Alife, autonomy of artificial agents is measured by their ability to generate novel behavior (Boden, 1996). Work in biology considers self-organization as well as individuation of self and other. Luck and d'Inverno's work (2002, in this volume) is complimentary to the biological perspective in that they take motivations central in understanding agent autonomy. They consider motivations as desires or preferences to generate or to adopt goals.

If we look at the definitions of "agent" in AI, we realize that the concept of autonomy is really foundational:

- Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed. (Maes, 1994)

- Intelligent agents are software entities that carry out some set of operations on behalf of a user or another program with some degree of independence or autonomy, and in so doing, employ some knowledge or representation of the user's goals or desires. (IBM Agent)

- Autonomous agents are systems capable of autonomous, purposeful action in the real world. (Brustoloni, 1991)

- *Autonomy*: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state. (Wooldridge and Jennings, 1995)

- An agent is an autonomous software entity that is self-contained and performs tasks on behalf of a user or user-initiated process (Gartner Group Agent)

- Software agents: share information with each other; *are able to work unaided*; learn how to improve their actions with experience (Crabtree, Wiegand and Davies Agent).

In other words, any definition of agent must be related with the concept of autonomy: there is no agent definition without a clear reference to autonomy.

The work of Barber, et al. (2002, in this volume), describes autonomy as "decision making control". Autonomy level is determined by the magnitude of the agent's contribution to a common decision. This is a part of their adaptive decision making framework.

5. AGENT-TO-AGENT AUTONOMY

The focus of agent-to-agent autonomy is issues related to interaction among agents at the peer level. This is in contrast to mixed initiative (see the next section) where humans play a supervisory role Agents can be considered to be objects (in the object-oriented sense), with the additional capability of initiating and making requests of each other and granting requests. As we've said earlier, an agent's ability for self-control and self-motivation is initiative. When this initiative is applied to delegation and dependence we consider it to be autonomy. Therefore, issues of autonomy and interaction are intertwined. Agents involved in interaction mechanisms that support giving and receiving of reasons must be endowed with autonomy.

Agents affect one another in unpredictable ways. Brainov and Hexmoor (2002, in this volume) explore influences among agents and how agent performances change due to these influences. They define a measure of autonomy based on the relative performance of agents. The definition captures the relative sense of autonomy from an observer point of view. However, performance is not a good choice of metric. Perhaps this definition can be applied to the independence of an agent and the resulting autonomy definition can measure how independence of an agent changes in the presence of different agents, i.e., using independence as a subjective measure. They conclude that the problem of finding a multiagent group with the maximum autonomy is NP-complete.

Boella and Lesmo (2002, in this volume) focused on norms among cooperating agents and suggested *anticipatory coordination*. Each agent predicts other agent actions and adjusts its autonomy accordingly. This is used by agents to regulate their own autonomy. Of course, autonomy is not limited to interactions without cooperation and it is useful in other types of interaction as in adversarial games.

Castlefranchi and Falcone (2002, in this volume), analyse autonomy as a concept strictly linked with other relevant notions-- the notion of dependence, the notion of delegation, the architecture of the agent, the notion of automaticity. In particular, they more deeply consider autonomy in collaboration, its relationships with the control, and the modalities of its possible adjustment.

6. MIXED-INITIATIVE AND ADJUSTABLE AUTONOMY

Mixed initiative defines a scenario where initiative is shared between a human user and a machine. In this paradigm human users are distinguished from agents and have authority over agents. Agents obey human users and attempt to elicit their preferences. Typical applications are in domains where agents take over tasks that are dirty, dull, or dangerous for humans. Human users remain in the loop to oversee and make subjective decisions. Cohen and Fleming (2002, in this volume) present an approach where agents weigh the perceived benefits of interaction with a human user against the perceived costs. Agents then adjust their autonomy accordingly. There are many domains where humans and agents must collaborate. One such application is joint human/agent planning (Burstein, et al 2000).

Franklin and McCauley (2002, in this volume) describe an autonomous agent IDA that converses with sailors in the U.S. Navy, using natural language in order to arrive at a job placement situation that is beneficial for the sailor and the Navy. The IDA agent is endowed with explicit motives (i.e., drives) that influence her actions. They suggest that ability to reason about autonomy enhances the IDA agent's ability to deal appropriately with novel and unexpected situations.

Myers and Morley (2002, in this volume) describe how human desires to delegate authority can be encoded in terms of policies. Agents are made to follow such authority governing policies and adjust their autonomies. Controlled by policies, agents in this system provide a sense of predictability and thereby increase user confidence. They plan to investigate nontrivial cases where a group (instead of a single agent) can be guided and policy conflicts can be resolved.

The work of Pynadeth and Tambe (2002, in this volume), describe a real-world situation where agents reason about human preferences for scheduling meetings. Agents adjust their initiative to interact with users. This work is focused on optimal policies that improve teamwork. They have developed a conflict resolution method they call "transfer of control" using Markov decision procedures.

Bradshaw, et al. (2002, in this volume), have outlined their experiences with implementing adjustable autonomy in space applications involving human-centered teamwork. This work uses a language for encoding astronaut activity, which accounts for resources, activities, patterns of and

emergent and routine human-machine interaction. The principal method of adjustable autonomy is encoding policies in a somewhat similar ways to Myers and Morley's work that capture preferences and deontologies of human participants.

7. CONCLUSIONS: AUTONOMY AND ITS WORRIES

Autonomy is a core characteristic of agents. However, since autonomy is a common sense notion, it needs to be made technical in order to lend itself to objective quantification and modeling. We are witnessing steps to systematically approach this idea in the agent community and particularly in the collection of papers in this volume. There are topics that are left out of this volume and many that are bases for formulating open problems. Several of the papers in this book deal with some of those foundational issues. Obviously this is only the opening of a necessary interdisciplinary debate. No one of these problems has been solved, but we hope that at least some of these issues have been clarified and relationships among them is untangled.

Autonomy invites critical investigations, from the theoretical as well as the philosophical points of view. Here we will give an incomplete list of open problems.

- How can we consider autonomy as a property of a collective set of agents? What are the autonomy issues when an agent interacts with groups of agents?
- How can autonomy be considered in organizations and institutions?
- How does natural language convey autonomy?
- What are linguistic autonomy clues?
- How is autonomy related to free will?
- Despite the multitude of possible levels and dimensions of autonomy, is it possible to arrive at a unified notion?
- Will the operational definition (computational and robotic) contribute to the clarification of the concept of autonomy as interesting for the philosophical, behavioral and social sciences?
- From the point of view of technical and application scenarios; will there really exist artificially autonomous machines?
- How much autonomy can there be from human users? Why is this an advantage or a necessity? Are there dangers in this perspective?
- How much can we trust autonomous artifacts?
- How can we maintain devices and protocols of control or of negotiation with artifacts?
- What is the relationship between autonomy and automaticity?
    - autonomy and unpredictability?
    - autonomy and independence?
    - autonomy and self-interestedness and self-motivation?
    - autonomy and freedom? Are these synonymous?
    - autonomy and norms?
    - autonomy and control?
    - autonomy and trust?
    - autonomy and resources and knowledge?
    - autonomy and power?
    - autonomy and personal goals?
    - and many more questions.

As you can see, the issue about *autonomous* agents opens relevant and challenging questions. But it is also a rather practical domain with concrete problems in human computer interaction, in human-robot Interaction, in computer-mediated cooperation, in electronic commerce and virtual organizations, etc. Natural language is rife with deontic notions. We tell one another our stance about our relationship with the world by cues for permission and obligation. "You may", "Please don't", "I will", I don't feel free to" bear information about the desired type of interaction. These statements are indirect indications of autonomy. As far as we know there is no systematic study that illustrates the relationship between deontic phrases and information about autonomies they embody.

The relationship between emotions and autonomy is not addressed. Naturally, they have constraining influence on one another but only indirectly. For example, if I experience "fear" or "joy", I might experience deontic factors that lead to a diminished or expanded autonomy and vice versa.

Several papers in the book deal with those practical and technical promising issues. However, it is also worth mentioning moral and political concerns related to machines, autonomy especially relevant when the autonomous-agent paradigm covers not only HMI and virtual environment, but decision support systems, computer mediated collaboration and organizations, and even our physical environment, etc. via the ubiquitous and disappearing computing.

Let's mention the "prophecy " of the famous criminal, "Unabomber." His mentality was criminal but his vision is rather realistic.

> "What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines make more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide."
>
> *Theodore Kaczynski* - the criminal, "Unabomber"

On the other hand, just one year ago Stephen Hawking, the noted physicist, has suggested using genetic engineering and biomechanical interfaces to computers in order to make possible a direct connection between brain and computers "so that artificial brains contribute to human intelligence rather than opposing it." The professor concedes it would be a long process, but important to ensure biological systems remain superior to electronic ones. "In contrast with our intellect, computers double their performance every 18 months," he told Focus magazine. "So the danger is real that they could develop intelligence and take over the world."

Those are not marginal ethical-political considerations on the autonomy of artificial agents.

Will we be able to build with intelligent-proactive-autonomous machines and environment the same relationship that we have established with biological active autonomous entities through the invention of agriculture or of pasteurization? Given the sociality of the relationship, will we be able to maintain a master-slave relationship between the dominant and the dominated?

8. REFERENCES

Boden, M.A 1996. Autonomy and artificiality. In Boden, *The Philosophy of Artificial Life*, Alife.

Bradshaw J. M., Sierhuis M., Gawdiak Y., Jeffers R., Suri N., Greaves M., 2001. Teamwork and Adjustable Autonomy for the Personal Satellite Assistant, In The IJCAI-01 Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents., Seattle.

Brustoloni, J. C. 1991. Autonomous Agents: Characterization and Requirements, Carnegie Mellon Technical Report CMU-CS-91-204, Pittsburgh: Carnegie Mellon University

Breazeal***,C. 1999. Robot in Society: Friend or Appliance?***Autonomous Agents Workshop on Emotion-Based Agent Architectures, Seattle.

Burstein, M., Ferguson, G., and Allen J. 2000. Integrating Agent-based Mixed-initiative Control with an Existing Multi-agent Planning System, In *the Proceedings of the 2000 International Conference on Multi-agent Systems* (ICMAS), July 2000.

Castelfranchi, C., Falcone, R., 1998. Towards a Theory of Delegation for Agent-based Systems, *Robotics and Autonomous Systems*, Special issue on Multi-Agent Rationality, Elsevier Editor. Vol. 24, pp. 141-157.

Dorais, G.A., Bonasso, R.P., Kortenkamp, D., Pell, B. and Schreckenghost, D. 1998. Adjustable Autonomy for Human-Centered Autonomous Systems on Mars, In *Proceedings of the First International Conference of the Mars Society*, Aug/98.

IBM Agent, http://216.239.51.100/search?q=cache:NPhIYjGFp5sC:www.cpe.eng.kmutt.ac.th/research/projects/2543/micropayment/micro3.htm+%22set+of+operation+on+behalf+of+a+user%22&hl=en&ie=UTF-8

Maes, P. 1994. Modeling Adaptive Autonomous Agents, *Artificial Life Journal*, C. Langton, ed., Vol. 1, No. 1 & 2, MIT Press.

Musliner, D. and Pell, B., 1999. Call for Papers, 1999 *AAAI Spring Symposium on Agents With Adjustable Autonomy*, March 22-24, 1999, Stanford University.

Wooldridge, Michael and Nicholas R. Jennings, 1995. Agent Theories, Architectures, and Languages: a Survey, In Wooldridge and Jennings Eds., *Intelligent Agents*, Berlin: Springer-Verlag, 1-22