Autonomy: A Nice Idea in Theory

Michael Luck* Mark d'Inverno†

* Dept. of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK
mml@ecs.soton.ac.uk
† Cavendish School of Computer Science, Westminster University, London W1M 8JS, UK
dinverm@westminster.ac.uk

1 Introduction

Autonomy is perplexing. It is recognisably and undeniably a critical issue in the field of intelligent agents and multi-agent systems, yet it is often ignored or simply assumed. For many, agents are autonomous by definition, and they see no need to add the tautologous prefix in explicitly considering *autonomous* agents, while for others autonomy in agents is an important yet problematic issue that demands attention. The difficulty when considering autonomy, however, is that there are different conceptual levels at which to reason and argue, including the philosophical and the practical.

The notion of *autonomy* has associated with it many variations of meaning. According to Steels, autonomous systems must be automatic systems and, in addition, they must have the capacity to form and adapt their behaviour while operating in the environment. Thus traditional AI systems and most robots are automatic but not autonomous — they are not independent of the control of their designers [7].

2 What is Autonomy?

A dictionary definition will tell us, among other things, that autonomy amounts to freedom of will (and we will add that it includes the ability to exercise that will). In short, this means that it provides the ability to exercise choice, which is particularly relevant in the context of goals and goal-directed behaviour, as in Castelfranchi's notions of goal (or motivational) autonomy [1]. In this view, autonomous agents are able to generate their own goals, to select between multiple alternative goals to pursue, and to decide to adopt goals from others (to further their own ends). Franklin and Graesser's definition of an *autonomous agent* as a system that pursues "its own agenda" reinforces this perspective [4].

Now, from a purely *conceptual* or theoretical point of view removed from practical considerations, autonomy can naturally be regarded as absolute, without dimension or measure of degree. Yet, this *strong view* of autonomy contrasts with much of the practical work with agents in which autonomy is taken to be the same as *independence*, a very distinctly relative notion. In what might be called this *weak view*, a non-autonomous agent either depends on others or is fixed (eg an automaton), while an autonomous agent can either be independent or depend on others. It is this last point that seems to suggest that autonomy is not the same as independence — an agent does not simply

lose its autonomy by virtue of depending on another for a particular goal; situations of dependence occur also for autonomous agents.

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

3 Autonomy through Motivation

For all the difficulty in pinning down autonomy, it is in our view key to understanding the nature and behaviour both of individual agents, and of interactions between them. In a series of papers, we have described and formally specified an extended theory of agent interaction, based on *goals* and *motivations*, which takes exactly this standpoint. The theory describes a framework for categorising different agents [5], and has been used as a basis for investigating aspects of the relationships between agents [6], providing an operational account of their invocation and destruction [3], as well as for reformulating existing systems and theories, including those relating to dependence situations [2].

In essence, autonomous agents possess goals that are *generated* within rather than *adopted* from other agents. These goals are generated from *motivations*, higher-level non-derivative components characterizing the nature of the agent that can be regarded as any desires or preferences affecting the outcome of a given reasoning or behavioural task. For example, *greed* is not a goal in the classical artificial intelligence sense since it does not specify a state of affairs to be achieved, nor is it describable in terms of the environment. However, it may give rise to the generation of a goal to rob a bank. The distinction between the motivation of greed and the goal of robbing a bank is clear, with the former providing a reason to do the latter, and the latter specifying what must be done.

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

4 Conclusion

The answer to whether we can control autonomy depends on the viewpoint adopted. In the strong view, it is by definition impossible to control autonomy externally. At the same time, however, we can design agents with appropriate motivations and motivational mechanisms that constrain and guide agent behaviour as a result of internal

imposition. In this way, control is *on-board*, and more and better processing of environmental information is required.

We must also question the *need* for autonomy. Certainly, there is value in the flexibility and robustness that autonomy can bring in a dynamic and open world, but many problems which merit an agent approach do not necessarily require autonomous behaviour. Indeed, the strong view of autonomy can be very dangerous if used for example in military applications for tank or missile control; independence with respect to a user or designer can often be bad. Thus, we also need to consider the kinds of situations to which autonomy is suited.

While we have offered an absolute theoretical viewpoint of autonomy as well as a weaker alternative that provides a practical realisation of it that is useful for many, it is important to understand the difference in purpose and context of these notions, and not to be dogmatic in practical situations. Clearly there is value in studying the general concept of autonomy, regardless of practical concerns, but we must also address ourselves to the practical imperative. It matters little what we call it (just as it matters little whether we call a program an agent) as long as it gives us the required robustness and flexibility we desire.

References

1. C. Castelfranchi. Guarantees for autonomy in cognitive agent architecture. In M. Wooldridge and N. R. Jennings, editors, *Intelligent Agents: Theories, Architectures, and Languages, LNAI 890*, pages 56–70. Springer-Verlag, 1995.
2. M. d'Inverno and M. Luck. A formal view of social dependence networks. In C. Zhang and D. Lukose, editors, *Distributed Artificial Intelligence Architecture and Modelling: Proceedings of the First Australian Workshop on Distributed Artificial Intelligence, Lecture Notes in Artificial Intelligence*, volume 1087, pages 115–129. Springer Verlag, 1996.
3. M. d'Inverno and M. Luck. Making and breaking engagements: An operational analysis of agent relationships. In C. Zhang and D. Lukose, editors, *Multi-Agent Systems Methodologies and Applications: Proceedings of the Second Australian Workshop on Distributed Artificial Intelligence, Lecture Notes in Artificial Intelligence*, volume 1286, pages 48–62. Springer Verlag, 1997.
4. S. Franklin and A. Graesser. Is it an agent, or just a program?: A taxonomy for autonomous agents. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings, editors, *Intelligent Agents III — Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*, Lecture Notes in Artificial Intelligence, 1193. Springer-Verlag, 1996.
5. M. Luck and M. d'Inverno. Engagement and cooperation in motivated agent modelling. In *Distributed Artificial Intelligence Architecture and Modelling: Proceedings of the First Australian Workshop on Distributed Artificial Intelligence, Lecture Notes in Artificial Intelligence, 1087*, pages 70–84. Springer Verlag, 1996.
6. M. Luck and M. d'Inverno. Plan analysis for autonomous sociological agents. In *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*, 2000.
7. L. Steels. When are robots intelligent autonomous agents? *Journal of Robotics and Autonomous Systems*, 15:3–9, 1995.