ISSN: 2832-790X
DOI: 10.33552/OJRAT.2024.02.000548
Online Journal of Robotics & Automation Technology
Iris Publishers

Mini Review



# Autonomy in AI and other Artificial Agents

**Paul Dumouchel***

*Department of Philosophy, University of Quebec in Montreal | UQAM, Canada*

***Corresponding author:** Paul Dumouchel, Department of Philosophy, University of Quebec in Montreal | UQAM, Canada

**Received Date:** April 26, 2024

**Published Date:** May 09, 2024

## Introduction

Ruffo [1] in a paper on Autonomous Weapons Systems (AWS) argues that autonomy should not be seen as a property of systems but as a meta-property. Her point is that being autonomous is not something which a system does, but rather a way of managing its different properties. Autonomy characterizes how the various capacities of the system are deployed, but it does not correspond to a property of the system as such. Take for example a military drone, it can fly from point A to point B, observe or identify a target, and perhaps destroy it. Throughout its mission it can be either remotely controlled or function in autonomous mode. In either case, its properties, what it can and cannot do, are the same, what changes is how its abilities are managed or controlled. Most AWS can function either autonomously or remotely controlled and in real life what takes place usually is a mixture of both. For certain things -travelling to its destination, identifying the target - the drone functions in autonomous mode, for others - especially firing - it is remotely controlled by a human operator.

Understood in this way, autonomy is not an additional characteristic which some weapon systems have, but a way of managing, of dealing with, in this case, their lethality. The above scenario also indicates that choosing one or the other mode of management corresponds to a human decision. A decision that can happen at different points in the conception and life of the system. Presently, many, perhaps most AWS can function in either mode. This is a decision that was taken at the moment of their conception. It could have been otherwise. The system was designed in such a way that there is a human in the loop and that the human operator can intervene and overrun, all or some, of the system's autonomous decision. In this sense, the extent to which a system is made to function only autonomously determines how it can interact with humans: operators, clients, citizens, enemy combatants. In consequence, and not surprisingly, autonomy tends to be viewed as a pure characteristic of a system when it is considered independently of the artificial agent's integration into a larger framework. When the purpose which the artificial agents serve is viewed as part of other larger objectives, its autonomy is understood relative to this environment. If a military drone is made to destroy enemy targets, the extent to which we will let it do that "by itself" is subject to various military and political considerations.

From this point of view, the opposite of autonomy is micro-management, preventing the system from responding by itself to any change in the circumstances or environment. This is a bad management strategy which is known to be a cause of rigidity and breakdowns in both material and social systems. However, even when they are not micro-managed employees remain employees and their behaviors are to some extent managed. The degree of freedom which they are granted is to be exercised in view of clearly stated objectives or purposes, for example, selling socks, pleasing customers, and keeping the inventory to date. These objectives are beyond the employee's domain of freedom. Their autonomy does not reach them. The same applies to any material or information system. It functions autonomously within well-defined parameters, in view of specific goals. The system's autonomy is defined as its ability to fulfill its function by itself. Failure to do that is not understood as an expression of the system's autonomy, for example its freedom to protest – as could be in the case of employees - but as a breakdown or malfunction. An artificial agent that is entirely determined by its agent function [2] the mathematical function that maps percepts onto actions, can only be autonomous within the limits defined by

its agent function. [3] argue that autonomy is not unidimensional but should be understood to involve at least two dimensions: self-sufficiency and self-directedness. First, self-sufficiency refers to a system's capacity to act independently, that is without the constant help and support of others, while self-directedness refers to the absence of outside control. An autonomous vehicle self-sufficiently involves, for example, the number of hours or the distance it can cover without requiring a recharge or a refill and also its ability to find and go by itself to its refueling station. The vehicle's ability to go from point A to point B without having to be remotely controlled, given complex and changing factors that make travel difficult, illustrates its self-directedness. The extent to which it is not subject to outside control,

Both dimensions indicate that autonomy is always relative to an environment, to a set of given circumstances. No system, whether natural or artificial, is perfectly self-sufficient. All need at least an energy source and a way to access it and they are subject to various physical constraints. Outside of the limits defined by their environment and constraints, the system will simply fail. Self-directedness is also constrained. To begin with an agent's choice of action is at least constrained by its physical limits. In the case of artificial systems, their self-directedness is evidently also constrained by the task or purpose we want the agent to serve. An artificial agent is only self-directed in relation to that objective. Any action decision that makes it more difficult, let alone impossible, for the system to achieve its goal constitutes a mistake. In this case we are not simply dealing with a system's adaptation to an environment, but to an adaptation to an environment in view of a particular goal.

There is a sense in which the autonomy of AI, of digital artificial agents, of disembodied cognitive systems, like an app, ChatGPT, or any chat box can only to be measured on the second dimension, that of self-directedness. Not because they are perfectly self-sufficient, but on the opposite, because they are not in any way at all self-sufficient. They do not have any autonomy over that dimension. They have no independence from the system in which they are embedded, in consequence there is nothing which they can do for themselves. Unlike a robot that can physically intervene in its environment, move in space to avoid a danger or to better accomplish the task it is charged to do, and whose capacity to help itself measures its self-sufficiency, disembodied AI systems are completely useless to themselves. They cannot change anything in the world unless it is done for them. For whatever they do, or whatever we want them to get done, they entirely depend on others, on humans or other artificial systems. On machines which are physical devices that can modify the world.

This is related to the fact that, as Woods argues, such autonomous systems are "not things at all, but instead are complex networks of multiple algorithms, control loops, sensors, and human roles that interact over different time scales and changing conditions." [4] This is also the case as he points out of autonomous road or air vehicles, for example. We nonetheless identify such robots as independent agents - "things" in that sense – because they are causally responsible for the realization of the task to which the system is dedicated: traveling from one place to another, attacking enemy positions. Their ability to carry out their mission in changing conditions constitutes a measure of their self-sufficiency. A dimension of autonomy which we can augment not only with the help of better sensors, alternative or redundant communications channels, etc. but also by making the robot less dependent on the network, increasing the onboard control.

This is not the case with purely digital agents. Whether it is an app that informs you of the time of arrival of the next bus, a system that evaluates job applications or one that commands train traffic, they do not exist outside of the network. This has consequences for their self-directedness also [1] reminds us that one of the reasons for the choice of autonomy as an alternative method of management is distance. When the target is far away in space from the command center and communication can be uncertain, autonomy is the preferred way of managing the system's capacity. Self-directedness and self-sufficiency are ways of resolving that difficulty. However, an agent which does not exist as an independent object in physical space is not subject to such contingencies and requires neither self-sufficiency nor self-directness because there is nothing to direct, no distance to cover, no environmental incident from which we need to protect the robot.

In consequence, such digital agents have zero autonomy on both dimensions. What we call their autonomy is encapsulated in the complexity of their agent function. This is not really surprising, though it may seem paradoxical. Once we realize that autonomy is always relative to the environment. It should be clear that digital agents, because they are not in the world [5] cannot be autonomous, at least not in this world where we live. Digital agents inhabit a model of the world, composed of data, to which they react, and they are autonomous within that environment to the extent that they can adapt to changes that take place while fulfilling their function. However, digital agents through the complex network of which they are part produce changes in the world. As long as they are driven by specific goals, like a GPS, their "action" remains constrained by those goals and what appears to us as their "autonomy" reflects our ignorance of their agent function.

Chatbots and LLMs like ChatGPT are to some extent in a similar situation. The problem seems different because they do not apparently have any particular objective. They are highly versatile and can respond to a very large range of queries. Furthermore, because they can develop new abilities which they have not been thought, they certainly manifest a greater level of self-directedness than more traditional artificial agents. However, the limits of using synthetic data to train them suggest that the horizon of any form of self-sufficient is still far away. Apart from their relevance to the question asked, there seems to be little constraints to which the agent's responses are subject. In consequence these answers do not always fall within the realm of shared knowledge (hallucination) or of accepted norms (bias). In such a case the environment in which the agent's actions are evaluated is not the set of physical constraints that limit a robot's ability to act, nor a particular objective to be accomplished, but a shared cultural and informational domain. Relative to that changing environment, the

question of their autonomy needs to be raised in a different manner which concerns the hidden cost in human labor involved in their performance [6].

## Acknowledgement

None.

## Conflict of Interest

No conflict of interest.

## References

1. Ruffo M (2020) La robotique militaire: possibilités d'emploi et enjeux éthiques in Ceulemans C, Dewyn m, Lambert D, Ruffo M, P Warnotte eds. Robotisation des armées, Paris: Economica pp. 19-38.
2. Russell S, Norvig P (1995) Artificial Intelligence a Modern Approach; Prentice-Hall: New York, NY, USA.
3. Bradshaw JM, Hoffman RR, Johnson M, D Woods (2013) The Seven Deadly Myths of Autonomous Systems. in IEE Intelligent Systems p. 2-8.
4. Woods DD (2016) The Risks of Autonomy: Doyle's Catch. in Journal of Cognitive Engineering and Decision Making, p. 2.
5. Dumouchel P (2023) AI and Regulations in AI 4(4): 1023-1035.
6. Casilli AA (2021) En attendant les robots enquête sur le travail du clic, Seuil, Paris, France.