

Autonomy and the Perspectives of Its Artificial Implementation [†]

Marcin J. Schroeder

Global Learning Center, Tohoku University, Sendai, Miyagi Prefecture, 980-8576, Japan;
schroeder.marcin.e4@tohoku.ac.jp; Tel.: +81-22-795-3246

[†] The Fourth International Conference on Philosophy of Information, Berkeley, California, USA,
2–6 June 2019.

Published: 14 May 2020

Abstract: Whichever definition of autonomy is used, it is usually formulated in a negative way by the absence, rather than presence, of the defining factors. Some definitions refer to the absence of external causes, physical determination, coercion or control. If positive factors are used, autonomy is associated with the shift from effective causes to final ones. Both approaches, the former of which is based on the elimination of determinism to secure free choice, and the latter of which is based on the replacement of determination by the past by determination by the future, are inconsistent with the scientific description of reality. This paper is an attempt to provide the positive, constructive characterization of autonomy consistent with the scientific view of reality, which can guide us in our search for its implementation in artefacts.

Keywords: autonomy; cause; interaction; intelligence; complexity; autonomous artefacts;

1. Introduction

The concept of autonomy has a long intellectual tradition with diverse definitions, typically formulated in the specific, restricted context of the relationship between human individuals and collectives. Sometimes autonomy is identified with the concept of free will, or the capacity to exercise free will, or alternatively it is considered a characteristic of human collectives corresponding to free will in individuals. Whichever definition is used, it is frequently formulated in a negative way by the absence, rather than presence, of the defining factors. We have many definitions referring to the absence of external causes, physical determination, coercion or external control. If positive factors are used, autonomy is associated with the shift from effective causes to final ones. Autonomy may require some predetermined morality (Kant), capacity for goal-oriented action, and so on.

Both approaches, the former of which is based on the elimination of determinism to secure free choice, and the latter of which is based on the replacement of determination by the past by determination by the future, are inconsistent with the scientific description of reality. If we want to consider the implementation of autonomy in the design of technological AI systems, we cannot afford such inconsistency between the comprehension of autonomy and the comprehension of the mechanisms of its implementation.

The second part of the paper following the introduction is an overview of the intellectual tradition in the study of autonomy, mainly in the context of the contrast between human agents, divine will and natural phenomena. The third part is focused on the issue that, no matter which definition of autonomy we prefer, no system, natural or artificial and devoid of autonomy, can be considered intelligent.

The main, fourth part is an attempt to find the positive, constructive characterization of autonomy, which can guide us in our search for its implementation in artefacts. The confounding

concept of cause is reconsidered. Cause was imported into philosophy from the human context and it was frequently associated with the scientific method, but it actually has only a metaphorical role in scientific inquiries. Thus, the main theme of the present paper presented in the fourth part is a proposal of the concept of autonomy, consistent with the scientific view of the world based on the concept of interaction instead of one-way action or causality, and formulated in terms of information.

2. Human Autonomy and Causal Relationships

The concept of human autonomy belonged to the earliest subjects of philosophical inquiry, going back to the stages of intellectual history in which philosophy was emancipated from religion in parallel to the emancipation of human beings from the omnipotence of the divine. The capacity to exercise human will started from the methods of “buying” the permission of the divine through religious offerings and consulting it about preferred courses of action with the use of divination. In the next step, this gradually evolved into the belief that the divine may delegate its power of making decisions to selected human individuals (sovereigns, warriors or priests).

The process of emancipation continued, and the autonomy of all human beings became the subject of reflection, although the degree in which this autonomy was considered possible varied, not necessarily because of the opposition between the omnipotence of the divine and the will of human beings, but because of the mechanisms governing the world. The limits could be related to the very broad idea of the cause. For instance, the consequences of humans’ earlier actions could restrict the effectiveness of later decisions, with this karmic influence crossing the division between incarnations.

The claims for representing the divine will by religious or political elites could be effectively challenged, but this did not eliminate the problem of conflicts within collectives and of the natural restrictions from the resistance of the environment. The action of human beings was opposed to natural causal relations in the environment independent from human or divine will. It was recognized that the change (motion) could be generated by humans, but also by natural non-human agents (e.g., Aristotelian motion as fulfillment of what exists potentially—local movement, generation and destruction and increase and decrease could be independent from human will *Phys.* 201^a–202^a). This brought into philosophical reflection the question about the initiation of causes.

The most influential in the Western intellectual tradition, with its source in Greek antiquity, was the Aristotelian distinction of the four causes: material, formal, efficient and final. The first two referred to the concept of the substance (that which actually exists), consisting of the combination of matter and form and expressed the view that for the existence of cause and its effect, the substantial characteristics are necessary. Matter and form provided the fundamental conceptual framework for the entire Aristotelian philosophy, and they did not introduce anything specific for the study of cause. This was the reason why they were easily eliminated from the discussion of the role of cause in the scientific method in modern times. The other two causes, efficient and final, directly addressed the characteristics of causal relationships. Francis Bacon tossed out the latter, which may seem strange, because his interest in the scientific method was mainly pragmatic.

The key point of the Baconian scientific revolution was the transformation of the two Aristotelian causes (efficient and final) into one concept, presented as the causal relationship between the cause and effect. The latter is nothing but the hidden final cause of Aristotle, as can be seen in his explanation of this concept. Bacon declared the elimination of the final cause, but actually it was just renamed as an effect. Whenever we want to get a desired effect (final cause), we have to bring into existence its cause (efficient cause).

The actual revolution of modern times was not in the concept of causal relationships, but in the concept of universal determinism, i.e., the assumption that the causal relationship in a natural context forms chains and that all events are links in these chains. The general idea was very old, present in the thought of the school of Democritus, but under the influence of Bacon, and especially of Newton’s mechanics presented in *Principia*, it became a key methodological concept in the

development of physics, not exactly in terms of the relationship between causes and effects, but of the deterministic relationship between the states of interacting physical systems.

Human autonomy, or if preferred, human free will, was questioned in the context of physical determinism. The main assault came in the beginning of the 19th century in the thought experiment of Laplace, in which a theoretical demon was conceived who could calculate the physical states (positions and momenta) of all the particles of the universe. The demon could then calculate all future states of the universe, which makes any effective human action impossible. The actual existence of the demon is irrelevant, as the argument just shows the inconsistency of mechanical determinism with the very idea of effective human autonomous action. The hope that quantum mechanical indeterminism could help proved unjustified, as human actions are at a macroscopic level. Of course, whatever threatens human autonomy can be used as an argument against the feasibility of autonomous artefacts.

In summary, the philosophical tradition of autonomy required the ability to initiate causal relationships, or their chains, without being causally influenced. No system whose state (or behavior) is determined by spatially or temporally external events or objects can be considered autonomous. Considering the determinism of classical physics, no system with components consisting of objects and mechanisms describable in terms of physics can be free from temporal determination by its earlier states, even if we assume that it is isolated, i.e., free from spatially external influence. This raises the question about the feasibility of designing autonomous artefacts.

3. Autonomy—Necessary Characteristic of Intelligence

Autonomy is a fundamental, although only necessary, condition for intelligent action or behavior. No non-autonomous system, human, natural or artificial, can be capable of intelligent behavior or action, because non-autonomous systems cannot act, but only react, to the external actions determining their state. If an artificial system cannot be autonomous, it cannot be intelligent. This makes the question about the feasibility of designing autonomous artificial systems critical for the status of artificial intelligence.

4. Autonomy as a Property of Sustainable Complex Systems

Before we can consider the positive definition of autonomy, it is necessary to clarify the status of the concept of cause or cause–effect relationships. In spite of the common belief among non-scientists, a cause–effect relationship is not reflected in formalisms of physical science theories. This fact was already noticed by Bertrand Russell in 1917: “All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word ‘cause’ never occurs...” [1]. Physicists refer to causality in the interpretation of physical theories, but not within these theories. The actual concept present, and fundamental, in physical theories is that of interaction.

Causality is just an expression of the presence of interactions which are responsible for some particular course of events involving two sides and dynamically altering the states of both of them, which with the attention of an observer restricted to one side, and with the selected direction of time, manifests as a causal relationship. This “interactivist” view of the irrelevance of causality in physical theories is criticized for being based on physics or physical sciences. However, the same can be found in non-physical sciences. When we say that bacteria cause a disease, we actually mean that bacteria interact with the cells of the organism, and in the result of these interactions, the entire organism changes its state. It is this crossing of the levels of the hierarchy in complex systems which makes the term “causality” convenient, but which also makes the concept of causality spurious.

A somewhat a similar situation was seen with Loschmidt’s Paradox, which was used as an argument against Boltzmann’s interpretation of the second law of thermodynamics [2]. The paradox was formulated in terms of the inconsistency between time-reversible mechanics and time-irreversible thermodynamics, arising when we have two levels of description in terms of microstates and macrostates. The paradox could be eliminated by giving the second law a statistical

character. The problem with using the same solution in the case of causality is that statistical interpretation of causality would destroy its primary feature of being a necessary relation.

The issue here is that time reversal exchanges the roles of cause and effect. Thus, the requirement of symmetry with respect to time reversal turns causality into nonsense. The time reversibility (T symmetry) of mechanics comes with the third principle of mechanics, introduced by Newton, which eliminates one-way action (necessary for causality) and enforces the dualistic view of interaction. Further development, enriching physics with electrodynamics involving interactions between electric and magnetic fields, required more general symmetry (PT symmetry) involving not only time inversions, but also parity (orientation) reversal. Finally, the discovery of nuclear forces and the concept of antimatter led to even more general symmetry involving, additionally, the reversal of charge (CPT symmetry). The entire development of physics was convoluted with the increase in the role of symmetry, including CPT symmetry.

Thus, instead of using the concept of causality, we have to find a description of autonomy in terms of interaction. The experience of making mechanics and thermodynamics compatible tells us that autonomy cannot be a feature of simple systems, which in their dynamics have to be symmetric with respect to CPT reversals. However, this symmetry can be broken, in the case of highly complex systems, although this symmetry violation may have only a statistical character, and in principle the symmetry can be restored if a sufficiently long time is considered.

Considering the fact that our objective is to implement autonomy in artificial systems and that we expect complexity as a characteristic of autonomous systems in order to avoid paradoxes, we should look for a conceptual framework for the study of autonomy based on the concept of information and its dynamics. The present author developed his theory of information, in which two always coexistent manifestations of information, selective and structural, are present [3]. The same conceptual framework was used in his article on complexity [4]. Now, the concept of autonomy, consistent with the scientific view of the world, can be formulated in terms of information and its dual selective and structural manifestations. The consistency is achieved by the assumption that information systems are open, and they can interact and practically always interact, at least with their intermediate environment, and therefore are always a subject of external influence, but that there is no one-way action involved but mutual interaction. Thus, unavoidable dynamic transformations change the components of the system, but not necessarily the system itself, understood as a structure. Here we have the presence of symmetry understood as invariance of the structure in spite of transformations changing its components.

Autonomy is understood here as a concept applicable only to complex systems equipped with structures satisfying the condition of structural stability. This means that the structure preserves its identity in interactions with the structured environment. Since structures are manifestations of information, which are dual to selective manifestations of information involved in choices, and every complex action consists of a chain of choices (selections), the preservation of the structure is associated with the sequence of selections which can be interpreted by an external observer as directed by some goal. However, it is the preservation of the structure which generates the goal, not the goal which generates the chain of selections. The preservation of the structure does not mean its static form. The initial structure may evolve to a higher level of complexity, under the condition that its original structural information is preserved.

When we ask the question about the feasibility of designing artificial systems which can be autonomous, the answer is, at present rather speculative, that this is possible. The necessary, but not necessarily sufficient condition, for achieving autonomy is an alternative form of computing based on interaction, not on one-way actions [5,6].

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Russell, B. *Mysticism and Logic, and Other Essays*; Unwin: London, UK, 1963.
2. Wu, T. Boltzmann's H theorem and the Loschmidt and the Zermelo paradoxes. *Int. J. Theor. Phys.* **1975**, *14*, 289–294.
3. Schroeder, M.J. Dualism of Selective and Structural Manifestations of Information in Modelling of Information Dynamics. In *Computing Nature, SAPERE 7*; Dodig-Crnkovic, G., Giovagnoli, R., Eds.; Springer: Berlin, Germany, 2013; pp. 125–137.
4. Schroeder, M.J. Structural and Quantitative Characteristics of Complexity in Terms of Information. In *Information and Complexity*; Burgin, M., Calude, C.S., Eds.; World Scientific Series in Information Studies, Volume 6; World Scientific: River Edge, NJ, USA, 2017; pp. 117–175.
5. Schroeder, M.J. From Proactive to Interactive Theory of Computation. In *The 6th AISB Symposium on Computing and Philosophy: The Scandal of Computation – What is Computation*; Bishop, M., Erden, Y.J., Eds.; The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, Institute of Psychiatry, King's College: London, UK, 2013; pp. 47–51.
6. Schroeder, M.J. Towards Autonomous Computation: Geometric Methods of Computing. *Philos. Comput. (Newsl. Am. Philos. Assoc.)* **2015**, *15*, 9–27.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).