This is the accepted manuscript of a paper published in *Human Affairs 30*(4), pp. 587-596, October 2020. Please use the version here for citation.

Artificial Life and 'Nature's Purposes': The Question of Behavioral Autonomy

Elena Popa

Abstract

This paper investigates the concept of behavioral autonomy in Artificial Life by drawing a parallel to the use of teleological notions in the study of biological life. Contrary to one of the leading assumptions in Artificial Life research, I argue that there is a significant difference in how autonomous behavior is understood in artificial and biological life forms: the former is underlain by human goals in a way that the latter is not. While behavioral traits can be explained in relation to evolutionary history in biological organisms, in synthetic life forms behavior depends on a design driven by a research agenda, further shaped by broader human goals. This point will be illustrated with a case study on a synthetic life form. Consequently, the putative epistemic benefit of reaching a better understanding of behavioral autonomy in biological organisms by synthesizing artificial life forms is subject to doubt: the autonomy observed in such artificial organisms may be a mere projection of human agency. Further questions arise in relation to the need to spell out the relevant human aims when addressing potential social or ethical implications of synthesizing artificial life forms.

Keywords: Artificial Life, biological functions, teleology, selected effect functions, behavioral autonomy

1. Introduction

Broadly defined as 'the synthesis and simulation of living systems' (Aguilar et al 2014: 2), Artificial Life (henceforth ALife) research took shape in the 1980s and led, among other things, to the current use of computational models in biology. Further advances in connection to robotics or synthetic biology are expected to have social consequences. This paper focuses on a particular question in relation to ALife, that of autonomy. As one of the uses of ALife has been epistemic, i.e. increasing the understanding of biological life, from this perspective synthesizing autonomous systems can be viewed as aiming to provide a better account of (natural) autonomy (Froese et al. 2007). Behavioral

autonomy, which this paper will discuss, should be distinguished from constitutive autonomy, or autopoiesis (as in Maturana and Varela 1980).

My analysis will have an epistemic focus, investigating how behavioral autonomy in ALife can shed light on autonomy in living systems and subsequently challenging a central assumption in ALife research, that of autonomy in living organisms and ALife being similar in kind. As behavioral autonomy would require that ALife entities behave in a way that would enable them to achieve specific goals or tasks, one important question arising is how to understand these goals, particularly how they are set. To address the former, I draw a connection to the question of functions and teleology in the philosophy of biology, with emphasis on the treatment of teleological notions within a naturalist framework. Regarding the latter, I point out a problem arising for ALife systems that does not hold for living organisms, namely of who sets the goals (section 2). I further use an illustration of a synthetic life form to argue that these goals are set in accordance with research agendas that come down to human interests (section 3). An important consequence of this is that what counts as autonomy in ALife is dependent on broader human goals.

The proposed analysis raises questions regarding how much can be learned from ALife regarding the autonomy of living systems that is not already projected, shedding doubt on the expected epistemic benefits. In broader perspective, my argument will help clarify further questions regarding the social consequences of the expansion of ALife and AI: if the autonomy of artificial systems is modeled on human goals, then potential social consequences should be investigated in relation to the goals of the human agents involved in the design or research program.

2. Artificial life, behavioral autonomy and teleology

This section discusses the distinction between behavioral autonomy and constitutive autonomy in ALife with focus on the former, then moves on to the question of teleology in relation to behavioral autonomy understood as acting to achieve goals set within a certain environment. I investigate this in relation to the naturalization of teleological concepts in the philosophy of biology concluding that there are significant differences between ALife and living organisms, and thus the naturalistic explanation used for biological organisms cannot work for ALife. Given that an explanation of ALife behavior in terms of ‘nature’s purposes’ (Allen et al. 1998) is unavailable, when discussing behavioral autonomy the question of whose goals ALife systems are acting on is unavoidable.

While autonomy has been closely connected to the goals of ALife research, the meaning of autonomy has not always been fixed. Froese et al. (2007) describe the goals of ALife in relation to autonomy as follows: ‘synthesize autonomous agents, and (...) through this process gain a better understanding of the generative mechanisms underlying autonomy in general’ (p. 455).

Subsequently, behavioral autonomy, focusing on external behavior, is distinguished from constitutive autonomy, focusing on internal organization (p. 456). In analyzing behavioral autonomy, the authors identify three categories (pp. 456-457):

(1) Autonomy in the context of engineering, concerned with building robots, referring to functioning without human intervention (as in Brooks 1991).

(2) Autonomy as acting to achieve goals or even set one's own goals in the context of a certain environment (as in Beer 1995; Nolfi and Floreano 2000) – this understanding relies on teleological notions.

(3) Autonomy as robustness and flexibility of behavior (as in Smithers 1992), connecting to the notion of self-organization and emergence.

The category (2) above will be the most relevant for my argument, as it adds more complexity to the question of autonomous behavior than sense (1), and raises further questions about teleology. Sense (3), while important, refers more to how an organism acts autonomously rather than providing conditions spelling out what autonomy is. One point to note about category (1) above is that it has been subject to criticism holding that the behavior is still dictated by the experimenter (Nolfi and Floreano 2000), so the system is not strictly speaking autonomous. After attempting to spell out the teleological notions used in sense (2), I argue that that goal directedness will also involve goals connected to a research program or more general human aims, but this is discussed at a more subtle level than in sense (1).

Having focused the discussion on behavioral autonomy, particularly in relation to setting and achieving goals, I am now going to explain the background for the claim that synthesizing ALife systems can provide a better understanding of behavioral autonomy. One issue here is whether the discussion of autonomy in ALife alongside living organisms by reference to functions and teleology is warranted. Another question is, given the aims of ALife research, what kind of knowledge is such investigation expected to generate regarding autonomy in living systems. Regarding how autonomy as applied to ALife can be commensurable with that of biological systems, it is worth noting that the discussion by Froese et al. (2007) focuses on 'approaches which do not treat the autonomy of living beings as qualitatively (though, perhaps, quantitatively) different from the autonomy of most artificial agents' (p. 256). If the two are taken to be similar in kind, then learning more about one could also shed light on the other. Nevertheless, as stated here, this appears to be more of an assumption of the research program than something supported by particular arguments. My discussion below will challenge this assumption by looking at how teleological notions underlie talk of behavioral autonomy in relation to both ALife and biological systems. Particularly, I will highlight that since teleological notions cannot be understood in the same way, the epistemic

ambitions of this research program, namely, a better understanding of autonomy in living systems, would also come into question. The discussion below explains that from a conceptual perspective referring to teleology and functions, while in section 3 I provide a further argument drawing from an example of ALife. Situating my view among philosophical debates on functions in biology and engineering, the view defended here goes against Kingma's (2020) discussion of synthetic biology that ascribes aetiological functions to *de novo* organisms, thus falling in line with the assumption I will criticize. I address a potential challenge stemming from Kingma's account at the end of section 4.

Teleological notions, particularly in relation to functional explanation, have been subject to debate in the philosophy of biology. From the perspective of the explanatory relation, the issue is that when attributing functions to traits the *explanandum* is placed at an earlier point in time than the *explanans*. For instance, if one were to explain the presence of a certain colour pattern in a mimicry butterfly in terms of enabling the butterfly to avoid predators, one can notice that the trait precedes the function, i.e. the colour pattern is present first, then the butterfly is able to avoid predators. Similar issues arise for any adaptive traits, including particular behaviors. One way of accounting for this would be to accept purposes, but in a naturalized form. This is what the selected effect (SE) theory of functions does (Wright 1973, Millikan 1989, Neander 1991): a trait has a function if it is a token of a type that has been selected in the past for performing the said function. In the case of the mimicry butterfly, a current butterfly has a particular colour pattern because it is part of a species that relied on that specific pattern for survival. In versions such as Wright (1973) or, more recently, Neander and Rosenberg (2012) selected effect functions rely on natural selection: as such there is no purpose by virtue of which certain traits are adaptive other than those associated with the process of natural selection, or 'nature's purposes'. A more recent version of selected effect functions by Garson (2017) expands the scope beyond natural selection: a trait 'need only have done something that allowed it to persist better (longer, more effectively) than some alternative trait within a population' (p. 524).

Narrowing this down to the question of behavioral autonomy, the naturalistic account above would explain particular behaviors by living organisms in terms of the contributions to the organisms' survival and reproduction. If similar teleological notions apply to ALife, the question is whether an analogous explanation would hold. As seen above, naturalizing functions avoids attributing particular goals or purposes to entities outside natural selection or other processes leading to the proliferation of certain traits. The same move is not available for autonomous behavior in ALife, however, as there is no natural selection process, no type from which tokens of adaptive behavior would belong to, thus the question is what ultimate goals would functions in

ALife serve. The same holds for Garson's theory of functions, even if the scope is extended beyond natural selection – there is a clear focus on biological functions, which are distinguished from artifacts. While beyond the scope of this paper, it is worth noting how Garson distinguishes between the two, with a further question regarding whether ALife functions would be closer to those of artifacts: 'beaver dams acquire their functions, in the first place, because individuals design, create, and use them; they don't get their functions in the same way that features of organisms do' (Garson 2019: 21).

Thus, the question of what goals would ALife behavior pursue still stands and to spell it out one would need to look at the purpose of synthesizing artificial life forms, determined by the research program. Thus, while an ALife system can set goals in connection to a specific environment where it functions, the ultimate goals are set by the experimenter. This is not as simple as a robot operated by a remote control, as in sense (1) above, but a common concern may be noted here: to what extent would autonomy understood as acting to achieve and set one's goals include the possibility of ALife systems setting goals that may not necessarily align with those of the research program? To illustrate the shortcomings of a naturalistic interpretation of teleological notions with regard to ALife with an example, Boden reports a case from Sims (1991) where 'unlike most GA [Genetic Algorithm] systems, the selection of the "fittest" examples is not automatic, but is done by the programmer', and 'the human being selects the images which are aesthetically pleasing, or otherwise interesting, and these are used to breed the next generation' (Boden 1998: 302). In this particular case, the selection is made by the programmer, and criteria such as aesthetic preferences are difficult to fit into a framework based on natural selection, or any framework that does not consider the other goals of the programmer, and what the aesthetic characteristics contribute to. Similarly, even if selection is automatic, questions regarding what the algorithm is supposed to select for arise. Thus, even if an evolutionary description is present, these processes are distinct from selection in biological organisms insofar they are driven by particular research agendas and preferences by the researcher.

One issue that arises by connecting this example to Garson's account described above is whether the use of an evolutionary algorithm would not yield into virtual entities that have selected effect functions. For instance, a programmer can provide a selection criterion that keeps certain virtual entities and disposes of others. My reply is that this differs from biological selection since it is all part of the design of the programmer: the selection of virtual entities with a particular trait is caused by them being in line with the criterion set by the programmer, and not of them being part of a type of entities that persisted because of an effect caused by the trait. At most, the design can be seem as a simulation (say, of an environment where a certain trait becomes crucial for survival). I

acknowledge the epistemic benefits of simulation, but in the context of behavior and autonomy, the selected entities would be mimicking the specific behavior, rather than behaving autonomously themselves.

Before articulating the argument, I will address another potential concern. One may point out that in the discussion above I have focused on contrasting functions of ALife systems with functions in biological organisms defined according to the selected effect theory. However, there are other theories of functions which have a similar way of accounting for traits of living organisms and ALife, thus supporting the idea that the difference is merely one of degree. My reply is that I have focused on selected effect functions because they do not seek to explain away teleological notions, but to naturalize them. Were one to adopt a theory that moves away from teleological notions, such as the causal role theory (Cummins 1975, 2010), then the explanation would take the shape of a system to which various traits make a contribution, without reference to evolutionary history. Cummins (2010) points out that while teleological explanations seek to answer a question regarding why a trait is there, on his approach functional explanations answer questions regarding how a trait contributes to a more complex process. Thus, the issue of purposes or goals would be dropped, and reference would be made to the workings of the system. Applying this to ALife, particular traits of synthetic organisms would be understood in terms of their contribution to higher level functions in a system. Once again, this system would be underwritten by human goals in a way that biological systems would not – while functional biology does not appeal to evolutionary explanations, evolution can be viewed as a wider framework for living organisms. The reason I chose selected effect functions for my argument is that this allows for a deeper exploration of teleology in relation to autonomy and goal setting in ALife, whereas a causal role approach could simply do away with teleological notions (and thus with sense 2).[1]

Summing up the earlier discussion, my argument can be stated as follows:

i. If the difference between behavioral autonomy in ALife and biological organisms is only one of degree, then similar types of explanations of particular behaviors in ALife and biological organisms should be available.
ii. In biological organisms functions (including of particular behaviors) can be explained by reference to their evolutionary history, while such explanations are not available for ALife because there is no evolutionary history.

1 To clarify this further, defenders of causal role functions can abide by Dobzhansky's (1973) dictum that 'nothing in biology makes sense except in the light of evolution', while still pointing out that functional biology does not appeal to evolutionary explanations aside from general picture claims like the one above, and as such, it does not use teleological notions in the sense of selected effect functions (see Griffiths 2009 for a discussion in relation to adaptation).

iii. Therefore, the difference between behavioral autonomy in ALife and biological organisms is not only of degree.

In this formulation, the argument questions the assumption regarding the extent to which autonomy in ALife is similar to that in biological systems.

This raises several questions, the first one being what would the major difference amount to if evolutionary history is not an explanatory tool for ALife, but evolutionary metaphors are used instead. My answer is that the difference consists in the involvement of particular goals and purposes associated with the research program in ALife autonomy, which is not present in living organisms. Thus, while one may explain traits in biological organisms in terms of what they were selected for, in the case of ALife, selection would amount to the question of what it was intended to do in accordance with a research agenda driven by human goals. Having spelled out this difference, the second question is to what extent is the goal of better understanding autonomy in living systems achievable by synthesizing ALife. As goals set and achieved by ALife depend on human goals, it appears that behavioral autonomy in ALife works as an extension of agency in humans. While a naturalistic perspective would be possible, spelling out the evolutionary benefits humans would gain, this does not dispel the worry that we may learn little beyond the human features of autonomy we project in ALife, such as the pursuit of particular goals. The presence of human goals points to a further implication of this argument: while ALife forms may act autonomously, the behaviors are driven by particular human interests. These interests cannot be overlooked when discussing social or ethical implications of ALife technology: traits in ALife systems are neither the result of blind variation, nor inevitable given interactions between the organism and its environment, but are underlain by particular human goals.

One objection that may be raised against the conclusion of the argument above is to point out that functions that result from artificial selection (in domesticated animals, for instance) are analogous to functions in ALife. Since no one would deny that artificial selection involves biological functions, the same would hold for ALife. This objection would rest on Kingma's account of selected effect functions in domesticated animals – for instance, domesticated cows producing more milk than their offspring needs allowed them to survive and reproduce with the mediation of humans (2020: 189). Kingma further argues against the claim that this process goes against natural selection by highlighting that the perspective may be shifted and cows can be seen as having 'successfully domesticated *Homo sapiens*, modifying the species by exerting a selective pressure on them to overcome the natural lactose intolerance in adults' (ibidem). In reply I hold that while artificial selection does involve biological functions, they are not the proper functions of the traits, but modified and/or co-opted according to human interests. In my view there is a central

difference between the two perspectives above: domesticating animals involves human intentions and working with traits that have been already subject to natural selection. Even if humans and cows co-evolved, there is no analogous story available about how humans came to digest milk as a result of the cows' intentions. As intentions are characteristic of humans, selecting traits by human intervention differs from 'enlistment': while certain traits are prevalent in certain species because of co-evolution, biological explanations of these traits would not refer to intentions. This point also applies to ALife functions – explaining behaviors in ALife would involve reference to human intentions – thus supporting my conclusion above.

3. The case of autonomy in xenobots

In this section I use a case study to provide further support to my argument above. By looking at the case of a synthetic life form and the description of its design in relation to fulfilling particular tasks, I show that interests by researchers or broader human needs underlie what autonomy is taken to be in ALife systems. I conclude by raising further questions regarding the development of ALife research and potential epistemic uses.

Kriegman et al. (2020) report the design of living systems (dubbed 'xenobots') using evolutionary algorithms and a cell-based construction toolkit. The cells used were taken from *Xenopus laevis* embryos in blastula stage, and they were manually engineered to match a previous in silico design to fulfill tasks such as locomotion, object manipulation, object transportation, and collective behavior (pp. 2-3).

Analyzing this case in relation to the discussion above, the question is to what extent is this example of ALife illustrative of behavioral autonomy. Given that there is no evolutionary history of the behavior of xenobots, but the behavior is meant to aid in performing specific tasks, one may ask in relation to what goals, or, in the light of the previous discussion, whose goals would the specific behavior follow. In my view, the answer lies in the description of the procedure: the in silico design is set on a certain 'desired behavior' (Kriegman et al. 2020) or 'target tasks' (Ball 2020). The goals here appear to be the scientists' in connection to a broader research program, and autonomy would mean the xenobots can fulfill these tasks with no intervention from the experimenter. The addition of a new layer of biological materials does not appear to add much to sense (1) of behavioral autonomy above. If the discussion is conducted at the level of sense (2) – this is possible because xenobots may prove capable to engage into a process leading up to successful performance of a task going beyond mere independence from human intervention - then the goals are the experimenters' or associated with the overall research agenda. Further support for this reading can be found in Kriegman et al.'s description of the significance of the project: 'others may use this approach to

design a variety of living machines to safely deliver drugs inside the human body, help with environmental remediation, or further broaden our understanding of the diverse forms and functions life may adopt' (2020: 1). This particular statement puts the research program in perspective: certain goals are clearly related to medical and environmental projects, both of which are closely tied to human interests. The issue of understanding life also connects to human interests, though the epistemic dimension is more prominent here, and it raises further questions that I will discuss below.

A question relevant for my investigation here raised by Ball (2020) in relation to Kriegman et al. (2020) is whether an engineered life form can be counted as an organism, noting its inability to reproduce, which would be part of how biological life forms are understood, and through which, I would add, evolutionary history can be used to naturalize functions. Under the hypothesis that reproduction could be incorporated in the design of ALife, Ball proceeds to ask 'are multicelled aggregates plastic enough to support totally different yet wholly viable lifeforms from the ones their genomes have evolved to create?' (2020: 265). Here, a clash between evolved biological function and functions attributed in accordance with specific research programs becomes apparent: the cells are engineered to fulfill roles different from those they evolved.

A further question to raise in the context of the present paper is whether this is sufficient for behavioral autonomy, or is it possible to take a further step and ask whether these lifeforms can perform tasks fully different from those they were designed to perform? A positive answer to this last question may enable the research program to shed light on biological autonomy, namely how ALife organisms can develop traits and behaviors outside the constraints imposed by the research program or human interests more broadly.[2] It is this sense that would go beyond the use of goals to describe ALife systems as autonomous insofar as they fulfill aims connected to human interests. I should note that I am referring to empirical possibility here, I am not claiming that this is technologically possible at present, nor that it is outside the scope of normative questions on its ethical implications.[3] Thus, in the current state, the contribution of synthesizing xenobots to understanding life more broadly can provide insights into biological issues such as how cells develop, whether organisms can be synthesized to reproduce etc, but not into behavioral autonomy.

2 An anonymous referee raises the question whether conceding the possibility of reproduction in ALife would undermine my claim above about evolutionary history. While this is subject to future empirical inquiry, from my discussion above the ability to reproduce would not necessarily yield into selected effect functions – see my considerations on artificial selection at the end of section 3: ALife may become equivalent to selective breeding.

3 The Sci-Fi computer game Mass Effect, for instance, illustrates such a possibility through a background story of synthetic life forms (though described as AI rather than ALife), called the Geth, revolting against their creators and taking over the planet (though deliberately allowing their creators to leave). Perhaps possibilities of synthetic life forms acting on their own goals can be imagined in relation to less threatening outcomes.

4. Conclusions

In discussing behavioral autonomy in ALife, I have looked at different senses of autonomy employed in robotics and AI research, focusing on the use of goals and teleological notions more broadly. By contrasting the naturalization of teleological notions in the philosophy of biology through the selected effect theory of functions with the use of teleological notions in ALife research, I have highlighted one major difference between the two: for ALife, goals are always described in relation to particular research aims or human interests. I have subsequently highlighted the importance of being clear about whose purposes would autonomous behavior in ALife organisms follow, especially for potential social issues that may arise in connection to AI and ALife. With regard to the epistemic ambitions of ALife research, I have provided a reason to doubt whether synthesizing ALife systems could shed light on behavioral autonomy in biological systems as, unlike in the case of the living world, ALife organisms function more like an extension of human agency. By looking at recent literature describing the synthesis of xenobots and their design and autonomy, I have provided further support for the central argument in this paper. This example also helps ask further questions about autonomy and the relation between biological functions as the result of evolution versus their uses in engineering new life forms.

The discussion here helps clarify the question of autonomy and the uses of metaphors in relation to teleology and evolution in ALife research, and this can provide more perspective in analyzing developments in synthesizing artificial life forms. In addition to providing conceptual clarity, a further suggestion is to narrow down the scope of expected epistemic gains in relation to concepts such as autonomy associated with related research programs. The question of human interests and their connection to concepts in ALife, teleology, and agency can help shape social or ethical discussions on ALife.

References

Aguilar, W., Santamaría-Bonfil, G., Froese, T., & Gershenson, C. (2014). The past, present, and future of artificial life. *Frontiers in Robotics and AI*, *1*, 8.

Allen, C., Bekoff, M., & Lauder, G. V. (Eds.). (1998). *Nature's purposes: Analyses of function and design in biology*. Cambridge, MA: MIT Press.

Ball, P. (2020). Living robots. *Nature Materials*, *19*(3), 265-265.

Beer, R.D. (1995) A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* 72(1-2), 173–215.

Boden, M. A. (1998). Autonomy and artificiality. *Cognitive Architectures in Artificial Intelligence: The Evolution of Research Programs*, *2*, 300-312.

Brooks, R.A. (1991) Intelligence without reason. In: Myopoulos, J., Reiter, R. (eds.) *Proc. of the 12th Int. Joint Conf. on Artificial Intelligence*, pp. 569–595. Morgan Kaufmann, San Mateo.

Cummins, R. (1975). Functional Analysis, *The Journal of Philosophy*, 72/20: 741-65.

Cummins, R. (2010). Neo-teleology. In Rosenberg A., Arp, R., *Philosophy of Biology. An Anthology*, 164-174, John Wiley & Sons.

Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. *Am Biol Teach* 35: 125–129

Froese, T., Virgo, N., & Izquierdo, E. (2007). Autonomy: a review and a reappraisal. In *European Conference on Artificial Life* (pp. 455-464). Springer, Berlin, Heidelberg.

Garson, J. (2017). A generalized selected effects theory of function. *Philosophy of Science*, *84*(3), 523-543.

Garson, J. (2019). *What biological functions are and why they matter*. Cambridge University Press.

Griffiths, P. E. (2009). In What sense does 'Nothing make sense except in the light of evolution'?. *Acta Biotheoretica*, *57*(1-2), 11.

Kingma, E. (2020). Functions and Health at the Interface of Biology and Technology. *Noûs*, *54*(1), 182-203.

Kriegman, S., Blackiston, D., Levin, M., & Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proceedings of the National Academy of Sciences*.

Maturana, H.R. & F.J. Varela (1980). *Autopoiesis and Cognition: The Realization of the Living*, (Boston Studies in the Philosophy and History of Science, 42), Dordrecht: Springer Netherlands.

Millikan, R.G. (1989). In Defense of Proper Functions, *Philosophy of Science*, 56(2): 288–302.

Neander, K. (1991). Functions as Selected Effects: The Conceptual Analyst's Defense, *Philosophy of Science*, 58(2): 168–184.

Neander, K., and Rosenberg, A. (2012). Solving the Circularity Problem for Functions. *Journal of Philosophy* 109: 613-22.

Nolfi, S., and Floreano, D. (2000): *Evolutionary Robotics: The biology, intelligence, and technology of self-organizing machines*. The MIT Press, Cambridge (2000)

Sims, K. (1991). Artificial Evolution for Computer Graphics. *Computer Graphics*, *253*(4).

Smithers, T. (1992) Taking Eliminative Materialism Seriously: A Methodology for Autonomous Systems Research. In: Varela, F.J., Bourgine, P. (eds.) *Proc. of the 1st Euro. Conf. on Artificial Life*, pp. 31–40. The MIT Press, Cambridge.

Wright, L. (1973) Functions. *Philosophical Review* 82: 139-168.

Asian University for Women
20 M.M. Ali Road,
Chittagong, 4000
Bangladesh
elena.popa@auw.edu.bd