Measuring Autonomy for Life-Like AI

Demyan Vakhrameev[1], Miguel Aguilera[1,2], Xabier E. Barandiaran[2] and Manuel Bedia[1]

[1]ISAAC Lab, Aragón Institute of Engineering Research, University of Zaragoza, Spain.
[2]IAS-Research Centre for Life, Mind, and Society, Department of Philosophy, UPV/EHU University of the Basque Country, Spain
dem.vakh@hotmail.co.uk

Abstract

Current success of Artificial Intelligence (particularly in the application of Deep Learning techniques) is bringing some of its methods closer to Artificial Life and re-opening old questions, social fears and envisioned applications. The concept of autonomy has long guided research and progress in Artificial Life. We explore how this concept can contribute to evaluate the autonomy of contemporary AI systems.

Why Autonomy matters for AL and AI

Artificial Life (AL) was partly born in contrast (sometimes in overt opposition) to Artificial Intelligence (AI) often named as GOFAI (Good Old Fashioned AI). AL came as an alternative to modelling cognitive phenomena as the abstract manipulation of explicitly encoded propositional knowledge (illustrated by expert-systems). Recent progress in AI (Lee et al., 2018), particularly in the fields of reinforcement (Vinyals et al., 2019) and deep learning (Sejnowski, 2020)[1], has brought both paradigms closer together, and, some might say, has partly shown that the early critiques of AL to traditional AI where well founded (Man and Damasio, 2019). Hybrid AL-AI approaches are now within the winning lot for successful applications. But what can AI still learn from AL? How does current AI score on its life-likeness? How can we make it even more autonomous?

Two main concepts have permeated the field of AL from its conception: the concepts of evolution and self-organization. This last one has often been re-conceptualized as autonomy to the extent that the very field of AL was once labelled as a "practice of autonomous systems" as the moto of ECAL'91. The property of autonomy, as a principle of living systems, has attracted the attention of AL modellers for years: how does a system become self-sustaining and self-governing? How does a *self* emerge in the first place? How do animated beings interact with their environments so as to become agents, owners of their actions, and not passive sufferers of external forces? Modelling autonomy has become a key challenge of AL as a discipline (Di Paolo, 2004; Froese et al., 2007): from the study of the emergence of metabolic networks and the origins of life, to the development of autonomous robotics (Barandiaran and Ruiz-Mirazo, 2008). And yet the last decade has witnessed an increase in the autonomy of Artificial Intelligent systems. How can we use existing progress on AL to assess the autonomy of current AI systems?

If there is somebody that has been sensitive of the problems and interactions between AI and AL approaches (since as early as the cybernetics era) that is Prof. Margaret Boden. In a 2008 article for a special issue on modelling autonomy from an AL perspective she emphasized the need to account for autonomy in a manner that was suitable to both traditions (AL and GOFAI) capturing three key characteristics of autonomy:

> "The first is the extent to which response to the environment is direct (determined only by the present state of the external world) or indirect (mediated by inner mechanisms partly dependent on the creature's previous history). The second is the extent to which the controlling mechanisms were self-generated rather than externally imposed. And the third is the extent to which any inner directing mechanisms can be reflected upon, and/or selectively modified in the light of general interests or the particularities of the current problem in the environmental context." (Boden, 2008, p. 2)

In the light of these three properties we define autonomy along three dimensions (modifying the order and part of the terminology by Boden): 0. Self-generation (equivalent to Boden's 2nd), 1. Self-organization, and 2. Self-control. Some of us and other authors have previously distinguished between constructive (or constitutive) and interactive (or behavioural) dimensions of autonomy (Barandiaran, 2004; Barandiaran and Moreno, 2008; Froese et al., 2007) based on previous conceptions of process closure and interaction closure (Christensen and Hooker, 2000; Collier, 2000). Our proposal is to further distinguish, within the constitutive or constructive dimensions of autonomy, between those aspect related to the very generations of the system (biological development, metabolic self-maintenance, autopoiesis, etc., Varela, 1979) and those related to the inte-

[1]Of particular interest to us are applications that control agents in virtual or real world scenarios with prominent success in complex video-games (OpenAI et al., 2019) and automobiles

grated organization of the internal processes in charge of behaviour (e.g. brain-body dynamics, robot-control circuits, etc.). This distinction is important if we are to apply autonomous measures to systems that are not autopoietic and can yet achieve significant degrees of autonomy.

We dedicate the rest of this paper to explore how the dimensions of self-organization and self-control can be applied to AI systems.

Autonomy in two dimensions

Dimension 1 - Self-organization

A system constitutes an individuality when it is organized into a unified whole that is irreducible into separate parts. This dimension of autonomy captures the way in which a unified/integrated self emerges out of the interaction among its parts. This notion has been captured by the idea of integrated information. Among different formulations of this idea, Integrated Information Theory (IIT) provides a way of measuring the irreducibility of a system down to its components (Tononi et al., 2016). Measuring a system's integration allows us to determine its holistic ability to affect its parts and processes, and thus its self-organizational level of autonomy (Aguilera and Di Paolo, 2019).

There are different proposals to capture integrated information, which we will label as Φ. A popular approach defines integrated information of a system as the information that is generated beyond the information generated by its parts. This can be measured by comparing the statistical distance of a probability distribution of the causal dependencies in the system, with the same distribution under the application of different partitions that disconnect the system into two independent parts (Oizumi et al., 2014). Applying this method to the transfer probability distribution of the states of a system yields

$$\Phi = I(S_{t+1}; S_t) = \min_{\text{cut}} \Big[D(P(S_{t+1}|S_t) || P_{\text{cut}}(S_{t+1}|S_t)) \Big] \tag{1}$$

Where D is a measure of statistical distance. Popular choices are the Kullback-Leibler divergence and the Wasserstein distance. Note that instead of S_t we could consider an arbitrary sequence of past states, but here we show just one state for simplicity.

Note that many different definitions of integrated information can be found in the literature (Oizumi et al., 2016; Mediano et al., 2019). Similar notions also have tried to calculate the balance between integration and segregation in a system measured using information theory or other methods like Granger causality (Tononi et al., 1994; Seth, 2011).

Dimension 2 - Interactive self-control

For our second dimension we nominate the capacity of self-control. To be interactively autonomous the system must show independence from - and to some degree control over - its environment, while remaining coupled with it.[2] For this purpose we propose Bertschinger's Non-Trivial Information Closure (NTIC) measure which tracks the system's control over its environment (Bertschinger et al., 2008; Chang et al., 2019). It is defined as the difference between the system's information about its own future, minus that information conditioned to the state of the environment.

$$NTIC = I(S_{t+1}; S_t) - I(S_{t+1}; S_t|E_t) \tag{2}$$

where I is a measure of information. A large value of NTIC implies that a system contains in itself information about its own future and that this self-predictive information contains the information about the effects of the environment in a system. That is, maximizing NTIC entails maximizing the predictive power of a system with respect to its environment, creating an asymmetric coupling between the two. Note that IIT and NTIC use different definitions of information (distance between original and partitioned system and mutual information). Further work could advance in unifying them under the same information measure (e.g. following Oizumi et al., 2016). Alternatively, one could consider capturing this dimension using notions as the minimization of sensory error from frameworks as Predictive Processing (Clark, 2013) and Active Inference (Friston et al., 2009). Finally, the control of an agent over the environment can also be considered from the notion of Empowerment (Salge et al., 2014).

Discussion and conclusion

The proposed measures can have a number of interesting applications. First, they can be used to assess or benchmark the autonomy of existing AI systems, and to systematically address issues of rights, responsibility, risks, etc. that are often associated to autonomous systems (particularly when these interact, as current AI systems increasingly do, within human environments). Second they can be used to steer (or avoid!) the development of AI systems towards increasing levels of autonomy. A crucial aspect of autonomy was purposefully left out of the scope of this abstract: the normative dimension. Work is still needed to investigate the relationship between proposed information-theoretic measures of autonomy and the emergence of norms within such systems.

Acknowledgements

XEB and MA acknowledge IAS-Research group funding IT 1228-19 from the Basque Government. XEB, DV and MA acknowledge funding from OUTONOMY research project ref. PID2019-104576GB-I00 by Spanish Ministry of Science and Innovation. MA was supported by MSCA grant 892715. DV acknowledges funding from the University of Zaragoza.

[2]Note that it would be insufficient for the system to simply be unaffected by the environment e.g. if its behaviour is governed by, lets say, a chaotic pattern generator

References

Aguilera, M. and Di Paolo, E. (2019). Integrated information in the thermodynamic limit. *Neural Networks*.

Barandiaran, X. E. (2004). Behavioral Adaptive Autonomy. A milestone on the Alife route to AI? In Pollack, J., Bedau, M. A., Husbands, P., Ikegami, T., and Watson, R. A., editors, *Artificial life IX: proceedings of the Ninth International Conference on the Simulation and Synthesis of Artificial Life*, pages 514–521, Cambridge, MA. MIT Press. ZSCC: 0000024[s0].

Barandiaran, X. E. and Moreno, A. (2008). Adaptivity: From Metabolism to Behavior. *Adaptive Behavior*, 16(5):325–344. ZSCC: 0000103.

Barandiaran, X. E. and Ruiz-Mirazo, K. (2008). Modelling autonomy: Simulating the essence of life and cognition. *Biosystems*, 91(2):295–304.

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2008). Autonomy: An information theoretic perspective. *Biosystems*, 91(2):331–345.

Boden, M. A. (2008). Autonomy: What is it? *Biosystems*, 91(2):305–308.

Chang, A. Y. C., Biehl, M., Yu, Y., and Kanai, R. (2019). Information Closure Theory of Consciousness. *arXiv:1909.13045 [q-bio]*.

Christensen, W. and Hooker, C. (2000). Autonomy and the emergence of intelligence: Organised interactive construction.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Collier, J. (2000). Autonomy and process closure as the basis for functionality. *Annals of the New York Academy of Sciences*, 901(1):280–290. ZSCC: 0000100.

Di Paolo, E. A. (2004). Unbinding biological autonomy: Francisco Varela's contributions to artificial life. *Artificial Life*, 10(3):231–233.

Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement Learning or Active Inference? *PLOS ONE*, 4(7):e6421.

Froese, T., Virgo, N., and Izquierdo, E. (2007). Autonomy: A Review and a Reappraisal. In Almeida e Costa, F., Rocha, L., Costa, E., Harvey, I., and Coutinho, A., editors, *Advances in Artificial Life*, volume 4648 of *Lecture Notes in Computer Science*, pages 455–464. Springer Berlin / Heidelberg.

Lee, J. H., Shin, J., and Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering*, 114:111–121.

Man, K. and Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1(10):446–452.

Mediano, P. A. M., Seth, A. K., and Barrett, A. B. (2019). Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy*, 21(1):17.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLOS Computational Biology*, 10(5):e1003588.

Oizumi, M., Tsuchiya, N., and Amari, S.-i. (2016). Unified framework for information integration based on information geometry. *Proceedings of the National Academy of Sciences*, 113(51):14817–14822.

OpenAI, Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., Pinto, H. P. d. O., Raiman, J., Salimans, T., Schlatter, J., Schneider, J., Sidor, S., Sutskever, I., Tang, J., Wolski, F., and Zhang, S. (2019). Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv:1912.06680 [cs, stat]*.

Salge, C., Glackin, C., and Polani, D. (2014). Empowerment–An Introduction. In Prokopenko, M., editor, *Guided Self-Organization: Inception*, Emergence, Complexity and Computation, pages 67–114. Springer, Berlin, Heidelberg.

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*.

Seth, A. K. (2011). Measuring Autonomy and Emergence via Granger Causality. *Artificial Life*, 16(2):179–196.

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461.

Tononi, G., Sporns, O., and Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91(11):5033–5037.

Varela, F. J. (1979). *Principles of Biological Autonomy*. Appleton & Lange.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.