Proceedings of Machine Learning Research 192:120–128, 2022 International Workshop on Self-Supervised Learning

The Holon System: Artificial General Intelligence as 'Work on Command'

Bas Steunebrink BAS@GENERALSYSTE.MS
Jerry Swan JERRY@GENERALSYSTE.MS
Eric Nivel ERIC@GENERALSYSTE.MS
General Systems Theory

Editor: Kristinn R. Thórisson

Abstract

Recent interest in the 'Large Language Models' of deep learning has led to widespread conjecture that artificial general intelligence (AGI) is thereby imminent. At the other end of the spectrum, it has also been claimed that 'general' intelligence cannot exist at all. In this extended abstract, we argue that both of these perspectives are misconceived. We provide a pragmatic definition of general intelligence, grounded in fundamental business and engineering requirements. We explain why a 'deployed regression model' (such as deep learning) cannot meet this criterion for *generality* of intelligence. We then proceed to describe the Holon system, designed and implemented to meet this criterion.

Keywords: Artificial General Intelligence

1. Introduction

The recent notable successes of machine learning (ML) has lead some to conjecture that it might be the appropriate technology for delivering general intelligence. The framework of machine learning has demonstrated undoubted economic value. However, we argue here that is nonetheless fundamentally at odds with any reasonable notion of intelligence. In a recent book (Swan et al., 2022) we argue that AI requires a fundamental change in perspective, mirroring that which took place in the philosophy of science in the mid-20th century (Popper, 1963). The essence of this change in perspective is to give primary emphasis to *compositionality*. Although this term is increasingly used within the machine learning community, this usage typically lacks the strong guarantees of the mathematical sense of the term. Informally, compositionality means that properties of aggregate structures can be reasoned about as a function of properties of their parts. A key requirement here is the ability to perform *reflective reasoning* — a system cannot be said to be compositional if it has no means of determining whether some desired property can be preserved. This has obvious implications for safety and alignment. In this extended abstract, we outline the principles underlying the design of Holon, a system which performs principled compositional reasoning, a reference implementation of which is described in Swan et al. (2022). This is achieved via hybrid symbolic-numeric inference mechanisms based on universal constructions from the mathematical discipline of category theory.

Philosophical considerations aside, intelligent machines are ultimately tools for implementing a new leap in *automation*, with generality being the metric of expected benefit. To that end, Holon delivers the following value proposition:

The Value Proposition for General Intelligence

For all practical purposes, *general intelligence* is a necessary property of a system which:

- Performs *'Work on Command'.*
- Can operate in environments which are not artificially constrained.
- Can adapt to novel (i.e., 'out-of-distribution') scenarios without incurring the manual labor/downtime of comprehensive re-training.
- Respects safety constraints.
- Is explainable and auditable.

The following sections provide a short and informal summary of key aspects.

2. 'Work on Command'

Since the very definition of intelligence and the nature (or even existence) of generality are the subject of much debate, it is helpful to start with an informal example. Consider tasking a system (which is currently tasked with assembling some widget X) as follows:

> *"Stop assembling X immediately: here's a specification of Y, and here are most of your old and a few new effectors. Now start assembling Y, avoiding such and such kinds of defects and wastage."*

We say that a system which can respond immediately to such a request is performing *'Work on Command'*, more specifically characterized as:

- The ability of a system to respond, at any time, to changes in task specification and/or operating environment.
- Changes can be both positive (goals to be achieved) and negative (constraints to be respected).
- The ability to leverage all relevant knowledge from prior tasks with little effort.

'Work on Command' affords a means of framing intelligence in the only practical context that matters: that of increased automation. The essential missing ingredient in automation via contemporary approaches is this *immediate responsiveness to change*. In contrast, contemporary machine learning yields a deployed artifact which simply performs a matrix-vector multiplication of the input vector on a matrix of weights. This matrix is fixed at the end of pre-deployment training, unchanging thereafter regardless of how many out-of-sample inputs it is subsequently exposed to. For any finite learning mechanism, this cannot accord with a meaningful definition of intelligence, which intuitively sees the importance

of appropriately timely responses to an ever-changing world. We therefore take the pragmatic stance that *generality of intelligence* is synonymous with the degree to which 'Work on Command' is possible. 'Work on Command' can be shown to generalize the framework traditionally used to describe Reinforcement Learning problems.[1]

'Work on Command' can be seen as a specialization of a large body of previous work on 'goal-directed[2] operation' (e.g., Akaishi and Hoshi (2017); Bhat and Mohan (2018); Fikes and Nilsson (1972)). For example, Ingham et al. (2006) describe a goal as "a constraint on the value history of a state variable over a time interval." The distinction is that 'Work on Command' additionally stipulates three things: specification, dynamic coupling, and responsivity. *Specification* of goals/constraints uses a language grounded in the state space, which is more general than Ingham's formulation, in that it can express relations in a hierarchical manner, including relations synthesized by the system itself and hence not known apriori to the system designer. The *dynamic coupling* of causal (i.e., task-agnostic) world knowledge and goals/constraints is to compute plans and control signals. *Responsivity* means the system responds immediately to change happening at any time (in both knowledge and goals/constraints) by redoing the coupling. Hence, while previous work on 'goal-directed operation' is similar in spirit, we use the term 'Work on Command' to refer to this additional functionality.

2.1. Second-Order Automation Engineering

In adopting the 'Work on Command' perspective, intelligence is naturally considered as the *process* of acquiring and adapting knowledge to serve time-bounded goals, in the presence of limited resources (Nivel et al., 2013). This recovers the intuitive notion of intelligence as the ability to make a dynamic trade-off between competing priorities.[3] By considering 'intelligence' and 'generality' as context-dependent *processes* rather than nouns, it becomes evident that those who argue that "even humans don't exhibit general intelligence"[4] are taking an overly-static perspective.

The ubiquitous task of Automation Engineering (AE) is the coupling of the physical world (chemical plants, assembly lines, logistics pipelines, smart cities, etc.) to a computational model which can be used for prediction and control. This typically requires considerable human expertise: in the past, it was the task of mathematicians to devise models (e.g., in the form of differential equations or constraints to be optimized over) which are then implemented computationally. More recently, Deep Learning has enjoyed considerable success in skipping the mathematical modeling phase and learning 'digital twins' of systems from vast quantities of operational data. However, it is typical that such data needs to be made amenable to the learning process, so there is still a requirement for human labor in the form of data science expertise. Despite this success, a deployed system trained via Deep Learning cannot accommodate change—whether in the problem to be solved or a significant drift in the distribution of inputs presented to it. Hence any significant change in the business requirements or operating environment means that the business will incur

1. See Swan et al. (2022), Chapter 6.
2. A.k.a. 'goal-oriented.'
3. Which includes at least *some* ability to discount the future cost of making the trade-off.
4. Online comment by Yann LeCun— accessed 9th August 2022.

costs: through the need for human expertise in retraining the system and/or the associated downtime.

In order to eliminate these costs and downtime (the latter being mission-critical in 'just in time' scenarios) it is necessary to *automate automation*, i.e., to enable control systems to self-reconfigure when they encounter such change. To this end, Second-Order Automation Engineering (2OAE) is our operational characterization of 'Work on Command'. 2OAE therefore necessarily imposes stringent requirements on the adaptive learning procedure, which should:

1. Adapt on-the-job to change in mission or environment.
2. Meet real-time task deadlines in real-world environments.
3. Generalize from and leverage existing knowledge.
4. Not incur 'catastrophic forgetting', in which learning a new task degrades competence at other tasks.

Given that both computational and physical resources (e.g., processing power and raw materials in an assembly-line scenario) are finite, an intelligent system must operate according to principles of 'bounded rationality' (Nivel et al., 2015), which requires that the system can perform a continual trade-off between solution quality and ability to meet deadlines. An intelligent system must therefore perform a 'continual refactoring' of its own knowledge and plans. It must balance its agency (i.e., the need to exert control over its environment to perform required tasks) against the limitations of its physical embodiment (e.g., computational resources and sensor-effector capabilities). This implies the following key properties of the learning architecture:

Self-interpretability

A learning process which is predominantly driven by sampling cannot scale to the long tails of real-world domains. Philosophically, sampling is in the 'empiricist' tradition of the scientific method. In contrast, what is needed is something closer in spirit to the mid-20th century revolution in the philosophy of science due to Karl Popper (Popper, 1963): the ability to introspect upon the *structure* of world models in order to reason beyond training examples. World models are thus continually revised in order to maintain *coherence* (Thórisson, 2021), both with respect to their internal structural relationships and how sensors and effectors relate to the external world. This continual maintainance, together with a bias towards 'simple, relevant' causal relationships (Steunebrink et al., 2013), means that these relationships are far more likely to be *generalized* statements about the world. By this means, there is a bias towards the abstraction of aspects of the world model which are relevant across recent tasks[5], which helps to address the 2OAE requirement to avoid 'catastrophic forgetting.'

Compositionality

The mathematical discipline for the study of compositionality is *category theory* (MacLane, 1971), which arose from the study of algebraic topology as a mathematical Rosetta stone, allowing properties of objects to be transferred between different reasoning contexts. It takes an operational perspective, in which entities are characterized by the manner in which they

5. More details on mechanisms for abstraction can be found in Chapter 9 of Swan et al. (2022)

interact with other entities, and in particular which structural properties are preserved via these interactions. Holon uses the methods of category theory to provide a principled approach to:

- Safety: the reasoning process can ensure that composition preserves safety properties (e.g., not entering some forbidden region of a factory floor, or ensuring that a controlled process remains at a certain setpoint).
- Generalization: given the ability to decompose a collection of observations into a collection of component parts, it is then more likely that these parts can be recombined to match unseen situations, whilst still retaining as many of the previously observed consistencies as possible. This then acts in support of 2OAE generalization requirements.

Endogenously Situated

The death of symbolic AI was heralded by Brook's 'Physical Grounding Hypothesis,' whereby it became common culture within the AI community that systems must be *situated* in order to gain meaning (Brooks, 1990). However, the appreciation of what this actually requires has lessened over time. A system is *situated* (Wang, 2009/06) when:

- It operates in the real world (i.e., complex, noisy, asynchronous environments).
- It has an end-to-end causal model of the sensor-effector mapping.
- This mapping is updated via feedback from the environment.

It is *endogenously situated* when the causal model also includes observations of the system's own *internal capabilities* (e.g., memory capacity, battery life, sensor sampling rate, lifting capacity, etc.). The combination of a causal feedback loop with an introspectable self-model means that prediction and control receive strong guidance from the joint constraints of the environment and the system's own representations of its reasoning. These constraints form a hierarchy, which then allows reasoning at a range of granularities and hence time-scales, thereby supporting the 2OAE requirement for real-time responsiveness.

2.2. Semantically Closed Learning

The requirements and desiderata above make it clear that what is needed for 2OAE is a system that can reason so as to "let its hypotheses die in its stead" (Popper, 1972), i.e., having the ability to internally envisage the consequences of an action rather than having to physically enact it by sampling the environment. This means that it must be possible for the system to reason directly about its representation of the world. This is achieved via proven methods of static analysis developed over decades of work in system verification. We use the term 'Semantically Closed Learning' (SCL) to refer to a system which:

- Can learn hierarchical structure which represents a compressed description of the world. This can be seen as a 'domain-specific language' pertinent to the task at hand.
- Can learn how to interpret that language for the purposes of prediction and control.
- Learns how to feed back information from the world in the vocabulary of that language.

More concretely, a *semantically closed learner* is a system equipped with a *stateful interpreter* for the learned representation language, such that:

- The next step in the state trajectory of the system is determined via the application of the interpreter.
- In the event of a prediction failure (e.g., unexpected success or wrong prediction), a 'repair' to the interpreter is achieved by updating interpreter state as a function of the discrepancy between predicted and actual states.
- As a result of this learning, the interpreter tends to be a better predictor.

Taken at face value, the above notion of 'repair' could be considered to be equivalent to traditional mechanisms of backpropagation. However, the essential distinction is that the repair can be mediated in a hybrid numeric-symbolic manner via a learned denotational semantics for the representation language, and hence be compositionally applied at increasingly hierarchical levels. The ability to respond to environmental surprises with representation change at such arbitrary scale can be seen as a form of *abductive* reasoning (for more detail, see Section 9.4 of Swan et al. (2022)).

In common with modern control theory, we adopt a state-space perspective. However (and in contrast to contemporary ML) the dimensions of the state-space can vary at any time: whether through the addition/deletion of sensors or effectors or through the system's own synthesis of new dimensions, for example to denote abstractions. Although the motivation for the term 'Semantic Closure' takes inspiration variously from Rene Thom's 'General Theory of Models' (Thom, 1972) and Patee's treatment of open-ended evolution (Pattee, 1995), we are pragmatically concerned here with a system that can perform useful work. Hence we take the perspective of 'AI as tool' rather than 'AI as organism': the more exploratory aspects of intrinsically-motivated learning are biased towards goal-relevance by user-imposed deadlines. For a more detailed explanation of the mechanics of SCL the reader is directed to Chapter 9 of Swan et al. (2022).

A Hybrid Architecture

There is increasing acknowledgement that symbolic approaches can complement the proven strengths of Deep Learning and there have been a number of recent approaches which attempt to 'get the best of symbolic AI and ML' (e.g., Csordás and Schmidhuber (2019); Paaßen et al. (2020); Franke et al. (2018)). In general, such approaches proceed quite literally, essentially re-implementing reasoning (and the necessary supporting representations) in the terms and components of ML. We proceed differently. Three aspects of SCL are of particular relevance to the reconciliation of ML and symbolic AI:

- Strong typing. Analogous to physics, in which different dimensions such as 'time' and 'electrical charge' are incommensurate.
- Fine-grained, open-ended, continual, and compositional inference.
- Emergent resource-aware and goal-directed attention.

We claim that these principles combine the strengths of both approaches: whilst SCL can be provided with prior domain knowledge in any desired form, it is not subject to the problems which plagued symbolic AI, since the ability to reflectively reason at the type level allows the sustained and progressive learning of invariants from the environment.

3. Conclusion

There has been much academic debate about the meaning (and even the very existence) of 'general' intelligence. We propose a practical, grounded definition via the ability to perform 'Work on Command' and describe its realization via the Holon system. In this extended abstract, we introduce the notion of 'Second-Order Automation Engineering' as the discipline of 'automating automation', and explain how this is facilitated via a situated, self-interpretable system which continually updates the compositional knowledge of its world model.

References

Rei Akaishi and Eiji Hoshi. Information seeking and simulation: Roles of attention in guiding a goal-directed behavior. *bioRxiv*, 2017. doi: 10.1101/104091. URL https://www.biorxiv.org/content/early/2017/01/29/104091.

Ajaz Ahmad Bhat and Vishwanathan Mohan. Goal-Directed Reasoning and Cooperation in Robots in Shared Workspaces: an Internal Simulation Based Neural Framework. *Cogn. Comput.*, 10(4):558–576, 2018. doi: 10.1007/s12559-018-9553-1. URL https://doi.org/10.1007/s12559-018-9553-1.

Rodney A. Brooks. Elephants Don't Play Chess. *Robot. Auton. Syst.*, 6(1-2):3–15, June 1990. ISSN 0921-8890. doi: 10.1016/S0921-8890(05)80025-9. URL https://doi.org/10.1016/S0921-8890(05)80025-9.

Róbert Csordás and Jürgen Schmidhuber. Improving differentiable neural computers through memory masking, de-allocation, and link distribution sharpness control. *CoRR*, abs/1904.10278, 2019. URL http://arxiv.org/abs/1904.10278.

Richard E. Fikes and Nils J. Nilsson. STRIPS: a new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3):189–208, 1972.

Jörg Franke, Jan Niehues, and Alex Waibel. Robust and scalable differentiable neural computer for question answering. In Eunsol Choi, Minjoon Seo, Danqi Chen, Robin Jia, and Jonathan Berant, editors, *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 47–59. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-2606. URL https://aclanthology.org/W18-2606/.

Michel D. Ingham, Robert D. Rasmussen, Matthew B. Bennett, and Alex C. Moncada. Generating requirements for complex embedded systems using state analysis. *Acta Astronautica*, 58(12):648–661, 2006. ISSN 0094-5765. doi: https://doi.org/10.1016/j.

actaastro.2006.01.005. URL https://www.sciencedirect.com/science/article/pii/S0094576506000257.

Saunders MacLane. *Categories for the Working Mathematician*. Springer-Verlag, New York, 1971. Graduate Texts in Mathematics, Vol. 5.

Eric Nivel, Kristinn R. Thórisson, Bas R. Steunebrink, Haris Dindo, Giovanni Pezzulo, M. Rodriguez, C. Hernandez, Dimitri Ognibene, Jürgen Schmidhuber, Ricardo Sanz, Helgi Páll Helgason, Antonio Chella, and Gudberg K. Jonsson. Bounded Recursive Self-Improvement. *CoRR*, abs/1312.6764, 2013. URL http://arxiv.org/abs/1312.6764.

Eric Nivel, Kristinn R. Thórisson, Bas Steunebrink, and Jürgen Schmidhuber. Anytime Bounded Rationality. In *Artificial General Intelligence*, pages 121–130. Springer International Publishing, 2015. doi: 10.1007/978-3-319-21365-1_13. URL https://doi.org/10.1007/978-3-319-21365-1_13.

Benjamin Paaßen, Alexander Schulz, Terrence C. Stewart, and Barbara Hammer. Reservoir memory machines as neural computers. *CoRR*, abs/2009.06342, 2020. URL https://arxiv.org/abs/2009.06342.

Howard Pattee. Evolving self-reference: Matter, symbols, and semantic closure. *Communication and Cognition - Artificial Intelligence*, 12:9–27, 1995.

K. R. Popper. *Objective knowledge: an evolutionary approach*. Clarendon Press, 1972.

K.R. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge classics. Routledge, 1963. ISBN 9780415285940.

Bas R. Steunebrink, Jan Koutník, Kristinn R. Thórisson, Eric Nivel, and Jürgen Schmidhuber. Resource-Bounded Machines are Motivated to be Effective, Efficient, and Curious. In Kai-Uwe Kühnberger, Sebastian Rudolph, and Pei Wang, editors, *Artificial General Intelligence - 6th International Conference, AGI 2013, Beijing, China, July 31 - August 3, 2013 Proceedings*, volume 7999 of *Lecture Notes in Computer Science*, pages 119–129. Springer, 2013. doi: 10.1007/978-3-642-39521-5_13. URL https://doi.org/10.1007/978-3-642-39521-5_13.

Jerry Swan, Eric Nivel, Neel Kant, Jules Hedges, Timothy Atkinson, and Bas Steunebrink. *The Road to General Intelligence*, volume 1049. Springer Studies in Computational Intelligence, 2022. ISBN 978-3-031-08019-7. URL https://doi.org/10.1007/978-3-031-08020-3.

René Thom. *Structural stability and morphogenesis - an outline of a general theory of models*. W. A. Benjamin, 1972. ISBN 0-201-40685-3.

Kristinn R. Thórisson. The 'explanation hypothesis' in general self-supervised learning. *Proceedings of Machine Learning Research, International Workshop on Self-Supervised Learning 2021*, 159:5–27, 2021. URL https://proceedings.mlr.press/v159/thorisson22b.html.

Pei Wang. Embodiment: Does a laptop have a body? In *Proceedings of the 2nd Conference on Artificial General Intelligence (2009)*. Atlantis Press, 2009/06. ISBN 978-90-78677-24-6. doi: https://doi.org/10.2991/agi.2009.44. URL https://doi.org/10.2991/agi.2009.44.