Emergence, Costs and Limitations of Self-Articulation

Vadim Bulitko

UNIVERSITY OF ALBERTA

August 18, 2025

Outline

- Introduction
  - team
  - context
  - target
- Levels-of-Computation Hypothesis
  - costs & limitations of articulation
- Testbeds & Preliminary Findings

Team

- Matthew Brown (UofA, neuroscience)
- Vadim Bulitko (UofA, CS / AI)
- Ramon Lawrence (UBCo, CS / databases)
- Shinichi Nakagawa (UofA, evolutionary biology)
- Roscoe Smith (UofA, CS / AI)
- Shway Wang (UofA, CS / AI)
- William Yeoh (WashU, CS / AI)
- Michael Youngblood (Filuta AI, CS / AI)

Context

- Neurosymbolic AI [Garcez and Lamb 2023]
  - flexibility and power of neural ML
  - explainability and portability of symbolic AI
- Program synthesis
  - per-problem algorithm design [Bulitko et al. 2022]
  - algorithm discovery [Stevens, Bulitko, and Thue 2023]
- Multi-agent systems
  - communicate among themselves [Sirota et al. 2019]
  - communicate to humans [Vasileiou and Yeoh 2023]
- **Downsides**
  - ML/synthesis/articulation algorithms are human-constructed
    - programmatic RL [Verma et al. 2019]

Our Target: *Emergent* Learning & Self-Articulating Agents

- Learn
  - individual learning
  - social learning
- Communicate
  - among themselves
  - with humans
- Articulate/explain their behaviour
  - to other agents
  - to humans
- **All components emergent** (i.e., not human-constructed)
  - learning
  - articulating
  - communicating

Hypothesis: Levels of Computation

- Critical task
  - agent can do it
  - agent cannot articulate how it does it
- For any cognitive agent a critical task exists
- Two agents belong to cognitive level i when
  - neither can articulate the other's critical task
- Level $i+1$:
  - agents at level $i+1$ can articulate critical tasks for agents at level i
  - smallest increase of complexity from i to $i+1$

Recursion Theory

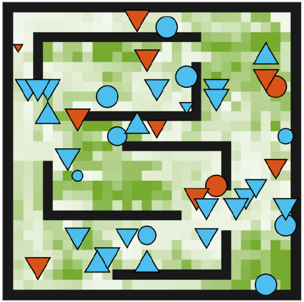
- Computability of functions [Rogers 1987]
- a Turing machine (TM) computes $\varphi : \mathbb{N} \to \mathbb{N}$ functions
  - all such functions can be integer-indexed: $\varphi_0, \varphi_1, \ldots$
- a set $W \subseteq \mathbb{N}$ is recursive iff a TM program can check membership in it
- a set $W \subseteq \mathbb{N}$ is recursively enumerable iff a program can enumerate its members
  - if a set $W \subseteq \mathbb{N}$ is recursively enumerable but not recursive then there exists φ_i
    - $\varphi_i(m) = 1$ when $m \in W$
    - $\varphi_i(m)$ does not stop when $m \notin W$
- What about the set $K = \{i \mid \varphi_i(i) \text{ halts}\}$?

Recursion Theory

- $K = \{i \mid \varphi_i(i) \text{ halts}\}$ is recursively enumerable but not recursive
- Now consider a TM with an oracle $A \subseteq N$ (denote it by TM^A)
  - on it $\varphi_i(m)$ computes normally but can query if $j \in A$ in the process
- Is TM^K more powerful than TM?
  - TM^K can compute everything that TM can
  - TM^K can also compute things that TM cannot
    - K becomes recursive for TM^K
- Analogy
  - agent doing a task ~ enumerating members of a set
  - agent articulating a task ~ checking membership in a set
  - critical task: enumerating members of K
  - cognitive level i ~ TM
  - cognitive level $i + 1$ ~ TM^K

A-life

- **Base task**: survival in A-life [Ackley and Littman 1991; Wilensky and Rand 2015]
  - 2D world with grass and agents
- **Control module**
  - feedforward ANN
  - specified by the agent's gene
  - evolved ANN architectures and weights
- **Articulation module**
  - neural formula synthesizer (FC or transformer)
  - externally developed
- **Interpretation module**
  - manually coded formula interpreter
    - to update the ANN policy
    - within life-time learning



A-life: Neural Agents

$$\begin{matrix} 0.233 & 0.732 & 0.875 \\ 0.003 & 0.206 & 0.549 \\ 0.485 & 0.399 & 0.439 \end{matrix} \rightarrow 2$$

$$\begin{matrix} 0.843 & 0.872 & 0.522 \\ 0.498 & 0.023 & 0.738 \\ 0.783 & 0.934 & 0.457 \end{matrix} \rightarrow 4$$

$$\begin{matrix} 0.196 & 0.085 & 0.382 \\ 0.332 & 0.001 & 0.249 \\ 0.032 & 0.118 & 0.489 \end{matrix} \rightarrow 1$$

$$\vdots$$

A-life: Formula Fitting

$$\begin{matrix} 0.233 & 0.732 & 0.875 \\ 0.003 & 0.206 & 0.549 \\ 0.485 & 0.399 & 0.439 \end{matrix} \rightarrow 2$$

$$\begin{matrix} 0.843 & 0.872 & 0.522 \\ 0.498 & 0.023 & 0.738 \\ 0.783 & 0.934 & 0.457 \end{matrix} \rightarrow 4$$

$$\begin{matrix} 0.196 & 0.085 & 0.382 \\ 0.332 & 0.001 & 0.249 \\ 0.032 & 0.118 & 0.489 \end{matrix} \rightarrow 1$$

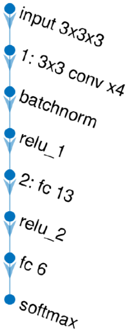
$$\vdots$$

$$\operatorname{argmax}\left(g_{\leftarrow}, g_{\uparrow}, g_{\rightarrow}, g_{\downarrow}\right)$$

A-life: Articulation

$$\text{argmax}\left(g_{\leftarrow}, g_{\uparrow}, g_{\rightarrow}, g_{\downarrow}\right)$$

Why Articulate?



- Self-reflection via articulation can be useful
  - neural ↻ symbolic learning [Verma et al. 2019]
  - enables knowledge-based bias
- Articulation is important for explainable AI
- Articulation enables knowledge transfer (e.g., parenting)
  - teacher: neural → symbolic
  - learner: symbolic → neural

$$\operatorname{argmax}\left(g_{\leftarrow}, g_{\uparrow}, g_{\rightarrow}, g_{\downarrow}\right)$$

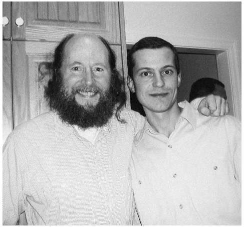
Articulation of Critical Tasks

- Agent articulating $\circlearrowright$ interpreting
  - neural $\circlearrowright$ symbolic
  - symbolizations can be simplifications/abstractions
  - symbolizations must be simplifications/abstractions for critical tasks
    - due to articulation/interpretation overhead
  - thus unable to fully symbolize neural knowledge
    - for survival (a critical task)
  - need additional experiential learning (neural)

- Human education
  - listening is not enough
  - learning via doing

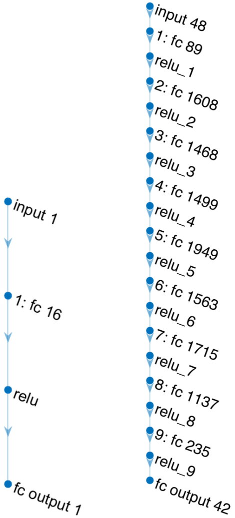
The Bitter Lesson

- No agent is able to articulate its own critical tasks
- Failure of AI based on human idea of human reasoning [Sutton 2019]
  - “We have to learn the bitter lesson that building in how we think we think does not work in the long run.”
  - “The second general point to be learned from the bitter lesson is that the actual contents of minds are tremendously, irredeemably complex; we should stop trying to find simple ways to think about the contents of minds, such as simple ways to think about space, objects, multiple agents, or symmetries.”



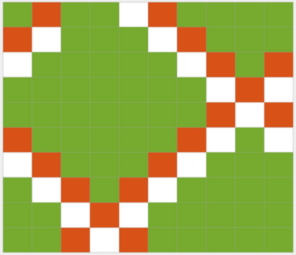
How to Articulate?

- program synthesis
  - no need for pre-training
  - slow
  - human engineered
  - unreliable
- neural distillers (FC or transformers)
  - fast
  - the same hardware: can be evolved (in principle)
  - massive pre-training
    - **0.31** $\times$ **10^6** training data (I/O pairs)
  - massive in size
    - FC: **49** versus **15** $\times$ **10^6** ANN weights
    - unlikely to emerge on the same evolutionary scale
    - would kill the agent via energy depletion



A Simpler Testbed: 1D A-life

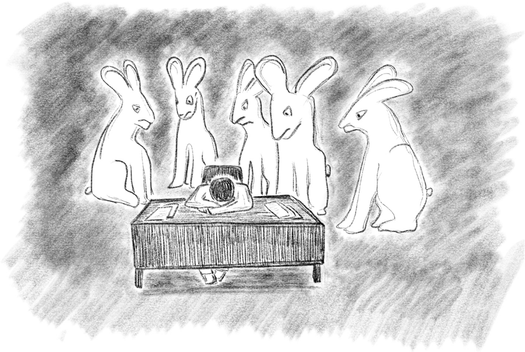
- **Base task**: survival in A-life
  - a binary 1D torus
  - agent sees grass left/right ($n = 2$ inputs, $2^n = 4$ states)
- **Control module**
  - what do **I** do in state s?
  - truth table ($2^n = 4$ rows)
  - policy $\pi : \mathbb{B}^2 \to \mathbb{B}$
- **Interpretation module**
  - what would **another agent** π do in state s?
  - truth table ($2^{2+2^n} = 64$ rows)
  - universal policy $\pi^U : \mathbb{B}^2 \times \mathbb{B}^{2^n} \to \mathbb{B}$
- Both are possibly small enough to emerge in evolution



Conclusion

- Self-explaining AI agents to *emerge*
- Costs and limits of self-explanation
- A hierarchy of computational levels

- `bulitko@ualberta.ca`



Acknowledgments

- Valeriy Bulitko
- Jonathan Schaeffer
- Evelyn Chiew
- Ethan Chung
- Emma Reid
- Dinara Shukayeva

- NSERC
- CERC
- DRAC

Bibliography

- Ackley, D. and M. Littman (1991). “Interactions between learning and evolution”. In: *Artificial life II* 10, pp. 487–509.
- Bulitko, V. et al. (2022). “Portability and explainability of synthesized formula-based heuristics”. In: *Proceedings of SoCS*. Vol. 15. 1, pp. 29–37.
- Garcez, A. and L. Lamb (2023). “Neurosymbolic AI: The 3 rd wave”. In: *AI Review* 56.11.
- Rogers, H. Jr. (1987). *Theory of recursive functions and effective computability*. MIT Press.
- Sirota, J. et al. (2019). “Towards procedurally generated languages for non-playable characters in video games”. In: *Proceedings of CoG*.
- Stevens, J., V. Bulitko, and D. Thue (2023). “Solving Witness-type Triangle Puzzles Faster with an Automatically Learned Human-Explainable Predicate”. In: *arXiv preprint arXiv:2308.02666*.
- Sutton, R. (2019). “The Bitter Lesson”. In: URL: `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`.
- Vasileiou, S. L. and W. Yeoh (2023). “PLEASE: Generating Personalized Explanations in Human-Aware Planning”. In: *Proceedings of ECAI*.
- Verma, A. et al. (2019). “Imitation-projected programmatic reinforcement learning”. In: *Proceedings of NeurIPS*.
- Wilensky, U. and W. Rand (2015). *An Introduction to Agent-Based Modeling*. MIT Press.