

# Towards a Checklist of AGI Implementation - Can a Critic Become a Solutionist?

Janne P. Hukkinen – inrobotico.com, Helsinki

## Describing AGI Agent & Environment

### Agency<sup>1</sup>

- **Spirit:** Virtual control/operating system of any agent (cells, animals, humans, families, cities, ecosystems, corporations, nation states, ...)
- **Agency:** Cybernetic control plant in feedback relation to environment
- **Sentience:** Agent discovers itself in and its relationship to the world
- **Consciousness:** Agent aware of own attention, able to control it. Creates coherent interpretation. *Maintains indexed memory for disambiguation, learning, and reasoning.* Mediates knowledge within mind.
- **Self:** self-image, 1st person perspective. Content modulated when agent turns intentions to actions. Between discovery of own existence and deconstruction of the self-representations.
- **Emotions:** content & expressions are **learned**, on top of low level bodily **valence**<sup>2</sup>
- **Mental models** of self & world create subjective reality
- **Generic Reward is enough** hypothesis<sup>3</sup> & Motivations<sup>4</sup>

### Environment

- Dimension scales have alternatives.  
**Observations** : discrete – continuous  
**Actions** : discrete – continuous  
**Time** : discrete – continuous  
**Dynamics** : deterministic – stochastic – chaotic  
**Observability** : full – partial  
**Agency** (others') : single – multiagency  
**Uncertainty** : certain – uncertain  
**Reality** : simulated – real-world  
• ...<sup>3,5,4</sup>

### Communication

- **Signaling**
- **Signal Combinations**
- **Symbols** are physical entities on sensory modality; labels for concepts
- **Language** as multi-level rule & symbol system: phonology, morphology, universal grammar

### World Knowledge, Representation

- What do we know about world structure and dynamics?
- **Motor control** of body, movement, locomotion
- **Mental models** of environment & self<sup>21</sup>
  - **Causal Systems**/Networks
- 3+D physical world (sensory modalities + time)
  - **Objects**
  - **Affordances**
- **Social** World (human culture), other agents
- **Abstract Concepts**
- **Energy**
- **Survival**
- **Resolution**<sup>4</sup> of time, space, information channel width, world knowledge, decision making, etc.

### Hardware

- **Sensors, Actuators, Signaling**
- **Motor Control**, multiscale time and space resolutions
- **Embodied Learning:** body constrains and modulates learning
- **Implicit Computation** by physical & mechanical properties of body

### Cognitive Capabilities

- **Perception**
- **Abstraction, Conceptualization, Objectification**
- **Learning**
- **Memory** (sensory, motor, experiential, episodic, procedural)
- **Mental simulation**
- **Reasoning, Planning**
- **Navigation**
- **Causality:** interaction between mental models

## Why AGI?

### Why Build?

- Complex system is **best understood** by modeling it. Building a system **reprioritizes** and explicates what we don't understand (mechanisms instead of narratives<sup>7</sup>)
- AGI agent needs to be run in the world for **alignment testing** with *world dynamics (aesthetics), which is extrapolated from highest level purposes of civilization*<sup>1</sup>
- White hat security: **Improve security and ethics** by trying to break/missuse a working system
- Not building does not **protect us from adversarial** and unethical entities using AGI systems against us.

### Why Checklist?

- Many AGI models exist, but have gaps. **Can you find any right now? How would you build AGI?**

### Definition, Criteria

- Human-level or super-human behavior and *adaptation with insufficient knowledge and resources*<sup>8</sup> in undefined environments and tasks
- **native** (system information content) vs. **performance intelligence**<sup>9</sup>
- Computational part of reaching goals adaptively<sup>9</sup>
- *Hypothesis: generic objective of maximizing reward is enough for AGI*<sup>3</sup>

### Goals

1. Minimize & explicate unknowns.
2. Help design & evaluation (of functionality, ethics, progress).

## AGI Design Thinking

### AGI Design Thinking: Modular “Designed Organization”

1. **Define application** requiring AGI (cognitive goal/task/problem)
2. **Empathize environment**, world-knowledge, and cognitive capabilities required (from human intuitive to explicit technical) by (1)
3. **Create descriptive functional system's architecture** (high-level intuition-pumped human-inspired design narrative aid)
4. Make an **inventory of algorithms** and hardware available
5. **Divide labor & orchestrate** computational modules
6. **Operationalize cognitive architecture:** specify software & hardware
  - goal/task
  - perception (world & self)
  - data processing
  - learning, cognitive scaffolding
  - orchestration
7. **Try to implement.**
8. **Iterate**

- **Key Problem: Orchestration:** How can we know beforehand whether a particular architecture actually works?

### AGI Design Thinking: Cybernetic “Constrained Organization”

1. **Identify purposes**<sup>1</sup>/ rewards<sup>3</sup>/goals on the highest level
2. **Empathize RICH environment** for agent(s)<sup>3</sup> to facilitate & constrain learning
3. Make an **inventory of algorithms** and hardware available. Evaluate with theoretical *Universal AI*<sup>5</sup> [hutter\_universal\_2005]
4. **Decide: 1 or >=2 agents**<sup>1,3,5</sup>
5. **Divide labor & orchestrate** *Society of Mind* for >=2 agents
6. **Operationalize agent(s):** specify software & hardware
  - purpose/goal, generic reward
  - environment (world & self)
  - perception
  - cybernetic agent
  - reinforcement learning
  - algorithms;
7. **Try to implement.**
8. **Iterate**

- Training in interaction with environment
- **Reward-is-enough hypothesis:** Rich environment and generic reward/goal is enough for AGI. When agent gravitates towards reward, sub-goals/skills are learned implicitly on the way.<sup>3</sup>
- **Key Problem: How to partition & scaffold learning space?** Scaffolding of learning, to monitor orchestration & transparency.

### Intuition Pump

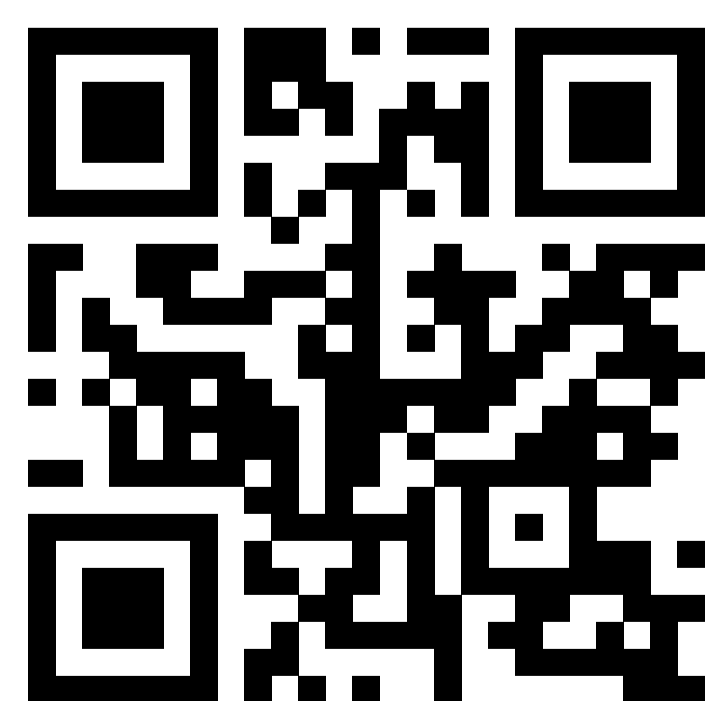
- Pros: Helps in designing architecture.
- Cons: Illusion of explanation. Only the tip of the iceberg of cognition and the world is visible.

### Open Problems

- Perception, Learning, Architecture<sup>1</sup>, Orchestration
- Values & priorities of civilization?<sup>1</sup>

## Cognitive Architecture of Modules/Agents

- A **set of modules/agents** comprise complete AGI agent (*society of mind*<sup>1,6</sup>)
- No sentience, self, consciousness, etc. for sub-modules/sub-agents
- **Divide labor** between modules/agents
- Orchestration
- Marrian computational levels
- **purpose**
- **algorithm**, 1 per module/agent
- **implementation/hardware**
- **Reward-is-enough** framework<sup>3</sup>



## References

- (1) Science, Technology & the Future. Joscha Bach - Agency in an Age of Machines, 2022.  
(2) Feldman-Barrett, L. *How Emotions Are Made: The Secret Life of the Brain*; Pan Macmillan, 2017.  
(3) Silver, D.; Singh, S.; Precup, D.; Sutton, R. S. Reward Is Enough. *Artificial Intelligence* 2021, 299, 103535.

- (4) Dörner, D.; Güss, C. D. PSI: A Computational Architecture of Cognition, Motivation, and Emotion. *Review of General Psychology* **2013**, *17* (3), 297–317.
- (5) Hutter, M. Universal Artificial Intelligence, 2016.
- (6) Minsky, M. *Society of Mind*; Simon; Schuster, 1988.
- (7) Rooij, I. van. Psychological Models and Their Distractors. *Nature Reviews Psychology* **2022**, *1* (3), 127–128.
- (8) Wang, P. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* **2019**, *10* (2), 1–37.
- (9) Legg, S. Machine Super Intelligence. PhD Thesis, Università della Svizzera italiana, 2008.