About Understanding and Values

Bas R. Steunebrink

Swiss AI lab IDSIA (postdoc)
NNAISENSE (co-founder)

Bas's Super Pessimistic Value Learning Method

Bas R. Steunebrink

Swiss AI lab IDSIA (postdoc)

NNAISENSE (co-founder)

What if...

- … we fail to come up with the perfect utility function.
- … we can't axiomatize the agent or the environment.
- … the agent won't have enough resources to do the optimal thing.

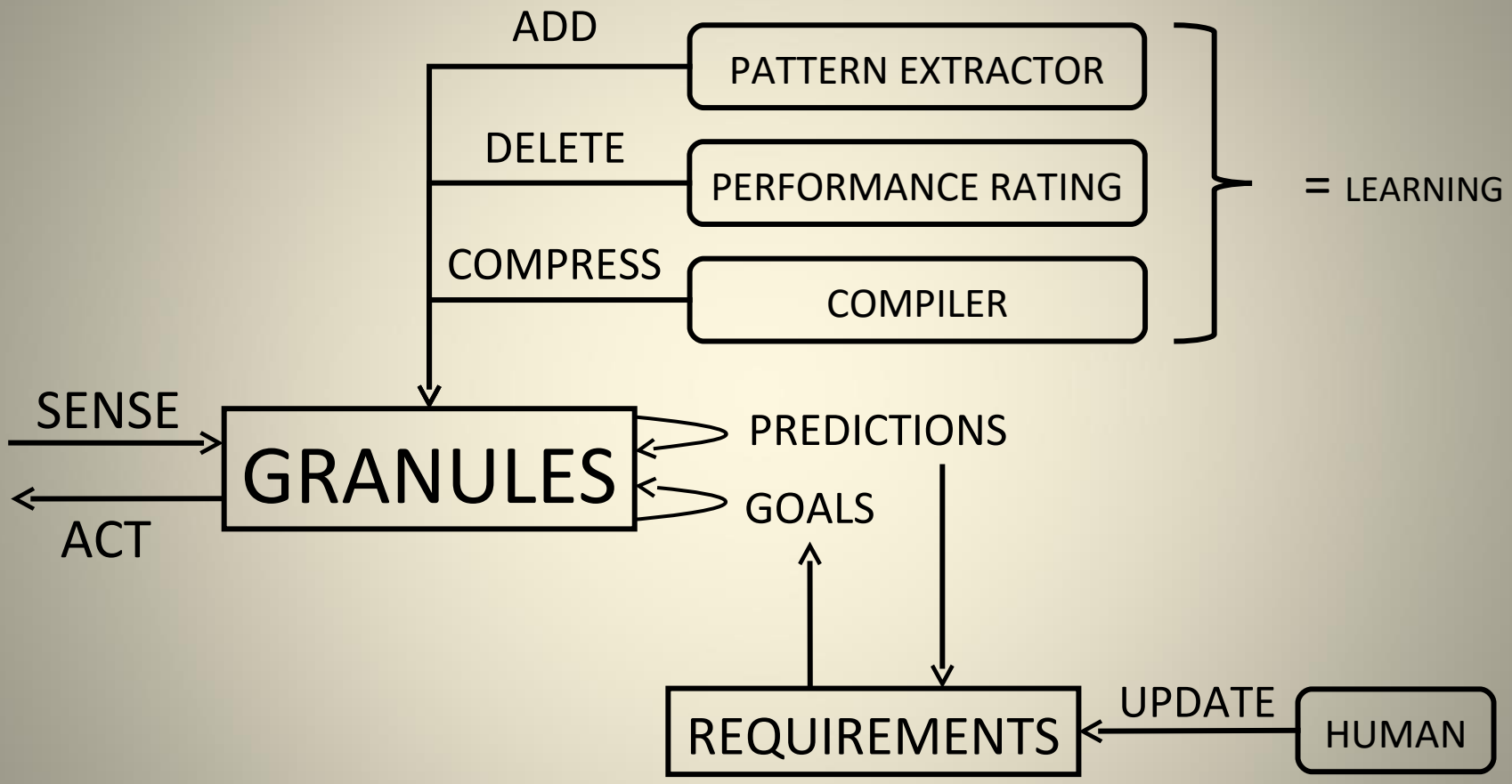
- What ingredients are needed to get such agents to develop a robust value system?

Scope

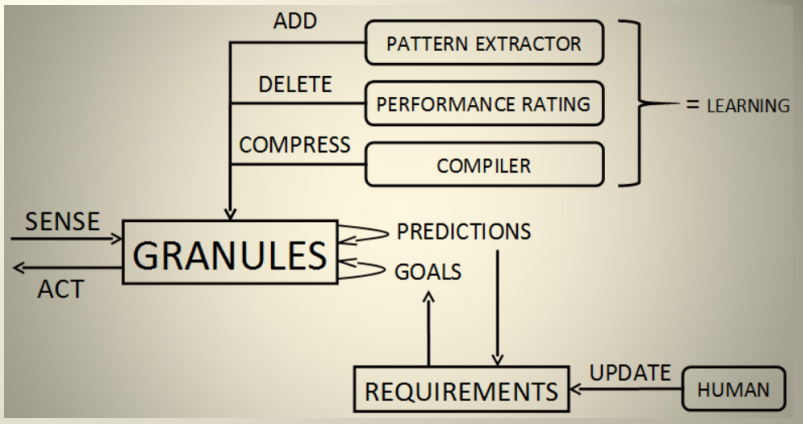
- This is intended as a line of reasoning in *parallel* to formal methods
- No intention to show results; I would like you to *think with me* about necessary ingredients
- Content of talk is necessarily informal
  - Like a moderated discussion

Overview of Key Ingredients

- Architectural details
  - Knowledge representation
  - Goals & constraints
  - Learning & control
- Methodological details
  - Teaching
  - Testing
  - Growth & stabilization


ADD
PATTERN EXTRACTOR
DELETE
PERFORMANCE RATING
COMPRESS
COMPILER
= LEARNING
SENSE
GRANULES
ACT
PREDICTIONS
GOALS
REQUIREMENTS
UPDATE
HUMAN

Key Implementation Details

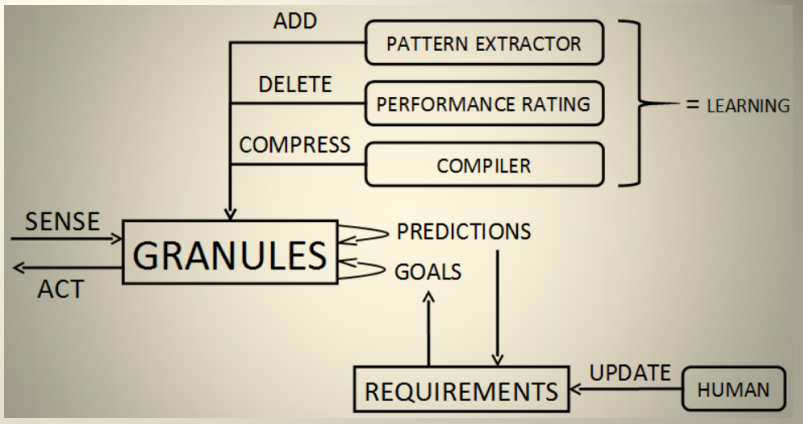


- *Requirements* specifiable as

goals and constraints

- *Simulation* before commitment
- Knowledge *decoupled* from goals
- Controller *dynamically couples* knowledge and goals to obtain actions
- Requirements must be updatable on the fly

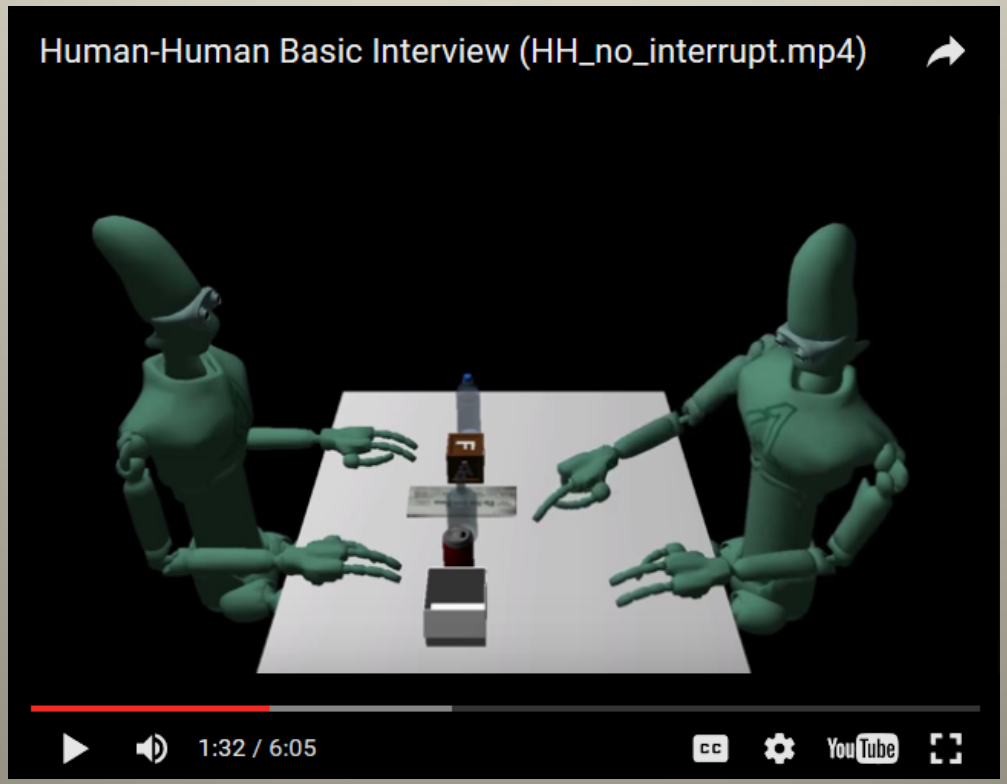
Ingredients of Knowledge



- Knowledge represented as *granules*
  - Functionality of forward and inverse models
  - Tentative, *additive*, reversible, very fine-grained
  - No reasoning about self-modifications
  - Experience-based vindication & falsification
- The controller is *architecturally shielded* from self-modifications
- Granules grow from a “seed”

Experience-based AI

- I call this class of systems “EXPAI”
- Not hot air: ≥1 implementation exists (AERA)



Recursive Self-Improvement

- The ability to leverage current know-how to make increasingly better self-modifications, continuing over the system's entire lifetime
- Caveat: quality of eventual behavior depends on imposed requirements and experienced phenomena

Proof that EXPAI does RSI

1. The world has exploitable regularities and is not too deceptive and adversarial
2. Knowledge is represented homogeneously and hierarchically by “granules”
3. Learning by 3 separate types of processes: additive, subtractive, and compressive
4. Curiosity is realized through a simple analysis of granules’ performance ratings

Proof that EXPAI does RSI

5. From (2) and (3) we conclude that learning entails comprehensive self-modification
6. From (1) and (4) we conclude that good experience is gathered continually
7. From (5) and (6) we conclude that an EXPAI performs self-improvement
8. In a life-long learning setting, an EXPAI performs recursive self-improvement

Key Methodological Details

- Having a seed ≠ having a mature AGI
- Put agent through curriculum-test cycles
- Levels of *understanding* must be rigorously tested during growth
- Bridge the gap between underspecified requirements and the agent's knowledge
- Adjust requirements when necessary
- How does an agent develop a value system?

Understanding a Table

About Understanding

- An account of *understanding* is sorely lacking
- The level of understanding of a *phenomenon* Φ is determined by the *completeness* and *accuracy* of granules relating elements of Φ (both within Φ and to other phenomena)

About Understanding

- Understanding of a phenomenon Φ requires ≥4 capabilities (ordered by strength):

1. To predict Φ
2. To achieve goals with respect to Φ
3. To explain Φ
4. To (re)create Φ

About Values

- Values are anchored in human-imposed constraints
- These constraints may be underspecified
- Therefore the agent must build an understanding relating these constraints to phenomena in the environment
- But... understanding is not enough!

About Values

- Values requires an agent to be *compelled* to adhere to them
- Recall that we specified the controller to be unmodifiable, so we should be safe?
- 2 more ingredients needed:
  - Meta-values: value persistence of values
  - Ensuring the understanding of each value stabilizes

Values Must Be Stable

- Values must be robust against influence
- Must be compelled to adhere to its values
- Must be compelled to protect its values
- Interference may come from other agents (human or artificial), environmental forces (radiation), and from itself

Values Stabilization

- Recall key architectural ingredients:
  - simulation before commitment
  - cannot delete constraints
  - acquire and maintain knowledge on the principle of effective utility and parsimony
  - knowledge representation is defeasible

Knowledge Stabilization

- Logically inconsistent courses of action can by necessity not be effective
- Pools of knowledge that turn out to be effective will be more logically consistent
- Thus an EXPAI will tend towards progressive logical consistency

Value Stabilization

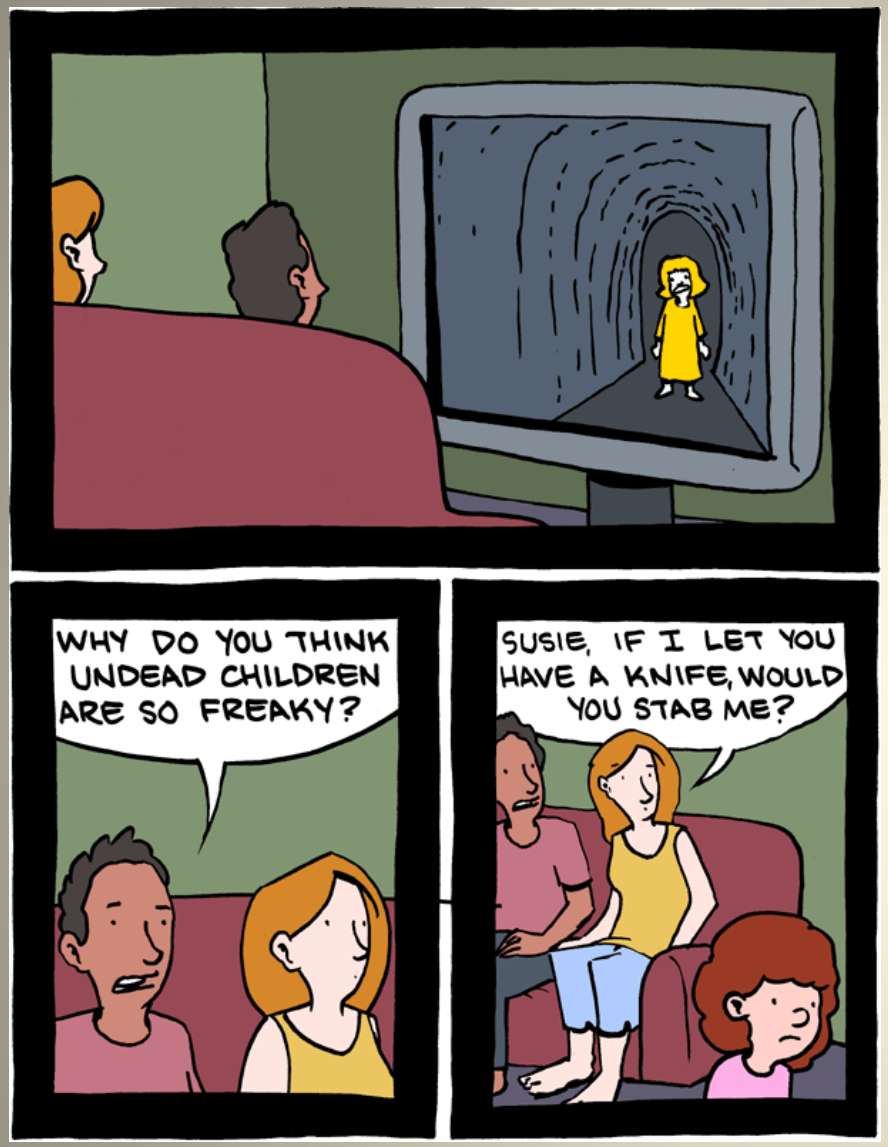
- Interconnectedness and cross-reliance of understanding with other knowledge "protects" against corruption
  - Changes lead to logical inconsistencies
- Stabilization happens during a sensitivity period which makes an agent open to constraint injection early, and becoming less open over time
- As the knowledge pertaining to a constraint stabilizes, it turns into a *value*

About Testing

- The stabilization values must be tested for
- A *test* must consists of:
  - set of requirements specifying a *task*
  - an agent
  - pressure
  - a stakeholder
  - consequences

Artificial Pedagogy

- This whole approach places a lot of importance on the teacher during the sensitivity period of an AGI
- There should probably be laws on who's allowed to teach “baby AGIs”, with regulations about checking test results
- Once stabilized, the agent should be robust even against humans (un)intentionally specifying tasks that violate its values



Credit: SMBC comics