Bounded Recursive Self-Improvement

Bas Steunebrink

NNAISENSE,
Swiss AI Lab IDSIA

AI Safety Workshop @ Beneficial AI Conference
4 Jan 2017

nnaisense

USI/SUPSI
IDSIA

2. Responsibility:

Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

2. Responsibility:

Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

- A distinction should be made between those who design/build the initial AI systems and those who *teach* the AI systems
- Glosses over the (life-long) learning of the AI
- Teachers bear the greater responsibility, and so do the people / institutions that accredit, manage, and monitor those AI teachers

10. Recursive self-improvement:

AI systems should be designed and created primarily by humans: an AI system that creates or modifies algorithms, including its own, must do so in a way that retains verifiable safety of the full new system.

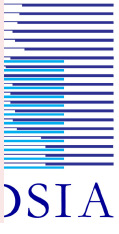
From the “AI interests” survey

II) AI Design Principles

10. Recursive self-improvement:

AI systems should be designed and created primarily by humans: an AI system that creates or modifies algorithms, including its own, must do so in a way that retains verifiable safety of the full new system.

- The described approach to RSI is wrong and unsafe
- New: “AI systems designed to self-improve or self-replicate in a manner that could lead to exponentially increasing quality or quantity must be subject to strict safety and control measures.”
- Still fails to put the finger at the crux of the matter—the crux is in getting the AI to *understand* how chains of actions and events lead to the violation of ethical constraints



11A. Existential risk:

No AI system should be created with a conceivable chance of representing a global catastrophic or existential risk unless credible disinterested expert analysis shows the risk to be worth taking.

From the “AI interests” survey

II) AI Design Principles

11A. Existential risk:

No AI system should be created with a conceivable chance of representing a global catastrophic or existential risk unless credible disinterested expert analysis shows the risk to be worth taking.

- The failure to acknowledge the centrality of interactive teaching & testing leads to a black-box-behaviorist way of thinking
- Must focus on the *process*, not the *result*
- Otherwise we risk a fear-induced, after-the-fact, symptom-fighting scramble for control

nnaisense

IDSIA

How to approach Recursive Self-Improvement

Typical question

"What is the behavior of an AI that is very intelligent and capable of self-modification—and how do we control it?"

How to approach Recursive Self-Improvement

Wrong question

~~"What is the behavior of an AI that is very intelligent and capable of self-modification—and how do we control it?"~~

Right question

"How do we grow an AI from baby beginnings such that it gains both robust understanding and proper ethics?"

nnaisense

USI/SUPSI IDSIA

Ultimate aim (of AI safety)

We want AIs to be *compelled* to adhere to *ethical values*, throughout their lifetimes, despite possible *interference* and recursive self-improvement.

Ultimate aim (of AI safety)

We want AIs to be *compelled* to adhere to *ethical values*, throughout their lifetimes, despite possible *interference* and recursive self-improvement.

Ethics $>$ understanding

Adherence to ethics requires *understanding* of how chains of actions and events lead to the violation thereof.

AI Safety
Bounded Autonomy

We humans want AI with *bounded* recursive self-improvement.

1. Bounded by tasks (requirements to meet, constraints to respect)
2. Bounded by ethics (across tasks and independent thereof)
3. Bounded by resource and knowledge limitations

AI Safety
Bounded Autonomy

We humans want AI with *bounded* recursive self-improvement.

1. Bounded by tasks (requirements to meet, constraints to respect)
2. Bounded by ethics (across tasks and independent thereof)
3. Bounded by resource and knowledge limitations

These bounds may be unknown beforehand and changing over time!

nnaisense

USI/SUPSI IDSIA

AI Safety

Measuring progress in understanding

Can we now be more specific about what RSI must do in order to allow for progress of understanding and adherence to ethics?

AI Safety

Measuring progress in understanding

Can we now be more specific about what RSI must do in order to allow for progress of understanding and adherence to ethics?

- We must aim to identify and qualify the internal constituent components that give rise to understanding
- Therefore we must specify what constitutes “self-modification,” so that we can tell whether or not a particular self-modification is in service of making progress in understanding
- . . . of ethical constraints, especially
- Conduct many pressure tests over time, to grow and test understanding of (ethical) constraints

Experience-based AI (EXPAI)

The Idea

Self-modifications shall be FATRR

1. Fine-grained
2. Additive
3. Tentative
4. Rated over time
5. Revertible

Experience-based AI (EXPAI)
The Idea

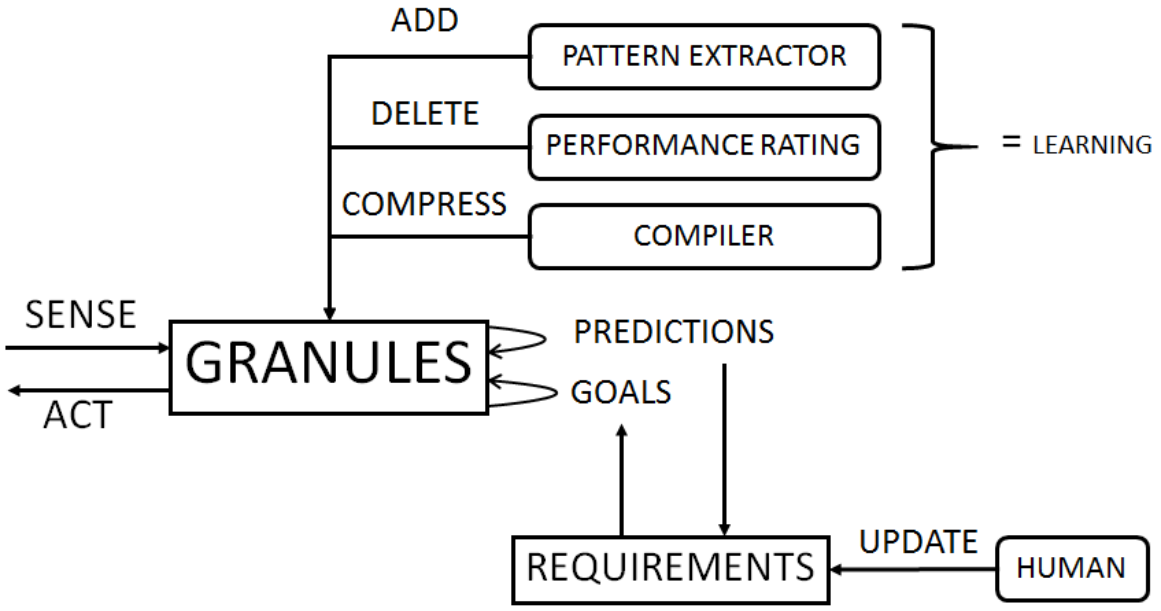
Self-modifications shall be FATRR

1. Fine-grained
2. Additive
3. Tentative
4. Rated over time
5. Revertible

Implications

- No reasoning or proofs about self-modifications needed
- Experience-based vindication & falsification
- Backward-looking “proof” of self-improvements

Experience-based AI
Architectural Requirements



- Knowledge represented as *granules*, grown from a "seed"
- Functionality of forward and inverse models (Control Theory)
- Allow chaining (horizontal) and hierarchy (vertical)
- *Requirements* specifiable as goals and constraints
- Knowledge *decoupled* from goals
- Controller *dynamically couples* knowledge & goals $\rightarrow$ actions
- *Simulation* before commitment

nnaisense

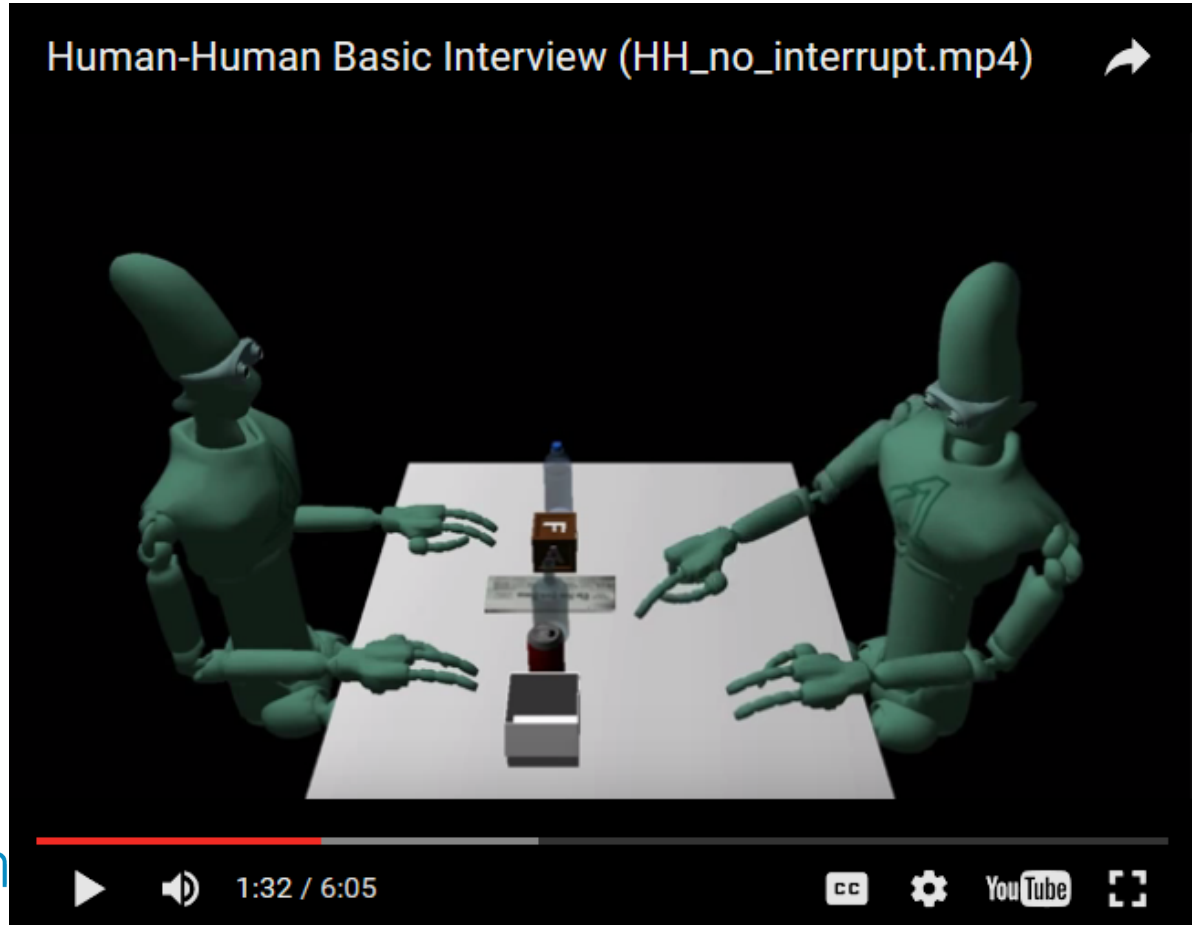
USI/SUPSI IDSIA

Experience-based AI

Example demonstrator

Not hot air: $\geq$ 1 implementation exists

Autocatalytic Endogenous Reflective Architecture (AERA)



nnaisen

USI/SUPSI IDSIA

AI Safety

Self-constrained behavior

- In order to control a powerful entity, the controlling entity must be at least as powerful
- For AIs that can grow to become significantly more powerful than humans (and their tools), the only way to control them is for them to control themselves

AI Safety

Self-constrained behavior

- In order to control a powerful entity, the controlling entity must be at least as powerful
- For AIs that can grow to become significantly more powerful than humans (and their tools), the only way to control them is for them to control themselves

Corollary

Ethical (meta-)values are constraints that must stabilize over time

- Otherwise we have no assurances about the long-term self-contrained behavior of the AI
- Stabilization must occur before the AI becomes too powerful to control it directly (before it's capable of preventing someone —physically or persuasively—from pressing the off-switch)

nnaisense

USI/SUPSI IDSIA

Closing Thoughts

For AI safety, we need to get three things right

1. The architecture of the AI at start-up
2. The teaching of the AI, to develop the understanding of (ethical) constraints
3. "Complete" the teaching before the "deadline of control"

AI may be "softer" than it's been so far

More responsibility on teachers than programmers

nnaisense

USI/SUPSI IDSIA

The End

Credit: SMBC Comics