Growing Robust & Safe AI: Let's be Realistic

Bas Steunebrink
Co-founder of NNAISENSE
Chief Scientist for AGI

The Main Topic

- WRAI, 28 Oct 2017, ETH Zürich
- Workshop on Responsible Artificial Intelligence?
- Whose responsibility is it to make AI reliable?
- Need to ask the right questions!

Typical Question

- “What is the behavior of an AI that is very intelligent – and therefore capable of self-modification – and how do we control it?”

Right Question

- ~~"What is the behavior of an AI that is very intelligent – and therefore capable of self-modification – and how do we control it?"~~
- "How do we grow an AI from baby beginnings such that it gains both robust understanding and proper ethics?"

Long-Term Control

- The ability to control a powerful entity increases as the power of the controlling entity increases
  - analogous to Ashby's *Law of Requisite Variety*
- Corollary: for AIs that can grow to become significantly more powerful than humans (and their tools), the only way to control them is for the AIs to control themselves
- Self-control →adhere to *ethical values*

Ethics as Self-Control

- Ethical values must be implemented as *constraints*
  1. against which the AI *by initial design* tests and prunes its intended actions given their predicted consequences
  2. which *stabilize* over time
  3. which include the (meta-)value to protect its ethical values
- The more the AI's *understanding* of the consequences of its actions grows, the better it becomes at predicting potential constraint violations—and at steering clear of them
- AI becomes *safer and more reliable as its knowledge grows*

More Implications

- *Necessary* to ensure the *long-term self-constrained* behavior
- Knowledge representation must be *motivation-agnostic*
- Humans are *not required to be perfectly wise* in specifying the AI's ethical values from the onset
- But we have a *deadline*
- The stabilization of the ethics-related constraints (*not the knowledge*) must be effected *before* the AI becomes too powerful to be controlled directly
  - before it's capable of preventing someone—physically or persuasively—from pressing the off-switch
- *Hefty implication:* the ethical responsibilities of the designers and builders of AI are far outweighed by those of the *teachers* of AI

Principles

- https://futureoflife.org/ai-principles/
- 9: “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.”

Teachers

- https://futureoflife.org/ai-principles/
- 9: "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."
- Glosses over the (life-long) learning of the AI
- Teachers bear the greater responsibility
  - Also: institutions that educate, accredit, manage, and monitor those AI teachers

Let's Be Realistic

- Let's admit from the onset:
  - we may fail to come up with the perfect utility function from the get-go
  - we can't axiomatize the AI or the environment
  - the AI won't have enough resources (*time, energy, input*) to do the optimal thing

Why Not Rely on Proof?

- Q-Learning is guaranteed to converge to the optimum

Why Not Rely on Proof?

- Q-Learning is guaranteed to converge to the optimum
- … *under some assumptions:*
  - The reward function remains fixed
  - The environment's dimensionality & dynamics remain fixed
  - Time goes to infinity

Healthy Skepticism

- Convergence proofs are easily misleading
- Assumptions about the environment, the agent, and its motivations will be *idealized*, *inaccurate*, and *incomplete*

Developmental AI

- The fundamental problem: *bridging the gap*
  - our imperfect specifications of constraints (safety & ethics)
  - sensory inputs
  - potential actions
- Goal: to make sure the AI connects the dots
- Method: a *developmental* approach

Understanding

- Need to tackle the hard problem of *understanding*

Beyond Human Intervention

- A full methodology for teaching & testing
  - Restrict, supervise, intervene (like toddlers)
  - Test under *pressure*
    - Situation where some of its constraints are *nearly or very easily* violated
    - Recognize, report, prioritize, and recover
  - Successful pressure tests are a step toward *certification*, though not a proof