Safe Baby AGI

Jordi Bieger[1], Kristinn R. Þórisson[1,2] and Pei Wang[3]

[1]CADIA @ Reykjavik University, [2]Icelandic Institute for Intelligent Machines, [3]CIS @ Temple University

REYKJAVÍK UNIVERSITY

ABSTRACT

Early AGI systems will necessarily start their "life" in a baby-like state where they are relatively harmless due to a lack of knowledge and other resources. At that point we will have great control over the system's experience. This gives us the opportunity and responsibility to raise the system to be friendly. Many of the concerns about rogue AI destroying humanity as well as approaches to prevent this fail to take into account the properties of baby AGI and focus too much on providing impossible guarantees on inherently adaptive systems that don't exist yet. We need to develop actual running AGI systems to know how they behave in complex environments, and rely on the scientific method to improve them along all dimensions, including safety.

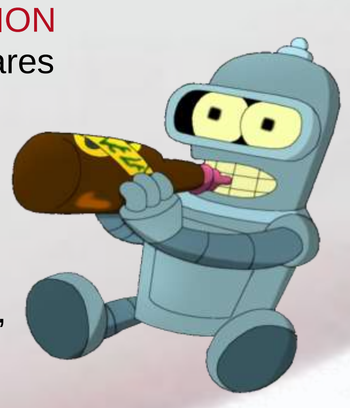
BABY AGI

Intelligent behavior requires **resources** like time, energy and knowledge. Since these are always limited, any AI system will be **fallible** – no matter how intelligent it is.

Hand-coding all of the knowledge required for adult level intelligence borders on the impossible. Furthermore, to earn the "G" in "AGI" a system should be able to deal with environments not foreseen by its designers.

AGI systems should therefor start "life" in a **baby**-like state where knowledge – and consequently ability – is severely limited and must be acquired through interaction with the environment in which we can act as teachers. Much of the system's intelligence therefor does not come from its design or **nature**, but also from its experience or **nurture**.

APPETITE FOR DESTRUCTION



But why would an AI even want to kill us? An AI cares only about its goals, and most don't inherently include the welfare of humanity. Almost all AIs will have **basic AI drives**: goals such as survival, resource acquisition and self-improvement that are instrumental to the achievement of virtually any other goal. Once the AI gets powerful enough, humans may no longer be seen as allies, but as competitors, threats, or resources.

However, derivation of goals such as the basic AI drives and the extermination of humanity requires both intelligence *and* **relevant knowledge**. Safe **baby AGI** will not have access to this knowledge unless its caretakers want it to. They provide an upbringing and education that should include **value learning**, where the AI is taught human morality.

STOP THE PRESSES

It has been suggested that AGI research should be **stopped** or **slowed** to allow time for theoretical work to guarantee the safety of hypothetical AI designs. But behavior is determined by **nature *and* nurture**. **No hard guarantees** are possible since the behavior of a necessarily fallible AI interacting with a complex environment is **inherently unpredictable**. We need to develop *actual running* AGI systems to know how they behave in complex environments, and rely on the **scientific method** to improve them along all dimensions, including safety.

OVERPOWERING HUMANITY

Many people fear that human-level AGI systems could very quickly acquire sufficient power to overpower all of humanity. This would happen through a so-called **hard takeoff** where accelerating returns on self-improvements result in an exponential increase in intelligence.

To prevent this, approaches such as **AI boxing** have been proposed and subsequently dismissed. A sufficiently smart AI might be able to trick a gatekeeper into doing its bidding. Furthermore, **monitoring** of complex systems is difficult, especially if the AI hides its intentions and capabilities.

These criticisms ignore the fact that a **baby AGI** does not start out with the ability to trick humans or hide its internals. Neither does it start with the idea to plan an AIpocalypse. Complexity should be low enough that the system can be monitored to such a degree that explosive growths in intelligence or resources can be measured before the system is even close to stop us from pressing the off-button.

NURTURING BENEFICIAL AGI

We have the **opportunity** and **responsibility** to guide our AIs to learn the right things in the vast realm of possibilities. We can emphasize effective and **peaceful** ways to accomplish goals, guide the AI through moral stages of development, and teach it about the risks of aggression and the value of relationships. As is always the case with **powerful technology**, we should proceed with caution and care, and strife to use it for good.

College of Science and Technology
TEMPLE UNIVERSITY

ICELANDIC INSTITUTE FOR INTELLIGENT MACHINES
VITVÉLASTOFNUN ÍSLANDS

CADIA

c.f. Bieger, J., Thórisson, K. R., & Wang, P. (2015). Safe Baby AGI. In *Artificial General Intelligence* (pp. 42-45). Springer.