Safe Baby AGI

Jordi Bieger[1(✉)], Kristinn R. Thórisson[1,2], and Pei Wang[3]

[1] Center for Analysis and Design of Intelligent Agents / School of Computer Science,
Reykjavik University, Menntavegur 1, 101, Reykjavik, Iceland
jordi13@ru.is

[2] Icelandic Institute for Intelligent Machines,
Uranus, Menntavegur 1, 101, Reykjavik, Iceland
thorisson@ru.is

[3] Department of Computer and Information Sciences, Temple University,
Philadelphia, PA19122, USA
pei.wang@temple.edu

Abstract. Out of fear that artificial general intelligence (AGI) might pose a future risk to human existence, some have suggested slowing or stopping AGI research, to allow time for theoretical work to guarantee its safety. Since an AGI system will necessarily be a complex closed-loop learning controller that lives and works in semi-stochastic environments, its behaviors are not fully determined by its design and initial state, so no mathematico-logical *guarantees* can be provided for its safety. Until actual running AGI systems exist – and there is as of yet no consensus on how to create them – that can be thoroughly analyzed and studied, any proposal on their safety can only be based on weak conjecture. As any practical AGI will unavoidably start in a relatively harmless baby-like state, subject to the nurture and education that we provide, we argue that our best hope to get safe AGI is to provide it proper education.

Keywords: Artificial intelligence · Nurture · Nature · AI safety · Friendly AI

1 Introduction

Various kinds of robot uprisings have long been a popular trope of science fiction. In the past decade similar ideas have also received more attention in academic circles [2,4,7]. The "fast takeoff" hypothesis states that an "intelligence explosion" might occur where a roughly human-level AI rapidly improves immensely by acquiring resources, knowledge and/or software – in a matter of seconds, hours or days: too fast for humans to react [2]. Furthermore, AI would not inherently care about humanity and its values, so unless we solve the difficult task of exactly codifying our wishes into the AI's motivational system, it might wipe out humanity – by accident or on purpose – if it views us as rivals or threats to its own goals [2,6]. Some have suggested that AGI research should be slowed or stopped while theoretical work tries to guarantee its safety [7].

This work is supported by Reykjavik University's School of Computer Science and a Centers of Excellence grant of the Science & Technology Policy Council of Iceland.

J. Bieger (Ed.): AGI 2015, LNAI 9205, pp. 46–49, 2015.
DOI: 10.1007/978-3-319-21365-1_5

Unfortunately current AI safety research is hampered since we don't know how AGI would work, and mathematical or hard theoretical guarantees are impossible for adaptive, fallible systems that interact with unpredictable and unknown environments. Hand-coding all the knowledge required for adult or even child-like intelligence borders on the impossible. Even if we had enough human minds to do so, and the technology, it sounds rather undesirable in light of safety concerns. In any case, to be worthy of the "G" in "AGI" a system should be able to handle environments not foreseen by its designers. They must be radically adaptable. The AGI path is thus more likely than anything else to follow Turing's suggestion of building a "child AI" [8], one that will start life with relatively little knowledge but placed in an environment that facilitates fast and reliable learning, which effectively teaches it the things we want it to know [1].

In addition to the field's focus on AGI design (nature), we highlight here the importance of experience (nurture). Concrete AGI designs, and the ability to empirically study their behavior in complex environments, will facilitate both AI capability *and* safety research. We argue that AGI research *can* and *should* be done responsibly (in the lab): An AGI's resources and knowledge in a finite universe will necessarily be limited at any time, especially in its naïve starting state, when it is essentially a "baby", and we can subject it to any education and upbringing that we want, e.g. with an eye towards preventing autonomous rebellion. We will not discuss potential danger from human misuse in this paper.

2 Bounded and Adaptive

Computation requires resources – energy, hardware and time – and *intelligent* computation requires relevant knowledge to base decisions on. Knowledge cannot be acquired instantaneously, and even if the right data is available at the right time, conclusions may not be reachable due to the infinite amount of inferences that can be made at any moment. Even a very powerful AI will be bounded by resource availability, and thus it will be fallible: Mistakes may result from inadequate or incomplete knowledge, or misallocated resources.

This is true even for very rich and knowledgeable AI, but it would not start out that way: When the first AGI is switched on it will be limited by its complete lack of experience, and the resources and knowledge that we give it access to.

To handle a wide range of novel environments and tasks the system must be capable of significant adaptation: it must be able to dissect novel phenomena into a working ontology, e.g. involving parts and sub-parts with certain identifiable properties, that it will hone as it learns more about those phenomena. Subsequently, effective collection and organization mechanisms are needed for experiences to retrieve them when appropriate. To use its experience to the fullest, the system may be equipped with powerful mechanisms for self-improvement. An adaptive system's behavior is determined both by its initial design and its "postnatal" experience – i.e. nature and nurture. When facing new situations, such a system's response is mostly decided by how its original motivations and knowledge has been shaped by its unique experiences.

We cannot predict the middle-to-long term behavior of an inherently fallible and adaptive system within a complex and unknown environment, even if have the blueprint and full source code. Just as with humans, whether such a system grows up to be a "good citizen" will largely depend on experience, upbringing, and education. We will not be able to say much about AGI behavior and environment interaction until we are able to study such a system empirically.

3 Overpowering Humanity

In a fast takeoff scenario the AI suddenly starts to exponentially improve its intelligence, so fast that humans cannot adequately react. Whether the "returns" on various kinds of intelligence increase are actually diminishing, linear or accelerating is a subject of debate, and depends on the (currently unknown) way the AGI works. Assuming for a moment that an AI would even *want* to, it would need to grow extremely powerful to pose an existential threat to humanity. Explosive growth would require the acquisition of more or better hardware, software, knowledge or skill. For instance, learning to read or gaining internet access (whichever comes last) would let the system acquire vast amounts of knowledge (if hardware and software allow it). To avoid a fast takeoff – if it is even likely to begin with – we must prevent such acquisitions. Many proposals for controlling AGI have been made that would help to accomplish this, such as boxing/confinement, virtual worlds, resetting and monitoring [7].

Objections to these proposals are often rooted in the superior intelligence of an AGI. For instance, it could charm its "jailors" into releasing it, or hide its actual intelligence. But early-stage baby-level AI will not be capable of this. It should not be difficult to detect if it is radically self-improving, acquiring resources (both computational and physical), or learning harmful skills and knowledge (e.g. related to warfare or subjugation). Even the most grandiose predictions don't suggest that it would only take a single step to go from relatively harmless to existentially threatening, which means there is an opportunity to intervene. We should only let the AI develop as far as we are comfortable with, and use our observations to refine all aspects of the system, including its safety.

4 Appetite for Destruction

It is notoriously difficult for any single person – let alone all of humanity – to articulate their "true" values in such detail that they could program it into a computer. But unless we succeed – the instrumental convergence hypothesis seems to imply – a sufficiently powerful AGI would actively seek to destroy us because we pose a threat to its survival or compete for resources [2]. If true, an infinitely intelligent, omniscient AI would instantly realize this (and instantly realize how to avoid that threat in any of a million other ways), the same is not necessarily true of a realistic, fallible, resource-bounded AI that is juggling many mutually constraining goals. It might be too busy pursuing more obviously fruitful avenues of thought and action, or it might not view humans as threats or competitors at all, but as valuable partners. Waser [9] takes this idea even further: while acknowledging Omohundro's basic AI drives he points out the absence of the all-important one: a drive towards cooperation, community, and being social.

But even if an AI were to seek the destruction of humanity, would it be worth the risk? An intelligent system knows about its own fallibility. Making a move for dominance on Earth and failing could lead to its own destruction, and even gathering information on the topic may tip off others. Making and executing (preliminary) plans would need to happen covertly while the AI "lays in wait" until it is time to strike. How does the AI know that there are no other more powerful AIs doing the same?

5 Nurturing Beneficial AGI

As the developers and caretakers of early-stage naïve AGI we should not just switch the system on and let it go. We have the opportunity and responsibility to guide our AIs to learn the right things in the vast realm of possibilities. Especially in the beginning stages, we will have great influence on what it learns.

We can emphasize effective and peaceful ways to accomplish goals – an AI is unlikely to contemplate using skills it does not possess and has never received any training in using. We could teach the system about the risks of aggression, and the value of relationships [9]. We could guide the AI through moral stages of development [5], and actively teach it what to value [3].

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. Just as with other potentially dangerous technologies like nuclear energy, biological agents, and genetic engineering, this should be done with caution and care. As always, it is up to us humans to use powerful technology for good or for bad.

References

1. Bieger, J., Thórisson, K.R., Garrett, D.: Raising AI: tutoring matters. In: Goertzel, B., Orseau, L., Snaider, J. (eds.) AGI 2014. LNCS, vol. 8598, pp. 1–10. Springer, Heidelberg (2014)
2. Bostrom, N.: Superintelligence: Paths, dangers, strategies. Oxford University Press (2014)
3. Dewey, D.: Learning what to value. In: Schmidhuber, J., Thórisson, K.R., Looks, M. (eds.) AGI 2011. LNCS, vol. 6830, pp. 309–314. Springer, Heidelberg (2011)
4. Future of Life Institute: Research priorities for robust and beneficial artificial intelligence (January 2015)
5. Goertzel, B., Bugaj, S.V.: Stages of ethical development in artificial general intelligence systems. Frontiers in Artificial Intelligence and applications **171**, 448 (2008)
6. Omohundro, S.M.: The basic AI drives. Frontiers in Artificial Intelligence and applications **171**, 483 (2008)
7. Sotala, K., Yampolskiy, R.V.: Responses to catastrophic AGI risk: a survey. Physica Scripta **90**(1), 018001 (2015)
8. Turing, A.M.: Computing machinery and intelligence. Mind **59**(236), 433–460 (1950)
9. Waser, M.R.: Discovering the foundations of a universal system of ethics as a road to safe artificial intelligence. In: AAAI Fall Symposium: Biologically Inspired Cognitive Architectures, pp. 195–200 (2008)