Chapter 60
Security Solutions for Intelligent and Complex Systems

Stuart Armstrong
Future of Humanity Institute, UK

Roman V. Yampolskiy
JB Speed School of Engineering, USA

ABSTRACT

Superintelligent systems are likely to present serious safety issues, since such entities would have great power to control the future according to their possibly misaligned goals or motivation systems. Oracle AIs (OAI) are confined AIs that can only answer questions and do not act in the world, represent one particular solution to this problem. However even Oracles are not particularly safe: humans are still vulnerable to traps, social engineering, or simply becoming dependent on the OAI. But OAIs are still strictly safer than general AIs, and there are many extra layers of precautions we can add on top of these. This paper begins with the definition of the OAI Confinement Problem. After analysis of existing solutions and their shortcomings, a protocol is proposed aimed at making a more secure confinement environment which might delay negative effects from a potentially unfriendly superintelligence while allowing for future research and development of superintelligent systems.

INTRODUCTION

With the likely development of superintelligent programs in the near future, many scientists have raised the issue of safety as it relates to such technology (Bostrom, 2006; Chalmers, 2010; Hall, 2000; Hibbard, 2005; Yampolskiy, 2011a, 2011b; Yampolskiy & Fox, 2012a, 2012b; Yudkowsky, 2008). A common theme in Artificial Intelligence (AI[1]) safety research is the possibility of keeping a superintelligent agent in a sealed hardware so as to prevent it from doing any harm to humankind. Such ideas originate with scientific visionaries such as Eric Drexler who has suggested confining transhuman machines so that their outputs could be studied and used safely (Drexler, 1986). Similarly, in 2010 David Chalmers

DOI: 10.4018/978-1-7998-0951-7.ch060

proposed the idea of a "leakproof" singularity (Chalmers, 2010). He suggested that for safety reasons, AI systems first be restricted to simulated virtual worlds until their behavioral tendencies could be fully understood under the controlled conditions.

This chapter is based on combined and extended information from three previously published papers: (Armstrong, 2011; Armstrong, Sandberg, & Bostrom, 2012; Yampolskiy, 2012a)*. We evaluate feasibility of previously presented proposals and suggest a protocol aimed at enhancing safety and security of such methodologies. While it is unlikely, that long-term and secure confinement of AI is possible, we are hopeful that the proposed protocol will give researchers a little more time to find a permanent and satisfactory solution for addressing existential risks associated with appearance of superintelligent machines.

In this chapter we will review specific proposals aimed at creating restricted environments for safely interacting with artificial minds. The key question is: are there strategies that reduce the potential existential risk from a superintelligent AI so much that while implementing it as a free AI would be impermissible a confined implementation would be permissible? The chapter will start by laying out the general design assumptions for the confined AI and formalizing the notion of confinement. Then it will touch upon some of the risks and dangers deriving from the humans running and interaction with the confined AI. The final section looks at some of the other problematic issues concerning the confined AI, such as its ability to simulate human beings within it and its status as a moral agent itself.

Motivation for AI Confinement

There are many motivations to pursue the goal of developing AI. While some motivations are non-instrumental, such as scientific and philosophical curiosity about the nature of thinking or a desire for creating non-human beings, a strong set of motivations is the instrumental utility of AI. Such machines would benefit their owners by being able to do tasks that currently require human intelligence, and possibly tasks that are beyond human intelligence. From an economic perspective the possibility of complementing or substituting expensive labor with cheaper software promises very rapid growth rates and high productivity (Hanson, 2001, 2008; Kaas, Rayhawk, Salamon, & Salamon, 2010). The introduction of sufficiently advanced AI would have profound effects on most aspects of society, making careful foresight important.

While most considerations about the mechanization of labor have focused on AI with intelligence up to the human level, there is no strong reason to believe humans represent an upper limit of possible intelligence. The human brain has evolved under various biological constraints (e.g. food availability, birth canal size, trade-offs with other organs, the requirement of using biological materials) which do not exist for an artificial system. Beside different hardware, an AI might employ more effective algorithms that cannot be implemented well in the human cognitive architecture (e.g. making use of very large and exact working memory, stacks, mathematical modules or numerical simulation), or employ tricks that are not feasible for humans, such as running multiple instances whose memories and conclusions are eventually merged. In addition, if an AI system possesses sufficient abilities, it would be able to assist in developing better AI. Since AI development is an expression of human intelligence, at least some AI might achieve this form of intelligence, and beyond a certain point would accelerate the development far beyond the current rate (Chalmers, 2010; Kurzweil, 2005).

While the likelihood of superintelligent AI is hotly debated, the mere possibility raises worrying policy questions. Since intelligence implies the ability to achieve goals, we should expect superintelligent systems to be significantly better at achieving their goals than humans. This produces a risky power differential. The appearance of superintelligence appears to pose an existential risk: a possibility that humanity is