Futures 72 (2015) 27–31

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

Linking simulation argument to the AI risk

Milan M. Ćirković [a,b,*]

[a] Astronomical Observatory of Belgrade, Volgina 7, 11000 Belgrade, Serbia
[b] Future of Humanity Institute, Faculty of Philosophy, University of Oxford, Suite 8, Littlegate House, 16/17 St Ebbe's Street, Oxford OX1 1PT, UK

ARTICLE INFO

Article history:
Available online 3 June 2015

Keywords:
Existential risk
Artificial intelligence
Simulations
Future of humanity
Risk analysis

ABSTRACT

Metaphysics, future studies, and artificial intelligence (AI) are usually regarded as rather distant, non-intersecting fields. There are, however, interesting points of contact which might highlight some potentially risky aspects of advanced computing technologies. While the original simulation argument of Nick Bostrom was formulated without reference to the enabling AI technologies and accompanying existential risks, I argue that there is an important generic link between the two, whose net effect under a range of plausible scenarios is to reduce the likelihood of our living in a simulation. This has several consequences for risk analysis and risk management, the most important being putting greater priority on confronting "traditional" existential risks, such as those following from the misuse of biotechnology, nuclear winter or supervolcanism. In addition, the present argument demonstrates how – rather counterintuitively – seemingly speculative ontological speculations could, in principle, influence practical decisions on risk mitigation policies.

1. Introduction

Bostrom (2003) has suggested that the posterior probability of our living in a computer simulation might be larger than naively expected. This conclusion rests on reasonable assumptions about the advances in information processing and simulation technology, as well as on important philosophical principles, such as Leibniz's principle of indifference. If we accept – under assumptions such as physicalism regarding minds – that sufficiently advanced simulation of an observer is another observer in her own right, we will in the fullness of time have observers in two categories: baseline physical, evolved ones and simulated ones. Even very limited experience of humans simulating physical objects such as bridges or airplanes or stars tells us that it is much cheaper in terms of resources to simulate an object than to construct it.[1] So, it is reasonable to allow for the possible future with cheap simulations and simulated observers possibly outnumbering the evolved ones by a large margin.

There are three possible conclusions: either (1) the human species is likely to go extinct before reaching a stage of capability for large-scale simulations of intelligent observers; or (2) any advanced civilization (human or posthuman) is

* Corresponding author at: Astronomical Observatory of Belgrade, Volgina 7, 11000 Belgrade, Serbia. Tel.: +381 69 1687200.
E-mail address: mcirkovic@aob.rs

[1] Even if we are not in position to construct the relevant objects, such as stars, it is still possible to try to replicate some of the aspects of relevant processes – like nuclear reactions in stellar cores – in both computer simulations and laboratory analogs (e.g., thermonuclear fusion reactors). The former is clearly and immensely cheaper.

http://dx.doi.org/10.1016/j.futures.2015.05.003
0016-3287/

extremely unlikely to run a significant number of simulations of their evolutionary history (or variations thereof – hereafter the "ancestor-simulations"); or (3) we are almost certainly living in a computer simulation. Obviously, accepting (3) would mean massive changes in our metaphysical outlook, although it might not, at first glance, present us with any new practical challenges.

However, the reasoning employed by Bostrom in reaching the trilemma does not take into account the possibly risky consequences of the very existence of the technologies necessary for running ancestor-simulations. Obviously, the explosive growth of our computing power, expressed through Moore's Law and similar generalizations (Kurzweil, 2005), as well as our capacity for simulating more and more complex systems, are facts of everyday life and it is not easy to perceive them as large, probably even existential, risk factors. However, there are multiple indications that, as far as increases in computing power and complexity go, we are dealing with the threshold phenomena in which reaching a range of critical values might result in large, possibly catastrophic shifts in the outcome. This is the major concern underlying contemporary discussions of the risk associated with the concept of artificial intelligence (henceforth AI risk).

The enormously increased computing capacities of future AI systems are at the core of several high-risk scenarios, which involve both the intrinsic unpredictability of the behavior of such systems and the systems' simulating powers, which are far in excess of our present-day simulating powers. Usually, these two aspects are dealt with separately, which might not be entirely justified. The present note deals with the *risk aspect of the simulation argument*, while showing how the central argument about enabling technologies could be further generalized. Once it is accepted that the enabling technologies carry a load of risk quite independently of the issue of simulations and observer-counting, there is a feed-back effect on the distribution of probabilities between the three possible outcomes of Bostrom's argument. This, in turn, brings about a rearranging of our priorities in dealing with the "traditional" existential risks vs. risks following from AI and the possibility of our living in a simulation. The present argument deals with the future of humanity, but it could be generalized to any set of technological civilizations in the universe at any given epoch.

2. A scenario

Consider the following scenario: the increase in computing power leads to viable whole-brain emulation and running human uploads. As far as complexity of both hardware and software go, this is an intermediate stage between the best present-day AI systems and envisioned superintelligent AI systems (in the further text denoted as AI++, following Chalmers, 2010[2]). AI++ systems are clearly a source of existential risk, for with their great power comes the lack of predictability following their superior cognition (Müller, 2014 and references therein). Therefore, efforts have been made to enable design of safe or "friendly" AI++ (e.g., Yudkowsky, 2008). The main difficulty stems from the fact that the conventional road to AI++ systems goes through self-improvement of lower-level AI systems, notably those equivalent to human intelligence at present, and possibly even much lower. This iterative procedure might occur in a self-accelerating mode and end up with AI++ "in a flash", i.e., before researchers, risk analysts, and policy-makers are able to ascertain the situation and gauge the relevant risks.

In order to highlight the complexity of the situation, let us first compare two extreme cases: (i) a world in which all human-level AI as well as AI++ are designed completely safe and sound. In such a world, there would be a huge amount of computing power available to everyone, including individual actors, and running detailed simulations of individual humans, as well as large-scale ancestor-simulations, would be cheap and easy. In this world, it is hard to avoid the conclusion that we are indeed living in a simulation, since the number of simulated observers would, in this world, vastly dominate in the total tally of all observers.

In contrast, we might wish to consider (ii) a world in which AI++ emerges rapidly, is extremely dangerous, and the probability of (post)humanity surviving its emergence is zero or sufficiently close to zero. In such case, there will be only evolved observers (up to the moment of the AI++ emergence) plus those simulated observers which would have been simulated prior to the moment of the AI++ emergence. In order to estimate the probability of our living in a simulation, we need to know the ratio between the two, or at least to gauge whether the interval between the advent of the technology of ancestor-simulations and the advent of AI++ is short or long. If that interval is very short, as suggested by rapid emergence of AI++, the measure of simulated observers will be small and, consequently, the probability of our living in a simulation would tend to zero.

The realistic case lies somewhere between these two extremes. But the very fact that *the magnitude of AI++ risk is related to feasibility and number of ancestor-simulations* should impose some constraints on the original simulation argument.

3. The argument

Consider the following set of premises:

1. Running ancestor-simulations will require computing resources of some minimal complexity C_{as}, to be conceived and executed at characteristic timescales t_{as}.

[2] For present purposes, AI++ is equivalent to what Bostrom (2014) dubs *superintelligence*.