

## **Short Abstract**

**Submission ID 99**

**Track 10**

**Title: Arms Control and Intelligence Explosions**

A number of commentators have argued that some time in the 21st century humanity will develop generally intelligent software programs at least as capable as skilled humans, whether designed *ab initio* or as emulations of human brains, and that such entities will launch an extremely rapid technological transformation as they design their own successors. The speed of such a 'Singularity' or 'intelligence explosion' would be so great that biological humans would lack time for extensive deliberation regarding or supervision of the process. Several authors have called for regulation to retard the pace of advancement in this field, allowing more time to ensure that any intelligent machines are safe and broadly beneficial, while various proponents of the Singularity hypothesis have replied that such attempts will fail because of competition between regulatory jurisdictions, sometimes making analogies to the failures of nuclear counter-proliferation efforts. This paper discusses some key considerations that distinguish the case of sapient software programs from the historical experience with nuclear weapons technology.

First, the rapid and competitive development of the technology poses unusually great risks of unintentional harm, harm which would affect all competitors. An arms race that lead to a trade-off of safety for speed might result in the creation of superintelligent beings indifferent to human welfare, while natural selection could result in an initially benign population evolving into extremely competitive replicators. Second, an intelligence explosion could result in extreme winner-take-all effects, as one power could lever an initial advantage to develop astronomically more capable intelligences and prevent others from duplicating its success. Such a capacity for unilateral dominance is historically unprecedented, even by the brief window of American nuclear monopoly. Third, the very hypothesized capabilities in computing, neuroscience, and artificial intelligence that could enable a Singularity would also provide powerful new means to enforce regulations and international agreements. In combination, these factors suggest that regulatory jurisdictions may find cooperative control of the development of software entities more desirable and more practically feasible than historical nuclear arms control efforts.

## **Extended Abstract**

**Submission ID 99**

**Track 10**

**Title: Arms Control and Intelligence Explosions**

### **1. Introduction**

Several prominent commentators have argued that in the 21st century humanity will develop software programs at least as generally capable as skilled humans, whether *ab initio* or via emulation of human brains, and that such entities will greatly accelerate technological transformation as they design their own successors. The speed of such a 'Singularity' or 'intelligence explosion' would be so great that biological humans would lack time for extensive deliberation regarding or supervision of the process. Several authors have called for regulatory restrictions on research into intelligent machines to allow

more time for deliberation and to ensure that the results are beneficial, while various proponents of the Singularity hypothesis (e.g. Kurzweil 2005) have replied that such attempts will fail because of competition between regulatory jurisdictions, sometimes making analogies to the failures of nuclear counter-proliferation efforts. This paper discusses some key considerations that distinguish the case of sapient software programs from the historical experience with nuclear weapons technology: unusually high risks of competitive development, an unprecedented winner-take-all potential, and the applicability of enabling technologies to the enforcement of a regulatory regime.

## **2. Global risks of superintelligent machines**

The idea of machine intelligences exterminating humanity is a cliché of fiction, but that status does not preclude a real risk. (Bostrom 2002; Friedman 2008; Hall 2007; Kurzweil 2005; Moravec 1999; Posner 2004; Rees 2003; Yudkowsky 2008). Setting aside anthropomorphic presumptions of rebelliousness, a more rigorous argument (Omohundro, 2007) relies on the instrumental value of such behavior for entities with a wide variety of goals that are easier to achieve with more resources and with adequate defense against attack. A variety of programming approaches risk creating such goals in an entity capable of self-modification (Yudkowsky, 2008). Similarly, evolutionary pressures among competing intelligences could select for monomaniacal focus on reproductive success (Bostrom 2005; Hanson 1994).

A nontrivial risk of global catastrophe from simply developing and testing a technology is an important qualitative difference from the historical situation of nuclear weapons: while some Manhattan Project scientists did consider the possibility that a nuclear test would unleash a self-sustaining fusion reaction in the atmosphere and render Earth uninhabitable (Konopinski *et al*, 1946), this was seen as extremely improbable. The closest analogue would be the risk of a nuclear exchange triggered by accident or miscommunication, which was historically an area of substantial superpower cooperation between nuclear powers in the establishment of 'hotlines' and communication between militaries.

However, a competitive race to develop advanced artificial intelligence would create incentives to trade off safety measures for speed of development, since the benefits of early development would be locally concentrated, while risks would be global, and the negative externalities would lead to a collective action problem: every competitor could benefit from a collective arrangement to slow research and take safety measures, provided that compliance was verifiable. The importance of these benefits would depend on estimates of the risks and the motives of officials, but they could plausibly provide an unprecedented incentive for a global regulatory regime.

## **3. Winner-take-all dynamics**

Even without the threat of catastrophic error, the development of controlled digital entities by any particular power group would remain a threatening prospect for others. For instance, cheap reproduction of skilled workers would be expected to drive wages below human subsistence (Hanson 1998). Human survival would depend on property rights respected by the new beings, or on income redistribution backed by them: Moravec (1999) suggests lowering the age of eligibility for social security payments to birth. A nation-state that developed such entities first could use them to make further advances, and gain sufficient military advantage to safely disarm rivals and secure its gains. The threat of such unchallenged economic and military dominance would be very different from nuclear arms races, and would more closely resemble an impenetrable missile defense enabling nuclear first strike without fear: a country seen to be developing such a capacity would invite preemptive strikes or threats thereof while possible. If visible advances in artificial intelligence convince world leaders that such a situation is imminent, they would have strong reason to coordinate in averting it without nuclear war, even at the cost of intrusive inspection.

The likelihood of such an overwhelming technological lead depends on the extent of positive feedback when artificial intelligences conduct research into improving their design. Hall (2007) argues against a rapid takeoff using a model of a single machine intelligence with a fixed dollar value of hardware improving its software, while Kurzweil (2005) projects that a 'Singularity' as he defines it will not occur for a number of years after the development of human-equivalent artificial intelligence. This paper will explore those models and show that projections of a slow takeoff are very sensitive to several uncertain parameters.

#### **4. Regulatory Mechanisms and Verification**

The feasibility of regulating artificial intelligence development can be considered at two levels: can states control research within their own borders, and can they form enforceable international agreements to coordinate their activities? Internally, advances in AI, especially pattern recognition, should greatly facilitate mass surveillance, e.g. wiretaps are now primarily limited by the labor costs of reviewing recordings (Friedman 2008). An illicit conspiracy to develop robust artificial intelligence would need to make groundbreaking scientific advances without leaving a trail of communication incriminating records. These and other considerations suggest that the primary obstacle to global regulation would be coordination between states, and assuaging concerns about secret research efforts in the style of the Manhattan Project. Since artificial intelligence research would not require any exotic components other than researchers and computers, verification would have to involve making surveillance of researchers and government officials directly available to other states, along the lines of Brin's (1998) *Transparent Society*. Alternatively, and less intrusively, the advances in neuroscience and brain-scanning technologies postulated by Kurzweil (2005) may enable powerful lie-detection techniques that could be applied to national and military leaders to verify claims in tandem with other measures.

Perhaps most importantly, if relatively unsophisticated artificial intelligences are developed, ones incapable of modifying their own designs easily, they could be used to monitor surveillance channels and enforce global regulations. As digital programs they could be copied at will, have their motives and behavior tested by all concerned parties, and be regularly reverted to a previously saved state to prevent goal drift. The technology most directly requisite for an intelligence explosion would also assist in postponing it.

#### **5. Examining regulation in full**

The creation of digital beings with superhuman intelligence would be in many ways an utterly unprecedented event. Analogies to historical failures of international arms control may not apply when the benefits of cooperation and the risks of uncoordinated activity are vastly greater, as suggested above. Regardless, if we take claims of potentially catastrophic transformative technologies seriously, then methods for mitigating risk deserve careful analysis rather than casual dismissal. An erroneous claim that there is no alternative to an arms race may become a self-fulfilling prophecy.

#### **References**

Bostrom, N. (2002). Analyzing Human Extinction Scenarios. *Journal of Evolution and Technology*, 9(1).

Bostrom, N. (2005). The Future of Human Evolution. In C. Tandy (Ed.), *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing* (pp. 339-371). Palo Alto: Ria University Press.

Brin, D. (1998). *The Transparent Society*. New York: Basic Books.

Friedman, D. (2008). *Future Imperfect: Technology and Freedom in an Uncertain World*. Cambridge: Cambridge University Press.

Hall, J.S. (2007). *Beyond AI: Creating the Conscience of the Machine*. Amherst: Prometheus Books.

Hanson, R. (1998). *Economic Growth Given Machine Intelligence*. Working Paper. <http://hanson.gmu.edu/aigrow.pdf>. Cited 11 Jan 2009.

Konopinski, E. J, C.. Marvin; Edward Teller (1946, declassified February 1973). *Ignition of the Atmosphere with Nuclear Bombs*. Technical Report Los Alamos National Laboratory LA-602.

Kurzweil, R. (2005). *The Singularity is Near*. London: Penguin Books.

Moravec, H. (1999). *Robot: Mere Machine to Transcendent Mind*. Oxford: Oxford University Press.

Omohundro, S. (2007). *The Nature of Self-Improving AI*. Paper presented at the 2007 Singularity Summit, San Francisco.

Posner, R. (2004). *Catastrophe: Risk and Response*. Oxford: Oxford University Press.

Rees, M. (2004). *Our Final Hour: A Scientist's Warning : how Terror, Error, and Environmental Disaster Threaten Humankind's Future in this Century - on Earth and Beyond*.

Yudkowsky, E. (2008). *Artificial Intelligence as a positive and negative factor in global risk*. In N. Bostrom and M. Cirkovic (Eds.), *Global Catastrophic Risks* (pp. 308-343). Oxford: Oxford University Press.