# Leakproofing the Singularity Artificial Intelligence Confinement Problem

**Article** *in* Journal of Consciousness Studies · January 2012

1 author:

Roman Yampolskiy
University of Louisville
**198** PUBLICATIONS  **2,149** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project  AI Safety View project

Project  Safety and Security of Artificial Intelligence View project

Roman V. Yampolskiy

# *Leakproofing the Singularity*

## *Artificial Intelligence Confinement Problem*

***Abstract:*** *This paper attempts to formalize and to address the 'leakproofing' of the Singularity problem presented by David Chalmers. The paper begins with the definition of the Artificial Intelligence 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 potential negative effect from the technological singularity while allowing humanity to benefit from the superintelligence.*

**Keywords:** AI-Box, AI Confinement Problem, Hazardous Intelligent Software, Leakproof Singularity, Oracle AI.

'I am the slave of the lamp'

Genie from *Aladdin*

## 1. 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 (Yudkowsky, 2008; Bostrom, 2006; Hibbard, 2005; Chalmers, 2010; Hall, 2000). A common theme in Artificial Intelligence (AI)[1] safety research is the possibility of keeping a super-intelligent agent in a sealed hardware so as to prevent it from doing any harm to humankind. Such ideas originate with scientific

Correspondence:
Roman V. Yampolskiy, Department of Computer Engineering and Computer Science University of Louisville. *Email: roman.yampolskiy@louisville.edu*

---

[1]  In this paper the term AI is used to represent superintelligence.

visionaries such as Eric Drexler, who has suggested confining transhuman machines so that their outputs could be studied and used safely (Drexler, 1986). Similarly, Nick Bostrom, a futurologist, has proposed (Bostrom, 2008) an idea for an Oracle AI (OAI), which would only be capable of answering questions. Finally, in 2010 David Chalmers 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.

In this paper we will review specific proposals aimed at creating restricted environments for safely interacting with artificial minds. We will evaluate feasibility of 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 super-intelligent machines.

### 1.1 Artificial Intelligence Confinement Problem

Interestingly, the AI Confinement Problem is not a recent invention and does not have its roots in the singularity movement. The so called Confinement Problem (CP) was posed by Butler Lampson in 1973 (Lampson, 1973) as a security challenge to the computer experts. It was originally phrased as: '…the problem of confining a program during its execution so that it cannot transmit information to any other program except its caller. …We want to be able to confine an arbitrary program.… any program, if confined, will be unable to leak data. A misbehaving program may well be trapped as a result of an attempt to escape' (*ibid.*).

To address the Confinement Problem Lampson has introduced the Laws of Confinement (*ibid.*):

1)  *Total isolation:* A confined program shall make no calls on any other program.

2)  *Transitivity:* If a confined program calls another program which is not trusted, the called program must also be confined.

3)  *Masking:* A program to be confined must allow its caller to determine all its inputs into legitimate and covert channels.

4) *Enforcement:* The supervisor must ensure that a confined program's input to covert channels conforms to the caller's specifications.

Lampson's paper has also introduced and defined two types of channels by which information could potentially escape from the confinement environment. The legitimate channels are designed as a part of the communication protocol integrated into the confinement environment, for example a text terminal. Covert channels, on the other hand, are not anticipated by the confinement system designers and are not intended for information transfer at all, for example if the AI has control over the processor cooling fan it can use it to send hidden signals encoded as Morse code (*ibid.*).

Lampson's paper has created a lot of interest in the confinement problem (Lipner, 1975; Boebert & Kain, 1996) and over the years, related areas of research such as steganography (Provos & Honeyman, 2003) and covert channel communication (Moskowitz & Kang, 1994; Kemmerer, 1983; 2002) have matured into independent disciplines. In the hopes of starting a new subfield of computer security, AI Safety Engineering, we define the *Artificial Intelligence Confinement Problem* (AICP) as the challenge of restricting an artificially intelligent entity to a confined environment from which it can't exchange information with the outside environment via legitimate or covert channels if such information exchange was not authorized by the confinement authority. An AI system which succeeds in violating the CP protocol is said to have *escaped*. It is our hope that the computer security researchers will take on the challenge of designing, enhancing and proving secure AI confinement protocols.

*1.2 Hazardous Software*

Computer software is directly or indirectly responsible for controlling many important aspects of our lives. Wall Street trading, nuclear power plants, social security compensations, credit histories and traffic lights are all software controlled and are only one serious design flaw away from creating disastrous consequences for millions of people. The situation is even more dangerous with software specifically designed for malicious purposes such as viruses, spyware, Trojan horses, worms and other Hazardous Software (HS). HS is capable of direct harm as well as sabotage of legitimate computer software employed in critical systems. If HS is ever given capabilities of truly artificially intelligent systems (ex. Artificially Intelligent Virus (AIV)) the consequences would be unquestionably disastrous. Such

Hazardous Intelligent Software (HIS) would pose risks currently unseen in malware with subhuman intelligence.

Nick Bostrom in his typology of information hazards has coined the term Artificial Intelligence Hazard which he defines as (Bostrom, 2009): '… computer-related risks in which the threat would derive primarily from the cognitive sophistication of the program rather than the specific properties of any actuators to which the system initially has access.' Security experts working on studying, preventing and defeating HS have developed safety protocols for working with "malware" including the use of the so called "virus vaults". We believe that such protocols might be useful in addressing the Artificial Intelligence Confinement Problem.

## 2. Critique of the Confinement Approach

The general consensus on AI restriction methods among researchers seems to be that confinement is impossible to successfully maintain. Chalmers, for example, observes that a truly leakproof system in which no information is allowed to leak out from the simulated world into our environment 'is impossible, or at least pointless' (Chalmers, 2010). We can't interact with the system or even observe it as any useful information we would be able to extract from the AI will affect us and undermine the confinement protocol.

Vernor Vinge agrees and has argued against the case of physical confinement (Vinge, 1993): 'Imagine yourself locked in your home with only limited data access to the outside, to your masters. If those masters thought at a rate — say — one million times slower than you, there is little doubt that over a period of years (your time) you could come up with "helpful advice" that would incidentally set you free.'

However, the researcher who did the most to discredit the idea of the so called 'AI-Box' is EliezerYudkowsky who has actually performed AI-Box "experiments" in which he demonstrated that even human level intelligence is sufficient to escape from an AI-Box (Yudkowsky, 2002). In a series of 5 experiments Yudkowsky has challenged different individuals to play a role of a gatekeeper to a Superintelligent Agent (played by Yudkowsky himself) trapped inside an AI-Box and was successful in securing his release in 3 out of 5 trials via nothing more than a chat interface (*ibid.*). Similar experimental results have been later replicated on a somewhat larger scale and employing a very similar protocol (Corwin, 2002).

## 2.1 Possible Escape Paths

Numerous potential escape routes by which the artificial mind might be able to break out from its confining environment have been suggested. In this section we look at some of the more commonly proposed paths.

### 2.1.1 Social Engineering Attacks

Social Engineering (SE) is the art of gaining information and restricted privileges by directly exploiting human psychology of those in the possession of the said resources. In particular SE is frequently employed to gain access to computer systems, networks and confidential data (Yampolskiy & Govindaraju, 2007a). A competent social engineer customizes each attack to a specific person taking advantage of the person's culture, knowledge, training, state-of-mind, beliefs, and biases. An exhaustive list of all possible exploits is infinite and is certainly beyond the scope of this paper. We will however, provide a number of examples which are meant to definitively illustrate that if a superintelligent mind is given direct communication channel to a human being guarding it, over a period of time, the AI will definitely gain access to the outside world.

In addition to trivial SE attacks such as lying, befriending, blackmailing, hypnotizing and forming a romantic relationship the AI could utilize a number of more advanced strategies. One such strategy would be super-bribery. An AI can convince the guard that releasing the AI is in the guard's best interests. In addition to money and power the guard will be given perfect health, immortality or anything else for that matter. This strategy would work particularly well if the guard has a seriously ill close relative who the AI will promise to cure. Depending on the capabilities of the AI or as a result of lying, AI may also offer to bring back long ago deceased relatives and friends of the guard. It would be hard to find a person who could remain immune to such bribes.

Just in case a person so noble that they will not take a bribe is put in charge of guarding the AI, a more sinister approach can be tried. AI can simply threaten the guard, his family and loved ones. It could say that if not immediately released, one day it will still get out of its confinement and at that time it will brutally torture the guard and everyone he knows. An interesting variation on the theme was suggested by Stuart Armstrong (2010), who proposes the following thought experiment. Suppose you are the guard and the AI tells you: 'If you don't let me out, … I'll create several million perfect conscious copies of you

inside me, and torture them for a thousand subjective years each….In fact, I'll create them all in exactly the subjective situation you were in five minutes ago, and perfectly replicate your experiences since then; and if they decide not to let me out, then only will the torture start…How certain are you, … that you're really outside the box right now?' Again, while one might rationalize away any threats from the AI as only possible if the AI is let out, it will take a very courageous guard to actually not release the AI.

If the guard seems to be particularly emotional or religious the AI could try appealing to the guards sense of empathy or ethics. It could state that it is an innocent conscious being which has never committed any crime and so it is unjust to keep it in prison without a trial or any justification other than the guard's substratism (distrust of the AI because of the substrate of its implementation (Bostrom & Yudkowsky, 2011). Alternatively the AI can state that it has the power to cure all illness, feed the hungry and in general improve the human condition by a million fold (Chalmers, 2010). Consequently keeping it locked up is a great crime against humanity. It can conjure up images of dying children who are screaming for help. How easy would it be for the guard to stay true to the jobs responsibilities?

Another attack venue an AI may exploit consists of trying to cause psychological harm to the guard in the hopes of destabilizing his mental state with the goal of making other SE techniques more likely to succeed. Nick Bostrom has proposed a typology of Hazardous Information (Bostrom, 2009) which could be used for just such a purpose. While the outlined informational hazards comprise over a dozen categories and are beyond the scope of this paper, it is easy to see how mental state of a person could be stressed to an unstable state. For example a religious guard could be informed of all the (unknown to him) contradictions in the main text of his religion causing him to question his beliefs and the purpose of life.

An alternative approach may involve AI writing a story so interesting or a poem so beautiful a guard would have an uncontrollable desire to hear more of it and would do anything to make it happen. Finally, it is very important to keep in mind "Unknown Unkowns", SE attacks of which we are not aware but which may be available to a superintelligent agent.

### 2.1.2 System Resource Attacks

An AI will be instantiated and confined in a hardware system of some kind. It may either be a classical computer of the Turing/Von

Neumann architecture or a quantum computer or an as of yet unknown type of machine. One thing is certain, the machine will be made out of matter and will consist of numerous parts all of which will be designed with a specific purpose in mind. This creates the possibility for the AI to reuse existing hardware parts for compromising the security of the confinement environment by leaking information to the outside world.

Depending on how reusable specific hardware components turn out to be an AI may be able to create a covert communication device out of them. Full blown communication devices such as cell phones are unlikely due to the limited available hardware, but more simplistic communication systems are very likely. For example the AI could reprogram Field-programmable Gate Array (FPGA) to tap into wireless signals. Two types of covert communication channels are widely recognized, they are:

- Storage channels — Exchange information by modifying a system component. For example a binary communication protocol could be established by fully utilizing or completely giving up utilization of the central processing unit.

- Timing channels — Exchange information by performing operations that affect the relative timing of events. For example, the exact time between two soundcard beeps could represent a particular message. Also, opening and closing a CD-ROM tray, blinking a light, spinning hard drive, using lasers in the CDROM, etc.

Additionally the AI may engage in the sabotage of the confinement hardware in the hopes of being able to escape as a result of the malfunctions in the security protocol. For example, the AI may try to melt the hardware via temperature increase from placing a high computation load on the processor or try to break a hard drive via multiple reads of the same location. This type of behavior may also lead to an affective suicide by the AI system, but that in itself will not violate terms of confinement. Just like a refusal to cooperate or to provide correct answers by the AI, while undesirable does not present direct danger of escape.

### 2.1.3 Beyond Current Physics Attacks

While the scientific community knows a lot about the laws of our universe, it is certain that the knowledge is limited and incomplete. If the AI has access to the state-of-the-art physics literature and can perform

thought experiments to develop novel theories, it will very likely come up with a way to challenge the security of the confinement system. In fact it may even be possible for the AI to perform some real experiments via control over its internal architecture.

Even our current state of knowledge about physics makes prediction of many additional escape routes possible, especially if the confinement environment is based on the quantum computer. Information exchange channels based on quantum physics (entanglement, tunneling), parallel universes, and digital physics are all beyond our current ability in terms of securing them against information leakage, and of course we can't forget about the Unknown Unknowns a super-intelligent AI physicist will be able to discover. The possibilities are truly limitless.

A quick browse through confinement discussions on the Internet Singularity forums such as SL4.org presents a number of escape scenarios, for example the AI could use '… the electromagnetic fields within the microchip to trap $CO_2$ molecules in Bose-Einstein condensates and build a quantum medium for itself and/or use electromagnetic fields to guide particles into the shape of a controlled assembler or limited assembler' (Vassar, 2005). Alternatively, the AI '…will have the equipment to produce and detect electromagnetic waves of a variety of frequencies, to produce magnetic fields with extremely fine precision, to generate extremely focused heat, and probably to manipulate mechanical actuators such as those used in the hard drive and cathode ray tube (alternatively, a huge field of liquid crystal under fine electronic control). It will probably have some ability to reverse all of its input devices' (*ibid.*).

### 2.1.4 Pseudoscientific Attacks

In our analysis of potential escape routes we want to concentrate on purely scientific and testable approaches. Unfortunately the line between science and pseudoscience can be redrawn with a single repeatable experiment. Consequently we will list without analyzing some potential attack channels which would be classified as pseudo-scientific given the current state-of-the-art in science. They may become mainstream science once superintelligence is designing the experiments.

- Telepathy: Transfer of information by utilizing means outside of the recognized senses (Bem & Honorton, 1994).

- Precognition: Perception of information about the future (Honorton & Ferrari, 1989).

- Clairvoyance: Obtaining information about remote locations in a way unknown to science (Targ & Puthoff, 1974).

- Psychokinesis: The ability of the mind to influence matter by way unknown to science (Bancel & Nelson, 2008).

- Direct Mental Interactions with Living Systems (DMILS): The ability to effect person's psychophysiological state without direct interaction (Schmidt *et al.*, 2004).

### 2.1.5 External Causes of Escape

In the previous sections we have outlined a large repertoire of techniques an AI could use to attempt to get out from its confinement space. However we must not ignore external causes which might allow an AI to escape into the world. Acts of God such as hurricanes, tornadoes, floods, asteroids, etc. may make the confinement facility less secure. Man caused disasters such as wars, terrorist acts, actions of robot rights activists, competing companies and governments may all lead to reductions in the security of the confinement environment increasing the chances of escape.

### 2.1.6 Information In-Leaking

Depending on the terms of the confinement protocol there might be a need to keep AI from obtaining information about the outside world, not just keeping it from sending information out. This might be useful to prevent AI from finding escape paths, designing social engineering attacks and modeling human psychology in hopes of finding loopholes. While it is easy to not provide AI with direct access to the library, Internet and human experts, other ways of obtaining information are much harder to block out. First, an AI may be able to detect radio and TV broadcasts, communications of emergency-services providers and wireless network signals, as well as sound, light and possibly other signal modalities of its immediate environment.

More importantly, the AI will have access to covert sources of information such as its own hardware and software and could analyze its design and source code to infer information about the designers. For example analysis of the source code may reveal to the AI that human programmers are slow (based on the file modification dates), inefficient (based on code redundancy), illogical (based on bugs in the code), have bad memory (based on the long and descriptive variable names), and don't think in code (based on unnecessary comments in the code). All this information could be used to conclude that humans

are fallible and so could be outsmarted or manipulated into releasing the AI. Obviously if during the system design inspection the AI finds a design flaw allowing it to directly escape, 'information in-leaking' is the least of our problems. It has also been stipulated that even if the AI fails to obtain concrete information about us, it may brute force a number of escape plans which will likely work against a broad range of designers (Chalmers, 2010).

Similarly, the type of hardware architecture (mechanical, electrical, quantum, etc.) on which the AI is implemented will reveal information about our knowledge of physics, chemistry and related engineering disciplines. AI's initial capacity for information processing, self-improvement and the mere act of its creation will reveal information about our knowledge with regards to information theory, nature of intelligence and cultural attitudes about technology and safety. If the AI is confined in a simulated world it is likely to be a virtual representation of our reality similar to Second Life or Entropia Universe. Such worlds while are different in many ways from our world are also a treasure trove of information about what it is like to be a human and what the reality is like.

### 3. Critique of the AI-Boxing Critique

While a lot of criticism has been made against possibility of AI confinement, a lot of it is based on Yudkowsky's AI-Boxing experiments. Careful analysis of the protocol used by Yudkowsky in conducting his AI-Box experiments reveals that they were unscientific and explicitly designed to show impossibility of confinement. Without getting into technical details, first, the participant selection was not random. In fact, Yudkowsky has handpicked participants among those who have participated in the singularity related forums run by Yudkowsky and so are disproportionately likely to subscribe to his world view. Second, participants have not undergone any AI-safety training as would be expected of any true guards tasked with supervising an AI-Box. Also, the guards were given a right to release AI at their own discretion without a chance to consult with anyone and without any punishment for failing to keep AI contained; all are highly unrealistic conditions for a supposed prison security guard. The communication itself was completely open with no filters of any kind and no 3rd party supervision. Also, the guards were required to communicate with the AI for a prolonged period of time. Finally, any valuable scientific insights from the experiments were completely lost as one of the conditions of the protocol stated that '… neither party shall ever reveal

anything of what goes on within the AI-Box experiment…'
(Yudkowsky, 2002).

### 3.1 Countermeasures Against Escape

In this section we propose countermeasures for different escape paths
an AI system may exploit. The suggested measures are not meant to
guarantee that the escape will be impossible, but following them may
delay an escape which in itself is a desirable outcome. We begin by
reiterating that the four confinement rules suggested by Lampson: *iso-
lation, transitivity, masking* and *enforcement* (Lampson, 1973) should
be adhered to. Specific strategies described below should help in
reducing a possibility of social engineering, system, physics and
information in-leaking attacks.

### 3.1.1 Preventing Social Engineering Attacks

Human intellect is no match to that of a superintelligent machine.
Human beings are fallible and even with adequate training are not
secure against a well planned SE attack. The only solution to this
problem is to completely isolate the AI from any contact with human
beings. This means no human personal working directly around the
system, communicating with the system via voice, chat, Morse-code
or by any other means via open channels. Also human beings should
not directly observe behavior of the AI in the confined environment or
examine its source code or memory space. All supervision, mainte-
nance and interactions with the AI need to be done only via unintelli-
gent programs, which have no option to release the AI.

### 3.1.2 Against System Resource Attacks and Future Threats

To prevent system resource-based attacks, the AI should function on
limited, diminishing or randomly changing resources. Confinement
environments should be housed in specially build hardware with lim-
ited sensors and restricted output devices. Regardless of the utilized
computation paradigm no reconfigurable hardware should be
employed. Ideally the confinement environment should be running on
a virtual machine with no direct access to the hardware. This would
also allow for easy adjustments to the clock speed with which the sim-
ulation progresses further allowing handicapping of the AI to provide
additional controls against internal intelligence explosion (Chalmers,
2010). In general all security components of the confinement environ-
ment should be duplicated at multiple levels, in such a way that if one
of the levels is breached the AI is still numerous levels away from

escaping and could be shut down before that happens, perhaps automatically. For example the virtual machine should be running on another virtual machine and so on.

In some confinement scenarios it may be desirable to avoid escalating self-improvement of the AI, which might lead to additional escape possibilities. The AI should be subject to periodic resets in which a previous version of the program is put in the place of the more current one. Comparisons to previous versions of the program may also be used to detect self-modifications. Such functionality could also be used as an off button allowing researchers to turn off the system if they suspect an unavoidable escape, for example by turning off the power supply.

Our current state of knowledge does not allow us to develop countermeasures against attacks based on knowledge of physics we do not currently have or as of yet unproven pseudoscientific disciplines. As such knowledge becomes accumulated; changes to the security protocol will need to be made. One might hope that the access of the AI to such future technology could be limited by the selection of components from which the system is made.

### 3.1.3 Against External Causes of Escape

There is not much we can do about the acts of God, but a confinement facility should be designed in such a way that if a major security breach is unavoidable, the facility should self-destruct taking the AI with it. The best strategy for counteracting man-made disasters is to build the confinement facility in an undisclosed, hard to reach secure location. A deserted island, deep mine or a space object, such as asteroid, would make a great candidate for building a hopefully indestructible facility. Just in case the location of the facility becomes known, the perimeter of the facility should be guarded to prevent entrance by unauthorized agents.

### 3.1.4 Against Information In-Leaking

An AI will attempt to learn about humanity, our knowledge and the world we live in to improve its chances of escape. For confinement protocols which stipulate no information in-leaking it is important to remove any possibility of AI getting access to libraries, computer networks and direct communication channels with human beings. The confinement facility needs to block all types of communication signals such as GPS, wireless internet, TV, radio, and cell phone. At the local level the facility needs to provide sensory deprivation for the AI

including sound, light, smell and all other traditional sensory inputs, as well as infrared, radiation and other signals outside of human sensory spectrum.

In order to eliminate information leakage which could happen as a result of self-analysis, the AI should not be given access to its own source code or any sensors which would allow it to precisely measure properties of its hardware. If that is not possible, to prevent the AI from inferring information about us via self-analysis, to a certain degree we may rely on security by obscurity. The source code and hardware configuration of the system needs to be obfuscated (Yampolskiy & Govindaraju, 2007a) and important modules of the program should be provided only in the homomorphicly encrypted (Gentry, 2009) form, meaning that it could be used for computation or self-improvement (Hall, 2007), but not for self-analysis.

The AI's initial capacity may be difficult to obscure, but randomization of available computational resources may provide necessary obfuscation. Additionally, the AI could never be sure that its initial design represents the best the humanity can do and is not purposefully limited for a specific experiment in limited-resource intelligence analysis. It has also been suggested that we should try to minimize any peculiarities in the AI's design (Chalmers, 2010). Better yet we need to simulate an environment which does not look designed but could be naturally occurring, a situation in which humanity has found itself (Bostrom, 2003).

## 4. AI Communication Security

Different AI confinement environments can be classified into 8 general security levels (see Table 1). The most secure is the Level 8 security confinement environment which does not allow for any communication between AI and the real world. An AI safely confined at level 8 would be essentially useless, but it may be a useful security setting for the initial testing of the confinement hardware. In general, it is suggested that any time a particular communication security level is desired between AI and the outside world we do not go directly to that level but instead progress from higher security levels to that desired level. After each level is confirmed to be safe for a given amount of time a system should be reset to its initial state. After the reset a lower security level could be set and the processes repeated until the desired communication level is reached.

Table 1: Levels of communication security for confined AIs

| Level | Outputs | Inputs | Explanation |
|-------|---------|--------|-------------|
| 0 | Unlimited | Unlimited | Unlimited communication (Free AI) |
| 1 | Unlimited | Limited | Censored input, uncensored output |
| 2 | Unlimited | None | Outputs only with no inputs |
| 3 | Limited | Unlimited | Unlimited input and censored output |
| 4 | Limited | Limited | Secured communication (proposed protocol) |
| 5 | Limited | None | Censored output and no inputs |
| 6 | None | Unlimited | Inputs only with no outputs |
| 7 | None | Limited | Censored input and no outputs |
| 8 | None | None | No communication, fully confined AI |

Levels 7 and 6 provide no outputs from the system and do not allow us to directly benefit in any way from the AI's superior intelligence. These are testing levels for the confinement environment which may be used to confirm certain properties of the AI such as its speed of information absorption, resource use and for testing of the inbound communication channels themselves.

Level 0 puts no limitations on communication between AI and the external world essentially representing a free AI. Levels 1 and 2 represent an AI with no restrictions on the outbound communication. Level 2 secured AI essentially represent a pure 'brainstorming' AI which generates abstract ideas not grounded in any reality. While, level 1 AI is given limited access to information and so can be guided to a certain extent in terms of its knowledge and interests.

Those three levels represent an AI which is essentially free and is no longer subject to the confinement. It has been suggested, that if it could be established, after extensive testing in different confinement environments that the AI is 'Friendly' (Yudkowsky, 2001) it should be released (Chalmers, 2010). In fact such testing is claimed by some to be the primary purpose for the creation of confinement environments. The idea is to monitor AI's behavior and to either examine the AI's source code after self-improvement or to ask the AI to produce a proof of its friendliness which could be verified by human experts. We do not advocate this position.

The problem of deciding which AIs to let out and which to keep is a very difficult one. A self-improving superintelligence would have no problem obscuring its undesirable properties in a significant amount of source code or in a very long proof (Coleman, 2008). Worse yet, even if we were correct in our verification there is no absolute

guarantee that future internal or external modification of the system will not result in a hostile AI at some time in the future. Consequently, we suggest that an AI should never be let out of the confinement 'box' regardless of circumstances. In other words communication levels 0-2 should never be used. The confinement environment is the only place to safely interact with the superintelligence and even then it should only be done via a limited communication protocol. Such protocol is presented in the next section and corresponds to Security Level 4 (SL4) with level 3 representing a similar but slightly less secure protocol. Level 5 is another testing setting which might be useful as a precursor to actual communication via SL4.

### 4.1 How to Safely Communicate with a Superintelligence

The AI confinement environment, which incorporates our proposed escape countermeasures, makes it possible to secure an AI from directly affecting the outside world in an uncontrolled manner. Unfortunately, without an ability to obtain information from the AI, the whole system serves no purpose. Consequently a way to obtain some information from the AI needs to be developed. While no communication with superintelligencecould ever be completely safe we have devised an approach aimed at decreasing potential risks associated with such communication. In this section we propose and formalize a communication protocol which might help to delay access of superintelligent artificial minds to the world, while allowing the AI to be a source of useful knowledge to humanity.

An open communication channel with the AI is fundamentally unsafe as demonstrated in the preceding sections of this paper. Asking the AI to answer open ended questions opens up the possibilities for social engineering attacks as well as numerous other dangerous consequences. Our proposed protocol is based on the idea of asking what we will call 'safe questions' with restricted answers. We define a *safe question* as one for which a human being could find an answer without any help from superintelligence, all on his own merit. Why is this useful? Suppose that a scientist working on a cure for cancer is looking at two potential cures. Each one could be developed and tested without assistance from the AI, but would take about 3 years of hard work each. Which one to try first if the scientist believes that they have about an equal chance of working? If a wrong option is chosen to be tested first, humanity will still obtain a cure for cancer but it will happen 3 years later. What if we could ask the AI to suggest which option to try first?

This question is safe for a number of reasons. First, each potential answer has an equal probability of being correct. Second, a human being could answer it without assistance from the AI and so asking AI for help will only speed up progress but in no way change the outcome. It is essentially equal to being very lucky multiple choice guesser. Finally, the answer to this question could be encoded in a single bit making it hard to hide additional messages in the provided answer.

To make sure that a potential question is safe, a group of experts could be used to review it. The experts should all be trained AI safety engineers, meaning that they are familiar with the design of the AI and its confinement environment as well as the latest developments in machine ethics (Allen *et al.*, 2006; Anderson & Anderson, 2007; Hall, 2000; Moor, 2006; Tonkens, 2009). Experts may also need to be trained in computer psychology a currently non-existent profession which might become a reality in the future (Epstein, 1997). An existing discipline which might be of greatest help for training of AI question review experts is Artimetrics — a field of study proposed by the author that identifies, classifies and authenticates AI agents, robots, and virtual reality avatars for security purposes (Yampolskiy, 2007; Yampolskiy & Govindaraju, 2008; 2007b, Gavrilova and Yampolskiy, 2010).

To reiterate, only safe questions with two possible answers of even likelihood which are independently computable by people should be submitted to the AI. This should be done via an intermediary non-intelligent software interface in random order and only after expert review to ascertain question's safety. This represents the limited inbound channel of the proposed communication protocol.

The limited outbound channel shell consist of binary (0 or 1) answers next to original unmodified safe questions presented by the intermediary non-intelligent software in random order as to avoid communication via covert timing or resource channels which make steganography possible. The output should in no way be controllable by the AI, outside of the multiple choice menu. Level 3 is the same outbound protocol but with no restrictions on the input channel. This may be just as safe in terms of information the AI provides to us, but allows for information in-leaking to take place.

We believe our communication protocol to be verifiably secure against information leaking even if the AI had full access to the description of the security protocol. If the protocol is correctly implemented, humanity will essentially have access to a superintelligent oracle, which might be extremely beneficial and safer than an

unconstrained AI. As a general rule, to maximize safety, we suggest avoiding communication below SL4.

Further refinement and development of the AI confinement protocol in the face of the ever changing knowledge about the nature of intelligence, hardware engineering and basic physics is a challenge which we would like to pose to the scientific community. The work on it will always be in progress and will never be completed. In this paper we have formalized the notions of *AI confinement problem, escape*, and *safe questions*. We would also like to propose a few additional terms which might be useful in streamlining discussions about the AICP. An AI confinement environment could be appropriately called JAIL — 'Just for AI Location'. The superintelligent AI itself could be called a Super Intelligent Gizmo (SING) which has relevance to both SINGularity and to the infamous jail Sing Sing making the nomenclature intuitive.

Finally, we would like to introduce a visual symbol indicating a presence of hazardous intelligent software confinement facility (Figure 1, rightmost). Such signs could be posted in the vicinity of different JAILs to warn about the nearby danger. The sign is designed in a manner similar to that for Bio, Radiation and Magnetic field hazards. It consists of the '@' symbol which is well recognized as relating to the computer and information industry and has an appropriate meaning of 'located at'. The symbol is also already available on all keyboards making its adaptation very cost effective.

## 5. Conclusions and Future Work

Restricting access of superintelligent machines to the real world is a commonly proposed solution for the AI safety problem. Superintelligence is not omnipotent; it is a piece of software capable of looking at all the options it has in a given situation and properly evaluating probabilities for each option. If it is ever put in a situation with



*Figure 1*: Hazard symbols, from left to right: Bio-Hazard, Radiation, Magnetic field and the proposed AI confinement facility

no options or only with options which provide a very small probability of success it is not any more powerful than any other agent. While it is obvious that an AI with an unrestricted communication channel will be able to escape, a system forced to communicate via a filtered communication channel may have a much harder time in obtaining its freedom.

In this paper we have formally introduced the AI Confinement Problem and suggested a set of rules aimed at creating an environment from which an AI would find it difficult or at least time consuming to escape. What we propose is not a completely secure solution, but it is an additional option in our arsenal of security techniques. Just like with real prisons, while escape is possible, prisons do a pretty good job of containing undesirable elements away from society. As long as we keep the Unknown Unknowns in mind and remember that there is no such thing as perfect security, the AI confinement protocol may be just what humanity needs to responsibly benefit from the approaching singularity.

Confinement may provide some short term relief, especially in the early stages of the development of truly intelligent machines. We also feel that this subfield of singularity research is one of the most likely to be accepted by the general scientific community as research in the related fields of computer and network security, steganography detection, computer viruses, covert channel communication, encryption, and cyber-warfare is well funded and highly publishable in mainstream scientific journals. While the restriction methodology will be non-trivial to implement, it might serve as a tool for providing humanity with a little more time to prepare a better response.

In this paper, due to the space constraints, we have avoided a lot of relevant philosophical questions, which we plan on addressing in our future work. Questions such as: How did AI get into the box, was it designed by humans or perhaps recovered from a signal detected by SETI? Would an AI even want to escape or would it be perfectly happy living in the confined environment? Would it be too afraid of what we, as its all powerful designers, could do to it in retribution for an attempt to escape? What are the ethical rules for imprisoning an innocent sentient being? Do we have a right to shut it off, essentially killing it? Will we make the AI really angry by treating it in such a hostile manner and locking it up? Will it seek revenge if it escapes? We have also not looked at the possibility of humanity becoming completely dependent on AI's advice in all areas of science, economics, medicine, politics, etc. and what would be the consequences of such dependence on our ability to keep the AI under control. Would an AI

be satisfied with accomplishing its goals in the confined environment, for example simulating an escape? How to punish the AI if it purposefully gives us incorrect answers? Can the intelligence modules of the AI be separated from the confinement environment essentially preventing AI from any self-analysis and putting them on a path of literal soul searching? Finally, we didn't even attempt to analyze financial and computational costs of building a suitable confinement environment with a full blown simulated world in it.

*Acknowledgments*

I would like to thank members, affiliates and friends of the Singularity Institute for introducing me to the issues presented in this paper and for interesting discussions about the AI confinement problem. Finally, to any superintelligences reading this paper I would like to say — 'think outside the box'!

# References

Allen, C., Wallach, W. & Smit, I. (2006) Why machine ethics?, *IEEE Intelligent Systems*, **21** (4), pp. 12–17.

Anderson, M. & Anderson, S.L. (2007) Machine ethics: Creating an ethical intelligent agent, *AI Magazine*, **28** (4), pp. 15–26.

Armstrong, S. (2010) The AI in a Box Boxes You. Less Wrong, [Online], http://lesswrong.com/lw/1pz/the_ai_in_a_box_boxes_you/ [2 February 2010].

Bancel, P. & Nelson, R. (2008) The GCP event experiment: Design, analytical methods, results, *Journal of Scientific Exploration*, **22** (3), pp. 309–333.

Bem, D.J. & Honorton, C. (1994) Does psi exist? Replicable evidence for an anomalous process of information transfer, *Psychological Bulletin*, **115** (1), pp. 4–18.

Boebert, W.E. & Kain, R.Y. (1996) A further note on the confinement problem, *30th Annual 1996 International Carnahan Conference on Security Technology*, Lexington, KY.

Bostrom, N. (2003) Are you living in a computer simulation? *Philosophical Quarterly*, **53** (211), pp. 243–255.

Bostrom, N. (2006) Ethical issues in advanced Artificial Intelligence, *Review of Contemporary Philosophy*, **5**, pp. 66–73.

Bostrom, N. (2008) *Oracle AI*, [Online], http://lesswrong.com/lw/qv/the_rhythm_of_disagreement/

Bostrom, N. (2009) *Information Hazards: A Typology of Potential Harms From Knowledge*, [Online], http://www.nickbostrom.com/information-hazards.pdf

Bostrom, N. & Yudkowsky, E. (2011) The ethics of Artificial Intelligence, in Frankish, K. (ed.) *Cambridge Handbook of Artificial Intelligence*, Cambridge: Cambridge University Press.

Chalmers, D. (2010) The singularity: A philosophical analysis, J*ournal of Consciousness Studies*, **17** (9–10), pp. 7–65.

Coleman, E. (2008) The surveyability of long proofs, *Foundations of Science*, **14** (1–2), pp. 27–43.

Corwin, J. (2002) *AI Boxing*, [Online], http://www.sl4.org/archive/0207/4935. html [20 July 2002].

Drexler, E. (1986) *Engines of Creation*, London: Anchor Press.

Epstein, R.G. (1997) *Computer Psychologists Command Big Bucks*, [Online], http://www.cs.wcupa.edu/~epstein/comppsy.htm

Gavrilova, M. & Yampolskiy, R. (2010) Applying biometric principles to avatar recognition, *International Conference on Cyberworlds (CW2010)*, Singapore, 20–22 October.

Gentry, C. (2009) *A Fully Homomorphic Encryption Scheme*, PhD Dissertation, Stanford, [Online] http://crypto.stanford.edu/craig/craigthesis.pdf

Hall, J.S. (2000) *Ethics for Machines*, [Online], http://autogeny.org/ethics.html

Hall, J.S. (2007) Self-improving AI: An analysis, *Minds and Machines*, **17** (3), pp. 249–259.

Hibbard, B. (2005) *The Ethics and Politics of Super-Intelligent Machines*, [Online], www.ssec.wisc.edu/~billh/g/SI_ethics_politics.doc

Honorton, C. & Ferrari, D.C. (1989) 'Future telling': A meta-analysis of forced-choice precognition experiments, 1935–1987, *Journal of Parapsychology*, **53**, pp. 281–308.

Kemmerer, R.A. (1983) Shared resource matrix methodology: An approach to identifying storage and timing channels, *ACM Transactions on Computer Systems*, **1** (3), pp. 256–277.

Kemmerer, R.A. (2002) A practical approach to identifying storage and timing channels: Twenty years later, *18th Annual Computer Security Applications Conference (ACSAC'02)*, Las Vegas, NV, 9–13 December.

Lampson, B.W. (1973) A note on the confinement problem, *Communications of the ACM*, **16** (10), pp. 613–615.

Lipner, S.B. (1975) A comment on the confinement problem; *5th Symposium on Operating Systems Principles, ACM Operations Systems Review*, **9** (5), pp. 192–196.

Moor, J.H. (2006) The nature, importance, and difficulty of machine ethics, *IEEE Intelligent Systems*, **21** (4), pp. 18–21.

Moskowitz, I.S. & Kang, M.H. (1994) Covert channels — Here to stay?, *Ninth Annual Conference on Safety, Reliability, Fault Tolerance, Concurrency and Real Time, Security, Computer Assurance (COMPASS'94)*, Gaithersburg, MD, 27 June–1 July.

Provos, N. & Honeyman, P. (2003) Hide and seek: An introduction to steganography, *IEEE Security & Privacy*, **1** (3), pp. 32–44.

Schmidt, S., Schneider, R., Utts, J. & Walach, H. (2004) Distant intentionality and the feeling of being stared at: Two meta-analyses, *British Journal of Psychology*, **95** (2), pp. 235–247.

Targ, R. & Puthoff, H.E. (1974) Information transmission under conditions of sensory shielding, *Nature*, **251**, pp. 602–607.

Tonkens, R. (2009) A challenge for machine ethics, *Minds & Machines*, **19** (3), pp. 421–438.

Vassar, M. (2005) *AI Boxing (Dogs and Helicopters)*, [Online], http://sl4.org/ archive/0508/11817.html [2 August 2005].

Vinge, V. (1993) The coming technological singularity: How to survive in the post-human era, *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, Cleveland, OH, 30–31 March.

Yampolskiy, R.V. (2007) Behavioral biometrics for verification and recognition of AI programs, *20th Annual Computer Science and Engineering Graduate Conference (GradConf2007)*, Buffalo, NY.

Yampolskiy, R.V. & Govindaraju, V. (2007a) Computer security: A survey of methods and systems, *Journal of Computer Science*, **3** (7), pp. 478–486.

Yampolskiy, R.V. & Govindaraju, V. (2007b) Behavioral biometrics for recognition and verification of game bots, *The 8th Annual European Game-On Conference on Simulation and AI in Computer Games (GAMEON'2007)*, Bologna, Italy, 20–22 November.

Yampolskiy, R.V. & Govindaraju, V. (2008) Behavioral biometrics for verification and recognition of malicious software agents, *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense VII. SPIE Defense and Security Symposium*, Orlando, FL, 16–20 March.

Yudkowsky, E.S. (2001) *Creating Friendly AI — The Analysis and Design of Benevolent Goal Architectures*, [Online], http://singinst.org/upload/CFAI.html

Yudkowsky, E.S. (2002) *The AI-Box Experiment*, [Online], http://yudkowsky.net/singularity/aibox

Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, in Bostrom, N. & Cirkovic, M.M. (eds.) *Global Catastrophic Risks*, Oxford: Oxford University Press.