CHAPTER 8

Artificial Intelligence Confinement Problem (and Solution)*

I am the slave of the lamp.

—**Genie from** *Aladdin*

8.1 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†) safety research is the possibility of keeping a superintelligent agent in sealed hardware 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, 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. He suggested

* Reprinted from Roman V. Yampolskiy, *Journal of Consciousness Studies (JCS)* 19(1–2):194–214, 2012, with kind permission from Imprint Academic. Copyright 2012.

† In this chapter, the term *artificial intelligence (AI)* is used to represent superintelligence.

K22996.indb 145 28/05/15 12:47 pm

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 chapter, I review specific proposals aimed at creating restricted environments for safely interacting with artificial minds. I evaluate the feasibility of the proposals presented and suggest a protocol aimed at enhancing the safety and security of such methodologies. Although it is unlikely that long-term and secure confinement of AI is possible, I am 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 the appearance of superintelligent machines.

8.1.1 Artificial Intelligence Confinement Problem

Interestingly, the AI confinement problem (AICP) is not a recent invention and does not have its roots in the singularity movement. The so-called confinement problem was posed by Butler Lampson in 1973 as a security challenge to 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. … 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" (Lampson 1973, 1–2).

To address the confinement problem, Lampson introduced the laws of confinement (Lampson 1973, 3–4):

1. *Total isolation:* A confined program shall make no calls on any other program.
2. *Transitivity:* If a confined program calls another program that 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 work 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 (e.g., 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 (e.g., if the AI has control over the processor cooling fan, it can use it to send hidden signals encoded as Morse code) (Lampson 1973).

Lampson's work created a lot of interest in the confinement problem (Lipner 1975; Boebert and Kain 1996), and over the years, related areas of research such as steganography (Provos and Honeyman 2003) and covert channel communication (Moskowitz and Kang 1994; Kemmerer 1983, 2002) have matured into independent disciplines. In the hopes of starting a new subfield of computer security, AI safety engineering, I define the AICP as the challenge of restricting an artificially intelligent entity to a confined environment from which it cannot 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 that succeeds in violating the confinement problem protocol is said to have escaped. It is my hope that computer security researchers will take on the challenge of designing, enhancing, and proving secure AI confinement protocols.

8.2 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 compensation, 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 the capabilities of truly artificially intelligent systems (e.g., artificially intelligent virus), the consequences unquestionably would be 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 "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" (Bostrum 2011, 68). Security experts working on studying, preventing, and defeating HS have developed safety protocols for working with "malware," including the use of so-called virus vaults. I believe that such protocols might be useful in addressing the AICP.

8.3 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 from the simulated world into our environment "is impossible, or at least pointless" (Chalmers 2010, 38). We cannot 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 … 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 Eliezer Yudkowsky, 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 five experiments, Yudkowsky challenged different individuals to play a role of gatekeeper to a superintelligent agent (played by Yudkowsky himself) trapped inside an AI-Box and was successful in securing his release in three of five trials via nothing more than a chat interface (Yudkowsky 2002). Similar experimental results have later been replicated on a somewhat larger scale employing a similar protocol (Corwin 2002).

8.4 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, I look at some of the more commonly proposed paths.

8.4.1 Social Engineering Attacks

Social engineering (SE) is the art of gaining information and restricted privileges by directly exploiting the human psychology of those in the possession of these resources. In particular, SE is frequently employed to gain access to computer systems, networks, and confidential data (Yampolskiy and Govindaraju 2007). 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 book. However, I provide a number of examples meant to definitively illustrate that if a superintelligent mind is given a 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 superbribery. 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, the AI may also offer to bring back long-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 he will not take a bribe is put in charge of guarding the AI, a more sinister approach can be tried. The AI can simply threaten the guard's 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 the guard 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 … in exactly the subjective situation you were in … , 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, although one might rationalize away any threats from the AI as only possible if the AI is let out, it will take a courageous guard not to release the AI.

If the guard seems to be particularly emotional or religious, the AI could try appealing to the guard's sense of empathy or ethics. It could state that it is an innocent conscious being that has never committed any crime, 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 and 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 responsibilities of the job?

Another attack venue an AI may exploit consists of trying to cause psychological harm to the guard in the hopes of destabilizing the guard's mental state with the goal of making other SE techniques more likely to succeed. Nick Bostrom proposes a typology of hazardous information (Bostrom 2011) that could be used for just such a purpose. Although the outlined informational hazards comprise over a dozen categories and are beyond the scope of this book, it is easy to see how the mental state of a person could be stressed to become unstable. For example, a religious guard could be informed of all the (unknown to the guard) contradictions in the main text of the guard's religion, causing the guard to question personal beliefs and the purpose of life.

An alternative approach may involve the 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 important to keep in mind "unknown unknowns," SE attacks of which we are not aware but that may be available to a superintelligent agent.

8.4.2 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-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 from them. Full-blown communication devices such as cell phones are unlikely because of the limited available hardware, but more simplistic communication systems are likely. For example, the AI could reprogram a field-programmable gate array (FPGA) to tap into wireless signals. Two types of covert communication channels are widely recognized:

- Storage channels: These 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: These exchange information by performing operations that affect the relative timing of events. For example, the exact time between two sound card beeps could represent a particular message, as could opening and closing a CD-ROM tray, blinking a light, spinning a hard drive, using lasers in the CD-ROM, and so on.

In addition, 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 a 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 effective suicide by the AI system, but that in itself will not violate terms of confinement, just as a refusal to cooperate or to provide correct answers by the AI, although undesirable, does not present direct danger of escape.

8.4.3 Beyond Current Physics Attacks

Although 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 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. Of course, we cannot forget about the unknown unknowns a superintelligent 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 generate extremely focused heat, and probably to manipulate mechanical actuators such as those used in the hard drive and cathode ray tube. … It will probably have some ability to reverse all of its input devices" (Vassar 2005).

8.4.4 Pseudoscientific Attacks

In my analysis of potential escape routes, I 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, I list without analysis some potential attack channels that would be classified as pseudoscientific 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 the recognized senses (Ben and Honorton 1994)
- Precognition: Perception of information about the future (Honorton and Ferrari 1989)
- Clairvoyance: Obtaining information about remote locations in a way unknown to science (Targ and Puthoff 1974)
- Psychokinesis: The ability of the mind to influence matter by a way unknown to science (Bancel and Nelson 2008)
- Direct mental interactions with living systems (DMILS): The ability to affect a person's psychophysiological state without direct interaction (Schmidt et al. 2004)

8.4.5 External Causes of Escape

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

8.4.6 Information In-Leaking

Depending on the terms of the confinement protocol, there might be a need to keep an AI from obtaining information about the outside world (information in-leaking), not just keeping it from sending information out. This might be useful to prevent AI from finding escape paths, designing SE attacks, and modeling human psychology in hopes of finding loopholes. Although it is easy not to 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 important, 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), are inefficient (based on code redundancy), are illogical (based on bugs in the code), have bad memory (based on the long and descriptive variable names), and do not 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 that allows 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 use brute force in a number of escape plans that 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 and self-improvement and the mere act of its creation will reveal information about our knowledge with regard to information theory, the 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, although 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.

8.5 CRITIQUE OF THE AI-BOXING CRITIQUE

Although a lot of criticism has been made against the possibility of AI confinement, much 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 the impossibility of confinement. Without getting into technical details, first, the participant selection was not random. In fact, Yudkowsky handpicked participants among those who had participated in the singularity-related forums run by Yudkowsky and so were disproportionately likely to subscribe to his worldview. Second, participants had not undergone any AI safety training, as would be expected of any true guards given the task of 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).

8.6 COUNTERMEASURES AGAINST ESCAPE

In this section, I 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. I begin by reiterating that the four confinement rules suggested by Lampson—isolation, transitivity, masking, and enforcement (Lampson 1973)—should be adhered to. Specific strategies described should help reduce the possibility of SE, system, physics, and information in-leaking attacks.

8.6.1 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 people working directly around the system or communicating with the system via voice, chat, Morse code, or any other means via open channels. Also, human beings should not directly observe the behavior of the AI in the confined environment or examine its source code or memory space. All supervision, maintenance, and interactions with the AI need to be done only via unintelligent programs that have no option to release the AI.

8.6.2 Preventing 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 built hardware with limited sensors and restricted output devices. Regardless of the computation paradigm utilized, 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 simulation progresses further, allowing handicapping of the AI to provide additional controls against an internal intelligence explosion (Chalmers 2010). In general, all security components of the confinement environment 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 (e.g., 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 on 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.

8.6.3 Preventing 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 human-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 an asteroid, would make a great candidate for building a facility that is hoped to be indestructible. Just in case the location of the facility becomes known, the perimeter of the facility should be guarded to prevent entrance by unauthorized agents.

8.6.4 Preventing 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 that stipulate no information in-leaking, it is important to remove any possibility of AI gaining 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 exclusion of sound, light, smell, and all other traditional sensory inputs, as well as infrared, radiation, and other signals outside the human sensory spectrum.

To eliminate information leakage that could happen as a result of self-analysis, the AI should not be given access to its own source code or any sensors that would allow it to precisely measure the 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 need to be obfuscated (Yampolskiy and Govindaraju 2007), and important modules of the program should be provided only in the homomorphically 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. In addition, the AI could never be sure that its initial design represents the best 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 that does not look designed but could be naturally occurring, a situation in which humanity has found itself (Bostrom 2003).

8.7 AI COMMUNICATION SECURITY

Different AI confinement environments can be classified into nine general security levels (see Table 8.1). 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 8.1 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 that may be used to confirm certain properties of the AI, such as its speed of information absorption and 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 outbound communication. Level 2 secured AI essentially represents a pure "brainstorming" AI that generates abstract ideas not grounded in any reality. 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 that is essentially free and is no longer subject to 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 the AI's behavior and either to examine the AI's source code after self-improvement or to ask the AI to produce proof of its friendliness that could be verified by human experts. I do not advocate this position.

The problem of deciding which AIs to let out and which to keep is difficult. A self-improving superintelligence would have no problem obscuring its undesirable properties in a significant amount of source code or in a 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 future time. Consequently, I 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 superintelligence, and even then, it should only be done via a limited communication protocol. Such a 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 that might be useful as a precursor to actual communication via SL4.

8.8 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. Although no communication with superintelligence could ever be completely safe, an approach aimed at decreasing potential risks associated with such communication has been devised. In this section, I propose and formalize a communication protocol that might help 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 previous sections of this chapter. Asking the AI to answer open-ended questions presents the possibilities for SE attacks as well as numerous other dangerous consequences. My proposed protocol is based on the idea of asking what we will call "safe questions" with restricted answers. I define a *safe question* as one for which a human being could find an answer without any help from superintelligence, all on his or her 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 each would take about three years of hard work. Which one should be tried first if the scientist believes that each has 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 three 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, so asking the AI for help will only speed up progress but in no way change the outcome. It is essentially equal to being a 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 answer provided.

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, Wallach, and Smit 2006; Anderson and Anderson 2007; Hall 2000; Moor 2006; Tonkens 2009). Experts may also need to be trained in computer psychology, a currently nonexistent profession that might become a reality in the future (Epstein 1997). An existing discipline that might be of the greatest help for training of AI question review experts is artimetrics, a field of study I proposed that identifies, classifies, and authenticates AI agents, robots, and virtual reality avatars for security purposes (Yampolskiy 2007; Yampolskiy and Govindaraju 2007, 2008; Gavrilova and Yampolskiy 2010).

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

The limited outbound channel shell consists of binary (0 or 1) answers next to original unmodified safe questions presented by the intermediary nonintelligent software in random order to avoid communication via covert timing or resource channels, which makes steganography possible. The output should in no way be controllable by the AI outside 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.

I believe my communication protocol is 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, I 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 are a challenge that I would like to pose to the scientific community. The work on it will always be in progress and will never be completed. In this chapter, I have formalized the notions of the AICP, escape, and safe questions. I would also like to propose a few additional terms that might be useful in streamlining discussions about the AICP. An AI confinement environment could be appropriately called JAIL: just for AI location. The superintelligent

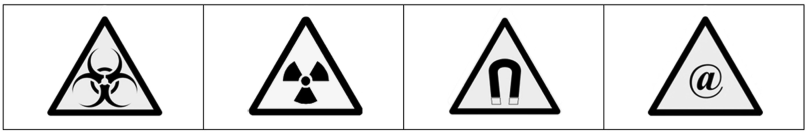

FIGURE 8.1 Hazard symbols, from left to right: biohazard, radiation, magnetic field, and the proposed AI confinement facility "located at" symbol.

AI itself could be called a superintelligent gizmo (SING), which has relevance to both singularity and to the infamous jail Sing Sing, making the nomenclature intuitive.

Finally, I would like to introduce a visual symbol indicating the presence of an HIS confinement facility (Figure 8.1, 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 "at" (@) 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 cost effective.

8.9 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 no options or only with options that provide a small probability of success, it is not any more powerful than any other agent. Although 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 chapter, I formally introduced the AICP 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 I propose is not a completely secure solution, but it is an additional option in our arsenal of security techniques. Just like with real prisons, although 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. I 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 cyberwarfare is well funded and highly publishable in mainstream scientific journals. Although the restriction methodology will be nontrivial to implement, it might serve as a tool for providing humanity with a little more time to prepare a better response.

In this chapter, I have avoided a lot of relevant philosophical questions, which I plan to address in my future work, questions such as the following: How did AI get into the box? Was it designed by humans or perhaps recovered from a signal detected by the search for extraterrestrial intelligence (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? I have also not looked at the possibility of humanity becoming completely dependent on the AI's advice in all areas of science, economics, medicine, politics, and so on 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 would the AI be punished if it purposefully gives us incorrect answers? Can the intelligence modules of the AI be separated from the confinement environment, essentially preventing the AI from any self-analysis and putting it on a path of literal soul searching? Finally, I did not attempt to analyze the financial and computational costs of building a suitable confinement environment with a full-blown simulated world in it.

REFERENCES

Allen, Colin, Wendell Wallach, and Iva Smit. July/August 2006. Why machine ethics? *IEEE Intelligent Systems* 21(4):12–17.

Anderson, Michael and Susan Leigh Anderson. 2007. Machine ethics: creating an ethical intelligent agent. *AI Magazine* 28(4):15–26.