Honeynets as a possible defensive strategy for aiding Artificial Intelligence Safety

Prasanna Kumar, Vikram Jothinagara

University of Auckland

pvik804@aucklanduni.ac.nz

Registration: 174355460

Abstract

Practical applications of Artificial Intelligence are increasing in our daily lives. There is an economic benefit in using such technologies. On one hand many icons like Elon Musk, Stephen Hawking, Bill Gates are advocating the dangerous side of Artificial Intelligence and the existential treat to human race. On the other hand, there has been a growing acceptance of Robots (a segment of AI) as caregivers in Japan which is bringing greater focus on robot rights. In such a trend of both fear and acceptance, from a security point of view, the logical step to take is to research more about AI safety and take steps to safeguard human interests. Yampolskiy presents that cybersecurity research can benefit AI Safety. In this paper, we explore one such concept in Cybersecurity i.e Honeynets and Honeypots, which may benefit the research of AI Safety.

What is Artificial Intelligence? The economic benefit to human society

Artificial Intelligence (AI) has been a study for more than fifty years and most of the concepts of AI applications were laid out by John McCarty, Marvin Minsky and Turing [1]. John McCarty was the first to have coined the term in 1955 and he defined it as "the science and engineering of making intelligent machines"[2]. From an applications point of view, AI can be seen as a domain that uses a combination of technologies, which can sense, comprehend, and act [1]. For example:

[1] Why artificial intelligence is future of growth? Mark Purdy and Paul Daughterty, Accenture Institute of High Performance, 2016

[2] https://en.wikiquote.org/wiki/Artificial_intelligence

- Sense: Image and audio processing. Systems that actively process images, sounds and speech. Facial recognition in a biometric solution is one such example.
- Comprehend: Systems that understand and analyse the data collected. For example, solutions that use of Cybots for penetration testing, online chats etc
- Act: Systems that take action in the physical world. For example, solutions such as auto-pilot, self-driven cars

Combining these abilities with access to cheap computer power and big data, daily applications of AI is bound to grow in the future and which can make a significant impact to human lives[3] .

From an economic point of view, AI is set to drive growth in at least three areas [3]. Firstly, through virtual workforce, where AI is used in automation of business processes and improve productivity. This will enable faster time to market, reduce errors, wastage of time and better utilization of resources. Secondly, by AI taking over the work that involves low value and repetitive tasks. Humans can then focus on high value work and move up the value chain. An example of effectively using labour and capital can be seen on how developed nations have outsourced low value work of services such customer support, software support, run the business support and manufacturing to India, Philippines, and China. The outsourcing nation can now refocus its labour and capital to high value work of R&D, innovation, design etc. Apple products continue to read “Designed in California, Assembled in China”. Lastly, AI will promote more innovation and collaboration across industries and promote new revenue streams for companies. An example could be how insurance companies and telecom companies collaborate to provide new services and policies to the users of driverless cars [3].

While, the applications of AI are transforming human lives, it also brings up questions on how to handle risk management, security, and ethics. Some of the questions, being debated are loss of jobs by AI, distribution of wealth created by AI, potential damage to human life and property, robot rights and singularity[4]. As this has a potential to affect daily lives, it is drawing professionals from non-technology fields such as lawmakers,

[3] Why artificial intelligence is future of growth? Mark Purdy and Paul Daughterty, Accenture Institute of High Performance, 2016

[4] https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

anthropologists, philosophers, theologists etc to provide inputs on the roles and rights of AI in the human society.

AI today is classified as narrow AI i.e. being able to perform in one domain (driving a car, sorting a mail etc). With the current pace of development in AI many scientists and technologists like Elon Musk, Stephen Hawking, Steve Wozniak, Bill Gates [5] believe there will be a time when the world will see a Generic AI or super intelligent system which will have the ability to switch multiple domains [1]. This super intelligent system has the potential to be very dangerous and could wipe out humanity in the long run. Roman Yampolskiy[6], is one such Computer Scientist who has vociferously advocated the security concerning the dangers posed by AI. He has written many articles on AI safety and engineering.

At the outset, we all understand the economic benefits of AI outweigh the negatives posed by it. However, it is imperative that such applications of AI technologies be designed for safety and safeguard human lives. In the event a super-intelligent system is created which may pose danger, human life should prevail.

AI Failures and Safety - deterministic and non-deterministic

In a system, it is important to understand the difference between reliability and safety [2]. M.G. Rodd explains that reliability is a measure about how often a system will fail and safety is concerned about what happens when the system fails [2]. An engineer will design a system with a primary focus on reliability and try to reduce the occurrences of malfunction. However, technology has limitations and every system will fail [2]. In the event of failure, solutions need to be developed that cope with the results of failure. These solutions that cope with the failure are termed as “safety” [2]. For example, if an airplane engine malfunctions, the event should not be disastrous and the plane should be able to glide to a safe landing [2]. Rodd goes further to explain that while reliability and safety are two different concepts, they are related by a common element of “determinism” [2]. For an engineer to design a reliable system and also ensure the safety of such a system, he/she should be able to determine what happens at all times and under all circumstances of the system in an operating environment [2].

[5] https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/
[6] http://cecs.louisville.edu/ry/

Regarding AI failures, Yampolskiy classifies them into two, one as the mistakes in learning phase and other in the performance phase [1]. He then observes the current trends of the failures and forecasts a list of possible failures to the future[1]. A super intelligent system(AGI) can possibly inherit a combination of these failures that may be dangerous to human race [1]. However, if one compares it to Rodd's explanation, these failures are merely the "reliability" of the system and not about "safety". There is no explanation on how these failures would be unsafe (in terms of human lives or property). For a moment, let us assume that these failures posed by Yampolskiy are disastrous, a safe AI system should also be deterministic as articulated by Rodd. For example, if a super-intelligent system is tasked to finding a cure for a deadly disease, it should not only achieve the said objective but also take the route, which is ethical and safe to humans. Hence, for a system to be deterministic, both the logical and the process to achieve the objective should be completely known and determined[2]. Applying these concepts, one way to define AI safety is - a system which is not only reliable but also be safe and deterministic i.e if a AI system fails, it fails to a safe option both logically and temporally [2]. While Rodd explains that determinism in real world is a misconception, it should at least fall within the acceptable safe limits (upper and lower bounds)[2].

In the article about Human Rights Vs Robot Rights by Jennifer Robertson, she explains how Robots have become integral part of Japanese society especially where they have been extensively used as caregivers to the old and the elderly[7]. In many cases, the robots were issued official document available to its citizens. It also suggests that, Japan is a unique society where its citizens are more comfortable with robots as caregivers than foreign caretakers. While, this scenario may not be applicable to other robot producing nations, it gives a possibility that in the distant future there is a high likelihood that more societies will anthropomorphise robots and provide more rights. While the current applications are deterministic, it is a possibility that such applications in the future may become non-deterministic (i.e free-will). In that scenario, AIs can act independently based on how it perceives the data – this would open up a whole new area of debates on machine morality,

[7] Human Rights vs Robot Rights: Forecasts from Japan, Critical Asian Studies 46:4(2104), 571-598 Jennifer Robertson

ethics, robot rights etc and how machines should be treated within the human society. It is this idea or notion, that if AIs are given freewill, then it would pose danger to humankind.

Yampolskiy puts forth, that machine ethics is wrong way to approach AI Safety, but he is actually taking a stand that Robot should not have rights and instead they should be perceived inferior in design to humans and should not be granted personhood in the society [3]. In other words, he is presenting that the design and applications of AI be deterministic in order to achieve a safe AI system.

Implementing "safety" means implementing security. There is never a 100% secure/safe system [1][2][3][4]. However, if the system is bound to be catastrophic, then organizations, governments and societies will need to invest and research more on AI safety. This does not rule out the fear from the entities and organization that may develop such technologies for malevolent purposes [1][3]. Yampolskiy suggests that AI Safety can be benefitted by research on cybersecurity [1][3]. He gives two main reasons [1]

- Comparing the objective of cybersecurity to the goal of AI Safety. The objective of cybersecurity is to eliminate or reduce the number of attacks that may compromise the system which is similar to the goal of AI safety i.e to eliminate any attack that can bypass the safety mechanism [1].
- the ecosystem of cybersecurity promotes information exchange between hackers and security experts and this enables to develop more secure systems [1].

When one looks at defensive strategies in implementing security [4], the most common concept that reflects the above two points in Cybersecurity are Honeypots and Honeynets.

Honeypots and Honeynets for AI Safety

The Honeynet Project mission statement reads "to learn the tools, tactics and motives involved in computer and network attacks, and share the lessons learned"[8]. Ray Kurzweil points out that "Intelligence is inherently impossible to control" in his book "The Singularity is Near: When Humans Transcend Biology, Viking Press, 2005" and so has Elon Musk, Stephen Hawking, Yampolskiy etc have shared similar thoughts. In such an event, one of the options humans can do is to develop strategies for counter intelligence for AI attacks. Apart from

[8] Mission statement from https://www.honeynet.org/about

satisfying the two reasons from Yampolskiy [2], Honeynets can also be viewed as a tool for developing counterintelligence against AI attacks and aid in the development of AI Safety [5].

Honeynets is a network of computer systems that help organisations to understand more about the attackers [5]. It is designed to attract black hats and once the black hat is in the honeynet, every transactions and activity including keystrokes, downloaded toolkits are recorded [5]. At the same time, any potential hostile activity intended by the black hat is controlled and monitored [5]. Honeypots are the building block of honeynet whose value lie in the unauthorized use by the hacker [5]. Unlike other cybersecurity strategies such intrusion detection system, honeypots do not have any specific objectives except in recording and controlling the activities of the black hat . This system of honeypots and honeynets help in gathering data which help in developing tools and strategies for prevention and detection of malicious activities [5]. Furthermore, Honeynets can be configured to low interaction to a high interaction system based on how much information and data gathering needs to be done [5]. One can further set up such systems in production type environment or a research type environment and apply virtualization for the whole set up to reduce infrastructure costs and operational support [5].

When comparing the defensive strategies to Lampson's framework of implementing security [4], Honeynets can be closely compared to "restrictive" defensive strategy, where it allows the bad guys in (black hats), sandboxing the whole environment and keeps them from doing damage to the system.

An AI safety system can be explored to have a similar system setup which emulates a honeynet environment, especially when the fear is that AI's actions can become non-deterministic. Such a system would be "bait" to see if AI can be dangerous or its intended actions are harmful or unethical. Of course, there will always be an area of unknown and this might for an area of data analytics to predict and comprehend the unknown data or actions to develop the right response strategies[9]. Since, honeynets are designed to observe how hackers and malware react in an environment, a similar "bait" environment can aid the AI safety community. Some of the benefits of Honeynets in AI safety could be

- act as early warning systems

[9] https://hortonworks.com/blog/cybersecurity-analytic-lifecycle/

- catch any malicious, malevolent, or unethical responses
- similar to “zero day defects”, any self-upgrade of AI capabilities can first be observed
- confuse AI
- apply data analytics to predict future actions

In an event, when one encounters the unknown or a real threat, one way to control the AI system could be to activate a runtime defensive strategy that is used in securing an application. Firstly, by isolation i.e. preventing the AI to access any other parts of the system, secondly provide the control back to the reference monitor (TCM) along with any defense in depth strategy i.e it would prevent the actions of the AI which would not comply to the policy and lastly encrypt the data which could be misused. This means, whenever AI moves into a non-deterministic scenario, give the control back to the reference monitor[4].

Researching honeynets and its applications for AI Safety could be an option to monitor any unintended actions of malicious AI, this would still have challenges of implementation in terms of hardware, software, network, design, capability etc. However, it does give a point of view on leveraging cybersecurity principles to develop AI safety.

Conclusion

Many Icons have voiced their concerns on Artificial Intelligence and the existential threat to human race, some have even advocated that AGI(Super-Intelligent system) research be monitored or even be made unethical [1] [3] on the lines of nuclear research and genetic engineering. At the same time, we have been witnessing a rise in acceptability of Robots- a segment of Artificial Intelligence in human societies. There is an increased focus on robot rights, machine ethics in the research circles[10]. We are also witnessing an economic benefit in engaging such technologies. In such a trend of both fear and acceptance, from a security point of view, the only logical solution to take is to learn more about AI in order to safeguard human interests i.e. lives and property. This reminds us of a famous saying from Sun Tzu, an ancient Chinese general, “if you know the enemy and know yourself, you need not fear the result of a hundred battles” [6]. Honeynets help improve counter intelligence in cybersecurity

[10] Human Rights vs Robot Rights: Forecasts from Japan, Critical Asian Studies 46:4(2104), 571-598 Jennifer Robertson , Euron Roboethics Roadmap Release 1.2 (Jan 2007)

and such a development and practical application of honeynets into AI Safety may help the research community to address the issues of fear and develop right prevention and detection mechanisms.

References

1. Roman V. Yampolskiy and M. S. Spellchecker. Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures, arXiv preprint arXiv:1610.07997, 2016
2. M.G. Rodd. Safe AI – is this Possible?. In: Engineering Applications of Artificial Intelligence Volume 8, Issue 3, June 1995, Pages 243-250
3. Roman V. Yampolskiy. Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach. In: Müller V. (eds) Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics, Vol 5. Springer, Berlin, Heidelberg, 2013, Pages 389-396
4. B. W. Lampson. Computer Security in the Real World In: Computer, Vol 37(6), June2004, Pages 37-46
5. David Watson. Honeynets: a tool for counterintelligence. In: Network Security, Vol 2007(1), 2007, Pages 4-8