THE FUTURE OF CYBER WARFARE

Jean-Jacques Halans
CSU May 21 2022

Introduction

Since we connected everything to the internet, we've entered the fourth industrial revolution (Mhlanga, 2020, p. 4). One of its pillars is Artificial Intelligence (AI). As we add technology to our daily lives, for example neural engines in our phones to speed up local AI and Machine Learning for image recognition, we also increase the surface for cyber-attacks.

In this essay I try to paint a picture how AI might play a role in future Cyber Warfare, but how it might differ from what we see in popular culture. I look at how technology shaped and still shapes humanity, and how government tries to keep up with these changes, often decades late. I highlight the different underlaying technologies that make up the AI and Machine Learning field, which is a very broad and complicated topic in and of itself.

I then show the broad impact, often hidden, AI already has on our daily lives, to demonstration the potential for disruption. I draw attention to the dangers of AI as what some see as a potential existential threat. Even beside the potential of cyber-attacks on AI, I show how today's AI-driven systems sometimes fail, by themselves without malicious external input, and why that might happen through vulnerabilities in the underlaying technology, as to point out how this may be exploited in a cyber-attack as well.

Eventually I describe several known attacks on AI and how they could be abused in targeted cyber-attacks of Cyber Warfare or Terrorism, further disrupting our daily lives that come to depend on these AI-powered systems as explained earlier in the piece, as well as how we potentially can mitigate these threats.

As The Art of War teaches us:

> "*Therefore the skillful leader subdues the enemy's troops without any fighting; he captures their cities without laying siege to them; he overthrows their kingdom without lengthy operations in the field.*" (Sunzi et al., 2007, Ch. 3 Attack by Stratagem)

Technology

Looking back at history, every technological advancement has been used to shape and progress humanity, as well as *abuse* it. Fire kept people warm, allowed people to cook their food, communicate over distance, but also burn down your enemies' villages. Or in fact, as a technology, looking at splitting atoms, to harness nuclear energy, and how that's been abused.

As we look at information and communication technology, in the 1960/70's phreakers would use so called Blue Boxes to make free calls, impersonating a telephone operator to set up conference calls and "party lines" (Coleman, 2020, p. 103). But all this sounds like people having fun, unlike the malware and ransomware we've seen crippling businesses since the start of the 21st century. With every technological evolution, the cyber-attack surface expands.

At the same time governments have tried to keep the legal frameworks up to date to technological changes, but often years if not decades late, repeatedly in the form of amendments of decades old legislation. An example in an Australian context is the *Telecommunications and Other Legislation Amendment (Assistance and Access) Act 2018* (or '*AA Act*' for short), which amends Australian surveillance legislation that includes the *Telecommunications Act 1997* and the *Telecommunications (Interception and Access) Act 1979* (Mann et al., 2020, para. 28). Note the passing of many years between those acts and amendments.

It's good to see that at least the EU is already looking at an Artificial Intelligence Act, and it is being discussed how to get to lawful and trustworthy AI (Smuha et al., 2021, p. 2).

Artificial Intelligence

Artificial Intelligence (AI) is everywhere, yet it is probably still in its infancy. When hearing about AI, you may think of Voice Assistants like OK Google, Siri, Alexa or Cortana, which have some smarts about recognising voice commands with varying degrees of success. But AI is already far more embedded in our daily lives, and most of it is invisible.

AI provides capabilities to develop smart and novel applications in healthcare, construction, transportation, finance and commerce, and the military, through its ability to analyse large datasets and use it to learn and model human behaviour. AI as such is just one component of the fourth industrial revolution which also includes robotics, ubiquitous linked sensors (IoT), V.R and A.R., blockchain and distributed ledger technologies, energy capture and storage and more (Mhlanga, 2020, p. 4).

Not unlike "cyber warfare", there is no single definition for "artificial intelligence" (Hassani et al., 2020, para. 1). There are many subtypes of AI technology of which the most common currently are: Machine Learning (ML), which identifies and analyses patterns to detect

associations in disparate datasets; Deep Learning (DL), which allows machines to make independent decisions using multilayer neural network (NN) models, like *multilayer perceptron* (MLP) or *radial basis function* (RBF); Natural Language Processing (NLP), which allows machines to analyse plain human language, as used in Voice Assistants; Computer Vision (CV), through which computers learn by analysing images, videos or video feeds to glean an understanding of the world (Kaul et al., 2020, Table 1).

1. In healthcare AI helps trawl through vast amounts of historical health data and discover new patterns and insights to make better clinical decisions. Numerous research studies even suggest that AI can outperform humans at fundamental healthcare tasks like identifying cancers, or setting up cohorts for clinical trials. (Davenport & Kalakota, 2019, Introduction)

2. In construction, AI can help identify risks in planning, considering machine up- and downtime, historical data of previous projects and weather patterns. It can create A.I.-powered generative architectural design alternatives (Rao, 2022, para. 2). AI can improve onsite construction safety, computing risk ratings through image recognition, as well as improve distributed labour and machinery planning. AI allows for the use of many more parameters and permutations that what previous scripting implementations would allow. And as a result, it improves the productivity and output of the entire construction industry, from design and planning to manufacturing and construction (Schober, 2020, Construction/execution).

3. In transportation, AI methods and computational-intelligence algorithms help traffic operators through vision sensing and modelling artificial transportation systems using weather predictions and behavioural changes (Wang, 2008, p. 8-13). As a simple example, when using your phone's navigation, it may alert you of traffic ahead and propose a reroute of your journey optimised for travel time.

4. In finance and commerce, AI improves the digital financial inclusion through risk detection, measurement and management, tackling information asymmetry, benefiting customer support through chatbots and recommender algorithms and facial recognition, and availing regulatory compliance through fraud detection and cybersecurity (Mhlanga, 2020, p. 10).

5. In the military, AI is seen as a force enabler, as propounded by Masakowski, (2019) "*During the twenty-first century, AI technologies will control information, people, commerce and future warfare*" (Masakowski, 2020, Preface). The potential of AI presents itself in all military domains (i.e. air, land, sea, space and cyber) and could be applied in reconnaissance and surveillance, threat evaluation, intelligence analysis and cyber security to name a few (Svenmarck et al., 2018). For example, AI enables unmanned autonomous drones either on their own or as fully autonomous drone swarms performing tasks like reconnaissance while communicating with each other, flying through dense forests (Zhou et al., 2022).

Dangers of AI

Popular culture is full of stories of AI taking over the world, or at least killing humans inhabiting that world. Think of WOPR in WarGames, HAL in 2001: A Space Odyssey or a time traveling Terminator. That is not yet the AI we have now, but it feels like that is what we are aiming for. As Jaan Tallinn from Future of Life Institute explains, "*Humans run about one million to one billion times slower than normal AI would*" (Tallinn, 2021, 8:55). Did AI engineers learn from the nuclear power scientists, with nuclear power being humanity's first existential threat? Do they not underestimate the potential dangers of AI?

All software inherently comes with bugs and vulnerabilities, none of it is 100% secure 100% of the time (except when powered down and unplugged?). It is all about managing the risk and AI is no different. But often these AI bugs, and potential vulnerabilities, manifest themselves differently to procedural applications. Not in how the application executes (and takes a wrong turn because of invalid input), the AI driven application actually works as intended, but the learned behaviour returns unexpected results given the data set it learned from.

There have been many examples of AI misbehaving over the years. From Microsoft's Tay millennial chatbot turning racists and genocidal in a matter of hours (Petri, 2016). Or Google's image recognition system incorrectly classifying some humans as gorillas (Keith, 2021, Abstract). Or the issue known as "*giraffing*", a term introduced by Melissa Elliot (Strickland, 2019, part 4 "Giraffing!"), where the Machine Learning tool trained to identify giraffes in pictures, but goes on identifying giraffes in pictures where there are none, because it trained with an over-representation of giraffes in images, but an under-representation of cases with no giraffes (Rhodes & McGrail, 2020, p. 93).

Or the inherent weakness of deep neural network machine learning to learn multiple tasks sequentially resulting in "*catastrophic forgetting*" when it absorbs new information. As new pathways are formed, the continual learning algorithm then occasionally "forgets" the previous tasks it was trained for (Kirkpatrick et al., 2017, p. 1).

AI is often considered as a black box, as the internal reasoning procedure of the machine learning models remain hidden from the user. As such, researchers have been working on *explainable artificial intelligence* (XAI) to create human-interpretable justifications for model decisions (Ebrahimi et al., 2021, p. 2).

AI tends to misbehave in these situations when the data set it learns from is incomplete or too small. In those cases, AI engineers turn towards *synthetic data*. Synthetic data is data that has been generated from real data through synthesis and has the same statistical properties as the real data, or is generated through existing models or simulations. Data in this context is not just structured data, but can also be unstructured text, like transcripts or notes and articles, or even images, for example portrait images of non-existing people (Emam, 2020).

So far, we've looked at how embedded AI is in our daily lives and highlighted just a few of the known issues encountered using AI, and potential solutions including explainable AI and synthetic data. Even if or when AI is used in conventional, kinetic warfare, how could it be abused through cyber warfare?

Attacking AI – The Future of Cyber Warfare

Probably the most fun (or annoying) way to "attack" AI is by calling out "Hey Siri" on a TV show or Zoom call, as phones, tablets, smart watches and smart speakers light up in living rooms. Good thing Siri isn't smart enough to do anything nefarious in the home. But what if a Voice Assistant in a smart home could be triggered silently through embedded sound to do something, like turn up (or down) the heat, run a bath, open the garage door, or unlock the front door?

AI and machine learning inherently don't generalize well. They are highly receptive to adversarial attacks where a tiny change in the input causes the deep neural network to stumble with high confidence (Jakubovitz & Giryes, 2018). Adversarial machine learning can be used to trick AI into classifying one image as another image by applying noise perturbations, while to a human, both images look exactly the same (Donnellan, 2019).

As such we are attacking the integrity of the AI system through a *Data Poisoning Attack*. By embedding malicious data into the training data, be it images or structured data, the model could learn wrong patterns (Jagielski et al., 2018). A related version is the *Transfer Learning Attack*. An AI model is trained on U.S. traffic signs and is distributed through an AI model repository. It also contains "backdoors" as it is trained to recognise "interference" stickers on traffic signs to change behaviour. The AI model is picked up by another user (victim) to train a Swedish traffic signs model using transfer learning. Through "data poisoning" and supply chain-based transfer learning, the "backdoors", changing behaviour on traffic sign recognition using stickers, are distributed globally (Gu et al., 2017, Ch. 5).

Another type of attack is the *Evasion Attack*, where attackers find pre-existing imperfections in the existing model that they, using finely tweaked inputs, can manipulate and exploit (Lohn, 2020, pp. 5-7). One such example is Tencent's Keen Security Lab where they were able to trigger a car's function, the windscreen wipers, not by using water, but by showing it some manipulated image. And far worse, they manipulated the vision based Auto Pilot system to get the car on Auto Pilot to veer into the oncoming traffic lane by putting an interference sticker on the road (Tencent, 2019, 1:26). The Tencent team didn't poison the vehicle's training data, they found flaws in what the AI model had learned and exploited that.

Applying this to cyber warfare, if an adversary can infer how a drone model recognises targets, they could develop specific camouflage techniques to counteract the drone's image classifier AI, or trigger functions on the drone to return home (or worse).

Furthermore, as we've seen the many ways AI is embedded in our daily lives, adversaries and terrorists could try and topple governments by attacking the Confidentiality of AI

systems and data by observing the AI's inputs and outputs, through *Model Extraction* and *Membership Inference Attacks* (Lohn, 2020, pp. 8-10). In healthcare or finance, a Membership Inference attack, by learning the specific attributes of data, could expose personal data, opening systems to identity theft. What if attackers could open bank accounts and flood AI lending systems with phony applications or trading requests (Vanderford, 2022, para. 17), siphoning the money abroad and undermining confidence in institutions?

Terrorist could target the use of AI in the pharmaceutical industry, as evident by Swiss Spiez laboratory's "Dr. Evil project" experiment where the laboratory reversed their platform's drug discovery goal, discovering 40.000 different lethal molecules similar in lethality to the VX nerve agent. There is in fact very little regulation or oversight in this area and most researchers have only limited awareness (Craig, 2022). Tools and datasets to repeat this experiment are publicly available.

Securing AI

As Abhishek Gupta, founder of the Montreal AI Ethics Institute, suggests in the Wall Street Journal: "*Machine-learning security is not just a combination of security and machine learning; it's a novel field.…When you introduce machine learning into any kind of software infrastructure, it opens up new attack surfaces, new modalities for how a system's behaviour might be corrupted*" (Vanderford, 2022, para. 8)

Be it AI on a smart phone, a car, an agricultural machinery or a military drone, each comes with inherit complexities and redundancies, which can pose a challenge to attackers. Yet these devices are then also connected to the Internet. And as with every Internet connected device, it needs an established, multilayered approach to defending against cyber-attacks. Something which even this year is still missing in many instances, as seen in the February 24th Viasat satellite attack, where the satellite modems received and installed firmware updates without verifying signatures (Halans, 2022, para. 14). But on top of that, AI also needs specific solutions to its inherent flaws mentioned earlier.

Secure multi-party computation (MPC) enables collaboration between different parties and their AI data sets. Through this federated learning framework (Yang et al., 2019), both parties can perform complex computations on the merged data set with full privacy protections in place, not revealing each other's underlaying data (Knott et al., 2021). In addition, a *publicly auditable secure computation* combined with an improved version of the SPDZ framework allows anyone to verify that the output is indeed correct through a transcript of the protocol (Baum et al., 2014). This helps establish transparency and trust, one of the challenges important to the military use of AI (Svenmarck et al., 2018, Ch. 4).

Differential Privacy provides strong privacy guarantees as it quantifies to what extent individual privacy in a statistical data set is retained while at the same time providing useful aggregate information about the data set (Geng et al., 2015). *Differential Privacy* could prevent *membership inference*.

Homomorphic Encryption (HE) allows for private AI training and prediction in a cloud environment using private encrypted data, without decrypting it. With *Homomorphic Encryption*, the order of encryption and computation can be switched around, as the same result is returned if encrypting first, then compute, or computing first then encrypt (Lauter, 2021).

When synthetic data and pre-trained models are used, it is imperative that the same lessons learned from securing the software supply chain are applied to machine learning security and its supply chain. Only trusted sources should be used, where repositories are accessed over secure channels and contain digital signatures that are verified, to prevent backdoored AI models.

Furthermore, governments should take note and enact regulations and oversight of some of these AI models and set up ethical oversight committees. Not everyone should have access to pharmaceutical, biological and chemical models.

Conclusion

The potential of AI is undeniable. But so is the potential of abuse. In this essay I tried to highlight the pain points of the current AI industry and the potential of abuse through cyber-attacks against AI now and into the future. One positive note is that there is a lot of ongoing research into this field to improve Machine Learning and plug the vulnerabilities.

But AI is still too much of a black box and more research seems to be needed to open this up, as to be able to simply explain why things happen as they happen, before they might happen. There are still too many surprises, and maybe we move too fast in this field.

Additionally, more research is required in the cyber defence field to protect AI systems from abuse, like investigating techniques for backdoor detection in machine learning. The Cyber Security industry needs to establish training resources on AI and its Cyber Defence, same as there is for SCADA, Cloud or Network security. As every sector in society applies AI, they all need training.

References

Baum, C., Damgård, I., & Orlandi, C. (2014). Publicly auditable secure multi-party computation. International Conference on Security and Cryptography for Networks,

Coleman, E. G. (2020). Phreaks, Hackers, and Trolls: The Politics of Transgression and Spectacle. In (pp. 99-119). New York University Press. https://doi.org/10.18574/9780814763025-010

Craig, J. (2022). *Widely Available AI Could Have Deadly Consequences*. Wired. https://www.wired.com/story/ai-dr-evil-drug-discovery/

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal*, *6*(2), 94-98. https://doi.org/10.7861/futurehosp.6-2-94

Donnellan, A. (2019). *Vaccinating machine learning against attacks*. CSIR. https://blog.csiro.au/vaccinating-machine-learning-against-attacks/

Ebrahimi, S., Petryk, S., Gokul, A., Gan, W., Gonzalez, J. E., Rohrbach, M., & Darrell, T. (2021). Remembering for the right reasons: Explanations reduce catastrophic forgetting. *Applied AI letters*, *2*(4), n/a. https://doi.org/10.1002/ail2.44

Emam, K. (2020). *Accelerating AI with Synthetic Data* (1st edition ed.). O'Reilly Media, Inc.

Geng, Q., Kairouz, P., Oh, S., & Viswanath, P. (2015). The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, *9*(7), 1176-1184.

Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Halans, J.-J. (2022). *Viasat Ukraine Case Study*. CSU.

Hassani, H., Silva, E. S., Unger, S., TajMazinani, M., & Mac Feely, S. (2020). Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future? *AI (Basel)*, *1*(2), 143-155. https://doi.org/10.3390/ai1020008

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018, 20-24 May 2018). Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. 2018 IEEE Symposium on Security and Privacy (SP),

Jakubovitz, D., & Giryes, R. (2018). Improving DNN Robustness to Adversarial Attacks Using Jacobian Regularization. In (Vol. 11216, pp. 525-541). Springer International Publishing. https://doi.org/10.1007/978-3-030-01258-8_32

Kaul, V., Enslin, S., & Gross, S. A. (2020). History of artificial intelligence in medicine. *Gastrointest Endosc*, *92*(4), 807-812. https://doi.org/10.1016/j.gie.2020.06.040

Keith, D. (2021). The precondition to humanizing AI - common sense knowledge. *Informaa quarterly official bulletin of the Records Management Association of Australia*, *37*(1), 40-44. https://doi.org/10.3316/agispt.20210601047677

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences - PNAS*, *114*(13), 3521-3526. https://doi.org/10.1073/pnas.1611835114

Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., & van der Maaten, L. (2021). Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, *34*.

Lauter, K. E. (2021). Private AI: machine learning on encrypted data. *Cryptology ePrint Archive*.

Lohn, A. (2020). *Hacking AI*. C. f. S. a. E. Technology. https://doi.org/10.51593/2020CA006

Mann, M., Daly, A., & Molnar, A. (2020). Regulatory arbitrage and transnational surveillance: Australia's extraterritorial assistance to access encrypted communications. *Internet policy review*, *9*(3), 1-20. https://doi.org/10.14763/2020.3.1499

Masakowski, Y. R. (2020). *Artificial intelligence and global security : future trends, threats and considerations*. Emerald Publishing Limited.

Mhlanga, D. (2020). Industry 4.0 in finance: the impact of artificial intelligence (ai) on digital financial inclusion. *International journal of financial studies*, *8*(3), 1-14. https://doi.org/10.3390/ijfs8030045

Petri, A. (2016). *The terrifying lesson of the Trump-supporting Nazi chat bot Tay*. WP Company LLC d/b/a The Washington Post. https://www.washingtonpost.com/blogs/compost/wp/2016/03/24/the-terrifying-lesson-of-the-trump-supporting-nazi-chat-bot-tay/

Rao, S. (2022). *The Benefits of AI in Construction*. Trimble. https://constructible.trimble.com/construction-industry/the-benefits-of-ai-in-construction

Rhodes, T., & McGrail, T. (2020). Successful application of AI techniques: A hybrid approach. *Transformers Magazine*. https://www.doble.com/wp-content/uploads/Successful-application-of-AI-techniques.pdf

Schober, K.-S. (2020). *How to increase efficiency over the entire lifecycle chain*. Roland Berger. https://www.rolandberger.com/en/Insights/Publications/Artificial-intelligence-in-the-construction-industry.html

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. *Available at SSRN*. https://doi.org/http://dx.doi.org/10.2139/ssrn.3899991

Strickland, E. (2019). *The Blogger Behind "AI Weirdness" Thinks Today's AI Is Dumb and Dangerous*. IEEE. https://spectrum.ieee.org/blogger-behind-ai-weirdness-thinks-todays-ai-is-dumb-and-dangerous

Sunzi, Giles, L., Babcock, J., & Aylward, D. (2007). *The art of war*. Ulysses Press.

Svenmarck, P., Luotsinen, L., Nilsson, M., & Schubert, J. (2018). Possibilities and challenges for artificial intelligence in military applications. Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting,

Tallinn, J. (2021). *Avoiding Civilizational Pitfalls and Surviving the 21st Century* [Interview]. https://futureoflife.org/2021/04/20/jaan-tallinn-on-avoiding-civilizational-pitfalls-and-surviving-the-21st-century/

Tencent. (2019). *Tencent Keen Security Lab Experimental Security Research of Tesla Autopilot* [Video]. https://www.youtube.com/watch?v=6QSsKy0I9LE&t=14s

Vanderford, R. (2022). *AI Experts Warn of Potential Cyberwar Facing Banking Sector*. The Wall Street Journal. https://www.wsj.com/articles/ai-experts-warn-of-potential-cyberwar-facing-banking-sector-11647941402?reflink=desktopwebshare_permalink

Wang, F.-Y. (2008). Toward a Revolution in Transportation Operations: AI for Complex Systems. *IEEE intelligent systems*, *23*(6), 8-13. https://doi.org/10.1109/MIS.2008.112

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.*, *10*(2), Article 12. https://doi.org/10.1145/3298981

Zhou, X., Wen, X., Wang, Z., Gao, Y., Li, H., Wang, Q., Yang, T., Lu, H., Cao, Y., Xu, C., & Gao, F. (2022). Swarm of micro flying robots in the wild. *Science Robotics*, *7*(66), eabm5954. https://doi.org/10.1126/scirobotics.abm5954