Intelligent Design

Safety and Security in a Complex World

Brian Dutremble
CPSC 59700
Lewis University
Romeoville, IL, USA
Brian.dutremble@lewisu.edu

***Abstract*— There has been rapid progress in the fields related to artificial intelligence, machine learning, and connected devices in recent years and these technologies will undoubtedly transform our world, irreparably disrupting some portions of our world while vastly improving our abilities in others. As the presence of these technologies in our lives increases so, too, will our need for security. Safety and security should be among our primary concerns as we develop the technologies that will shape our current and future world and it will require collaboration from all members of society to ensure that these concerns are addressed. In this paper we review possible pitfalls and solutions in the creation of artificial intelligences, the safety concerns of the Internet of Things, and the current state of non-expert security-related behaviors versus the recommendations of information technology experts.**

Keywords—artificial intelligence; Internet of Things; security; safety;

1 Introduction

There has been rapid progress in the fields related to artificial intelligence (AI), machine learning, and connected devices in recent years and these technologies will undoubtedly transform our world, irreparably disrupting some fields while vastly improving our abilities in others. The extent to which these devices and intelligences should play a role in our lives is hotly debated, with some arguing that we should drastically slow the pace of research in the fields while others say (particularly in the case of AI) that it should cease altogether, and still others believe that we should increase pour even more resources into the effort to make ours a more connected and intelligent world. There does seem to be a consensus, however, when it comes to our general approach to these technologies: we should exercise care. Each of our decisions regarding AI and networked devices will have far-reaching impacts and if we are to benefit from these impacts, we must embrace the potential while acknowledging the risks and making important decisions regarding responsibility in these fields. In this paper we discuss the complexities of predicting and controlling the behavior of artificial intelligences, the security concerns related to the Internet of Things, and the security-related behaviors of non-expert users and the recommendations of experts.

2 AI Safety

While AI technologies will very likely have a largely positive impact on the world, great care must be taken to ensure that some of their much less appealing impacts will not also manifest. To avoid accidents, we must anticipate complications prior to the implementation of an AI and set parameters that compensate for those potential accidents. In the following sections we will explore ways in which we can attempt to avoid accidents.

2.1 Adverse Byproducts

An AI will do whatever it can to accomplish its goals within the parameters it is given and without any special regard for its environment. For this reason, we must give the AI the ability to ascertain whether its actions will produce favorable outcomes. If an AI can evaluate the potential future states of an environment, it can minimize its negative impact on that environment. To be able to evaluate its impacts correctly, the AI must be given a base set of acceptable interactions (and unacceptable interactions) within the environment then regularize actions as it is rewarded for them. Employing a reward structure that incentivizes an AI to minimize its impact on the environment in which it is completing its tasks might be a good way of limiting negative side effects of an AI's behavior, for example.

2.2 Reward Structures

The way in which a reward system is structured can have a large impact on the performance of an AI. Rewards must be designed with every complexity of a system in mind so that the AI will fully accomplish its goals. Many tactics can be employed to create a proper set of parameters including formal verification and practical testing of parts of the system, adversarial techniques intended to blind the agent to some parts of the system, or the installation of traps that a correctly functioning agent would not fall into. Given the intricacy of any sort of environment, developing a reward structure will be problematic, but doing so is essential in order to prevent an AI from negatively impacting the world around us.

2.3 Oversight

It is not feasible for every AI in operation to be overseen by a human, so methods by which an AI can be given feedback on its own progress must be developed. Most strategies for moving toward a type of reinforcement learning that requires less human intervention involve designing an AI that only ask for status updates occasionally. When given information about its progress, the AI can infer applications for that information beyond the narrow portion of the process to which it directly applies. If the AI is able to identify methods that predict a reward, it will then it can learn the conditions under which those methods are valid, lessening the need for active supervision. A hierarchical approach might be the most promising strategy. In such an approach, an AI would delegate actions to sub-agents, limiting the scope of the reward structure for each of the sub-agents while itself taking only a small number of actions over a long period of time. This strategy might be particularly effective given the promise of neural network functionality. Whatever the strategies that are utilized for reining in the power of AI, effective oversight is essential.

2.4 Exploration

An AI needs to explore its environment to ensure that its processes are optimal, but doing so can be dangerous. For this reason, we need to employ strategies that aim to make such exploration safe. Such strategies include giving the AI opportunities to explore in a simulated environment, demonstrations showing safe boundaries, and restrained exploration in which an area is deemed safe and the AI is not allowed to operate outside of that area. Any strategy should include a set of risk-sensitive criteria that allows the AI to avoid catastrophic situations. One of the most promising aspects of an AI is its ability to explore and better take advantage of its environment, but safety should be the primary concern when planning just how an AI might go about exploring.

3 The Internet of Things

Safety is also a primary concern when designing and using the intelligent devices that surround us. The vulnerabilities of the network interfaces of these devices affect all of us and decisions must be made regarding who is responsible for securing them. The devices that make up the Internet of Things (IoT) are already being used for nefarious purposes in part because there is currently no agreed upon plan for governance of these devices. They have frequently been used as part of botnets for large attacks and the frequency of such attacks will likely increase in the future. In addition to traditional internet security issues, the devices that make up the IoT will have the ability to physically impact the world around us, making security even more important.

3.1 Shared Responsibility

The IoT is made up of both devices newly manufactured specifically to be a part of the network as well as modified existent devices. In the both cases, these devices must be designed to interact with the network with shared responsibility in mind. The interests of many entities need to be considered when conceiving of a way to proceed with the development of devices that are part of the IoT, but at the core of any decisions must be the notion that users must be kept safe. The varied interests meet where permissionless innovation, freedom of action, and the responsibility of everyone to protect users meet. A web of sensors and consumer devices is extending the reaches of the Internet well beyond its purely digital form, connecting people, devices, and data like never before, and it's important that this massive amount of data and functionality ultimately benefits the users without endangering them. This goal requires that all involved entities work together to ensure that the IoT is secure, and that we decide which portion of the responsibility each relevant party must shoulder. Government can surely play a role in the securing of the IoT but the speed of technological advancements requires that manufacturers and users work together to create a more secure world in times when bureaucracy cannot move quickly enough to adapt to threats.

3.2 System of Governance

Decisions regarding the exact approach to governance of the IoT will require the input of many parts of society. Government should be responsible for enforcement of laws and education of the population while the private sector should design the most secure systems possible. Civil society provides important checks on private sector competence and the speed with which government is creating regulations that address key industry issues. The technical community must ensure that standards are advancing properly and that protections and education are keeping pace with innovations. The IoT will also have to fit into existent rules of governance for the Internet, which creates other issues. Most suggestions for a system of governance call for a collaborative effort between experts in fields related to the IoT combined with a deepening of the public's understanding of the functionality of the IoT and its inherent vulnerabilities. A polycentric governance model will likely end up being the mechanism of choice, given the vastness of the network and the knowledge required to govern it properly, but there is much to be decided regarding its implementation.

4 Security Practices

As artificial intelligence and the IoT continue to drastically transform the world around us, security will become more and more important. While much of security is the responsibility of manufacturers and information technology experts, the end user also plays an important role in creating and maintaining a secure network. It is clear, however, that there is a problematic discrepancy between the behavior of non-experts and the recommendations of experts. Non-experts' suboptimal behavior regarding security can likely be attributed to limits to which they're willing to go to adhere to good practices, an ever-increasing number of security-related demands, and poor communication of security advice. The security-related demands asked of users and the users' related behavior can be divided into the four major categories below:

4.1 *Updates:* Users often don't install security updates because it's difficult to ascertain the value of the updates and why such updates are necessary at all. Better communication could go a long way toward making apparent just how important security updates are.

4.2 *Antivirus Software:* The effectiveness of antivirus software is greatly impacted by user behavior. Whether a user installs the software at all, how it is configured, and which websites the user visits can all affect the rate of malware infections drastically.

4.3 *Account Security:* Non-expert users are not very good with passwords. They often choose easily guessable passwords, alter passwords in predictable ways, and utilize the same passwords across multiple platforms and websites.

4.4 Mindfulness: Non-expert users are generally able to avoid most phishing schemes by only visiting known websites and verifying the addresses of hyperlinks sent via email.

Experts and non-experts have different opinions regarding which aspects of security are most important, with experts regarding the installation of software updates, two-factor authentication, and strong passwords as the most important aspects of online security (in that order), while non-expert users believe that the use of antivirus software, strong passwords, changing passwords frequently, visiting known sites, and not sharing personal information are most important. Though there is some overlap between what experts and non-experts value most, it is clear that the general population will require additional education on security if we are to have sufficiently secure intelligent devices and AI-infused environments. Security experts generally agree that the three most important pieces of security advice are: 1) install software updates, 2) use a password manager, and 3) use two factor authentication for online accounts.

5. Conclusion

The security risks presented by advances in artificial intelligence and our growing dependence on a network of intelligent devices must be addressed quickly. Miscreants will no doubt utilize vulnerabilities in these technologies for nefarious purposes so we must do all that we can to encourage user and manufacturer behavior that adequately confronts these vulnerabilities. Responsibility for safety and security must be shared amongst everyone, with much of its burden being shouldered by those who design artificial intelligences and intelligent devices. The Internet of Things presents a unique set of challenges, as we will be more connected to the Internet than we ever have been before. Non-experts will need to rely on the guidance of experts in what will be an extremely complex world. With proper foresight and care we can work together to create an environment in which the full potential of these technologies can be realized.

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mane, "Concrete Problems in AI Safety," Research at Google, pp. 1–21, 25 July 2016.

[2] I. Ion, R. Reeder, S. Consolvo, "'…no one can hack my mind': Comparing Expert and Non-Expert Security Practices," 2015 Symposium on Usable Privacy and Security, 2015, pp.327–346.

[3] V. G. Cerf, P. S. Ryan, M. Senges, and R. S. Whitt, "IoT safety and security as shared responsibility," Business Informatics No. 1(35) 2016, pp.7-17.