Research cybersecurity best practices in AI development we might incorporate, robust AI, adversarial defence

Summary (Harrison)

Provide a brief overview of the report, including the purpose, key findings from case studies, and conclusions.

In this report we will be researching cybersecurity's best practices in relation to AI development. Within that, we will be exploring the incorporation of robust AI and adversarial defence. The way this report will achieve that will be by exploring examples/case studies to see effective cyber-security's practices and reviewing relevant literature that will also assist in completing this research report. Additionally, we will explore literature in how to incorporate AI robustness and protect the AI from adversarial threats. Finally, we will provide a conclusion that will display key findings and how we can implement the findings of these reports.

Introduction (Harrison)

Background

In the research found, it has been made clear that the source of David Barrera (2022, pg.32) that cyber security relies on data security and privacy as a key concept of supporting best practices. Ensuring that data encryption, access controls and data anonymization are all established processes that will ensure cyber security is practicing at an effective level. In relation to robust AI, it has been highlighted by the source of Casey (2020, pg.30) as early security development is a foundational step is of high priority. The reason that is, is because it is ensuring security as a foundational level of AI development, it will help and assist the robustness building up, reducing the likelihood of any vulnerabilities. Furthermore, incorporating robust AI has been shown to be something ideally would be implemented at the beginning, adversarial defence on the other hand is something that should be trained, as this is a type of attack that is centred around deceiving the AI model to make mistakes and errors. This all will be explored more in-depth within the following steps of the reports.

Purpose of the Research

The clear purpose of this report is to identify and integrate cyber-security practices (that are considered the best) in relation to the development of AI, whilst enhancing the robustness of AI and improving their defence against possible attacks from adversarial defences.

Literature Review (Harrison)

The literature surrounding the best practice of cyber-security includes the source of Liu (2019, pg.43) harps on the importance of community engagement, and how releasing updates and notes about the software and how it may affect the user's (relating to personal data and privacy). Additionally, Liu (2019, pg.44) discusses how important adversarial training is. Possible implementation of Adversarial training includes

Case Studies/ Examples of Successful best practices (Use of AI)

Case Study 1: Google red team (AI training/development team)

- This case study is about the team at google integrating AI systems in into their systems and products, in this process they have developed robust practices to ensure they have an extensive security system. This is where the 'red team' comes into effect, red team, is an internal team that will try and break down the AI system or try and pass their own AI system.
- This is implemented in a way that would involve having a team that is trained in AI development and cybersecurity practices
- The results and overall effectiveness of the google red team has been significant, not only by their development of more secure AI models, but also by highlighting and resolving any possible critical vulnerabilities. Furthermore, the red team has strengthened the robustness of AI and implemented a strategic incident response strategy that could be used if there was a problem rising. Finally, this case study highlights how an individual company and team, can change the standards of how all companies and product providers should act, this red team has led to similar teams being created at different companies.
- This effective team is not without problems, the first key challenge to having a successful team like this is that the team must be well trained and competent in what they are doing. Additionally, ethical considerations must come into play when referring to the best cyber security practices, as teams like this must, report to ethical boundaries and legal constraints.

Example 2: Dual Audit Strategy

- To ensure that the company is following laws and regulations as well as adhearing to ethical standards surrounding AI, a Dual-audit strategy is an appropriate measure.
- The implementation will allow a company to identify vulnerabilities as well as areas for improvement. The method also provides an unbiased, outside perspective that members from the inside may not discover. Furthermore, if the application was to be used in a regulated industry such as health, the external auditors will be more experienced and equipped to ensure that the application meets the relevant industry standard. Additionally,

whilst investigating areas for improvement, Ideas based on these findings may potential be developed.
- Results include Overlooked vulnerabilities, holistic perspective, discoveries which lead to further implementations.
- Challenges faced may include a difference in opinion. Solutions suggested by the external party may not be aligned with the company and its values. Furthermore, it may not be a solution which requires attention. When sourcing an external auditor, it is best advised to source a reputable and experienced firm. Using an external auditor result increase the number of individuals that are exposed to the company's information, which may lead to a data breach. An audit may also contain a list of risks which are prioritised, this may lead to indifferences based on what the company themselves would like to focus on.

Example 3: Data Santization (investigating further, new card created)

An implementation process which ensures safe practice regarding the handling of data, and user privacy.

Implementation consists of identifying the needs of the application. Points at which data is collected, and areas which would require sanitization. This may consist of altering data obtained, one example being redacting emails to provide anonymity for users. The process would also require testing, to confirm that the process is coherent and working effectively sanitize the data during collection.

Results include improved overall security posture, strong alignment with compliance and regulation, improved AI robustness.

Challenges which may occur is the implementation of sanitization into the model include

- Adjusting to new trends and techniques
- Ability to process large amounts of data; the process may take long, essentially delaying the email process between sender and recipient
- Integrating the process into an existing infrastructure which may ause implications to other processes.

Discussion (Harrison) + (Shannon)

Key Findings

Key findings of this report post case studies and literature reviews is that implementing best cyber security practices must include legal and ethical considerations, this has featured heavily in all academic sources as a priority. Additionally, the research around robust AI has explained that there is a need for

In addition to legal and ethical considerations, we are not only factoring the company and its position. By failing to adhere to compliance and regulation can result to several outcomes. Decommissioning of the application, failing to meet certain standards may result in the app not

being available in the market. Fines are also a result of non-compliance, depending on the severity and nature of the violation penalties may occur. Lastly, the importance of an incident response plan is crucial. In the chance of a breach/attack failure to provide steps to mitigate the breach may result in the loss of data. This will tarnish the application and would alter trust with existing users and future customers.

Conclusion (Harrison) + (Shannon)

Summary of Key Points

The key takeaways from this report have included the need for ethical considerations of user's data, and the research highlighted specific measures (access control, encryption etc.)

In addition to this, another key finding is the need to incorporate robustness AI in the early stages of AI and program development. The reason behind this, is because the early it is implemented, the least opportunities for vulnerabilities are created.

Moreover, the research conducted in this report has now led into an investigation on Data security, with the use of AI. As collaborators, we will be delving further into a particular method of data security known as data santization.

References

List all sources cited in the report.

- Barrera, D. (2022). Security Best Practices: A Critical Analysis Using IoT as a Case Study. *ACM Transactions on Privacy and Security*, 26(2). doi:https://doi.org/10.1145/3563392.

- Casey, E. (2020). Digital Transformation Risk Management in Forensic Science Laboratories. *Forensic Science International*, 316, p.110486. doi:https://doi.org/10.1016/j.forsciint.2020.110486.

- Liu, H. (2019). Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. *Applied Sciences*, 9(20), p.4396. doi:https://doi.org/10.3390/app9204396.
-
-