Exploring Mechanisms for Uncovering the Misuse of AI as a Service

A Part III Project Proposal

Sumaiyah Y. Kola (*sk940*), Clare College

Project Supervisor: Dr Jatinder Singh

Abstract

Artificial Intelligence as a Service (AIaaS) offers pre-built AI models to customers looking to leverage AI capabilities with minimal cost and risk. Given the generic functionality of the AI models available, there is a large scope for misuse. This project involves generating realistic traces of AIaaS data from analogous datasets to indicate possible scenarios of misuse. I will then devise mechanisms for uncovering possible AIaaS misuse though machine learning-based detection models. The models will uncover 'signals' and 'signatures' in usage data that could indicate potential bad behaviour.

1 Introduction, approach and outcomes (500 words)

Artificial Intelligence as a Service (AIaaS) characterises the increasingly popular practice of third parties offering the outsourcing of AI technology. Consequently, it is important to ensure that customers use these AI services appropriately. Currently, AIaaS allows customers to leverage AI capabilities, without the frequently required computational expenses, in-house expertise. Additionally, service providers offer little oversight. This project is motivated by the recent work [4].

Given the generic nature of the pre-built AI models, and the lack oversight from service providers, there is a large scope for them to be used for problematic purposes. A comprehensive facial recognition model could be misused for population surveillance or intruding on the privacy of a chosen individual. The same object recognition software could be used to recognise fruit in the store and military vehicles in a war-torn nation. With the rise of AIaaS technologies, there is an incentive for service providers to protect their reputation and manage their potential obligations and liabilities.

This project involves investigating the 'signals' and 'signatures' of the usage patterns of AIaaS that could suggest potential bad behaviour. Firstly, I will generate realistic traces of AIaaS data from analogous datasets. After exploring potential use cases and problematic scenarios, I will generate synthetic data around this. I will then build machine learning-based detection models to uncover possible misuse.

While there is a lack of publicly accessible data (usage logs) from AIaaS providers, a dataset containing real-world usage data has recently been made public by Microsoft ([3], [1]). It contains customer usage of Microsoft's Azure Function as a Service (FaaS). In terms of mapping a usage landscape that is grounded in reality, we may consider this dataset reasonably analogous to AIaaS customer usage. Google has also made public data gathered in an operational data center over a period of time [7]. These datasets about cloud usage behaviour provide a starting point. After devising some problematic scenarios influenced by real events [6], I plan to synthesise data required to train the misuse detection models. Past research investigates the synthesis of data for related tasks, such as the detection of fraud and network intrusion ([5], [8]).I could also deploy my own cloud server to get an indication of real

usage logs. I intend to implement a number of these synthetic data generation techniques, contrasting the various approaches and their complexity.

I plan to implement and evaluate a variety of ML-based misuse detection techniques. Initial research reveals two primary approaches to uncovering patterns of misuse in the data; clustering and classification. ML methods, such as Naive Bayes, can be used to train classifiers that detect particular problematic situations. Clustering algorithms, such as a Gaussian Mixture Model, can group similar behaviour patterns into clusters. These can then be used to identify anomalous behaviour patterns. In addition to detecting misbehaviour from usage logs, I will explore detection techniques that can be applied on-the-fly, building behaviour profiles to detect suspicious situations. Detection methods will be evaluated according to their performance and complexity; if the computational cost is high, it is unlikely to be adopted by service providers.

2 Workplan (500 words)

This is a research project and as such a proportion of the project plan will be dedicated to studying and exploring existing approaches.

Before embarking on the misuse detection models, I will explore data synthesis techniques, performing a comparative evaluation of the different approaches available. Once complete, I will explore a handful of detection techniques, and conclude with an evaluation of the approaches available.

I have listed the key project milestones below.

2.1 Milestones

- 30.12.20: Determining the scope of the project and determining concrete misuse cases to explore
- 17.01.21: Analysis of data available data synthesis techniques
- 01.02.21: Structured plan explaining which misuse-detection techniques to implement
- 14.03.21: Data generated
- 11.04.21: Misuse detection mechanisms generated
- 25.04.21: Project results collected and processed
- 09.05.21: Dissertation draft complete
- 23.05.21: Dissertation complete

In order to track progress, I will work in two-week sprints with a clearly stated milestones. Slots marked slack time are intentionally empty and to be used as buffer slots to accommodate any unexpected delays. Initial time chunks focus on research with implementation to follow.

07.12.20 - 20.12.20 [Christmas vacation]

Research existing AIaaS products available. Come up with concrete use cases and misuse scenarios that can be explored.

Milestone: Work on an initial plan of the scope of the project in terms of misuse scenarios and the type of data required, e.g. the misuse of facial recognition software for surveillance, the misuse of speech-to-text software

21.12.20 - 03.01.21 [Christmas vacation]

Time off for Christmas and New Year

04.01.21 - 17.01.21 [Christmas vacation]

Explore the existing cloud usage datasets ([1], [2]). Research available data synthesis techniques. Set up extra computational resources

Milestones: Exploratory analysis of cloud usage datasets and data synthetics techniques

18.01.21 - 31.01.21 [Lent Term]

Research existing based machine learning techniques, deciding which to implement. These will include classification-based techniques, clustering-based techniques and any other potential techniques that would be appropriate for the misuse scenarios I am investigating.

Milestone: Clear understand of available machine-learning based detection services with a clear plan for which to implement

01.02.21 - 14.02.21 [Lent Term]

Slack time

15.02.21 - 28.02.21 [Lent Term]

Begin implementation of 2-3 data generation techniques using the work of ([5], [8]) as a starting point. Maintain documentation of these techniques and their positives and limitations.

01.03.21 - 14.03.21 [Lent Term]

Continue generating data

Milestone: Data generated

15.03.21 - 28.03.21 [Easter vacation]

Begin implementation of misuse detection techniques. Maintain documentation of techniques that can be utilised for the write-up

29.03.21 - 11.04.21 [Easter vacation]

Milestone: Misuse detection mechanisms generated

12.04.21 - 25.04.21 [Easter vacation/Easter term]

Evaluate core elements of the project and collect results

Milestone: Final project results collected, diagrams and tables drawn

26.04.21 - 09.05.21 [Easter term]

Complete first draft of the dissertation

Milestone: Hand in dissertation draft

10.05.21 - 23.05.21 [Easter term]

Improve dissertation, incorporating changes based on supervisor feedback. Proofread dissertation and finalise any changes. Hand in final copy

Milestone: Dissertation complete

References

[1] GitHub Microsoft Azure. Azure/azurepublicdataset: Microsoft azure traces. `https://github.com/Azure/AzurePublicDataset`. (Accessed on 11/17/2020).

[2] GitHub Google. cluster-data/clusterdata2011_2.md at master · google/cluster-data. `https://github.com/google/cluster-data/blob/master/ClusterData2011_2.md`. (Accessed on 11/17/2020).

[3] Ori Hadary, Luke Marshall, Ishai Menache, Abhisek Pan, Esaias E Greeff, David Dion, Star Dorminey, Shailesh Joshi, Yang Chen, Mark Russinovich, et al. Protean:{VM} allocation service at scale. In *14th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 20)*, pages 845–861, 2020.

[4] Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. Monitoring misuse for accountable 'artificial intelligence as a service'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 300–306, 2020.

[5] Emilie Lundin, Håkan Kvarnström, and Erland Jonsson. A synthetic fraud data generation methodology. In *International Conference on Information and Communications Security*, pages 265–277. Springer, 2002.

[6] Sean Mcgregor. When ai systems fail: Introducing the ai incident database - the partnership on ai. `https://www.partnershiponai.org/aiincidentdatabase/`, November 2020. (Accessed on 11/20/2020).

[7] Pascale Minet, Eric Renault, Ines Khoufi, and Selma Boumerdassi. Analyzing traces from a google data center. In *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pages 1167–1172. IEEE, 2018.

[8] Marvin Newlin, Mark Reith, and Mark DeYoung. Synthetic data generation with machine learning for network intrusion detection systems. In *European Conference on Cyber Warfare and Security*, pages 785–XVII. Academic Conferences International Limited, 2019.