TR 99:2021
(ICS 35.020)

TECHNICAL REFERENCE

Artificial intelligence (AI) security – Guidance for assessing and defending against AI security threats

Singapore Standards Council

TR 99:2021
(ICS 35.020)

TECHNICAL REFERENCE

Artificial intelligence (AI) security – Guidance for assessing and defending against AI security threats

Published by Enterprise Singapore

ISBN 978-981-5042-50-4

Contents

Foreword

This Technical Reference (TR) was prepared by the AI Security Working Group set up by the Technical Committee on Artificial Intelligence under the purview of the Information Technology Standards Committee.

This TR serves as the first comprehensive standard development to address the security requirements and defences against increasing security and privacy threats of AI systems. The objectives are to establish best practices for AI security, design attributes and security metrics, to raise the security and privacy assurance of AI products.

This TR is a provisional standard made available for application over a period of three years. The aim is to use the experience gained to update the TR so that it can be adopted as a Singapore Standard. Users of the TR are invited to provide feedback on its technical content, clarity and ease of use. Feedback can be submitted using the form provided in the TR. At the end of the three years, the TR will be reviewed, taking into account any feedback or other considerations, to further its development into a Singapore Standard if found suitable.

In preparing this TR, reference was made to the following publications:

1. ISO/IEC TR 24030:2021 Artificial intelligence (AI) – Use cases

2. ISO/IEC TR 24028:2020 Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence

3. Model AI Governance Framework – Infocomm Media Development Authority (IMDA) & Personal Data Protection Commission (PDPC)

Attention is drawn to the possibility that some of the elements of this TR may be the subject of patent rights. Enterprise Singapore shall not be held responsible for identifying any or all such patent rights.

NOTE

1. *Singapore Standards (SSs) and Technical References (TRs) are reviewed periodically to keep abreast of technical changes, technological developments and industry practices. The changes are documented through the issue of either amendments or revisions. Where SSs are deemed to be stable, i.e. no foreseeable changes in them, they will be classified as “mature standards”. Mature standards will not be subject to further review unless there are requests to review such standards.*

2. *An SS or TR is voluntary in nature except when it is made mandatory by a regulatory authority. It can also be cited in contracts making its application a business necessity. Users are advised to assess and determine whether the SS or TR is suitable for their intended use or purpose. If required, they should refer to the relevant professionals or experts for advice on the use of the document. Enterprise Singapore and the Singapore Standards Council shall not be liable for any damages whether directly or indirectly suffered by anyone or any organisation as a result of the use of any SS or TR. Although care has been taken to draft this standard, users are also advised to ensure that they apply the information after due diligence.*

3. *Compliance with a SS or TR does not exempt users from any legal obligations.*

Artificial intelligence (AI) security – Guidance for assessing and defending against AI security threats

0 Introduction

AI adoption has accelerated over the past few years. According to Gartner’s CIO surveys, the percentage of organisations that have deployed AI has increased five-fold, from 4% in 2018 to 19% in 2020. Furthermore, in 2020, 25% of organisations have committed to deploying AI solutions within the next one year. In another survey, half of the respondents reported that their organisations have adopted AI in at least one business function.

The accelerated adoption of AI is due to the significant benefits it brings to organisations. For companies which adopt AI, on average across all industries, the majority experience revenue increases and cost reduction. These encouraging results should spur organisations to explore further applications of AI within their organisations.

Between 2018 to 2019, the percentage of AI capabilities embedded in high-risk applications have increased in the healthcare, transport, and logistics industries, as well as in the financial services industry. These applications share the characteristic of being heavily regulated due to the potential risks and impact they may have on society.

Given the probabilistic nature of an AI system, prediction errors can occur. When AI systems are deployed to high-risk applications, additional precautions and stringent audits are necessary, as a malfunctioning AI can result in disastrous consequences.

For example, in autonomous driving, a simple manoeuvre of turning a left corner required human assistance so as not to bump into a railing. When encountering rarely observed vehicles, such as emergency vehicles, accidents have been reported. These accidents can be a result of natural degradation, or can also be a result of AI security threats, the latter being the focus of this Technical Reference (TR). Examples of AI security threats in the context of autonomous driving are physical and digital adversarial attacks (e.g., projections) that can tamper the autopilot's predictions, potentially leading to an accident [1].

These AI security threats extend from autonomous driving to other high-risk industries, e.g., healthcare, finance, or the public sector. In healthcare, misleading disease diagnoses can result in wrong medical prescriptions. These risks show that additional quality assurance measures are required with a special focus on cybersecurity-related threats against AI, identified by the World Economic Forum as being one of the top ten global risks.

In response to the increasing usage of AI in high-risk industries, new AI design and development frameworks have been presented by numerous industry players. This is followed by regulation, standardisation initiatives and guidelines, e.g., the AI Act published by the European Commission or first standards from the International Organization for Standardization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE). These first initiatives aim to promote the consideration of ethical and functional principles, such as fairness or robustness, while minimising the attack surface(s) exposed to the growing number of AI threats.

So far, the initiatives which include "security" outline the set of attack types, against which an AI system should be secured. Some regulations go as far as proposing exemplary mitigation, e.g., adversarial retraining which serves as one out of many undefined defence techniques. However, none provide a comprehensive overview, as to how AI applications can be defended against AI security threats. In addition, there is no consensus on how the level of security can be assessed. This TR aims to address this gap, by contributing the following:

(a) An in-depth explanation of AI threats which AI systems can encounter;
(b) A comprehensive high-level overview of the state-of-the-art security defences;
(c) First assessment measures for evaluating the security of an AI system enabling comparison between ever-advancing defence approaches;
(d) Four use case studies featuring real-world high-risk AI applications aligned with the national AI strategy of Singapore, which serves as validation and guide for further adoption.

1 Scope

This TR serves as guidance on the early lifecycle of an AI system, namely the design stage, for the management of organisations. Therefore, this TR aims to empower management to understand the potential security risks and to establish the relevant design choices. Developers can follow the design choices by integrating techniques related to the outlined baseline security measures in clause 5.

The focus of this TR is on cybersecurity risks which can have legal, social and economic impacts as follows:

(a) Legal impact: AI applications that impact how the law is applied to natural or unnatural entities (e.g., immigration decisions approving or rejecting visa applications);
(b) Social impact: AI applications that can cause physical or mental harm (e.g., autonomous vehicles driving in public environments);
(c) Economic impact: AI applications which actions can result in significant monetary losses or restrict consumers from essential private services (e.g., credit scoring or insurance premium applications).

Those applications with cybersecurity risks of less impact are subject to an economically driven choice. Namely, understanding the additional workload of integrating defences versus the benefits of security, e.g., a recommendation for an entertainment program which can incur damage in public perception when being compromised.

Examples of areas of applications include the following:

(a) Infrastructure (transport);
(b) Education (auto marking, adapted learning);
(c) Healthcare (diagnosis, assisted surgery);
(d) Human Resources (HR) (hiring, scoring);
(e) Essential private services (loans, insurance);
(f) Essential public services (border control);
(g) Law enforcement (surveillance);
(h) Law application (fact aggregation);
(i) Cybersecurity (malware detection securing high value).

Further applications can be found in ISO/IEC TR 24030:2021.

2 Normative references

There are no normative references in this standard.