ASSURANCE FOR MACHINE LEARNING SYSTEMS

ASSURANCE FOR MACHINE LEARNING SYSTEMS

BY

MILAD HASSANI, Ph.D.

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTING AND SOFTWARE

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Master of Applied Science

© Copyright by Milad Hassani, June 2021

All Rights Reserved

Master of Applied Science (2021) McMaster University

(computing and software) Hamilton, Ontario, Canada

TITLE:	Assurance for Machine Learning systems
AUTHOR:	Milad Hassani Ph.D. (Computer Science), McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Richard Paige
NUMBER OF PAGES:	??, ??

Lay Abstract

A lay abstract of not more 150 words must be included explaining the key goals and contributions of the thesis in lay terms that is accessible to the general public.

Abstract

Abstract here (no more than 300 words)

Contents

List of Figures

List of Tables

Notation, Definitions, and Abbreviations

Notation

$A \leq B$	A is less than or equal to B

Definitions

Challenge

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
AV	Autonomous Vehicle
ASIL	Automotive Safety Integrity Level
MDE	Model Driven Engineering

Chapter 1

Introduction

With new developments in Artificial Intelligence (AI) and ML, a growing number of research projects in this field and many companies have started utilizing these methods. ML methods are also used in many safety critical applications such as Autonomous Vehicles (AVs) and healthcare applications. Therefore, it is very important to have a clear perspective of the safety of such methods in these applications.

In some applications, an erroneous outcome of the ML model has a harmful impact on many lives, for example in medical diagnosis **?**, loan approval **?**, autonomous vehicles **?**, and prison sentencing **?**. Despite the numerous research papers in this subject, there is still a need to delve deeper and understand the behavior of ML systems in safety critical applications.

One major drawback in using ML algorithms is that they are often treated as a black box and hence, using safety procedures for these methods is sometimes inapplicable **?**. In a review of automotive software safety methods **?**, an analysis of ISO-26262 part-6 methods was performed with respect to safety of ML models. This assessment shows that about 40% of software safety methods do not apply to ML

models **?**.

Safety specifications often assume that behaviour of a component is fully specified. Since the training sets used in ML methods are not necessarily complete, they violate this assumption, and some parts of the specification becomes not applicable to the ML components **?**. Most widely used ML frameworks such as Tensorflow **?** and Theano **?** employ a model driven approach in problem solving. Although model driven engineering approach has been successful in safety critical applications such as Automotive industry, the ML models cannot be guaranteed to operate in a safe manner.

There are two approaches with respect to ML and safety, first is to study safety of ML methods, algorithms, and processes and the second is to use ML methods to improve pre-existing safety assurance procedures. We will initially follow the first approach and review the literature for the methods applied to standardize and measure the safety of ML methods.

There are inherent performance metrics related to ML methods, such as accuracy and robustness, which can affect their applicability in safety critical applications. ML models can also be dependent to the domain they are trained **?**. In addition, other perturbations such as noise, natural and imaging artifacts can cause ML models to function less accurately **?**.

Assurance cases have been successfully used in various industries to describe why a system can be trustfully used for a specific application **?**.

A recent definition of safety assurance case is described in **?** as

> "A structured argument, supported by a body of evidence, that provides a compelling, comprehensible and valid case that a system is safe for a

given application in a given environment"

A structured argument is a **?**

"connected series of statements or reasons intended to establish a position...; a process of reasoning."

Reasons used in a structured argument can be considered as premises in logical terms and a conclusion can be drawn based on them. **?**. this might need more expansion as to what are some of the examples of these premises and the assurance cases.

To obtain assurance for ML systems it is essential to understand the ML lifecycle. This lifecycle follows a spiral process model**?** and is comprised of four stages **?**

- Data Management(DM)
- Model Learning
- Model Verification(MV)
- Model Deployment

Chapter 2

Safety and assurance

In **?**, five major research problems associated with unsafe behaviour of ML models is presented. They can be summarized as

1. Avoiding Negative Side Effects: How to ensure that the model will not disturb the environment while pursuing its goals, e.g. can a cleaning robot knock over a vase because it can clean faster by doing so? Can we do this without manually specifying everything the robot should not disturb **?**?

2. Avoiding Reward Hacking: How to ensure that the model does not avoid situations to achieve a higher reward. For example, if we reward the robot for achieving an environment free of messes, it might disable its vision so that it won't find any messes, or cover over messes with materials it can't see through, or simply hide when humans are around so they can't tell it about new types of messes **?**.

3. Scalable Oversight: How to ensure the model respects the parts of the objective function that are expensive to evaluate and makes a safe approximation of these

parts. For example, in the cleaning robot example, if the user is happy with the cleaning quality is an expensive objective function, but it can be approximated to presence of any dirt on the floor when the user arrives **?**.

4. Safe Exploration: How to ensure that the ML model explorations are safe. For example, the robot should experiment with mopping strategies, but putting a wet mop in an electrical outlet is a very bad idea **?**.

5. Robustness to Distributional Shift: How to ensure that the model performs robustly if the environment shifts from the training environment. For example, strategies a cleaning robot learns for cleaning an office might be dangerous on a factory workfloor **?**.

Chapter 3

Machine Learning

Chapter 4

Literature Review

Chapter 5

Conclusion

Every thesis also needs a concluding chapter