ASSURANCE FOR MACHINE LEARNING SYSTEMS

Abstract

Abstract here (no more than 300 words)

Contents

List of Figures

List of Tables

Notation, Definitions, and Abbreviations

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
AV	Autonomous Vehicle
ASIL	Automotive Safety Integrity Level
MDE	Model Driven Engineering
PDF	Probability Distribution Function
RL	Reinforcement Learning
GSN	Goal Structuring Notation

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 behavior 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 **?**, Caffe **?**, Pytorch **?** 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 **?**.

In this report we will first explore basics of ML in Chapter **??**. Then in Chapter **??** we review definition of safety and how assurance cases are structured. Finally in Chapter **??** we survey the literature on ML assurance and identify some of the open challenges in this area.

Chapter 2

Machine Learning

In this chapter we will start by definition of ML in the literature, and continue with definitions of learning categories such as supervised, unsupervised and reinforcement learning. Next, we explore how data is managed in a given ML problem. Finally, we review how performance of ML methods are measured.

2.1 Definition

Machine learning algorithms can extract patterns and learn from data **?**. A brief definition of learning can be given as **?**

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

A task is the main objective of using an ML algorithm. For example, in an autonomous vehicle, driving the car is the task. A task is not the process of learning.

Learning is used as a means to achieve an ability to accomplish a task **?**. With developments in ML methods, they have been applied to different tasks, some examples of tasks are classification, regression, transcription, machine translation, denoising **?**.

The performance measure is used to quantify how successfully a task is accomplished, equivalently, number of erroneous outputs could be used as a way of indicating a method's performance.

Based on the above-stated definition, the ML algorithm undergoes an experience in the process of learning. This experience is generally classified into **unsupervised**, **supervised** and **reinforcement** learning.

2.2 Learning types

Unsupervised learning finds the properties of the overall structure of the dataset. Clustering as an example of unsupervised learning, finds clusters within a dataset and assigns each data-point to one of them.

In supervised learning, on the other hand, data-points that the learning algorithm experiences have a label. This label acts as a guide for the ML algorithm. The term supervised arises from the fact that the labels instruct the algorithm what to do. Labels are unavailable in unsupervised learning and the ML system is responsible to make sense of the data independently **?**.

Reinforcement learning (RL) algorithms experience an environment instead of a fixed dataset. The algorithm should learn how to maximize a reward function by taking an appropriate action **?**. The learner discovers this appropriate action by trying different actions and observing the value of the reward function. Actions not only affect the immediate reward, but can also change next actions' rewards. Trial

and error search and delayed reward are two main characteristics of RL.
The learner, also known as the agent in RL terms, should have the capability to sense the state of the environment, take actions that can alter the state and also have a goal to reach by taking actions. These three aspects are included in the reward function used by the agent **?**.

2.3 Data in ML

Evidently, ML algorithms need data to learn and function. A dataset can be described as a **design matrix**. Every row in the matrix contains an example, also known as data-point, and each column is a feature. Iris dataset is one of the first ones used in statistics and ML **?**. This dataset is comprised of 150 examples which have 4 features each. One example corresponds to one individual plant. Sepal length, sepal width, petal length and petal width are recorded as features of each plant **?**. This means that if X is the matrix, we can say $\mathbf{X} \in R^{150\times 4}$

The ML model will ultimately be deployed and used in a real world situation, hence, we are interested in how well an ML model performs on the data it has not seen before, this is also known as **generalization**. A portion of dataset is therefore not used in the training process and reserved as a **test set**. The data used in the training process is accordingly referred to as the **training set ?**.

In some cases, the training and test datasets might be limited in size and to have a better generalization, it is necessary to use as much of the data for training as possible. In other words, there will be less data available to estimate the performance of the model. One solution for this situation is **cross-validation**. The entire dataset is split into S subsets. In each run, $S-1$ subsets are used for training and one

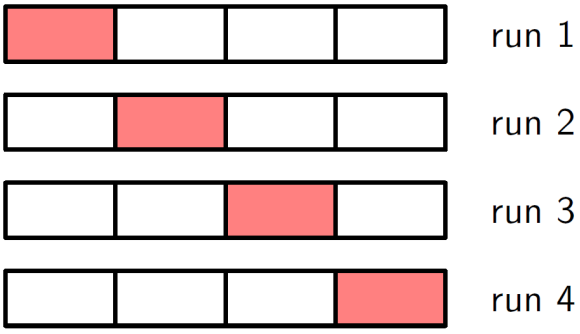

Figure 2.1: Cross validation for S=4 **?**.

		Predicted Class		
		C-	C+	
True class	C-	Tn	Fp	Cn
	C+	Fn	Tp	Cp
		Rn	Rp	N

Table 2.1: A confusion matrix

remaining subset is the test set. For the next run, a different test set is selected **?**. Figure **??** shows selection of subset for $S = 4$.

2.4 Performance measures

In classification tasks, *confusion matrix* is a way to demonstrate differences between predicted and true classes **?**. Table **??** shows the structure of a confusion matrix.

In Table **??** *Tp* and *Tn* represent true positives and true negatives respectively. *Fp* and *Fn* are in the same manner the count of false positives and false negatives respectively. *Cp* and *Cn*, therefore, are the total number of positive and negative examples. Finally, *Rp* and *Rn* denote total number of predicted positives and negatives, respectively **?**.

A variety of performance measures can be calculated from the confusion matrix,

e.g., accuracy, precision, sensitivity and specificity **?**. Accuracy is often considered as a performance criteria which is simply the fraction of correctly classified samples to total samples, i.e., $\frac{Tp+Tn}{N}$. It is also possible to obtain the same information by calculating the *error rate*.

The Receiver Operating Characteristic (ROC) curve helps to find a pareto-optimal point between true and false positive rates as the decision threshold changes **?**. Each point on the curve represents the Tp (vertical axis) and Fp (horizontal axis) for a decision threshold. The area under ROC curve (AUC) is, therefore, a measure the sensitivity of the model to changes in operating conditions. If AUC value is at maximum, i.e., one, it can be concluded that the $P(Fp) = 0$ and $P(Tp) = 1$ even when the operating conditions change.

2.5 Neural networks

In this section we briefly review the structure of neural networks and their building blocks. As the name suggests, neural networks are fundamentally a collection of entities called *neurons*, which can hold small units of data. Here we consider a simple neural network, depicted in Figure **??**,consisting of a hidden layer and an output layer. Hidden layers are the ones not directly accessible from the outside world, i.e., not the input or output layers.

Using the notation in **?**, if we have D input variables, x_i, we can calculate linear combinations a_j such that

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}$$

Here, $1 \leq j \leq M$ where M represents number of neurons (nodes) in the hidden layer.

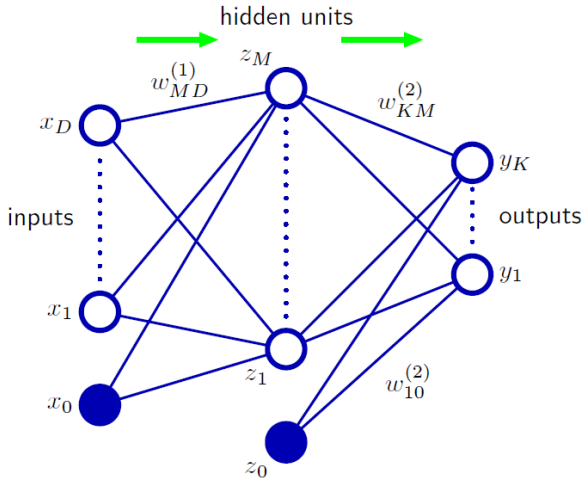

Figure 2.2: Structure of a two layer neural network. **?**.

The superscript (1) corresponds to the first layer of the network, i.e., the hidden layer in this example. $w^{(1)}_{ji}$ are called weights and $w^{(1)}_{j0}$ are biases for the hidden layer. The a_j values are referred to as *activations*. Next, we use activation function, h, to calculate z_j, such that

$$z_j = h(a_j)$$

The activation function is required to be differentiable due to the differentiation in the learning process. z_j values are then used to compute the linear combinations a_k in a similar manner to the previous layer, i.e.,

$$a_k = \sum_{j=1}^{M} w^{(2)}_{kj} z_j + w^{(2)}_{k0}$$

a_k is the output layer activation, therefore, similar to the hidden layer, an activation function, σ, will be used to reach final output values of the network, y_k.

$$y_k = \sigma(a_k)$$

Should I include the training process for neural networks? It takes more space. At least 1-1.5 pages.

2.6 Trends in ML research

With rapid development of various frameworks and libraries in ML they are increasingly easier to be applied in different applications. It is possible to develop smaller models which in turn makes applications such as IoT **?** more feasible. Prevalence of ML raises important questions in its fairness **?**, privacy **?**, and safety **?**. The term *responsible ML* covers the social impacts of using ML in everyday life. Moreover, the desire to have faster, better performing ML systems has lead to areas such as quantum ML **?**. As a matter of fact, quantum ML has been one of the initial motivations for experiments which lead to the famous quantum supremacy **?**.

Chapter 3

Safety and assurance

In this chapter, we start with basic definition of safety. Next, we delve into safety assurance cases and how they are structured. Finally, we review a representation method for assurance cases called Goal Structuring Notation (GSN).

3.1 Definitions

3.1.1 Safety

Safety is defined in ISO 26262 **?** as:

> Absence of unreasonable risk.

An unreasonable risk is a **?**:

> Risk judged to be unacceptable in a certain context according to valid societal moral concepts.

Various safety standards have been developed for different industries and activities. Some examples are ISO 26262 for functional safety of road vehicles, DO-178C for

aerospace industry, ISO 8124 for safety of toys, ISO 7164 for healthcare organization management.

3.1.2 Assurance

Assurance is defined in ISO 15026 to be **?**

> Grounds for justified confidence that a claim has been or will be achieved.

Assurance is, therefore, the grounds on which the users of a system can rely on its functionality. It is specially important for systems with complexity, such as ML, to give assurance to the users before they start utilization. The level of this assurance is closely related to the level of dependence or trust needed from the users' side. Adequate evidence and arguments need to be present to justify the safety and reliability of the system. The basis for this justification is achieved with reducing uncertainty in measurements, observations, estimations, predictions, information, inferences or effects of unknowns **?**.

3.2 Safety Assurance Case

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 **?**. The purpose of using an assurance case is to communicate a clear, comprehensive, defensible argument that a system is safe to be used in a particular context **?**. Assurance cases are comprised of five basic components: claims, arguments, evidence, justifications and assumptions. The most common use of assurance cases is to give assurance about system's functionality and properties to the parties which were not involved in the process of developing the system **?**.

Assurance cases reason in a subjective manner, as compared to the logical proofs which consider an absolute truth. In other words, assurance cases are useful because the full range of a system's properties are not always representable in a logical formalization. Also, assurance cases may sometimes be disproved because the underlying logical theory used in them is not relevant **?**.

Since assurance cases are considered artefacts, they inherit quality related properties of them such as: the structure of its content, semantic features such as completeness, creation and maintenance. The conclusions of the assurance case should also be stated clearly with clear level of uncertainty **?**.

```
  For hazards associated with warnings,
the assumptions of [7] Section 3.4
associated with the requirement to
present a warning when no equipment
failure has occurred are carried
forward. In particular, with respect to
hazard 17 in section 5.7 [4] that for
test operation, operating limits will
need to be introduced to protect against
the hazard, whilst further data is
gathered to determine the extent of the
problem.
```

Figure 3.1: Problems associated with textual representation **?**.

3.3 Goal Structuring Notation

When the safety assurance case is more complex in nature, textual representation suffers to express the case in a clear and understandable way. Figure **??** shows an example of such problem where the English structure of the argument is hard to understand. Having multiple cross references is specially difficult to capture in text **?**.

The Goal Structuring Notation(GSN) is a graphical notation for safety argumentation. A GSN explicitly represents elements of a safety argument and the relationships among these components. For example, how requirements are supported by claims or how claims are supported by evidence or how the case has a defined context **?**. Figure **??** depicts basic building blocks of a GSN with example instances of each element.

3.4 An example of a GSN

The goal structure is used to show how goals (claims about the system) can be split into sub-goals successively until the sub-goal can be directly supported by available evidence. Figure **??** represents an example of a GSN.

In this example, "Control System (C/S) logic is fault free." is one single top level

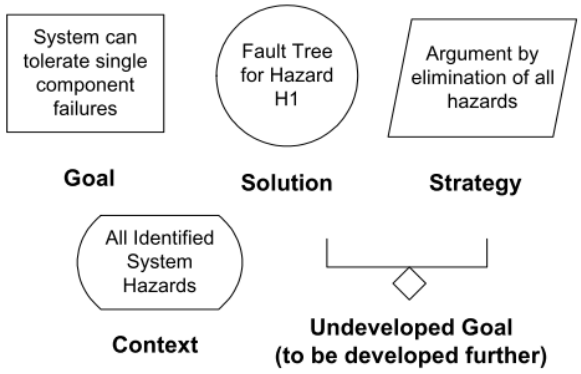

Figure 3.2: Basic elements of a GSN [?].

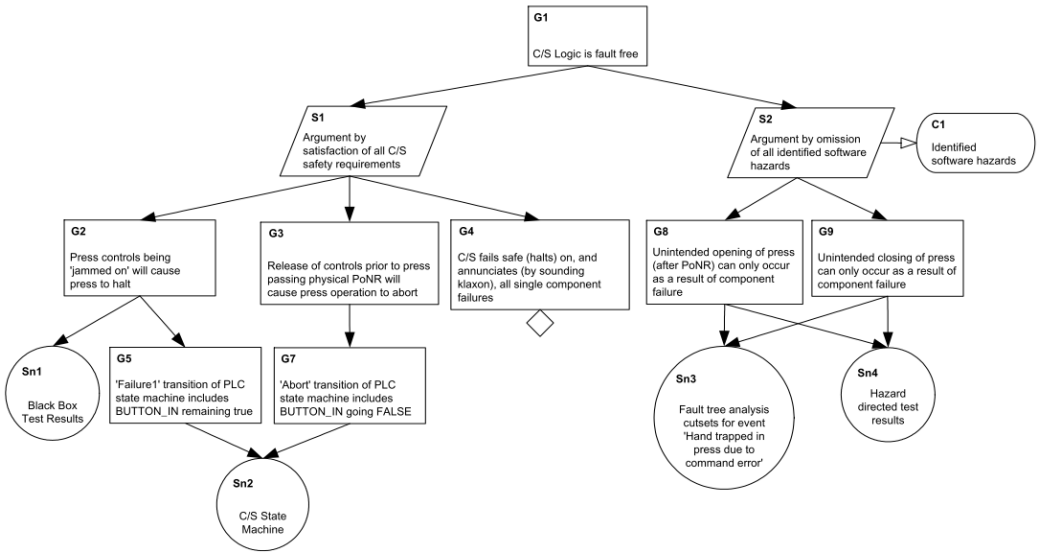

Figure 3.3: An example of a goal structure [?].

goal. The main goal is then divided to two sub-goals through strategies $S1$ and $S2$. These two strategies are then supported by five sub-goals $G2 - G4$ and $G8 - G9$. In a goal structure, there will be a stage where the sub-goals can be directly supported by solutions. In this example, sub-goals $G8 - G9$ are supported by $Sn3 - Sn4$ and there is no need to break down the goals further in this branch **?**.

3.5 Trends in safety research

Safety assurance should constantly evolve and adapt to the new paradigms in industry. With advent of Industry 4.0, new risks and challengers arise in workspace and occupational safety. Safety 4.0 is a response to these new challenges **?**. There is a growing number of research projects on using the recent advances in ML technologies to enhance safety, some of which is referred to as "safety informatics **?**". In addition, 5G technology has raised concerns about the long term health consequences. Critiques are collecting evidence that 5G may result in skin cancer and the millimeter wave radiation can ultimately affect the nervous system **?**. Unmanned Aerial Vehicles (UAV), or commonly known as drones, are rapidly spreading in industrial usage and thus their safety and privacy challenges are of interest **?**.

Chapter 4

Literature Review

In this chapter we will briefly review some of the literature about the safety of ML methods and identify major research questions in this area. Today, ML is used in variety of applications with varying safety requirements. This diverse portfolio includes smart phones **?**, cars **?**, surgical equipment **?**, construction industry **?** and many more. A fault in an ML system, e.g. a misclassification, has different repercussions in each application. An out of focus image taken by a camera can be easily remedied, but a malfunction in a surgical equipment could be fatal and result in an irreversible situation. Therefore, assuring safety is an essential part of the design process for these applications.

One major issue in safety assurance for ML is guaranteeing that the training data is complete and relevant. Data used in the operational stage is by definition relevant, however, training data may not reflect all possible situations that the learning algorithm needs to be exposed to. On the other hand, the operational environment may change and the training data may diminish in relevancy. As discussed in detail later in this chapter, assuring completeness and relevancy is challenging and may not

always be feasible.

As this part is mostly concerned with RL I should put it in a separate subsection

In **?**, five major research problems associated with unsafe behavior 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 work-floor.

4.1 Machine Learning lifecycle

To obtain assurance for ML systems it is essential to understand the ML lifecycle and how to analyze safety in each step. In this section we will first introduce these steps and review some of the safety measures for each step. This lifecycle follows a spiral process model, i.e., the stages are iteratively repeated to actively reduce risk **?**. ML lifecycle is comprised of four stages **?**

4.1.1 Data Management

This stage involves collecting, preprocessing, augmenting and initial analysis of data. The training and validation datasets are also prepared in this step. From assurance perspective, the data collected in this step should be

- Relevant: The dataset should be relevant to the desired functionality of the final model. For example a dataset of handwritten letters in Japanese language cannot be used for English language. In many cases, a pre-existing dataset will be used to train the model. This dataset should be obtained from trusted sources with an encrypted transmission medium. An attacker can add irrelevant samples into the training dataset to make the final model behave in a specified way for particular inputs **?**. The irrelevant data injected in the dataset is called a backdoor **?** Despite early efforts in detecting backdoors in face recognition **?**

finding a general solution is still an open challenge.

- Complete: The features of a dataset should not have unintended correlations that can confuse the classifier. For example, if a classifier is trained on pictures of wolves and huskies, and all wolves have snow in the background, it may be concluded that snow in the background means a wolf **?**. In this case the dataset is not complete because it does not include pictures of wolves with different backgrounds. Exploratory Data Analysis (EDA), is an important step in examining completeness of a dataset. It is also important to identify how the data points are distributed in the input sample space, measures such as gap ratio **?** can be used to evaluate uniformity.

- Balanced: For classification problems, it can happen that one class has significantly more data-points in the training set than the others and thus, classifier has more exposure to that class.

- Accurate: This property considers factors like sensor accuracy, correctness of data collection and processing method. In the case of supervised learning, labels' accuracy is important. The data collection process should be documented to identify potential inaccuracies **?**.

4.1.2 Model Learning

In this stage of the ML lifecycle, the type of the model and its hyper-parameters are selected. For some ML applications, the dataset is very large or the model structure is complex and therefore, the learning process needs considerable amount of computational power. In these cases, it is reasonable to take advantage of a previously trained

model and adapt it to our needs by re-training only the parts that are different from the other application. This process is called *transfer learning* **?**. If there is a need to do transfer learning, it will be decided at this stage and finally the learning process starts using the train dataset obtained in previous stage. In order to have a clear view of the model in safety related aspects, the final model should be **?**

- Performant: As a requirement for a safer model, it should have a justifiable performance according to the measures introduced in **??**.
- Robust: The model should be able to perform as well on the unseen data as the training data,i.e., generalizes well to be considered robust. Data augmentation is one of the techniques to increase generalization **?**.
- Reusable: Using transfer learning can help to use the assurance evidence of the original model, provided that the transfer learning is performed in the right context for the source and destination models. However, reusing models comes with the risk that safety issues propagates to the destination model too
- Interpretable: This property shows how much the decisions made by the model are explainable and thus helps to analyze the safety of such decisions.

4.1.3 Model Verification and Validation

The black swan problem expounds one of the major challenges in validating ML models. A system or a person could incorrectly conclude from abundance of training data samples that common observations are true **?**. A model which has only seen white swans, may infer that all swans are white and ignore the fact that there are black swans **?**. One major challenge is thus making sure that the model works well, i.e., satisfies its

requirements, on the data it has not seen before which is also known as generalization. If the model fails in this stage, the process will go back to Data Management or Model Learning steps. Model verification involves requirements encoding, test-based verification and formal verification. The verification stage should be **?**

- Comprehensive: Model verification should ensure that all the requirements of the system and also intended goals of the previous stages of ML lifecycle, i.e., data management and model learning, are covered.
- Contextually relevant: Verification process should be relevant to the intended use of the ML model. For example an ML model used in autonomous vehicles, we are more concerned about how changes in the environment will affect model's performance and thus, how robust is the model with changes in weather rather than the changes in image quality.
- Comprehensible: Verification results should be understandable for the users. Requirement violations should be clearly expressed in such a way that the cause of it can be identified and fixed **?**. Ideally, the results should also include any black swan biases present in the model **?**.

4.1.4 Model Deployment

Preparing the ML model to be used in the final application. Activities in stage includes integration, monitoring and updating. To assure safety of this stage of ML lifecycle, the ML model should have the following properties

- Fit-for-Purpose: The difference in hardware can cause performance differences between ML stages. Also, each distinct hardware setting where a model is

deployed can affect model's performance. For a model to be fit for purpose, the performance seen in the previous stages should be carried over to the deployment phase.

- Tolerable: The system should be able to tolerate occasional incorrect outputs of the ML model. To accommodate this, the host system should be able to identify the incorrect outputs and to replace them with a safe value so that the system continues the normal processing activities.

- Adaptable: Deployed models are in many cases needed to be updated due to variety of reasons including operational, legislative or environmental changes. This property indicates how safe is the process of updating.

4.2 Open challenges in ML assurance

Using an ML component in a system poses several challenges in each step of the ML lifecycle. In the data management step, further research is needed to guarantee security of data and its fitness for the purpose. Although a vast amount of research has been conducted in the model learning stage, there is still a need to further study hyper-parameter selection. In addition, with recent successes in transfer learning, there is still need for more research in assuring safety in this area. Furthermore, safety assurance requires ML models to be reusable and interpretable. Model verification assurance is mainly accomplished using test-based and verifications. However, there is still a need to develop methods to encode model requirements into proper and formal tests. In model deployment stage, there is no explicit equivalent for updating models in software engineering world, therefore, there is a need to devise assurance

methods for adaptable safety-critical systems **?**.

In some applications requirements for a safe ML system reinforce each other. For example, accuracy in data management stage will most likely result in more performant model. However, in some cases, there is a trade-off between requirements, an explainable model is probably more exposable to cyberattack **?**. In spite of attempts to address this issue **?**, more research is required to adapt these concepts to ML.

Chapter 5

Conclusion

In this report we first started with the basidc definitions and principals of machine learning and safety assurance concept and eplored some of the fundamentals in both areas in Chapter **??** and Chapter **??** we also glanced currently trending research in these areas. Finally, in Chapter **??** we reviewed some of the literature in assurance of ML systems and some of the open challenges and research questions in this field.

State some of the major open challenges I found,