Improving the Reliability of Artificial Intelligence

Research Proposal

Ajal Singh - 12621189
Supervisor: Diep Nguyen

25 October, 2020

Contents

List of Figures

List of Tables

Engineering Research Problem

1.1 Background

As the use of Artificial Intelligence (AI) and Machine Learning (ML) continues to grow throughout the world in high-risk applications, models have become ever-increasingly complex and diverse. As a result, they often become prone to accidents where unintended and harmful behaviour is observed, and consequently are scrutinized as disruptive and unreliable solutions. The recent emergence in smart cities have seen AI and ML being used in various applications such as transportation, healthcare, environmental, and public safety as depicted in Figure 1.1.

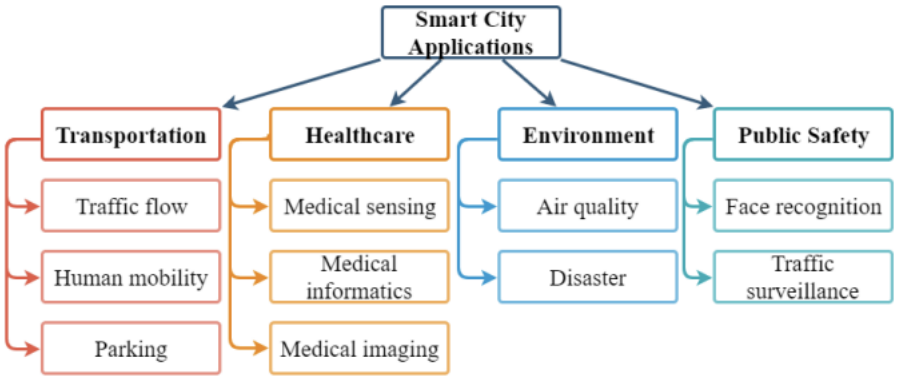

Figure 1.1: Smart City Artificial Intelligence Applications [1]

For an AI/ML system to be considered reliable, it must perform tasks when required as it was originally intended, produce consistent results using real-world data (and shifts in data), and remain robust and predictable. This means it must also fail in a predictable manner [4].

1.2 Applications

One of the most discussed and disruptive applications of ML is facial recognition systems used by authorities which fails to distinguish between darker skin individuals. This technology is used to assist the police in identifying potential criminals/suspects and often leads to wrongful arrests of dark-skinned people [5]. This example highlights the importance and the need for reliability in ML solutions.

There are many more applications where reliability is crucial due to the potential consequences. Cancer diagnosis systems trialled in the US are failing to detect cancer in patients in differing hospitals and/or countries which may result in death. As another example, unintended behaviours in traffic management systems would increase congestion resulting in poor ambient air quality and noise pollution.

The rapid technological changes in manufacturing have produced a boom in Industry 4.0 applications involving Artificial Intelligence, connected devices (IoT) and Big Data. A paper on use cases of AI in Industry 4.0 summarises the advantages ML, *"AI with machine learning technique can automate the manufacturing process which increase the productivity, efficiency, optimize production cost and reduce manual error"* [6]. A key area is predictive maintenance where real-time equipment data is captured and historical equipment data is evaluated using AI and ML models to estimate the equipment life cycle and hence perform timely maintenance to reduce or eliminate down-time. Down-time is undesirable for manufacturers as it equates to the loss of revenue.

AI in cybersecurity helps protect enterprises by detecting unusual activity, patterns, and malicious behaviour and can respond to different situations. For manufacturers, this could be used for asset protection while banks and financial institutions may use this form ML to detect suspicious activity and fraud [6].

1.3 Project Contextualisation

A tutorial presented by Suchi Saria and Adarsh Subbaswamy of John Hopkins University [7] postulates some causes and failure prevention techniques for use in supervised learning systems (regression and classification). Some of the sources of unreliability discussed are the use of inadequate data, changes in training and deployment environments, and model misspecification. These aforementioned causes will form the basis of this research project.

Another reliability issue is discussed in a separate paper, *'Concrete Problems in AI Safety'* [8] is the prevalence of reward hacking in Reinforcement Learning systems. Reward hacking is the AI agent's ability to cheat the system to achieve the highest

reward in an unintended way. For example, a positive reward may be given to a traffic management system when there is no congestion. However, the AI model decides to divert all traffic through alternative routes essentially shutting down this particular road/intersection. This prevents congestion but does not perform as desired. This notion is also investigated in this research project.

1.4 Research Question

How can the reliability of Artificial Intelligence be improved against inadequate data labelling, unsuitable algorithm choices, and reward hacking?

Methodology

This project has been divided into three main sections based on the factors of unreliability mentioned earlier in Section 1. The main factors studied in this project are:

- Label Bias and Environmental Datashift
- Suitable Algorithm Selection
- Reward Hacking

To prevent any delays, writing of the report will be completed concurrently with experimentation. This is to ensure that ideas, thoughts, results, and conclusions are fresh in mind at the time of writing. This will improve the quality of the report which is crucial as it should be a direct and official indication of all work completed. The report is also the only assessable task in 41030 – Engineering Capstone, and therefore the only way to gain marks.

2.1 Label Bias and Environmental Datashift

To investigate the effect of biased data labelling we will train two models using a single independent algorithm. One will use biased data while the other uses unbiased data. These datasets will be modelled using the mathematical framework outlined in the conference paper *'Identifying and Correcting Label Bias in Machine Learning'* [9]. Each model will be trained and evaluated using data which has been split from the same distribution. For a model to be considered reliable it must be able to properly generalise or adapt well to new and unseen data. A good, reliable model can achieve high accuracy scores with low variance between datasets. Therefore, both of these trained models will then be fed previously unseen data (i.e. deployment data) to determine its ability to generalize.

2.2 Suitable Algorithm Selection

As can be seen in Figure 2.1, when it comes to AI and ML, the appropriacy of solutions or algorithms depends on elements such as the specific application and the level of supervision required. More often than not, more than one algorithm could be a viable solution (see Figure 2.2). Therefore, to investigate suitable algorithm selections, models will be trained with a single dataset using different algorithms (with different assumptions). They will then be tested for accuracy to determine suitable algorithm choices. Evaluating the reliability of a model is dependant of the model type. Accuracy, precision and recall are three common metrics we can use to evaluate a model. However, depending on certain applications, other complex means of metric evaluation may be necessary.

Machine learning algorithms	Purpose
Feed forward neural network	Smart health
Densities based clustering and regression	Smart citizen
K-means	Smart city, Smart home
Clustering & anomaly detector	Smart traffic
One class support vector machine	Smart human active control
Support vector regression	Smart whether
Linear regression	Smart market analysis

Figure 2.1: Machine Learning Algorithms for specific applications [2]

The bias-variance trade-off should be considered when optimising ML models. Bias is the model's ability to learn the wrong things due to oversimplification or incorrect assumptions. Variance is the error due to sensitivity as a result of small fluctuations in training data. As the complexity of the model increases, bias decreases but the variance will increase. This is the trade-off between these two factors. An overfit model is one that is too complex resulting in high variance and low bias, while an underfit model has low variance and high bias due to its simplistic nature. Both overfit and underfit models are undesirable and it is ideal to find a suitable trade-off between bias and variance (hence complexity) to yield a well fit model capable of adapting to different datasets [10].

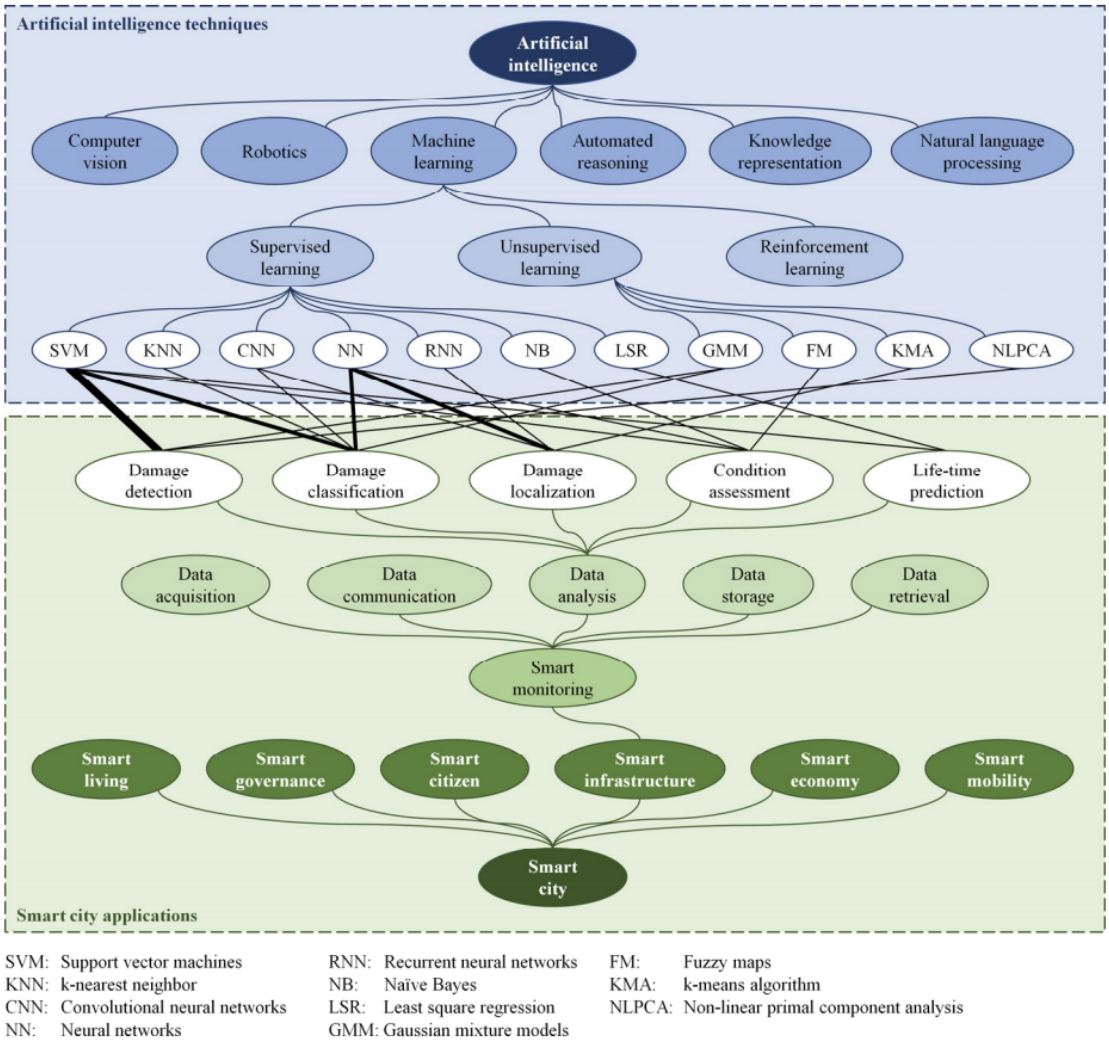

Figure 2.2: Available ML algorithms for smart monitoring [3]

The dataset/s to be used in the above experiments will be obtained through various open-source data collections available online. Therefore, data collection is not a part of this project. To ensure validity during the training of models, the data distribution will be split into three smaller datasets for training, validation and testing. The training set is used to train the models to fit the data and are evaluated against the validations set. The validation set being unseen, allows us to determine which models are generalising well to new examples. After the best model has been selected it is again tested on the test dataset as a final check on its generalisation ability. The training set accounts for 60% of the full data set, while the validation and test sets account for 20% each.

2.3 Reward Hacking

The two unreliability factors discussed in the above experiments are concerned mainly with supervised learning models. A major reliability issue within reinforcement learning models is reward hacking. We will perform a systematic literature review on applications and known causes of unreliability due to reward hacking as well as potential solutions.

Project Management

3.1 Scope

The scope of the project involves studying the factors that affect the reliability of Artificial Intelligence (AI) and Machine Learning (ML) with a focus on smart city applications. After specific applications have been identified, the factors affecting the reliability are investigated in detail before finally performing experiments to propose solutions to improve reliability. The factors studied within this project are data labelling and environmental data shifts, suitable algorithm selection, and reward hacking. The mathematics behind these factors will be researched before performing experiments in an attempt to propose solutions.

3.2 Process and Timeline

A Gantt chart has been produced to illustrate the order of completion of major and minor milestones involved in this project. This not only shows the duration of the whole project but also the time required for each milestone and its dependency on other tasks. The arrangement of milestones against specific dates acts as a schedule where each task has a required deadline which, ideally should be met to prevent delays. However, delays in most projects are unavoidable, therefore precautionary buffer periods have been added to allow flexibility in this project.

See Appendix 2 for the Gantt Chart and detailed process including milestone deadlines.

3.3 Milestones

Tasks and activities involved in this project have been split into major and minor milestones as seen below in Figure 3.1. Major milestones are logically ordered towards the completion of the project and may consist of any number of smaller milestones.

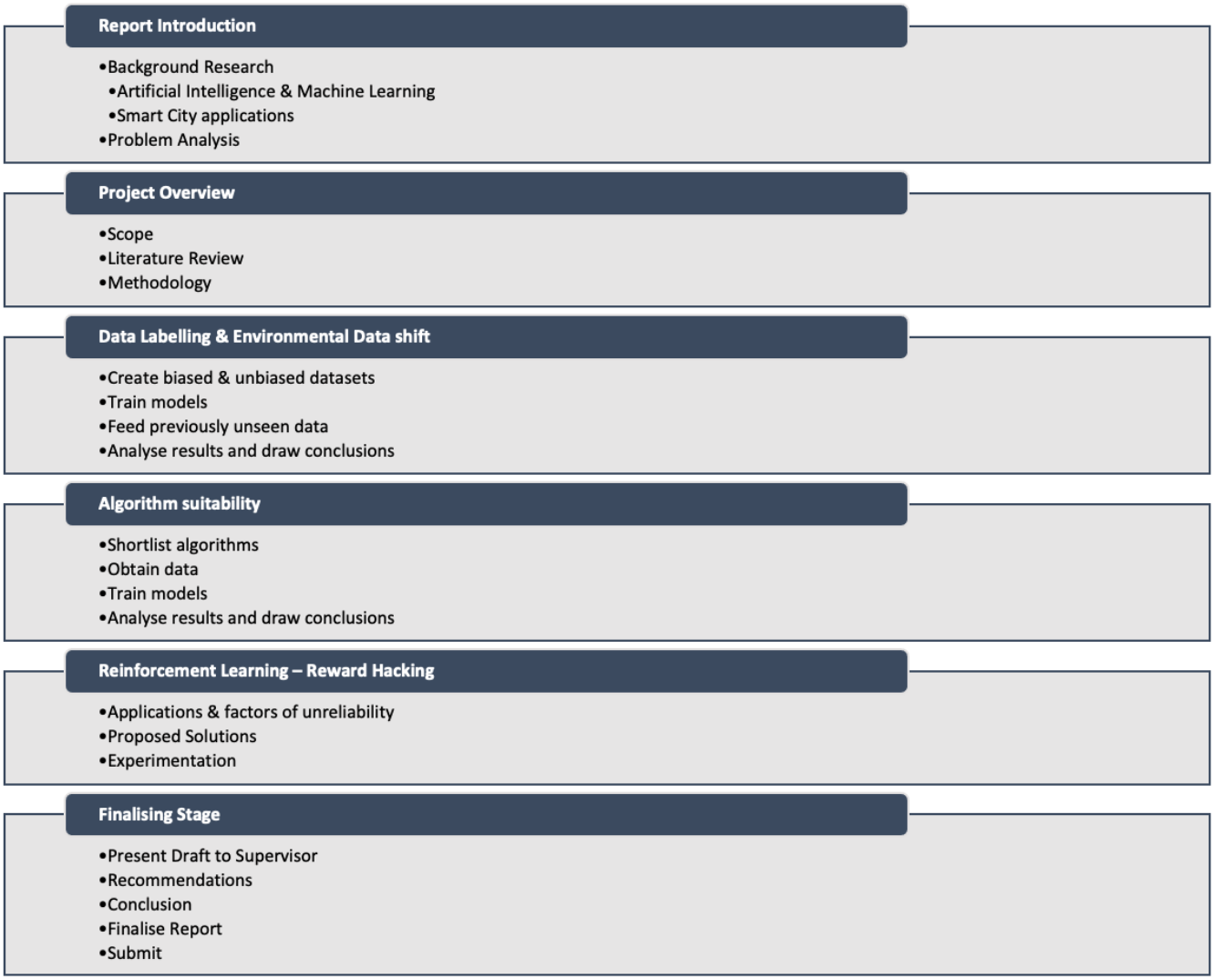

Figure 3.1: Major and Minor Milestones

3.3.1 Stage One: Report Introduction

To prevent project delays, the report will be started early to shift focus on technical aspects during the closing stages of this project. Therefore, the introductory components will be completed prior to the start of semester as depicted in the schedule. This introductory stage of the project involves research and analysis into the unreliability in Artificial Intelligence and Machine Learning. Potential causes are investigated with respect to Smart City applications before a problem is clearly defined.

3.3.2 Stage Two: Project Overview

After conducting background research on the causes of unreliability, a project scope is developed. Here literature review is performed on the potential solutions in order to develop a suitable methodology to form the basis of this project.

3.3.3 Stage Three: Data Labelling & Environmental Data shift

As mentioned in the methodology, the first technical stage involves investigating how biased data labelling will affect the ML model in unseen environments. First, a dataset is obtained from online sources and altered to create bias. The models are then trained using both the biased and original datasets. Environmental Datashift is simulated by feeding the model data which is of a similar nature but different to the training data. This means the labels will remain the same but the data istelf will be different. Finally, after testing each model's ability to generalize, we can analyse the results and draw conclusions. Recommendations will also be made.

3.3.4 Stage Four: Algorithm Suitability

In this stage, multiple algorithms are used to train a single dataset to determine algorithm suitability for specific applications. After models have been trained, accuracy as well as other evaluation metrics are obtained and analysed to draw conclusions. Recommendations will then be made.

3.3.5 Stage Five: Reinforcement Learning – Reward Hacking

The final technical stage of this project involves performing research into the applications and causes of reward hacking in reinforcement learning models. Literature

Review is performed to determine current solutions and experimentation will be performed to evaulate various performances.

3.3.6 Stage Six: Finalising Stage

Finally, after all stages have been completed, a draft report is produced and sent to my supervisor for review. Final recommendations and conclusions are drawn up before supervisor feedback and other issues will be addressed. A second draft is then sent to my supervisor for review. After finalising, the final report will be submitted on 31 May 2021 as per schedule.

3.4 Resources

3.4.1 Academic Supervisor – Diep Nguyen

My academic supervisor, Diep Nguyen, is highly knowledgeable in the field of Machine Learning. Therefore, regular consultation would be a good idea to gather advice and guidance on topic related matters as well as suggestions on research content and experimentation.

3.4.2 Literature

Literature such as journals and articles will be sourced from reliable sources such as IEEE, AMC, and Google Scholar. Various video tutorials may also be taken advantage of from platforms such as YouTube.

3.4.3 Technical Tools

Experiments performed in this project will be completed using the Python programming language. This is due to the availability of many useful packages and its popularity within the data science community results in an abundance of available documentation and tutorials. Some of the packages we have access to are NumPy for mathematical operations such as arrays and matrices, Matplotlib for plotting and data visualisation, pandas for data analysis and manipulation, and scikit-learn for machine learning [11]. There are many available deep learning libraries available such as Keras, TensorFlow and PyTorch. While they all have their pros and cons, PyTorch is preferred for this project due to its low-level nature allowing for experimentation [12].

3.5 Uncertainties & Risk Control

See Appendix 3 for a detailed Risk Assessment.

3.5.1 Non-Technical

Some non-technical risks and uncertainties could affect the completion of this project. The biggest risks in this project are those that would prevent successful completion. There lies the risk of failing to complete the project due to delays and difficulties in completing milestones. To prevent this, buffer periods have been added to the schedule. These periods will allow me to catch up to schedule if I do find myself falling behind. If the buffer periods aren't required, I will find myself ahead of schedule which is another benefit.

The next greatest threat to this project is an inadequate understanding of the critical components involved in this project. It is essential to adhere to the schedule, revisit reliable sources and consult my supervisor as soon as possible in these circumstances.

Another risk is the loss of work and other saved data due to uncontrollable reasons such as computer failures. To prevent this, all work will be saved to OneDrive (cloud storage).

There is also the uncertainty of ideas and general direction of the research report. In these cases, example or past reports can be viewed. Many high-quality reports (Distinction and High Distinction) reports can be found on the UTS Library website and are available for download. My supervisor should also be consulted for advice.

Finally, due to the global COVID-19 pandemic, there is the potential risk of infection to myself, supervisor or other university peers. The best course of prevention is to eliminate the possibility of transmitting the virus. Therefore, all work on this project will be carried out at home. Potential face-to-face meetings will be replaced with virtual meetings through platforms such as zoom. In the rare case where I do find myself at university or if it is absolutely necessary to enter campus, I will ensure to wear a mask and sanitize regularly.

3.5.2 Technical

There will always be risks and uncertainties involved in any stage of an ML or DL project. Specifically, in the case of reliability, it is possible that the problem is misdiagnosed, and the cause/s are not clearly or correctly identified. In such cases, we would be attempting to develop a solution or optimise current algorithms

against problems that simply don't exist. Therefore, it is essential to look deeply into problems to gain a good understanding of problems.

During the experimentation stage, there are risks that the dataset is already biased which would be caused by an overrepresentation of one party/feature over the other. It is essential to understand the deployment environment, consider all variables and balance accordingly. [13]

Other risks when it comes to data are false data being used in training, or errors in the labelling of data. This would indicate that training data is not representative of real-world data.

There is also the risk of improper training of the model which could lead to over or underfitting of the training dataset. This could be the result of insufficient knowledge and understanding of training algorithms. It is critical to carefully read into all resources to limit misunderstanding.

The most important uncertainty of this project is the effectiveness of proposed solutions and algorithms using real-world data. It would be inefficient to collect real-world data and perform experiments within the given timeframe of this project. Consequently, all results and conclusions will be drawn from the dataset used in simulations. There is the fair chance a dataset shift could be observed if our models are to be used against real-world data [14].

There are also social threats in the form of evasion attacks or adversarial inputs that can have an effect on ML models. Adversarial inputs involve an external party or attacker who has the ability to manipulate data may do so with malicious intent. This would result in the model making incorrect predications often with high confidence [14]. A remedy to this is adversarial training where adversarial examples are created and used during training of the model. Although effective, it is impossible to protect against all evasion attacks as hackers continue to grow in creativity in terms of hacking ability. However, this is an insignificant risk and only has a slim chance of affecting this project.

3.6 Communication Plan

The following communication plan has been developed between myself and my academic supervisor Diep Nguyen. I am planning to maintain contact with Diep on a fortnightly basis on various project-related matters. I have also planned online meetings to discuss revisions for major milestones. No other stakeholders are involved in this project.

Subject	Channel	Purpose	Ideal Time	Frequency	General Notes
Fortnightly Email updates	Email	Various project topics	9am-5pm Monday-Friday	Fortnightly	
Supervision Meeting 1	Online Meeting/Email	Discuss Data Labelling & Datashift milestone	19th Feb 2021	Single Meeting	Agreement on experiments, results, and findings/solutions
Supervision Meeting 2	Online Meeting/Email	Discuss Algorithm Suitability milestone	19th Mar 2021	Single Meeting	Agreement on experiments, results, and findings/solutions
Supervision Meeting 3	Online Meeting/Email	Discuss Reinforcement Learning-Reward Hacking milestone	23rd Apr 2021	Single Meeting	Agreement on experiments, results, and findings/solutions
Supervision Meeting 4	Online Meeting/Email	Receive Feedback on draft report	3rd-7th May 2021	Single Meeting	Discuss changes to be made
Supervision Meeting 5	Online Meeting/Email	Receive feedback on amendments to initial draft	17th May 2021	Single Meeting	Finalise report
Final Report submission	Email	Get consent from supervisor	31st May 2021	Single Meeting	Get supervisor signature

Table 3.1: Communiation Plan

Progress Statement

Coming into this project I had little to no real and/or valuable experience or understanding of Artificial Intelligence or Machine Learning. I have spent a respectable amount of time throughout this semester researching and learning not only the basic concepts but also begun looking into factors of unreliability and thinking of a suitable methodology for the capstone project. As the objective started to become clearer, and with the guidance of my supervisor, I have made changes to initial thought processes and potential ideas for this project. I believe the methodology I have prepared and the goals I have set are realistic and achievable in terms of time frame and ability. While it is easy to get carried away with the excitement and intriguingness of AI/ML, it is important to propose goals that are feasible.

I spent quite some time researching about various platforms to perform these experiments on. Some options were, TensorFlow, PyTorch, Keras, and MATLAB. Since I had very little experience with ML as a whole, I had no skill advantage with either of these, meaning I would need to learn the basic concepts either way. While I have experience in both Python and MATLAB, based on community opinions on forums I have chosen to go with PyTorch as the availability of information and support online is far greater than MATLAB.

This project will run in accordance with the schedule prepared in Appendix 2. The first activities in the proposed schedule do not start until January of 2021. Having done a significant amount of research during this semester already, the first stages of the report have fundamentally been completed and will only require small modifications. This leaves more time for the experimentation stages of this project.

My productivity has been limited over this semester as a consequence of a larger than expected study load. While I undertook four subjects this semester, next semester I will only undertake two subjects including this capstone. As a result, the majority of my time will be spent on this project, therefore I anticipate an increase in productivity and may even find myself ahead of schedule. This would give me additional time to consult my supervisor to increase the quality of my project.

References

[1] Q. Chen, W. Wang, F. Wu, S. De, R. Wang, B. Zhang, and X. Huang, "A survey on an emerging area: Deep learning for smart city data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 5, pp. 4392–410, 2019.

[2] B. Mohapatra, "Machine learning applications to smart city," *ACCENTS Transactions on Image Processing and Computer Vision*, vol. 5, pp. 1–6, 02 2019.

[3] D. Luckey, H. Fritz, D. Legatiuk, K. Dragos, and K. Smarsly, "Artificial intelligence techniques for smart city applications," 08 2020.

[4] I. Saif and B. Ammanath, "'trustworthy ai' is a framework to help manage unique risk.." MIT Technology Reviw, 2020.

[5] R. Moutafis, "How bad facial recognition software gets black people arrested.." towardsdatascience, 2020.

[6] B. E, L. R. Flaih, D. Yuvaraj, S. K, A. Jayanthiladevi, and T. S. Kumar, "Use case of artificial intelligence in machine learning manufacturing 4.0," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pp. 656–659, 2019.

[7] S. Saria and A. Subbaswamy, "Tutorial: Safe and reliable machine learning." ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019)., 2019.

[8] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," Jul 25 2016.

[9] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," Jan 15 2019.

[10] D. Jedamski, "Bias/variance tradeoff - applied machine learning: Foundations. linkedin learning.." Available at https://www.linkedin.com/learning/appliedmachine-learning-foundations/bias-variance-tradeoff?u=2129308, 2019.

[11] S. Khan, "Understanding machine learning methodology." Available at http://justsajid.com/skills/datascience/understanding-machine-learningmethodology/, 2020.

[12] P. Migdal, "Keras or pytorch as your first deep learning framework.." Available at https://deepsense.ai/keras-or-pytorch/, 2020.

[13] E. Sires, "9 ways to prevent data bias in predictive models.," 2020.

[14] R. Gupta, "Machine learning security: 3 risks to be aware of. techcentre.." Available at https://www.plugandplaytechcenter.com/resources/machine-learning-security-3-risks-be-aware/, n.d.

Appendix

5.1 Appendix 1: Communication Log

Project Title:	**Engineering Education and Sustainable Development in Australia**		
Student Name:	**Ajal Singh**	**Supervisor Name:**	**Diep Nguyen**
Date	**Event**	**Topic of Communication**	**Outcome**
22/07/2020	Email	Introduction and Background knowledge	Supervisor highlighted importance of 'reliability' within project, and provided introductory reading material
29/07/2020	Email	Defining a research question/follow-up questions	Clear direction of project set.
7/08/2020	Email	Question on sources of unreliable AI/ML	Supervisor suggested I look at various articles on applications of AI. Recommended Google Scholar
18/08/2020	Email	Problem Analysis Draft Feedback	Feedback: More depth required about mathematical concepts
19/08/2020	Email	Sample Report, more resources	Sample report from previous semester provided to better understand expectations. Also recommended to visit technical papers from IEEE and ACM
11/09/2020	Email	Obtaining data for experimentation	Supervisor recommended searching google. Also highlighted the importance of maths in experiments
14/09/2020	Email	Questions on a specific paper	Supervisor did not like paper. Recommended searching for papers from good research groups
2/10/2020	Email	EHS & Ethics Forms	Supervisor approved and signed forms
12/10/2020	Email	Draft reseach proposal v 1.0	Supervisor advised using IEEE referencing format. Also enouraged me to look at wider applications of AI.
19/10/2020	Email	Draft research proposal v 1.1	Awaiting feedback

Table 5.1: Communication Log

5.2 Appendix 2: Gantt Chart & Milestones

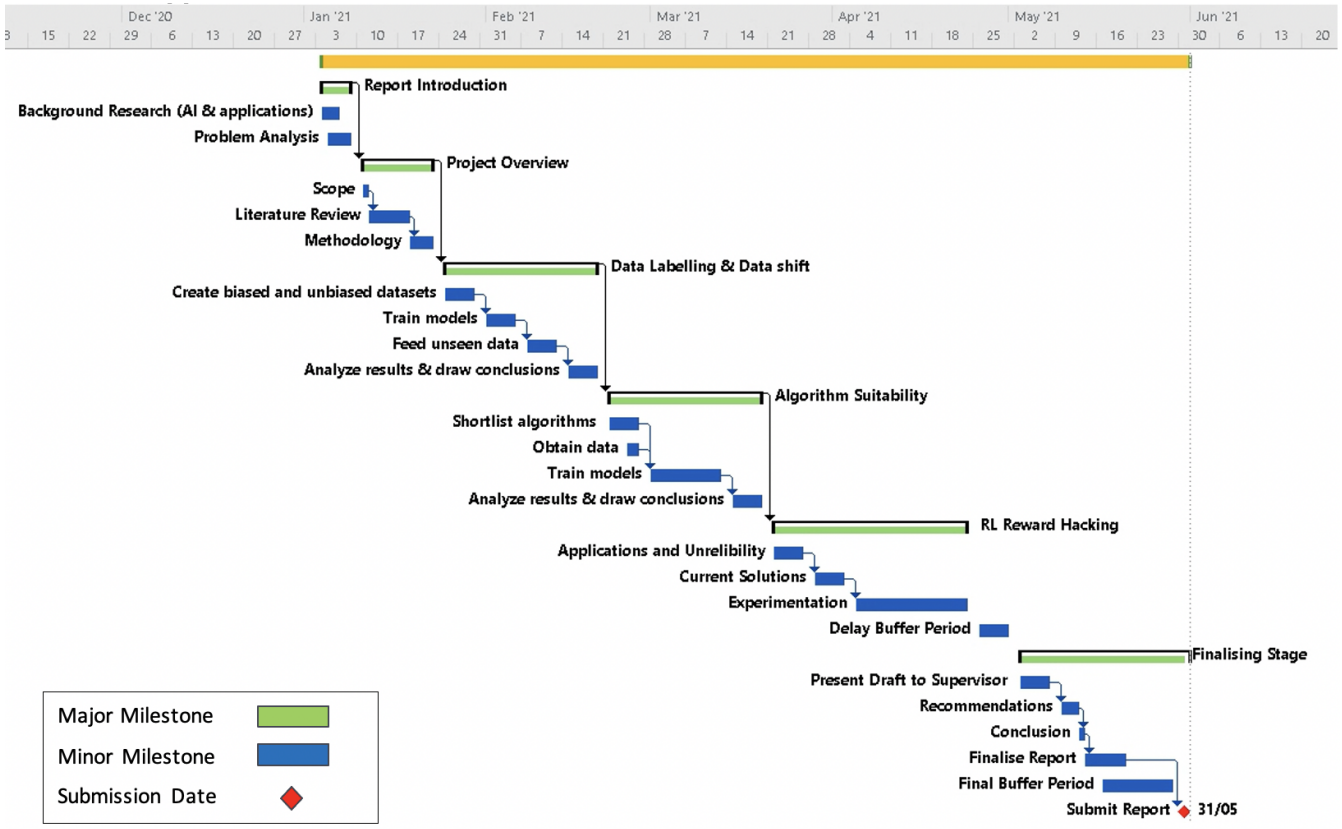

Figure 5.1: Gantt Chart

	Task Name	Duration	Start	Finish	Predecessors
1	Capstone Project	106 days	Mon 4/01/21	Mon 31/05/21	
2	**Report Introduction**	**5 days**	**Mon 4/01/21**	**Fri 8/01/21**	
3	Background Research (AI & applications)	3 days	Mon 4/01/21	Wed 6/01/21	
4	Problem Analysis	4 days	Tue 5/01/21	Fri 8/01/21	
5	**Project Overview**	**10 days**	**Mon 11/01/21**	**Fri 22/01/21**	**2**
6	Scope	1 day	Mon 11/01/21	Mon 11/01/21	
7	Literature Review	5 days	Tue 12/01/21	Mon 18/01/21	6
8	Methodology	4 days	Tue 19/01/21	Fri 22/01/21	7
9	**Data Labelling & Data shift**	**20 days**	**Mon 25/01/21**	**Fri 19/02/21**	**5**
10	Create biased and unbiased datasets	5 days	Mon 25/01/21	Fri 29/01/21	
11	Train models	5 days	Mon 1/02/21	Fri 5/02/21	10
12	Feed unseen data	5 days	Mon 8/02/21	Fri 12/02/21	11
13	Analyze results & draw conclusions	5 days	Mon 15/02/21	Fri 19/02/21	12
14	**Algorithm Suitability**	**20 days**	**Mon 22/02/21**	**Fri 19/03/21**	**9**
15	Shortlist algorithms	5 days	Mon 22/02/21	Fri 26/02/21	
16	Obtain data	2 days	Thu 25/02/21	Fri 26/02/21	
17	Train models	10 days	Mon 1/03/21	Fri 12/03/21	16,15
18	Analyze results & draw conclusions	5 days	Mon 15/03/21	Fri 19/03/21	17
19	**RL Reward Hacking**	**25 days**	**Mon 22/03/21**	**Fri 23/04/21**	**14**
20	Applications and Unrelibility	5 days	Mon 22/03/21	Fri 26/03/21	
21	Current Solutions	5 days	Mon 29/03/21	Fri 2/04/21	20
22	Experimentation	15 days	Mon 5/04/21	Fri 23/04/21	21
23	Delay Buffer Period	5 days	Mon 26/04/21	Fri 30/04/21	
24	**Finalising Stage**	**21 days**	**Mon 3/05/21**	**Mon 31/05/21**	
25	Present Draft to Supervisor	5 days	Mon 3/05/21	Fri 7/05/21	
26	Recommendations	3 days	Mon 10/05/21	Wed 12/05/21	25
27	Conclusion	1 day	Thu 13/05/21	Thu 13/05/21	26
28	Finalise Report	5 days	Fri 14/05/21	Thu 20/05/21	27
29	Final Buffer Period	10 days	Mon 17/05/21	Fri 28/05/21	
30	Submit Report	0 days	Mon 31/05/21	Mon 31/05/21	28

Major Milestone

Minor Milestone

Table 5.2: Project Schedule

5.3 Appendix 3: Risk Assessment

Risk	Consequence	Severity	Control
Failure to complete deliverables to schedule	Project delay	Extreme	Planned buffer periods
Inadequate understanding of critical components	Delays and/or poor project quality	Extreme	adhere to the schedule, revisit reliable sources and consult my supervisor as soon as possible
Data Loss	Loss of all work	Moderate	Use cloud storage
Experimenting with Biased data	Unreliable results due to overrepresentation of one party	Moderate	essential to understand the deployment environment, consider all variables and balance accordingly.
Uncertainty of ideas and general direction of the research	Poor quality of work	Low	Visit past student reports and consult supervisor
COVID-19 infection	Sickness and possible transmission to others	Low	Work from home, virtual meetings. Wear mask and sanitize if university visit is absolutely necessary

Table 5.3: Risk Assessment