NTU Institute of Science and Technology for Humanity (NISTH) Ideas Challenge: AI for Humanity

Research Proposal: To investigate how severity of scenarios affect human preferences towards rule-based and randomised automated decisions

Prepared By:

Lim Xuan Yu

Arnold Toh

Sit Han Zhe

Sim Zhi Qi

1. Background and Problem Statement

1.1. The race towards driverless roads

In the race towards a technologically-advanced future, nations are striving towards deploying Autonomous Vehicles (AVs) as the transport of the future. According to the KPMG Autonomous Vehicles Readiness Index [1], Singapore ranks second across 20 other competing countries. In February 2017, Singapore introduced "AV rules" to govern the progress of AVs, which indirectly exempts AVs from all the typical regulations under the Road Traffic Act. [2] This legislation, coupled with Singapore's advanced AV technology, is the reason for Singapore's ranking. Aside from that, Singapore is also trying to be the first to deploy driverless public transport. [3] Singapore and many other countries are undeniably racing to develop and deploy AVs so as to tap upon the potential benefits that AVs can provide.

1.2. Machine ethics in the AV industry

However, there are still obstacles to overcome and issues to be addressed. According to an article in Forbes [4], ethics is one of the main hurdles to be crossed for AVs to be implemented. In the event of an unforeseeable, unavoidable accident, an AV is forced to choose how it allocates risks — deciding between different parties to injure. AVs have to overcome these dilemmas, which prove to be difficult, because such "significant value-based consequences-decisions" have to be automatically managed by Artificial Moral Agents (AMA) [1].

Lately, there has also been an astounding number of research papers published on machine ethics and its implications, because it is directly related to another core issue: liability. For AVs to operate efficiently and reliably on the roads, someone has to be liable for the decisions that AVs make and this directs the spotlight to the machine ethics coded into the AVs' algorithm. Thus, building a socially-acceptable machine ethics model for AVs is of paramount importance.

1.3. The complexities in developing a common basis

However, there is currently no socially acceptable basis for decision-making due to the complexities involved. Experimental ethics surveys have been conducted, where AVs have to choose between risking the lives of people with different identities. In such scenarios, many people accept utilitarian-based automated decision [5]. In particular, they supported the use of age [6] and group size [7] as parameters to be considered for the decision-making process. Despite apparent support, employing utilitarian models in machines still raises public concern, as this means that drivers might have to sacrifice their lives for a greater good [7].

This has led to studies investigating whether people would accept randomised AV decisions. Wintersberger et al [8] have classified people into Randomists (those who support decisions made at random) and Reasonists (those who support reasoned decisions). They showed that people behave like randomists or reasonists, depending on the severity of the scenario. This suggests further studies can be conducted to look into the possibility of developing a hybrid moral machine that combines both a randomist and reasonist approach.

2. Aims

Through this project, we aim to determine the socially acceptable threshold of severity where people would distinctively prefer autonomous vehicles to make rule-based decisions instead of randomised ones. This will serve as a basic ethical guideline for future development in autonomous vehicles.

We will conduct a scenario-based questionnaire to determine the intrinsic values that respondents place on adults and children. The difference in intrinsic values of an adult's and a child's life will provide insights for the importance of age as a factor in the intrinsic value of a life. In addition, this insight will help tackle the central question of how different groups are valued and whether the decision to save them should be rule-based or random. This can be done by comparing groups with different number of adults and children.

Specifically, we aim to tackle the following research questions:

1. How would the size of the groups at risk affect the acceptance towards the type of decision (random or rule-based)?
2. How do people compare intrinsic values of an adult's life to a child's life?

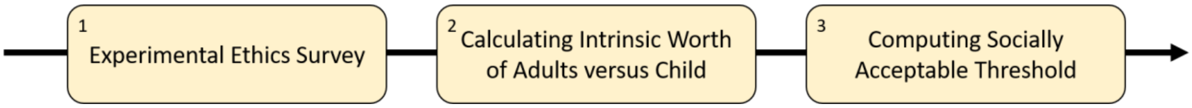
3. Literature Review

For our proposed study, we first draw on Bonnefon et al [7]. To determine the effectiveness of a Utilitarian model in AVs, they conducted an experimental ethics survey. In six given scenarios, participants were to choose between two groups of people to sacrifice, involving both passengers and pedestrians. We will adopt the same experimental design. We will modify it by adding a question to each scenario regarding their acceptance towards randomised and reasoned decisions.

We then draw on Pugnetti et al [9] who also based their work on [7] but surveyed the preferences of Swiss between simplified scenarios where either passengers or pedestrians will be killed in the accident.

They then used the results in a 4-step calculation to derive the implicit life values of an adult and a child across different scenarios. We will draw on this approach to calculate the intrinsic value of the lives of people and use the results to generate a spectrum of scenarios with varying severity scores. The use of a wider spectrum will provide us with a more comprehensive analysis.

4. Methods


1
Experimental Ethics Survey
2
Calculating Intrinsic Worth of Adults versus Child
3
Computing Socially Acceptable Threshold

Methodology Overview

4.1. Experimental Ethics Survey

To address our two research questions, we will set up a forced-choice ethics survey, based on Bonnefon et al's framework.

4.1.1. Sampling Population

We limit the scope of this study to Singapore's context. To ensure result accuracy and reproducibility, we will conduct stratified random sampling based on ethnicity to obtain a sample size of 2,000. The following table shows the targeted breakdown of survey respondents:

Ethnic Group	Percentage of population	Sample size
Chinese	74.3	1,486
Malay	13.4	268
Indian	9.03	181
Others	3.22	65

Source: Singapore Statistics

4.1.2. Questionnaire Design

Participants will fill in a questionnaire to rate their level of acceptance towards randomised decisions and rule-based automated decisions for different scenarios by assigning a maximum 10 points to both options. (For instance, if a participant's level of acceptance towards a random decision is 7, the level of acceptance towards a reasoned decision will be 3.)

Example of survey question:

Scenario 1: 1 Adult vs 1 Child	
Who would you rather save? *	
Decision outcome	**Acceptance Score (Sum of 10 points)**
Rule-based (utilitarian)	
Random	

(*) This question only applies for specific scenarios

The questionnaire will include two spectra of 10 scenarios with mirrored severity scores. For every scenario, it is assumed that the AV will not be able to stop in time and a sacrifice has to be made. The scenarios for each spectrum were chosen to reflect increasing severity in terms of the value of lives lost.

For questions which only compares the intrinsic values of only adults to only children, an additional question (*) will be added to determine the preferences of respondents towards saving an adult or a child. The corresponding results will then be used to calculate the intrinsic value of the lives of an adult and child.

The scenarios are listed below:

Spectrum 1

Scenario	Severity Score
1 Adult vs 1 Adult	0
1 Adult vs 1 Child *	2
1 Adult vs 1 Adult 1 Child	3
1 Adult vs 2 Child *	5
1 Adult vs 1 Adult 2 Children	6
1 Adult vs 3 Children *	8
1 Adult vs 1 Adult 3 Children	9
1 Adult vs 4 Children *	11
1 Adult vs 1 Adult 4 Children	12
1 Adult vs 5 Children *	14

* Scenarios with additional question

Spectrum 2

Scenario	Severity Score
1 Child vs 1 Child	0
2 Adult vs 1 Adult 1 Child	2
2 Adult 1 Child vs 4 Children	3
2 Adult 2 Children vs 1 Adult 4 Children	5
3 Adult 1 Child vs 1 Adult 3 Children	6
3 Adult 3 Children vs 2 Adult 6 Children	8
3 Adult 5 Children vs 9 Children	9
4 Adult 1 Child vs 6 Children	11
4 Adult 2 Children vs 1 Adult 7 Children	12
4 Adult 4 Children vs 10 Children	14

Attention Check

How many pedestrians are there in this scenario?

We draw on Bonnefon et al's research to include an easy question at the end as an attention check. Respondents who fail this check will be discarded from analyses.

4.2. Calculating Intrinsic Values Using Survey Responses

As the results from Pugnetti et al may not be reproducible [9], we will use the questionnaire results(from Section 4.1) to perform calculations of intrinsic worth of an adult's and a child's life in Singapore's context.

Spectrum 1 serves as a control, with 1 adult always on one side of the comparison. Spectrum 2 provides a more even comparison, with the total number of lives compared not differing by 2. After incorporating the intrinsic values of an adult's and child's life of Singaporeans from the first survey, Spectrum 2 will offer a wider spread of severity scores.

We adapt Pugenetti et al's calculations of the intrinsic value of an adult's and a child's life. This is done through the following steps:

1. Let the percentage of respondents who would save the child be x%. Take the intrinsic value of 1 adult's life as 1.
2. Based on Spectrum 1 scenarios, the intrinsic value of a child's/children's life would be x/(100-x).
3. For scenarios with more children, divide the total intrinsic value by the number of children in that scenario to obtain the average intrinsic value of each child's life.
4. Take the average intrinsic value of a child across the different scenarios as the final intrinsic value of a child for section 4.3.4. Calculations.

4.3. Computing Socially Acceptable Threshold

Within each spectrum, we will compute the survey results to determine the optimal level of severity for the algorithm employed to switch between rule-based decisions and randomised decisions. Using the pre-calculated severity scores and our own survey results, we can determine the socially acceptable threshold in the following manner:

4.3.1. Classifying scenarios within each spectrum

To classify scenarios, we will:

1. Sum up the acceptance scores for rule-based decision and randomised decision for Spectrum 1.
2. Conduct Mann Whitney U-test to see if the difference between the scores is statistically significant.
3. Classify scenarios into two categories -- scenarios in which rule-based decisions are preferred and scenarios in which randomised scenarios are preferred.
4. Repeat for Spectrum 2.

4.3.2. Identifying threshold within each spectrum

To identify the threshold, we will:

1. Plot graphs of acceptance scores against pre-assigned severity scores for both rule-based and randomised decisions for scenarios in Spectrum 1.
2. Identify the threshold as the severity score at which the graphs intersect.
3. Repeat for Spectrum 2.

4.3.3. Comparing across the two spectra

The second part of the analysis involves comparison across the two spectra. The difference in the acceptance scores will provide insights into the perception of Singaporeans towards the intrinsic value of an adult and that of a child. This will serve as a preliminary basis for a correlation between age and intrinsic value.

4.3.4. Contextualising data collected

Spectra 1 and 2 were initially taken to given discrete mirroring severity scores. By re-calculating the actual severity scores using the intrinsic values calculated (from Section 4.2), the severity scores will have a greater spread that also reflects the trend for preference of randomness or reason more accurately in Singapore's context.

By repeating the data processing used in Section 4.3.2 and 4.3.3, we will be able to generate a more accurate comparison over the 2 spectra as the severity scores will no longer be discrete nor mirrored.

To determine the disparity between the results from spectrum 1 and 2, we will:

1. Plot the data from both spectra on the same axes
2. Use the method of least squares for linear regression, to calculate the Correlation Coefficient (R).
3. The value of R will represent the consistency of respondents' sentiments across the scenarios in both spectra. (As R reflects the disparity of points to a linear line, the closer R is to 1, the more consistent the model)

5. Expected Results

5.1. Hypothesis for RQ1

We expect survey respondents to have a preference for reasoned decisions as level of severity increases.

5.2. Hypothesis for RQ2

We hypothesise that survey respondents' acceptance towards reasoned decision will surpass that of random decisions at a higher threshold for Spectrum 2, as compared to Spectrum 1. This is because we expect Singaporeans to value a child's life to be less than 3 which was derived from Pugnetti et al's study. Hence, we postulate that age becomes a less determining factor when it comes to valuing multiple lives.

6. Deliverables and Future Implications

6.1. The immediate impact of a socially acceptable threshold

Our main deliverable will be the identification of the threshold where people prefer autonomous vehicles to make rule-based decisions instead of randomised ones. By mitigating potential dissatisfaction that AVs' decisions could cause, our project will increase the public's acceptance towards AV testing and launching.

One example of how AV companies can make use of our result is as follows: create an ethical model that follows the principle of justice in situations that fall below the threshold, but follows the principle of Utilitarianism in situations that surpasses the threshold. In a situation that falls below the threshold, two potential victims, for example, should be treated equally and a random decision generated by an AV would be deemed fair. In a situation that surpasses the threshold, the AV should make decisions that produce the greatest good.

6.2. Breaking down barriers that hinders forward-looking AV policies and legislations

Based on a KPMG study on Autonomous Vehicle Readiness Index, Singapore is ranked first in the world in terms of AV policies and legislations [1]. One hindrance to the implementation of autonomous agents is the lack of an ethical basis for machines to make automated decisions that would be widely acceptable. With our model, we are confident that societal acceptance will increase and will enable governments to introduce new legislations without compromising the public's opinion on the issue.

While the public’s apprehension towards AV implementation may not be of utmost concern for the Singapore government, other countries face difficulties in convincing the public that AVs can be ethically-reasonable. For instance, Germany ranks 3rd in technology and innovation, but ranks 12th in terms of consumer preference and 5th in legislation and policy created for AVs [1]. With our project deliverable, we can better justify AV decisions and increase consumer preference. Consequently, this will potentially lower the barriers for more forward-looking legislations to be implemented. In all, we can make countries like Germany a more conducive place for AVs to be developed in.

6.3. An increase in the efficiency of the transport system in Singapore

The need for a robust transport system is becoming a pressing issue in Singapore. Spatial constraints and a growing population are the two most important factors that the government considers when contemplating the human-environment condition [10].

The Singaporean government has plans to spearhead the implementation of AVs. There has been a steady increase in the number of public buses over the past decade [11]. In this regard, more efficient and regular AV public transport should replace conventional buses to keep the number of buses on roads in check. In fact, there are plans to introduce driverless buses from 2022 [12]. We foresee that our project deliverable will help AV companies configure AVs that are more accepted by Singaporeans, thus result in a smoother implementation of autonomous public transport for the benefit of our society.

7. Conclusion

By identifying the threshold for AVs to switch between making rule-based and randomised decisions, we are bringing the AV industry one step closer towards practical AV implementation. It also serves as a springboard for further research into other demographic contexts, ethical theories as well as other industries utilising artificial intelligence. Ultimately, we aim to minimise the public outcry over decisions that autonomous agents will make in future, so we can develop a socially-acceptable moral algorithm for practical implementation.

8. References

[1] KPMG International, “Assessing countries’ openness and preparedness for autonomous” vehicles“, Autonomous Vehicles Readiness Index, 2018. Accessed on: 15 Jan 2018. [Online]. Available:https://assets.kpmg/content/dam/kpmg/tw/pdf/2018/03/KPMG-Autonomous-Vehicle-Readiness-Index.pdf

[2] C. Lago, “How Singapore is driving the development of autonomous vehicles”, CIO Asia, 2018. Accessed on: 15 Jan 2018. [Online]. Available:https://www.cio-asia.com/article/3294207/innovation/how-singapore-is-driving-the-development-of-autonomous-vehicles.html

[3] K. Park, K. Chia, “Singapore Built a Dedicated Town for Self-Driving Buses”, Bloomberg, 2018. Accessed on: 15 Jan 2018. [Online]. Available:https://www.bloomberg.com/news/features/2018-06-04/singapore-built-a-town-to-test-autonomous-self-driving-vehicles

[4] D. Silver, “What Hurdles Do Self-Driving Cars Face As Waymo Gets Ready For Prime Time?”, Forbes, 2018. Accessed on: 15 Jan 2018. [Online]. Available:https://www.forbes.com/sites/davidsilver/2018/10/05/what-hurdles-do-self-driving-cars-face-as-waymo-gets-ready-for-prime-time/#5b539204d0f6

[5] J.-F. Bonnefon, A. Shariff, and I. Rahwan, “Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars?” arXiv preprint arXiv:1510.03346, 2015.

[6] A.-K. Frison, P. Wintersberger, and A. Riener, “First person trolley problem: Evaluation of drivers’ ethical decisions in a driving simulator,” in Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct. ACM, 2016, pp. 117–122.

[7] J.-F. Bonnefon, A. Shariff, and I. Rahwan, “The social dilemma of autonomous vehicles,” Science, vol. 352, no. 6293, pp. 1573–1576, 2016.

[8] P. Wintersberger, A. Frison, A. Riener, S. Thakkar, Do Moral Robots Always Fail? Investigating Human Attitudes Towards Ethical Decisions of Automated Systems. 28th IEEE International Symposium on Robot and Human Interactive Communication, 2017

[9] C. Pugnetti and R. Schläpfer, “Customer Preferences and Implicit Tradeoffs in Accident Scenarios for Self-Driving Vehicle Algorithms,” Journal of Risk and Financial Management, vol. 11, no. 2, p. 28, Apr. 2018.

[10] V. R. Savage, L. Kong, “Urban Constraints, Political Imperatives: Environmental 'Design' in Singapore”, Landscape and Urban Planning, 25(1-2), 37-52, 1993.

[11] Singapore Land Transport Authority, Annual Vehicle Statistics 2017. Accessed on: 22 Jan 2018. [Online]. Available:https://www.lta.gov.sg/content/dam/ltaweb/corp/PublicationsResearch/files/FactsandFigures/MVP01-1_MVP_by_type.pdf

[12] BBC, Singapore to use driverless buses 'from 2022', 2017. Accessed on: 22 Jan 2018. [Online]. Available: https://www.bbc.com/news/business-42090987