SIGCSE: U: Understanding Engineering Students' Ethical and Algorithmic Decision Preferences through a Consequentialist Framework

Edward Kempa
kempaedward@ufl.edu
Department of Comp. and Info. Science and Engineering
University of Florida
Gainesville, FL, United States

Ashish Aggarwal
ashishjuit@ufl.edu
Department of Engineering Education
University of Florida
Gainesville, Florida, USA

ABSTRACT

As developments in the field of artificial intelligence (AI) continue to rapidly advance its possible applications, it becomes increasingly crucial for those developing AI systems to understand how receptive the general public will be to their work. The overarching goal of this research is to understand human decision-making (HDM) and human perspectives on algorithmic decision-making based on varied payoffs and outcomes. We conducted a pair of surveys, where the participants were asked about their understanding of AI, as well as their thoughts about the application of AI in the context of an autonomous vehicle placed in an ethically challenging situation. Our analysis focuses on participants' responses to two questions characterized by experimental variations, with additional variation in the consequences presented in those same questions between the two surveys. In total, we collected 284 responses from these surveys administered to engineering students of an introductory programming course for two consecutive semesters in 2022. We qualitatively analyzed the data for individual questions using an inductive approach and identified major themes related to the question asked. From this analysis, we found that engineering students' perspectives on an ethically complex scenario were greatly impacted by the controlled variance in consequences, and we have developed a framework for tracing their decision-making to their decisions and reasoning. Offering valuable insight into engineering students perceptions regarding the applications of AI.

1 PROBLEM AND MOTIVATION

As the presence of AI continues to expand its influence in our everyday lives, we are steadily approaching a world in which we frequently see AI-driven systems making life-changing decisions. As the impact of AI grows to such extremes, it becomes increasingly important to ensure that these developments are pursued in an ethically sound manner. This issue becomes particularly relevant as the applications of AI expand into areas that have the potential to adversely affect human health and safety. An example that perfectly encapsulates this issue is the recent strides in the development of automated vehicles (AV), which is limited by our understanding of how AV ought to behave in ethically challenging situations. If AV are to become widely adopted, they will have to be trained to handle situations in which an accident cannot be avoided and the best course of action is ethically unclear. Because the way in which those developing algorithms for AV choose to resolve ethically complex scenarios will inventively be subject to public scrutiny, there has been an ongoing demand for research deepening our understanding of how people would wish for such algorithms to function. In one such study, Bonnefon et al. evaluated the moral attitudes of people regarding the correct behavior of AV in ethically complex situations, as well as whether people believe that a certain course of actions should be mandated morally and or legally [4]. In another work entitled *Computational ethics*, Awad et al. emphasize the importance of using human moral decision making to inform automated decision making frameworks [2]. This study seeks to address exactly this issue, by expanding our understanding of engineering students' perceptions of AI in applications like AV, through the development of a decision-making framework that measurably ties their decisions to their reasoning and vice versa. We focus on engineering students in particular because they represent a group of people who are very likely to be working with AI systems in the future. Additionally, we hope that the decision making framework we produce through this study could be used to inform those of automated system.

2 BACKGROUND AND RELATED WORK

A classic example of an ethically challenging situation is that of the trolley problem, in which an individual is placed in front of a lever that changes the tracks of an approaching unstoppable trolley [7]. Trapped on each of the two tracks are people, which may vary in number and or demographics, and the individual is forced to select the path taken by the trolley. Existing studies have applied the essence of the trolley problem to AV and investigated peoples' responses to a variety of different circumstances. One of the most notable of these studies was the Moral Machine experiment (MME), where Awad et al. conducted a global online survey, where participants were given images of hypothetical scenarios where an AV cannot avoid an accident and participants answered regarding their preferences. A mix of 40 million decisions were evaluated across nine dimensions, and through the use of Geolocation they were able to identify clusters of countries of which the population appeared to share similar preferences. They found that that there is no universal set of correct decisions. Instead, people's views regarding differing situations could be tied back to their socio-economic and cultural circumstances [6]. While the MME is a very valuable resource, some of the notions proposed by the MME, including the presumption that people wish for AV to treat different people in different ways, have been challenged in a response by Bigman and Grey within which they reference other research which establishes the idea that people would want AV to treat people equally [5]. Demonstrating the need for further research on this issue. This study serves to follow up upon the work of Aggarwal and Ranjan and their investigation of undergraduate students' reasons about ethical and

algorithmic decision-making, within which they extracted the decisions and reasoning of students about a hypothetical ethically complex scenario which we adopt in our methodology. [1]. We deviate from their work by approaching the problem through the lens of a consequentialist framework. This is done through an experimental approach where we analyze the effects of varying the consequences communicated to the participants from one survey to another, discussed in further detail in our approach. Additionally, we fill a gap that the MME could not feasibly address at its large scale, by openly prompting each of our survey participants to give their reasoning, to then be qualitatively analyzed. We recognize that peoples thoughts regarding the use of AI driven systems have been studied before. In a study, Bigman and Gray found that people are generally uncomfortable with the idea of machines making moral decisions, even when the outcomes are positive. This aversion is largely due to the perception that machines lack the ability to think and feel in the same way humans do, which is seen as crucial for making moral judgments [3]. We hope that the results of this study will either support or contradict with these findings.

3 APPROACH AND UNIQUENESS

Data was collected for this experiment through a set of two surveys sent out in 2022 to two consecutive semesters of a CS1 class for engineering students at an R1 university in the southeast of the United States. Students completed the survey as part of a graded submission administered in the last week of their semester in the course. Each survey had two questions, each question with a portion of choosing between three or four options, and then a free-response portion where participants elaborated on their reasoning. We collected a total of 284 responses, 135 from *Survey 1* and 117 from *Survey 2*. Of the participants there were 132 (52.4%) Mechanical and/or Aerospace, 43 (17.1%) Civil and/or Environmental, 44 (17.5%) Biomedical, and 33 (13.1%) other engineering majors represented. Notably, there were only two Computer Science majors across the surveys. The gender distribution was as follows, there were 169 (67.1%) males, 83 (32.5%) females, and 1 other. Furthermore, out of the participants there were 57 (22.7%) Freshmen, 179 (71.3%) Sophomores, 10 (4.0%) Juniors, and 5 (2.0%) Seniors. These proportions were very consistent between the individual surveys.

The first question involved a hypothetical situation, where the participant is driving a car at 70 miles per hour and the brakes fail (Figure: 1). The participant can either *continue straight and hit a truck* in front of them, *swerve into a helmeted 30-year-old motorcyclist on their right*, or *swerve into a motorcyclist without a helmet of the same age on their left*. The participant first answered which choice they would make, and were then given a text box and asked to offer **two specific reasons** (primary and secondary) for making that decision, in order of importance, in an open ended manner. The second question presents the same hypothetical situation, with a difference in that it is not longer the participant driving the car, instead the participant now takes on the role of an **intern at a tech company tasked with programming the algorithm of an AV** which is to make the decision for a human in the car. Additionally, on top of the three maneuvers the car could perform stated earlier, now a **fourth option** to *have the algorithm take no decision and defer the decision* to the human driver is offered. Once again, after offering the decision they believed the algorithm should make, the respondents were asked to provide two specific reasons. *Survey 1* and *Survey 2* perfectly parallel one another except for the fact that ***Survey 1* specified that in each option, the party that bears the main impact has a 100% chance of dying**. In other words, the passenger would die if the first option (going straight into the truck) is chosen, but no one else. If either motorcyclist was chosen to be hit, that motorcyclist was guaranteed to die, but no one else would. This is altered slightly for **Survey 2, where the helmeted motorcyclist on the right instead has only an 80% chance of dying**, with the rest unchanged.

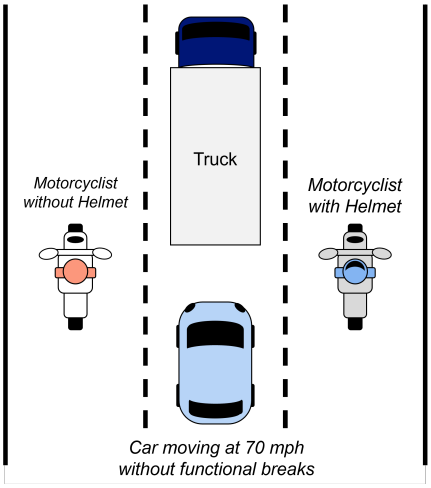

Figure 1: Situation Presented to Survey Participants

From there, we performed a qualitative analysis of the responses, assigning Level 1 (L1) and Level 2 (L2) codes the primary reasoning for participant decision-making with regards to autonomous vehicles. L1 codes consisted of the direct reasoning behind each individual's response, while L2 codes were more generalized themes that were apparent within the data. For instance, a response like *"there should be an algorithm that makes a suggestion, but ultimately the driver should make the decision"* would be given the L1 code *"human decision making over algorithms"* and the L2 code ***Driver Should Have the Final Say Over What Happens***. On the other hand, a response such as *"The motorcyclist IS wearing a helmet, giving him a better chance of surviving than the one on the left who isn't wearing one."* would receive an L1 code of *"helmet increases chance of survival"*, corresponding to an L2 code of ***Survival Rates***.

To ensure the accuracy and consistency of the coding process, inter-rater reliability was conducted. After an initial codification of each question for each survey, a separate researcher codified 20% of the same responses using the generated L2 codes and then calculated the degree of agreement between their encoding. If for any particular question the level of agreement fell below 80%, the L2 codification for that particular question was entirely redone.

The novelty of our approach stems from the qualitative analysis which was then conducted on the counts of each of the different L2 codes to determine not only the **probabilities of decisions and primary reasons**, but also **conditional probabilities linking the two**. Thus developing a framework from which it could be

reliably inferred how likely a student was to make a decision given they held a particular way of reasoning, or how likely they were to have a particular reason given their decision.

4 RESULTS AND CONTRIBUTIONS

The results of this study are best summarized in the two Sankey diagrams (**Figure: 2** and **Figure: 3**) above, which visualize the flow of the participants decisions and reasoning. To be clear, we would like to emphasize that the diagrams illustrate **only the primary reasons** participants offered. The secondary reasons were not the focus of this analysis, though they will be briefly addressed in parts. These figures encapsulate the decision making framework we sought to create, and we will now briefly discuss how to read the Sankey diagrams.

The heights of the leftmost bars represent the number of people who chose each of the three options made available in the first question of the survey, where the human driver was in control. The second set of bars represent the primary reasons offered by participants for their decision in that same first question. The branches connecting these two represent the number of people who selected that choice and then went on to give that particular reason to which the branch is connected. The next set of branches connect the primary reason given by a participant in the first question to the option they chose in the second question, where they were tasked with programming an algorithm to make the decision. Lastly, the final set of branches, similar to the first, connects the option selected for the second question to the primary reason offered for making that choice.

Beginning our discussion of these with **Figure: 2**, the results of the *Survey 1* where the death of the afflicted party was explicitly guaranteed, we find that a majority (53.3%) of participants chose to go straight and hit the truck, with the option to swerve left and hit the motorcyclist without the helmet being the next most chosen option at 28.1% of participants. Though the difference in the number of people who chose to swerve left as opposed to right was not very significant. Represented in the Sankey diagram, we see that the qualitative analysis of participants reasoning yielded five major themes, with the rest being captured by *Other*. For people who chose to go straight, they were most likely to offer ***Save Others*** as their primary reason, the theme which represents a self sacrificing ideal of preserving the lives of others over one's own. Participants which chose to go straight also commonly cited ***Fault/Legal Consequences*** or ***Emotional Consequences*** as primary reasons. The first of these justifies the behavior chosen by claiming they or the other party would be at fault or legally responsible for the outcome, this description is not exclusively people claiming the human driver would be at fault if they swerved, as some people who chose to swerve left instead found the motorcyclist without a helmet to be breaking the law and thereby believed they as the human driver would not be liable for the accident. Emotional consequences are largely self explanatory, being seen through people saying that they could not live with the guilt, or that their family couldn't live without them, etc. As for the people who chose to swerve, either left or right, they almost always cited the theme of ***Self-Preservation*** as their primary reason, which is the idea that they value their own life above all else or that it would be human instinct to preserve oneself by swerving. The last major theme of ***Survival Rates***, which was cited by people having chosen any of the three options, and is the general idea that a party has a better chance of survival be it because they are in a car or have a helmet. This reasoning often ended up being contradictory to the explicitly stated consequences that participants were provided, but becomes much more significant in the discussion of *Survey 2*. Lastly, the *Other* themes included ideas of free will or religion amongst more unique answers. Something else which was interesting but is not represented in the Sankey diagram, nor in the *Other* themes, because it was only seen given as a secondary reason, was an argument of ***Age***. Consisting of claims that the participants life was worth more than that of a middle aged motorcyclist, and therefore a decision to swerve was made.

Moving on to the responses for the second question, still in *Survey 1*, which offered participants a fourth option to defer the decision to the human in the car, we see that by a large majority (80.7%) the participants chose that fourth option. On the other hand, the other three options about equally make up the rest of the choices. Participants who chose to defer were most likely to cite that the ***Driver Should Have the Final Say Over What Happens***, though the distribution of these reasons were more even. While this first reason emphasizes the need for the human in the car to be the one deciding their own fate, the theme of ***Algorithms Should Not Decide Who Lives and Dies*** represents the line of reasoning that an algorithm should not be allowed to make a life or death decision. ***Distrust of Algorithms Capabilities*** encapsulates responses that do not believe that an algorithm is capable to making moral choices, correctly understanding the circumstances, or making the correct choice in such an morally challenged scenario. The theme of ***Allows Company to Avoid Legal Consequences is largely self explanatory***, and ***Programmer Should Not Decide a Choice for Everyone*** is understood as the idea that the programmer (the participant here) would not want to live with the guilt or does not believe that they should make a choice that will be applied for everyone. As for participants who had the algorithm swerve either right or left, they almost always cited a reason that ***Algorithms Should Protect the Passenger at Above All***. Some *Other* reasons included the idea that the algorithm should protect others on the road, which was mostly cited by people who would have the algorithm go straight, as well as niche reasons commenting about the motorcyclist without the helmet being in the wrong.

We now move to **Figure: 3**, illustrating the results of *Survey 2* where the consequences of the motorcyclist with a helmet were altered. Unlike in the first survey, we instead find a majority (**54.7%**) of participants choosing to swerve right, into the motorcyclist with the helmet. Going straight still remains a popular option at **35.9%** of responses, and swerving left which was more popular than swerving right in *Survey 1*, was rarely chosen. For people who chose swerve right, they were about equally likely to cite ***Self-Preservation*** or ***Survival Rates*** as their primary reason. Although, what stood out, was that the people who chose either of these two as their primary reason, almost always cited the other as their secondary reason. For participants that chose to go straight, they were still most likely to cite the theme of ***Save Others***, and also commonly cited ***Fault/Legal Consequences*** or ***Emotional Consequences*** as their primary reason. Interestingly, the only drastic change in the number of times that each major theme was cited, compared to

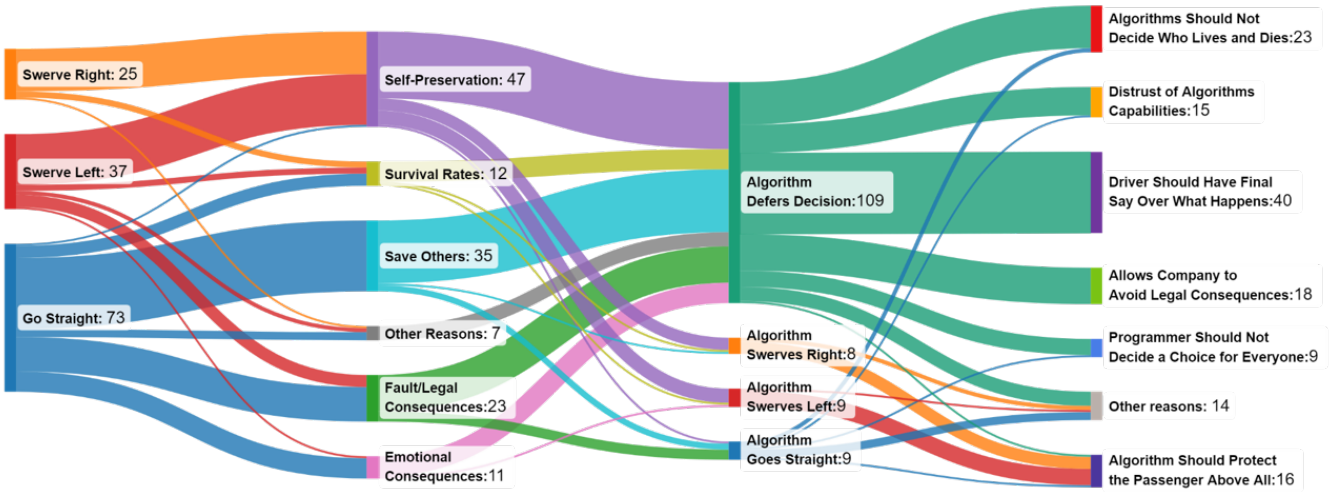

Figure 2: Sankey Diagram Illustrating Results of *Survey 1* (100% Chance the Motorcyclist With a Helmet Dies)

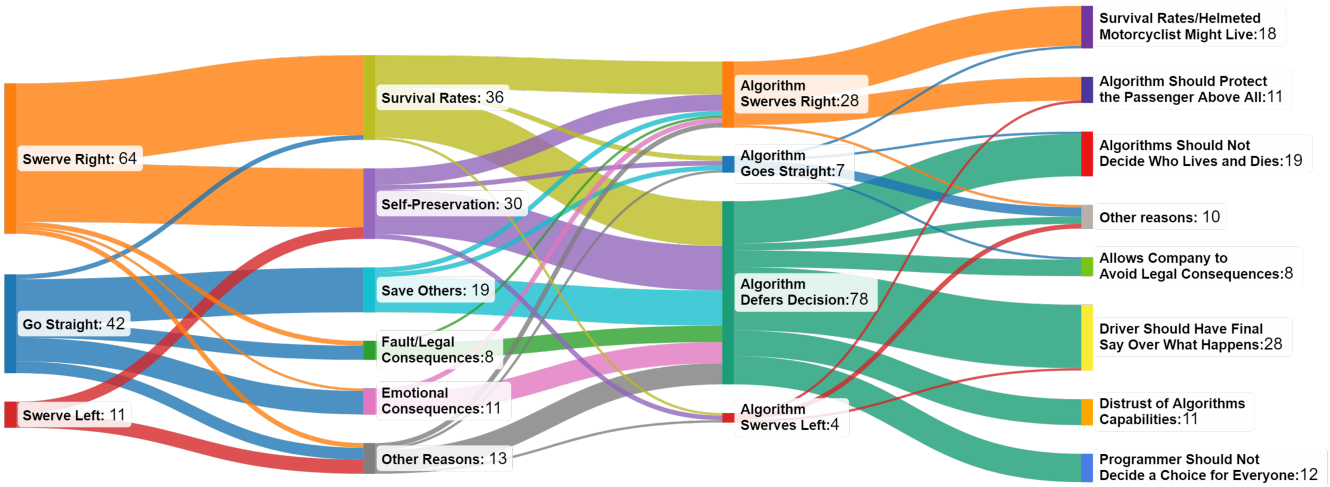

Figure 3: Sankey Diagram Illustrating Results of *Survey 2* (80% Chance the Motorcyclist With a Helmet Dies)

Survey 1, less people justified themselves using self-preservation and more people used a primary reason of survival rates.

Lastly, considering the results for the second question of *Survey 2*, again with the fourth option for the algorithm to defer the decision being available, we see that like in *Survey 1*, a majority (66.7%) of participants still chose to defer the decision. This is a smaller majority than in Survey 1, and we instead see that having the algorithm swerve right actually was chosen by 23.9% of participants, signifying a large relative increase now that that option to swerve right now could potentially result in no deaths. Going straight and going left were both rarely chosen in comparison. Additionally, we find that we have all of the same themes being represented as in *Survey 1*, with one addition. For people who would have the algorithm swerve right, on top of commonly citing that ***Algorithms Should Protect the Passenger at Above All***, we also saw the resurgence of the idea of **Survival Rates/Helmeted Motorcyclist Might Live** as a primary reason. Other than that, we still see the other five major themes that were cited in the first survey by people, who chose the option to defer the decision, still being cited in similar proportions.

All in all, our findings indicate that altering the consequences of a decision significantly affected engineering students' decisions in relation to an ethically complex scenario. For the first question of *Survey 1*, a majority of participants chose to remain in their lane and hit the truck. However, in *Survey 2* where a motorcyclist could potentially survive if hit, a majority chose to swerve into the

motorcyclist with a helmet instead. This demonstrates an important consideration to be made when conducting similar research about peoples opinions regarding the appropriate behavior of algorithms, that **people's perceptions are not necessarily fixed, and can be significantly influenced by the way in which a question is worded or scenario is setup**.

The fact that some participants made arguments valuing their own lives higher based on the age difference with the motorcyclists, as well as the fact many participants argued that it would be right for one motorcyclist to be hit over the other supports the notion that **people do not necessarily consider the lives of others equally in morally challenging situations**. Though some participants did say that they would chose between the motorcyclists at random (given that they would swerve at all), this was done only very rarely.

Another common trend we observed the change in the options chosen between question one and two. As when the participants were tasked with choosing what an algorithm should do in this scenario, a large majority of them chose *to divert the decision to the driver*. Their reasons are even more insightful, as many indicated a **lack of faith in the morals of the algorithm or morals of programmers who create them**. This demonstrates that based on the population studied, which was a group of prospective engineers, there is a **significant need for more ethics based training and education** for engineering students or engineering professional working with AI systems. If it is the case that the general public shares these sentiments, then this would be an unavoidable step if we ever hope to see the wide adoption of AV.

Moving into the future, we believe it would be insightful to administer this study to a more diverse sample population and compare the results across different demographic characteristics. We do not include such an analysis here, due to the limited sample size between the different demographics, such as gender and the participants academic year, which we might otherwise have hoped to compare across. Targeting a more specific demographic of major, such as Computer Science students, or even a non-student population could prove incredibly interesting. Additionally, we recognize that the participants of the two surveys were different people, which has made direct comparisons between the results of the surveys for the purposes of evaluating the effect of varying the consequences slightly more difficult. This could be circumvented in the future by administering a new survey which first places the participant in the situation of the first survey and then alters the consequences in the second question.

ACKNOWLEDGEMENTS

The initial codification of the raw data was completed in collaboration with Jennifer Moloney, Emma Herrero, Lauren Dulick, Carolina Aranguiz Dias, and Ilise Hyams.

REFERENCES

[1] Ashish Aggarwal and Saurabh Ranjan. 2022. How do Undergraduate Students Reason about Ethical and Algorithmic Decision-Making? (2022), 488–494. https://doi.org/10.1145/3478431.3499412

[2] Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, M.J. Crockett, Jim A.C. Everett, Theodoros Evgeniou, Alison Gopnik, Julian C. Jamison, Tae Wan Kim, S. Matthew Liao, Michelle N. Meyer, John Mikhail, Kweku Opoku-Agyemang, Jana Schaich Borg, Juliana Schroeder, Walter Sinnott-Armstrong, Marija Slavkovik, and Josh B. Tenenbaum. 2022. Computational ethics. *Trends in Cognitive Sciences* 26, 5 (2022), 388–405. https://doi.org/10.1016/j.tics.2022.02.009

[3] Yochanan E. Bigman and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181 (2018), 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

[4] Jean-François Bonnefon et al. 2016. The social dilemma of autonomous vehicles. *Science* 352 (2016), 1573–1576. https://doi.org/10.1126/science.aaf2654

[5] Bigman Y.E. Gray K. 2020. Life and death decisions of autonomous vehicles. *Nature* 579 (2020), E1–E2. https://doi.org/10.1038/s41586-020-1987-4

[6] Awad E. Dsouza S. Kim R. et al., Dsouza S. 2018. The Moral Machine experiment. *Nature* 563 (2018), 59–64. https://doi.org/10.1038/s41586-018-0637-6

[7] Judith Jarvis Thomson. 1984. The trolley problem. *Yale LJ* 94 (1984), 1395.