Exploring Morality through Data Science Students

Mahati Gorthy
CMSC320: Introduction to Data Science

March 24, 2025

1 Introduction

Morality isn't always simple. We like to think we know right from wrong, but the way we judge situations is often influenced by our experiences, backgrounds, and surroundings. This report looks at how data science students at the University of Maryland judge moral scenarios—specifically when deciding if someone is being a "jerk."

The dataset consists of survey responses collected from two semesters: Fall 2024, which includes students from both the Max's and Fardina's sections, and the 2025 and 2024 datasets, which includes students from Max's section. By analyzing these responses, I explored patterns, including differences based on age and how external factors (such as the time of day and day of the week) may influence responses. Some interesting findings include age shifts in harshness: Younger students in 2025 were more judgmental than their older counterparts, a reversal from 2024. Relatability: While people sometimes judged more leniently when they identified with a scenario, this wasn't always the case. External influences on morality: Participants tended to be harsher in their judgments late at night and in the middle of the week. Differences between sections: Students from different class sections exhibited different levels of harshness.

Overall, this report aims to highlight how complicated moral decisions can be and the mix of factors that influence the way we judge others.

2 Background

The data comes from surveys where students rated social scenarios as "not a jerk," "mildly a jerk," or "strongly a jerk." The scenarios covered personal conflicts, ethical choices, and financial decisions. The survey also collected demographic details like age, school year, and political affiliation, helping me analyze how different factors influence responses.

3 Methodology

3.1 Data Cleaning

1. Handling Missing Values: A small number of responses had missing values in ”Am I a Jerk” related questions. Since these cases were randomly distributed and didn’t systematically affect any specific group, I dropped them.

2. Standardizing Responses: I cleaned up categorical values by removing extra spaces, converting everything to lowercase, and fixing spelling mistakes (”sophmore” to ”sophomore” and ”famale” to ”female”). Additionally, I standardized these values: ”mildly a jerk” to ”mild jerk,” ”strongly a jerk” to ”strong jerk,” and other responses like ”don’t know / it’s complicated” to ”unsure.”

3. Converting Responses to Numbers: To make the data easier to analyze, I converted the morality words into numerical values: 0 for ”not a jerk”, 1 for ”mild jerk”, and 2 for ”strong jerk”. This allowed me to calculate averages, compare groups, and run tests.

4. Gender Correction: When comparing or combining datasets with gendered scenarios (Max Dataset and Fardina Dataset), I made sure that the columns being analyzed matched and consistently referenced the same gender and pronouns.

4 Findings

4.1 Do different ages view morality differently?

One of the first questions I explored was whether moral judgments change with age. Additionally, I compared the difference in responses between 2024 and 2025. To do this, I compared the mean ”jerk scores” of two age groups (17-20 and 21+) across the 2024 and 2025 surveys. In this case, ”jerk score” refers to adding up all values (+0 for ”not jerk”, +1 for ”mild jerk”, and +2 for ”strong jerk”).

I found that in 2024, the older group (21+) judged people more harshly than the younger group (mean jerk score: 0.72 vs. 0.68). But in 2025, things flipped—now, younger students were harsher (0.83), while the older group became more lenient (0.62). This reversal is

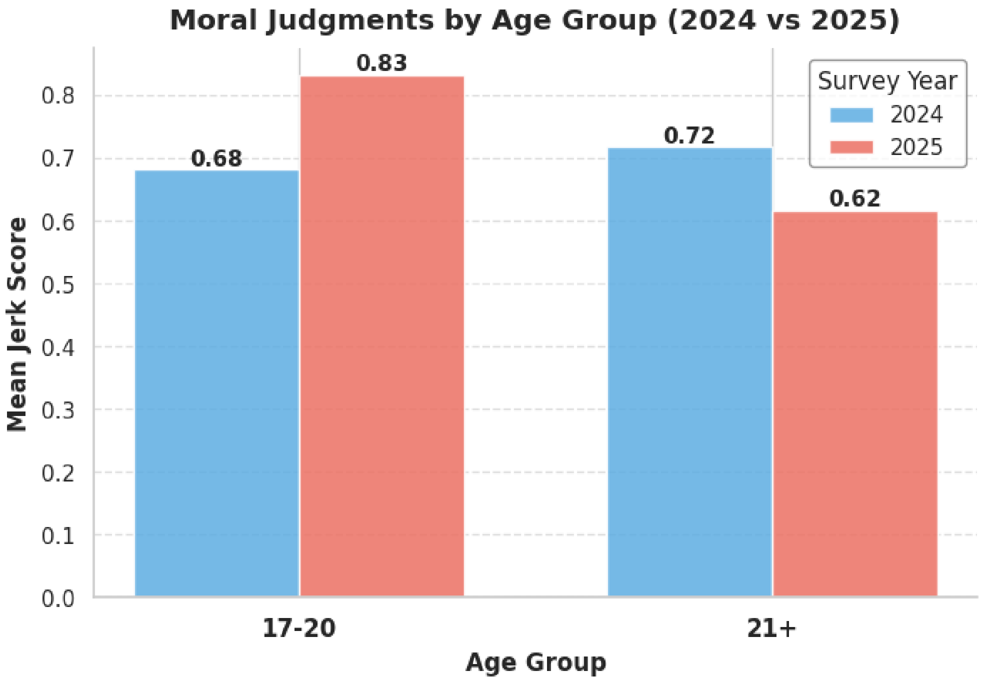

Figure 1: Moral Judgements by Age Group (2024 vs 2025) Bar Chart

interesting. It could reflect changing social attitudes, differences in who took the survey each year, or even other factors like current events or personal experiences.

I wanted to see if these differences were statistically meaningful. To test this, I ran a two-sample t-test comparing 2024 and 2025 responses. The null hypothesis is that there is no significant difference between the two years. If the p-value is below 0.05, we reject the null hypothesis.

For the 17-20 age group, the p-value, 0.2051, is greater than the significance level of 0.05, thus we fail to reject the null hypothesis. For the 21+ age group, the p-value of 0.4791 is much greater than 0.05, meaning we fail to reject the null hypothesis.

Since neither result was below 0.05, we can't conclude a real difference in moral judgments between 2024 and 2025–it might just be a random variation.

4.2 Are people more forgiving when they "see themselves" in the scenario?

This question explores if people judge situations more leniently when they can personally relate to them. To test this, I manually assigned gender and age labels to each scenario and gave each respondent a similarity score based on how closely their demographics matched the scenario: +1 point if their gender matched, +1 point if their age was within a year of the scenario's age, and +0.5 if their age was within 2 years. Then, I compared these similarity scores to their jerk score.

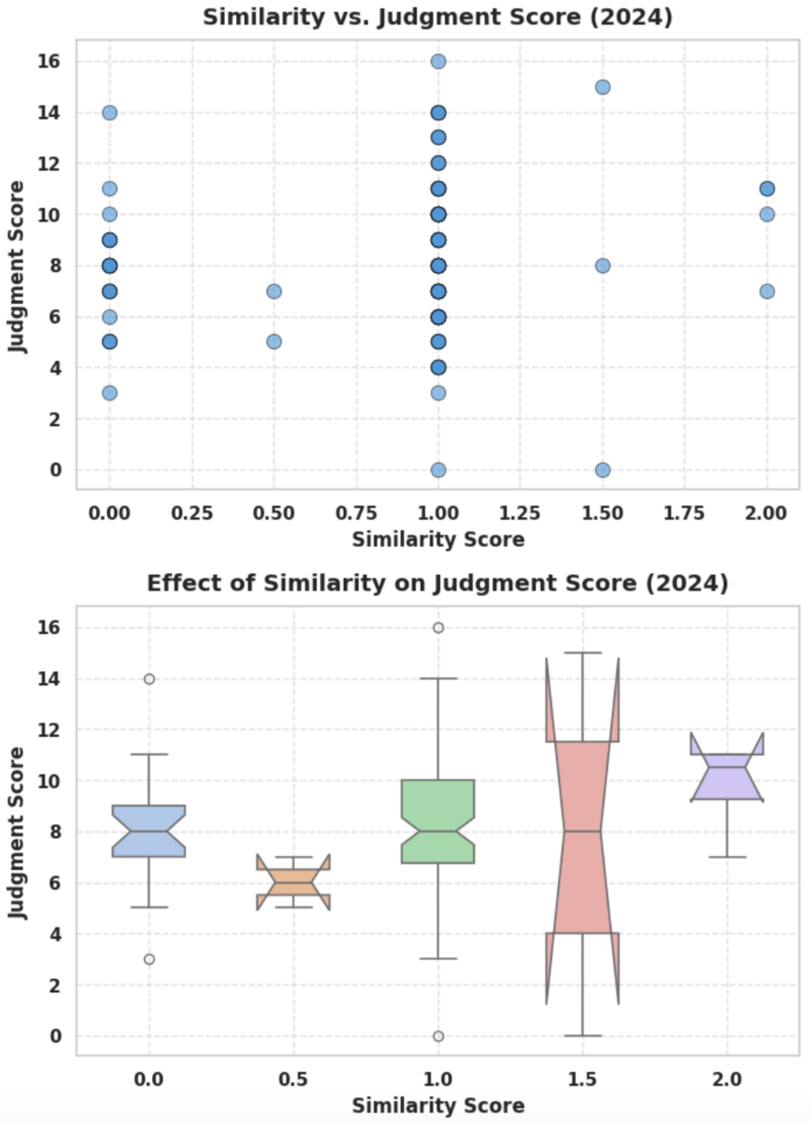

Figure 2: Similarity vs. Judgement Graphs (2024)

In 2024 (Figure 2), the results show that when similarity is low, judgment scores vary but tend to be moderate. As similarity increases, responses become more scattered—some people judge more harshly, while others are more lenient. The box plot highlights that at times, higher similarity levels lead to greater variability in judgment scores. The 2025 graphs (Figure 3) show that judgment scores vary widely at both low and high similarity levels. Like 2024, there isn't really a clear pattern in how similarity affects judgments. While some respondents were more lenient when they related to a scenario, others remained just as critical. This could mean that relatability may influence judgments in some cases, but

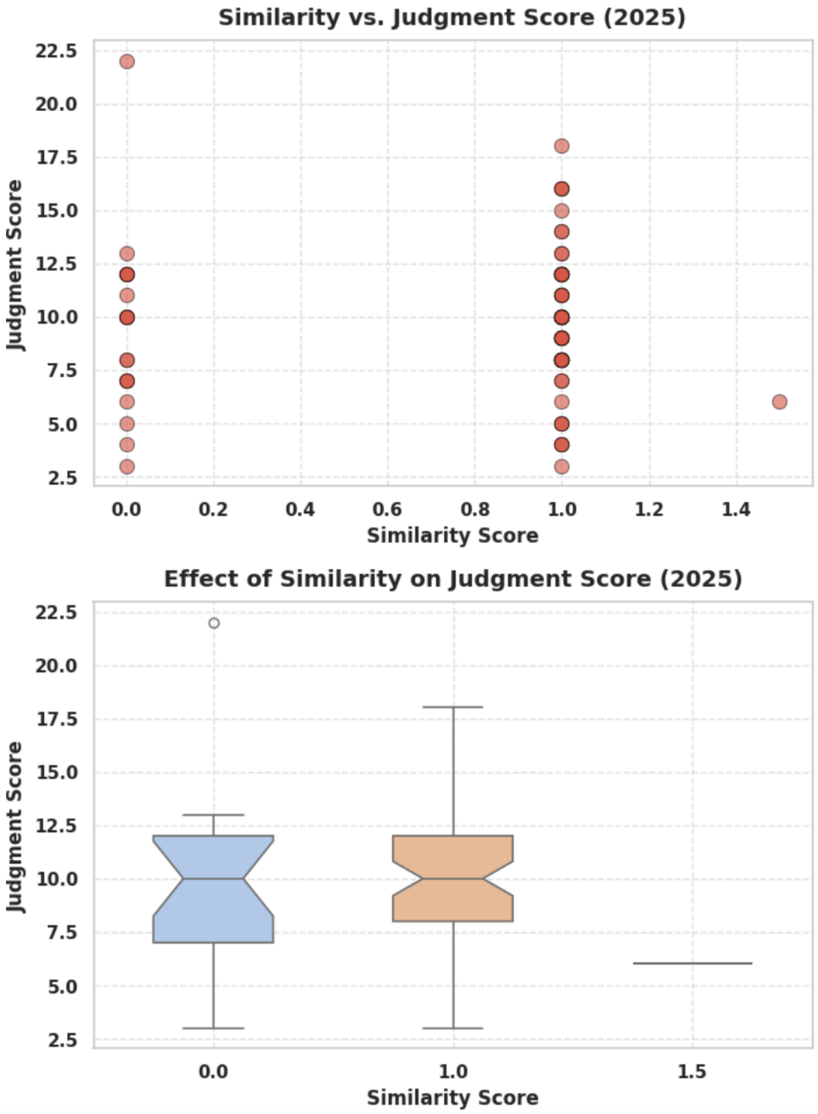

Figure 3: Similarity vs. Judgement Graphs (2025)

not in a consistent or predictable way.

4.3 Is there a time effect on people's judgments?

Next, I looked at whether the time students took the survey influenced how harsh their responses were. For this, I focused on data from Max's section.

Using Figure 4, the graph shows fluctuations in judgment severity throughout the day. The highest scores appear just after midnight, peaking around 12 AM. Harshness then declines steadily through the morning, reaching the lowest point around 11 AM to 12 PM. However, there is a sharp spike in harsh judgments around 2 PM, followed by a decline. Evening judgments are inconsistent, with another increase around 8 PM before dropping

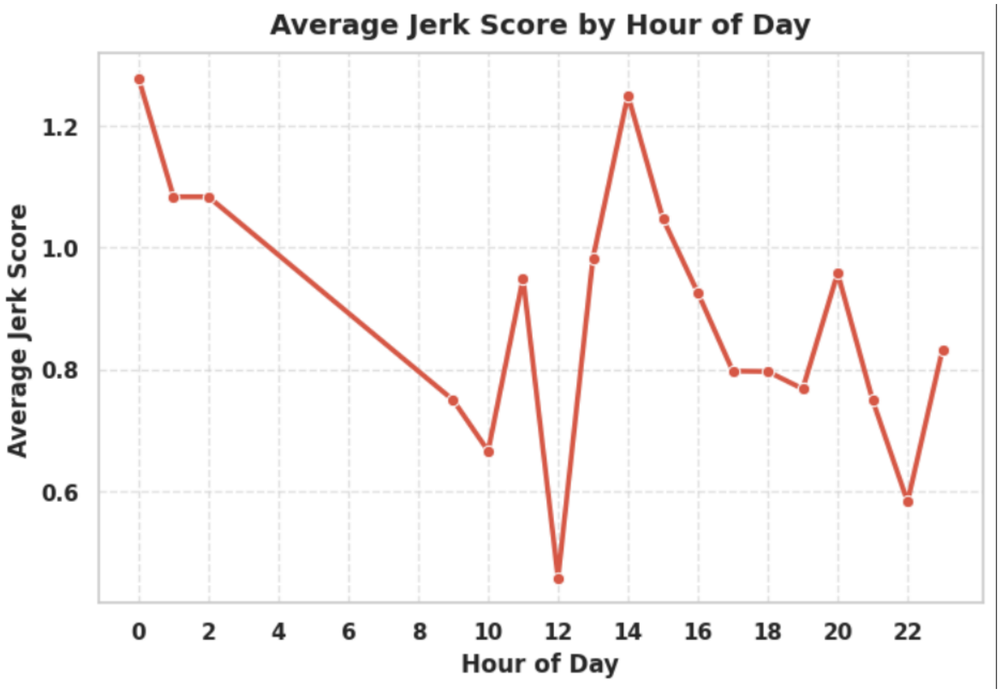

Figure 4: Average Jerk Score by Time

again late at night.

These findings suggest that people tend to be more critical in the early morning and mid-afternoon, while judgments are more lenient in the late morning. Anything may influence this like mood, tiredness, and daily routines.

4.4 Is there an effect of the day of the week on people's judgments?

This question looks at whether people's judgments vary by the day of the week. I converted the response timestamps into days and grouped them by response day.

The graph (from Figure 5) has some changes in mean jerk score across the week. Tuesday-Thursday were the most judgmental days, with Wednesday being the harshest. Saturday was the most lenient. Monday and Sunday had moderate harshness.

This makes sense—people may be more stressed midweek, leading to harhser evaluations. By the weekend, they might be more relaxed (or just less invested in showing judgment).

4.5 Is there an effect of the day of the week on judgments in relation to gender?

After addressing the previous two questions, I focused on analyzing the role of gender. I grouped the dataset by both gender and response hour, then calculated the average jerk score

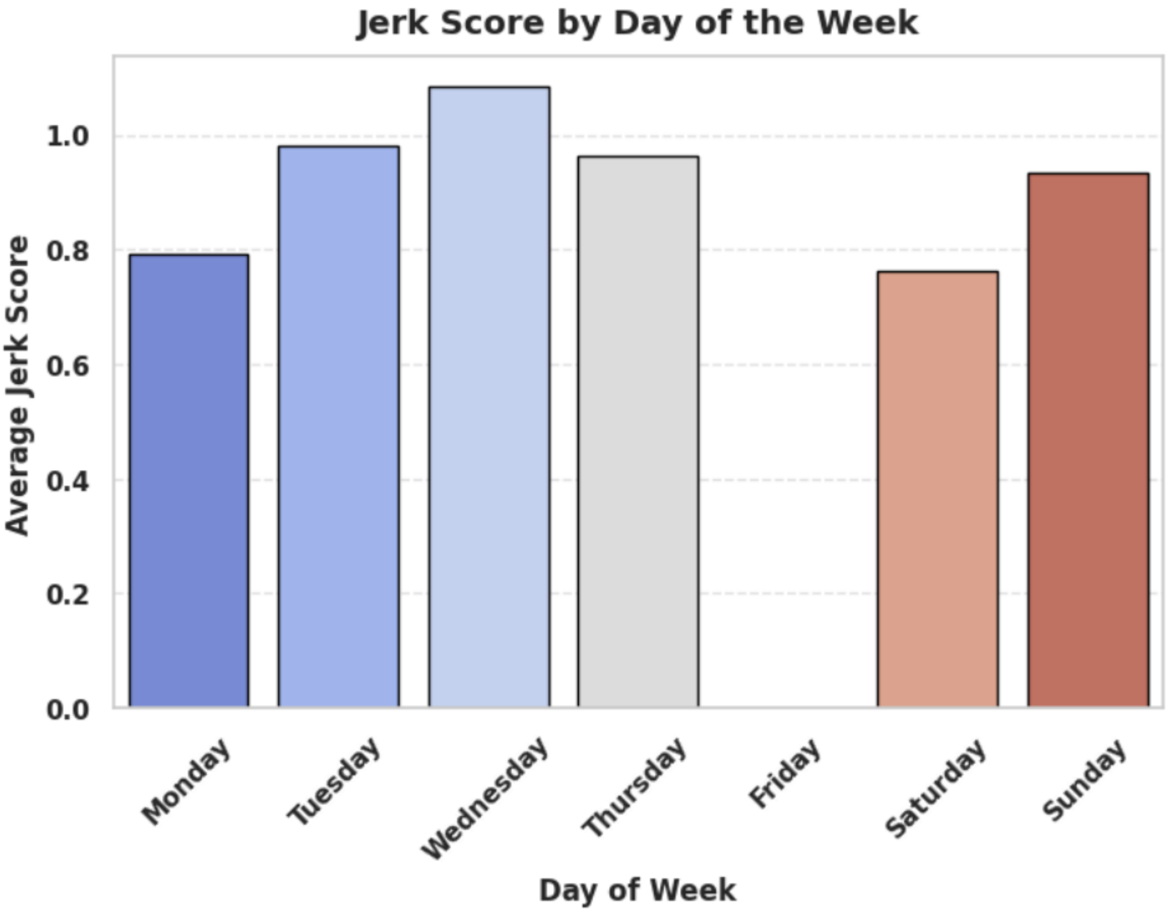

Figure 5: Average Jerk Score by Day of the Week

for each combination. To visualize the relationship between all three variables, I created a heatmap.

The heatmap (from Figure 6) shows different patterns. Males and females have a similar trend of higher judgment scores occurring in the late night hours. Females show a really high harshness score of 1.42. Some of the lowest scores, 0.25 for males and 0.5 for females, occurred around 10 AM - 1 PM. The afternoon and evening hours have fluctuations, with another peak appearing around evening hours (5 PM - 8 PM) for some gender groups.

There are not as many responses from non-binary/other respondents and those who preferred not to say their gender, making it difficult to draw any conclusions. The non-binary/other group shows a spike in harshness around 5 PM.

There may be some relationship between gender and time of day as they interact with response harshness, with late night responses generally being the harshest across most groups.

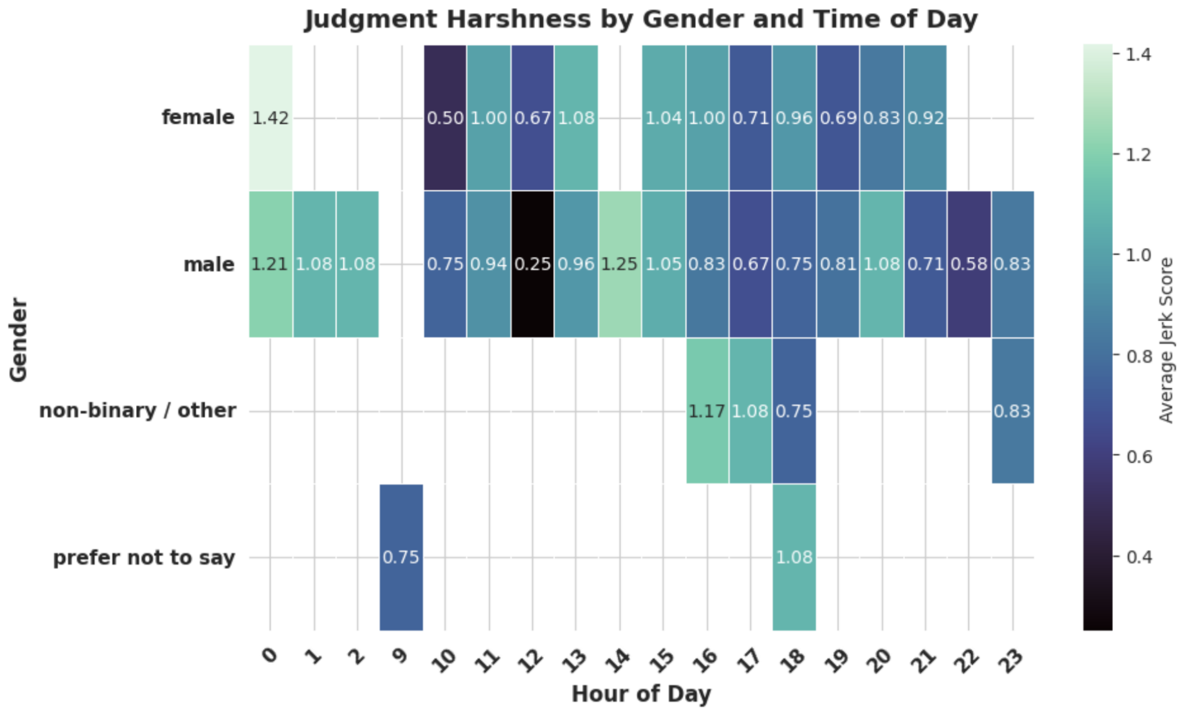

Figure 6: Average Jerk Score by Day of the Week

4.6 Which Scenarios Were Most Controversial?

This question allowed me to look at the scenarios that sparked major disagreement and which had clear consensus. I categorized responses based on extremity, where extreme responses included ”strong jerk” and ”not a jerk,” and moderate responses included ”mild jerk.” Then, the percentage of extreme responses was calculated by dividing the number of strong reactions by the total number of responses.

Additionally, I combined both Max’s and Fardina’s sections (focusing only on similar columns) to have more data. Scenarios with the highest percentage of extreme responses were the ones with the most consensus. Scenarios with a balance between strong reactions and moderate responses were considered the most controversial.

Using Figure 7, the most controversial scenario was “Girlfriend’s Knee Pain”, where some people saw the boyfriend’s comment as a valid concern, while others thought it was unnecessary and harsh. Other debated scenarios included ”Flight Help for Sister” (do you owe family help?) and ”Trust Fund Disagreement” (should financially privileged people split expenses evenly?).

On the flip side, the most agreed-upon scenarios discussed more clear violations. Scenarios such as ”Lost Cat Reward,” where a person withheld a found pet until they received a promised reward, probably triggered strong judgments. Similarly, ”Child Support Investi-

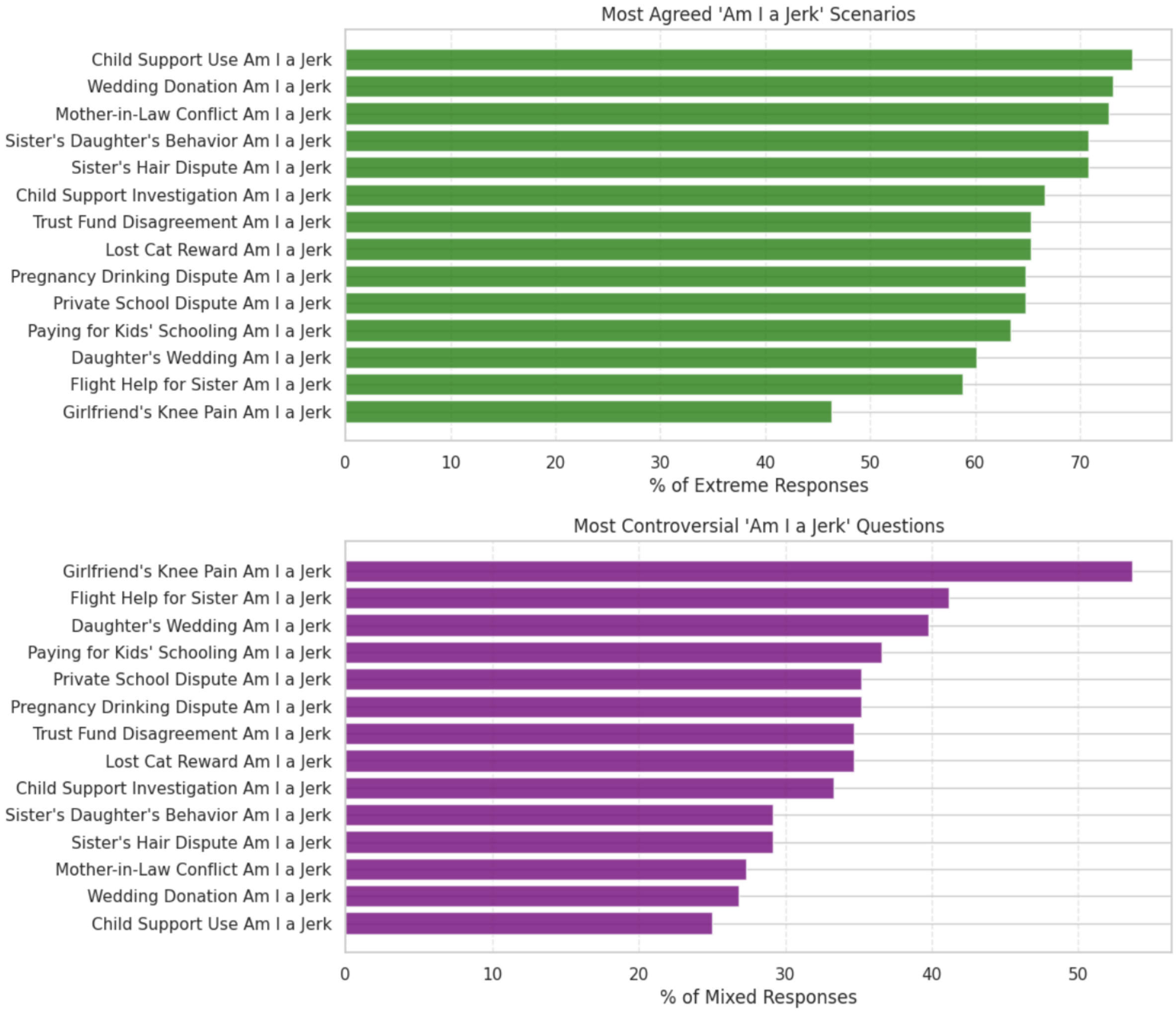

Figure 7: Controversial and Agreed Upon Scenarios

gation," where a parent repeatedly pursued legal action to increase child support payments, had strong opinions.

Moral judgments tend to be most divided when the scenario involved conflicting ideas, like individual rights versus family obligation, or financial independence versus fairness. On the other hand, scenarios where one party clearly acted inconsiderately usually led to more agreement.

4.7 What is the difference in harshness between Max's class and Fardina's class?

This question analyzes the differences in harshness between the two classes. I calculated the percentage of "strong jerk" responses for each scenario that both datsets had. Then, I found the differences in harshness to identify which scenarios exhibited the greatest disparities in moral judgment. Additionally, I graphed an overall class comparison of harshness.

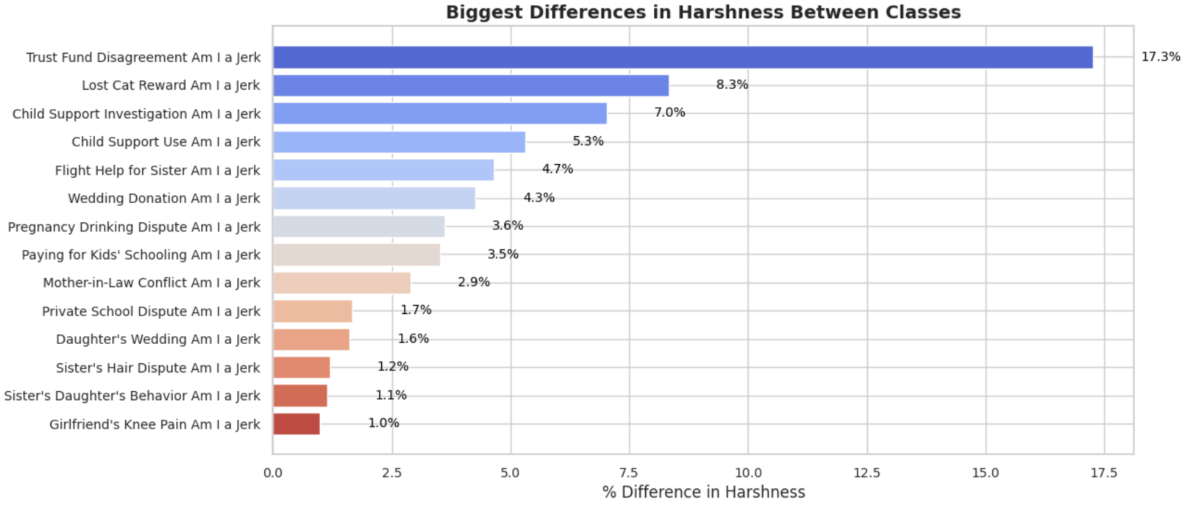

Figure 8: Difference in Harshness Between Classes

Looking at specific scenarios (Figure 8), the biggest difference in harshness was seen in the Trust Fund Disagreement scenario, where Max's Class was 17.3% harsher than Fardina's Class. This means they were much more likely to view the trust fund recipient's decision to split expenses equally with their lower-income partner as unfair. Another difference was in the Lost Cat Reward scenario (8.3%) and the Child Support Investigation scenario (7.0%).

On the other hand, some scenarios had nearly identical judgments across both groups. The smallest differences appeared in cases like Sister's Daughter's Behavior (1.1%) and Girlfriend's Knee Pain (1.0%), meaning that when it came to issues involving medical responsibility or childcare, both classes had the same idea.

The graph (Figure 9) comparing the two classes shows overall differences. Max's Class was more likely to label someone as ”strongly a jerk” than Fardina's Class, with an average harshness rating of 26.1% compared to 22.8%. This could mean that participants in Max's Class tended to be more critical or judgmental.

The trend in overall harshness shows that different groups may have different preferences when evaluating scenarios. Max's class may be stricter while students in Fardina's class are more lenient. These findings show how different groups can interpret ethical situations in varying ways.

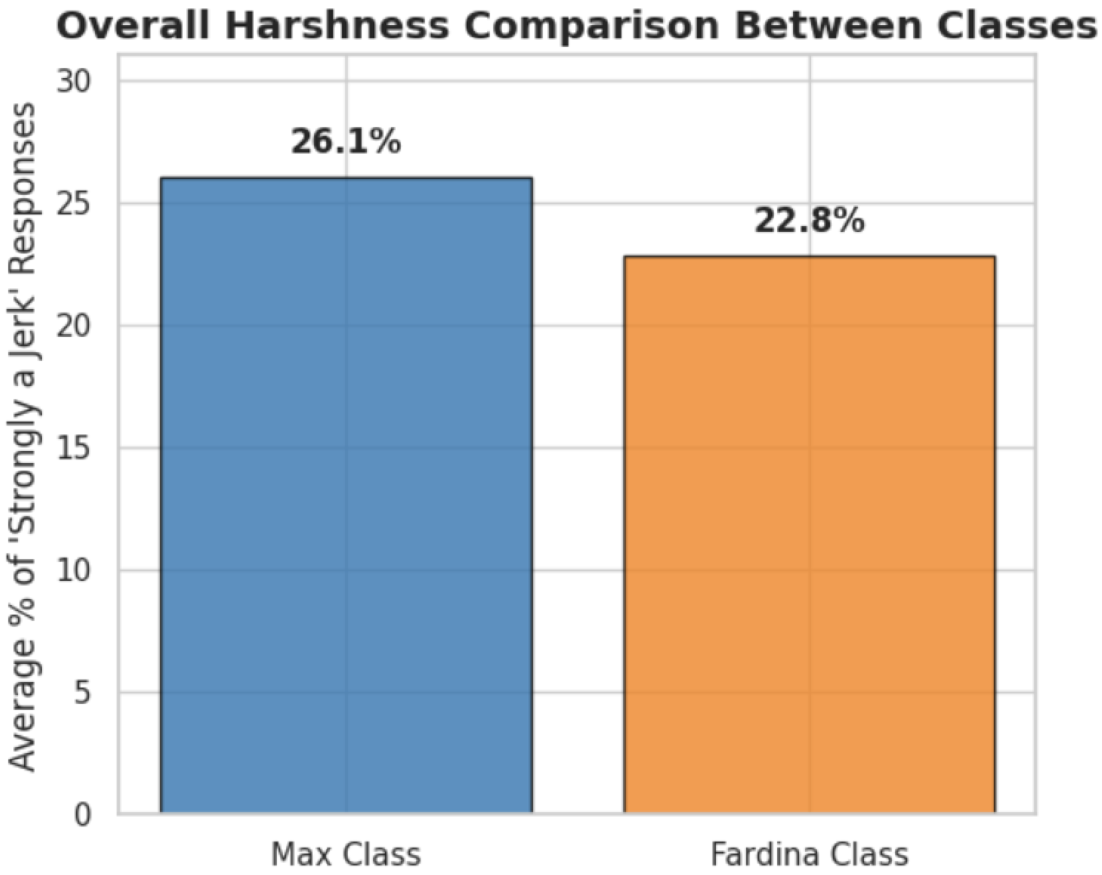

Figure 9: Overall Harshness Between Sections

5 Conclusion

At the start of this analysis, I expected to find clear evidence or statistical findings that certain trends exist within the datasets. But, the results showed very subtle trends.

The data showed that while moral judgments changed slightly between age groups, the changes were not statistically significant. Interestingly, while younger participants in 2025 judged more harshly than their older counterparts, this was not a universal trend across both years.

The assumption that I had that people are more forgiving when they see themselves in a scenario turned out to be only partially true. While there were some signs that similarity led to leniency, it was inconsistent, and some participants stayed just as critical regardless of how much they related to the scenario.

Time of day and day of the week had a more pronounced impact on moral judgment, with late night responses being the most severe. Furthermore, the differences between Max's and Fardina's classes suggests that there are subtle differences between sections which may be caused by registration dates or preferences for professors.

I think the most interesting takeaway is that morality isn't rigid—it shifts depending on the environment and a person's background or opinions when making judgments.

6 Future Work

While this report looks at how data science students navigate tough scenarios, there are several other routes this research can take. Expanding the dataset to include students from different majors or backgrounds could show whether these judgment patterns hold up across different areas of education. Additionally, tracking the same participants over time could show whether individual perspectives change based on life experiences, shifting norms, time of day, or age.

7 Appendix

Datasets