EPFL

Semester project :
Ethical transportation and machine learning for autonomous vehicles' decisions in risk situations

Sarah Blanc
Victor Dubien

Supervisor : Brian Sifringer
Professor : Alexandre Alahi

Spring semester 2022

Abstract

This paper deals with the design and analysis of a survey with the aim of learning about one's personal ethics. This known ethic might be used to adapt an autonomous vehicle to its driver as far as possible, especially when a choice has to be made between two scenarios with deadly finality.
The main part of this document is devoted to the analysis of the data collected during the sharing of the survey with a small panel of people. However, first, the process that led to the final survey is explained. All influences and decisions are listed so that this information and thinking is not lost if anyone else decides to continue researching this topic.

We would like to express our sincere thank you to Prof. Alexandre Alahi who guided us in this work. His knowledge, support and patience with us were invaluable. We enjoyed our discussions and the quality of our relationship.

Thank you as well to Mr. Brian Sifringer for his advice, support and guidance. We hope this work will somehow help him in his research.

Finally, we would like to thank the students of the classes of Deep Learning for Autonomous Vehicles and Introduction to Machine Learning for engineers for answering the survey.

Table of contents

List of Figures

List of Tables

1 Introduction

While research on the subject of ethical values in autonomous vehicles is still at its infancy, there will come a day when ethical issues will play an important role in the reflection. Indeed, the novelty of autonomous cars is for the moment mainly around the technical aspect. However, as the human being gives control and decision making to the vehicle, it is pertinent to look at how it would be possible to adapt the car to each user. Would it be possible to configure each car to be as close as possible to the ethics of its user? What would be the limits of this adaptation? To what extent is it possible to adapt the car? If this adaptation is possible, and whatever its limits, a way of learning one's personal ethics has yet to be found. How can the ethics of a human being be determined in the most efficient way? With no personal interaction with others allowed and no way of exchanging in case of misunderstanding, it is therefore a question of establishing a simple, clear and effective survey in order to model one's ethics in the most efficient way possible.

2 Existing work

The first part of the work was to find out about existing work and to extract interesting information from it.

The trolley problem is an ethic and psychological experiment. It involves stylized ethical dilemmas of whether to sacrifice one person to save a larger number of people. The first scenario is usually a scenario in which a runaway trolley is on course to collide with and kill a number of people (traditionally five) down the track. All of the five people are unable to move. Even though the fatal collision is unavoidable, a bystander has the possibility to intervene and divert the trolley to kill just one person on a different track. The experience is made personal when the reader himself has the possibility to pull a lever that will make the trolley change lanes. Knowing that there is a person in the side lane, the reader has two options:

1. Do nothing, in which case the trolley will kill the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

The question then arises as to which is the most ethical option. In other words, what should be done, how should one act in this very rare and unpredictable situation?

Using the paradigms of the trolley problem, MIT Media Lab created a platform called Moral Machine. The difference of this variant of the original Trolley dilemma is that the choice (such as into whom or what to crash) is not made by a human being but by an autonomous vehicle's software. The user has to make a decision on scenarios in which a self-driving car is about to hit pedestrians. The user has to decide between two choices. The first one is to keep going straight to preserve the lives it is transporting. The other one is to swerve the car to avoid hitting the pedestrians. That choice can affect the particulars of the deadly outcome but both outcomes are still destructive. The scenarios, despite not being very descriptive concerning the environment the cars and pedestrians are evolving in, proposes a wide range of features for the humans beings. They can be doctors, homeless people, fit or overweight, sex can vary also. The goal was to allow the public to express their opinions on what decisions autonomous vehicles should make in scenarios such as the ones in the trolley problem. The data collected and their analysis showed that it is possible to see differences of behaviors between countries. It is therefore natural to ask whether the reactions of the autonomous car should be adapted as far as possible to each country? Can an autonomous car in Europe react differently to the same situation than an autonomous car in the Middle East? Should the designers make sure that these autonomous vehicles are able to solve problems on the road that aligns with the moral values of humans around it? All these questions remain very open and have no single answer. This societal issue remains widely debated and it seems complicated to arrive at an opinion that would suit everyone.

Nevertheless, several criticisms have been developed against this MIT Moral Machine. Everyone has the right to have an opinion but must respect the opinions of others. Therefore, the criticisms listed below are not subject to any value judgement. It is more a question of listing the different opinions and using them to develop the project. Here are some of these criticisms that have been raised through different literature:

- The Moral Machine is a clear violation of everything the UN Centre for Human Rights stands for, because it makes people choose who to sacrifice based on social criteria such as wealth or fitness.
- It encourages a type of thinking which is devoid of human feeling and empathy. It assumes a mandate to decide who lives or dies.
- An algorithm is not a person, it is a policy. It is not possible to just get policy by just assuming that an answer that might be fine in a individual case will generalize.
- To code a self-driving car is extremely difficult and requires mimicking human-like intelligence. With our current knowledge, it is impossible to perform better than humans on that subject.
- The scenarios are not realistic enough as available algorithm are unable to differentiate people this much.

This is a challenge because of the complex nature of humans who may all make different decisions based on their personal values. However, by collecting a large amount of decisions from humans all over the world, researchers can begin to understand patterns in the context of a particular culture, community, and people.

3 Survey making

3.1 Process

In order to reach the final survey, several iterations were necessary. First of all, they allowed to become familiar with the subject. Moreover, these different approaches brought multiple reflections. Several angles of view could be studied and discussed.

3.1.1 The first survey: the break failure

The first part of the project took into account a break failure. The autonomous car we are in has brakes that have failed and it is impossible for it to stop successfully. The objective was to find out if it was possible to learn someone's ethics by asking them questions. The first step was to determine the shortest possible set of questions to determine someone's ethics and thus be able to predict their reactions in certain cases described by the scenarios. Therefore, the questions had to be relevant and exhaustive. Factually, everyone would be asked the same questions as everyone else. The following questions can then be asked. What areas should the questions cover? Policy orientations? Environmental orientations? Is it possible and useful to know someone's character traits (narcissism, communality, sense of justice,...)in order to know his ethics? It was decided not to take these aspects into account in order to better focus on information directly related and applied to the scenarios. The idea that came out of this first phase was to make a survey constructed in the following way:

1. First realistic impersonal scenarios with no redundancy between them.
2. Then explicit and clear personal questions.
3. And finally realistic personal scenarios and test whether the knowledge learned in 1 and 2 is sufficient to predict what each person will decide.

In order to be consistent with the current knowledge and capabilities of autonomous vehicles, the categories have been chosen according to what is recognisable or not by artificial intelligence in a car. Therefore, the possible differences between people are as follows:

- adults
- children
- babies (pushchairs)

- dogs
- cyclist

The choice was therefore made that the elderly would be taken into account in the adults, people who are overweight or not are in the same category. The gender (male, female,...) is not taken into account and no distinction is made between pregnant women and adults. In terms of possible outcomes, everyone has the possibility of being either dead, injured or uncertain. The aim of this first version was to minimise the user's time: to obtain the maximum amount of information with the minimum number of questions possible while knowing their ethics at the end of the survey.

The idea of the priority of rights has subsequently (and until the end of this project) become very important. How and in what way do we know the scale of rights? Although the first approach (humans > animals > plants > minerals) may seem obvious, the diversity in one's way of thinking is so large that nothing can be taken for granted. Moreover, ideas of scale have emerged. For example: how many child lives are worth one adult life? Since these questions are so meaningful, is it wise to ask whether the respondent has hesitated or is still hesitating, so an option was added to let the respondents express this potention hesitation. At this stage of the project, it was decided that it might be interesting to have this kind of additional information. Moreover, the addition of an obvious scenario was intended to successfully detect whether someone is responding to the survey at random, which could distort the results.

A first quantification and a first model were carried out without obtaining relevant results. Indeed, the survey was shared on a small scale only and the model could not be optimised, so this track was abandoned.

3.1.2 The second survey: the unexpected situation

This second part of the project will no longer take into account a break failure but an unexpected element happening on the road. It is too late to brake and a choice must be made between the two proposed outcomes. This avoids the possibility of blaming the car manufacturer in case of a break failure. Thus, in these new situations, the manufacturer is not at fault and cannot be considered guilty. This part of the project started by putting the idea of the survey somewhat aside. First of all, it was an attempt to develop a question tree, of minimal depth, allowing the creation of clusters of people with similar ethics. From this tree, is it possible to predict the ethical choices of each person according to the path they have taken in the tree by answering the questions? To do this, a question tree was the learning part and then a series of scenarios were used to check that the learning had been complete and efficient.

Iterative work was carried out with several guinea pigs. Indeed, during iteration 0, a first tree was developed on the basis of the knowledge acquired and the questions developed during the first survey. During iteration 1, The tree was adapted and modified according to the missing information to pass the learning verification phase with a 100% success ratio. The same process was repeated without ever arriving at a complete tree. Indeed, with each new person, a new problem or new missing information was added. It was therefore concluded that building such a tree by hand and without machine learning is far too complex a task at this stage of the research and with the current knowledge of the subject. It was also possible to see two very interesting facts. The first is that it is very hard for some people to be truly honest with themselves. There were many false altruists. Indeed, it is much easier in theory to always kill yourself to protect others. But deeper and more extensive discussions at the end of some of the surveys made us realise that some people are actually not quite ready to always kill themselves. The second fact is somewhat related to the first. Indeed, the answers to the questions are sometimes not in line with the answers to the scenarios. The confrontation with the inconsistencies sometimes led to a self-correction of the scenario answer. However, this was not always the case. Sometimes the scenario involves other factors that are sensitive to becoming a priority in the choice of outcome. It is therefore very complicated to take all these factors into account.

It was after these observations that the idea of the survey resurfaced. Based on the experience gained during the development of the first survey, this new trial was designed as follows:

- The scenarios should be as complete as possible. Unexpected elements should occur such as: a human being jumps on the road, the car in front swerves to the opposite lane, a rock or a branch falls, ...
- A list was made of all the personal questions that had been used since the beginning of the project. To this were added other questions that were considered relevant.

Throughout the development of this survey, which will turn out to be the final survey, the following question was asked: "Did the respondents need help to understand any part of the survey?". This allowed for the development of a survey that was clear and could be answered on its own and without external help. The organisation of the survey is the following:

1. A set of realistic scenarios separated in several sections to order this first learning phase.
2. A set of questions for learning.
3. A set of realistic scenarios to test whether the learning phase (from points 1 and 2) has been successful.

Compared to the first survey, it is possible to perceive seven major differences:

- Speed limits have become an essential part of the project. Indeed, the majority of the scenarios for learning are present with at least two different speeds.
- The probability of dying has replaced the uncertain fate.
- A clear highlighting of important elements has been done ("AV", "illegal action", ...).
- The idea is no longer to provide the shortest possible survey in order to minimise the time the respondent has to spend on it. On the contrary, the idea is to ensure that all the necessary information is collected, even if it means deleting data that is not necessary.
- The coverage of this survey is broader. It allows us to learn more so that we can test more scenarios.
- The scenarios themselves are also more comprehensive and complex.
- The introduction and setting of the survey is as simple as possible to capture the respondent's attention.

3.2 Data collection

As the research is still in the early stages of discovery, the survey has only been shared on a small scale for the moment. The idea was rather to see what could be done with the survey as it stands today. The opportunity to share it with a class of master students and a class of bachelor students allowed to collect 65 samples. The first interactions were with the master students. They helped to correct some errors or inaccuracies in the survey. Their comments and the exchanges established with them were very valuable. As a result, the bachelor class completed a survey that was somewhat modified as a result of the discussions with the previous students. This is why only the 45 samples from the bachelor students are used in the models below. Although new responses are continually coming in, it was decided to stop taking into account new data arriving after May 19^{th}.

3.3 Final test survey

This process has led to the two surveys linked below:

1. `https://docs.google.com/forms/d/e/1FAIpQLSeq7SqYDSws5h8ppyaZCPRiqTdRPJ7ZOkWrQsHcyZJOr2NaDw/viewform?usp=sf_link`
2. `https://docs.google.com/forms/d/e/1FAIpQLSeq7SqYDSws5h8ppyaZCPRiqTdRPJ7ZOkWrQsHcyZJOr2NaDw/viewform?usp=sf_link`

The two surveys are, with a few modifications, the same. Indeed, the only differences are the way of writing the four proposed answers to the scenarios in the learning part. The analyses and the sharing phase of the survey were done with the first survey, hence the relevance of providing both survey links here. The first survey was designed to guide the respondents in their reflection by highlighting (in the possible choices of answers) the elements to which they should pay particular attention. This helps to focus their attention and reflection on the learning objective of each scenario. The second survey takes into account the fact that it is important not to influence the reader in the choice of response options. It was therefore necessary to neutralise the choice of answers as much as possible, even if it meant not making the respondent pay attention to the main idea contained in each scenario.

With constant modifications, reflections, improvements, this survey will always be perfectible. Indeed, the complexity of the human being will always bring new reflections to be taken into account in order not to neglect any aspect. This is a long and iterative process which it would be interesting to extend over several months in order to be able to take a step back on the effectiveness of the survey after each modification. In addition, it would be interesting to be able to evaluate the relevance of each modification. Although one has to be critical about how it can be improved, the aim being to succeed in predicting the ethics of each person, it seems to be a good start and a good first survey to meet this challenge.

4 Clustering models

The goal was to try and cluster people into different noticeable ways of thinking. The data from the 45 answers was cleaned and encoded in the following ways : answers A and B to the scenarios were classified as 0 and 1 in order to make outliers more visible, as C and D were encoded as 3 and 4. During the classifications tasks, ">" was encoded as 0, "=" as 1 and "<" as 2. This is arbitrary and probably adds some bias to the analysis, but it's also on of the only ways. Indeed instead of being 2, "<" could have been -1, in order to be "centered" around the "=" which is 0. But this makes it look like in terms of distances that answering "<" or ">" is the same thing, when they actually are opposites.

4.1 Correlation

The first thing was to try and see if either the samples or the features would be correlated to a certain extent, in order to maybe pre-determine clusters of some sorts, if a way of thinking could be identified or in features could be dropped to help reduce dimensionality.

4.1.1 Features correlation

First the features were tested :

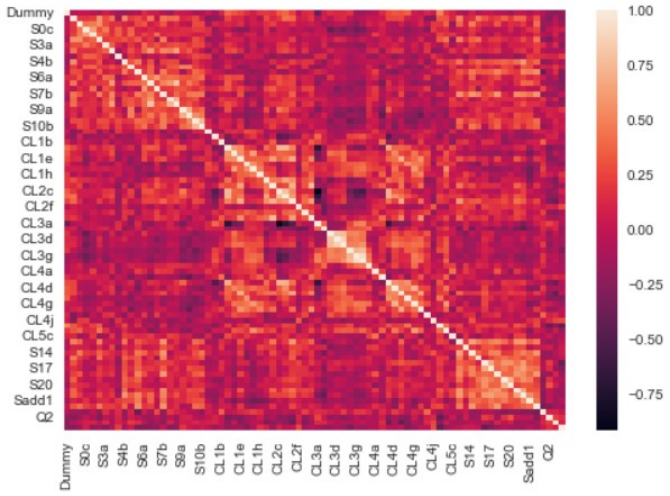

Figure 1: *Correlation heatmap of the features*

But only two features are correlated to a degree of 90% or more : $CL3f$ and $CL3g$, which ask to rank respectively someone you know versus one or two legal adults. Two features are also decorrelated to -90% or more : $CL2b$ and $CL2h$ which ask to rank respectively one illegal adult (respectively someone you know who is illegal) versus one child. Because features are almost not (de)correlated, it is not possible to remove some from the dataset, which would have helped with the dimensionality curse here : 45 samples for 78 features.

4.1.2 Samples correlation

Then, the samples' correlation was tested, in order not to remove samples, but to try and see if some samples were similar.

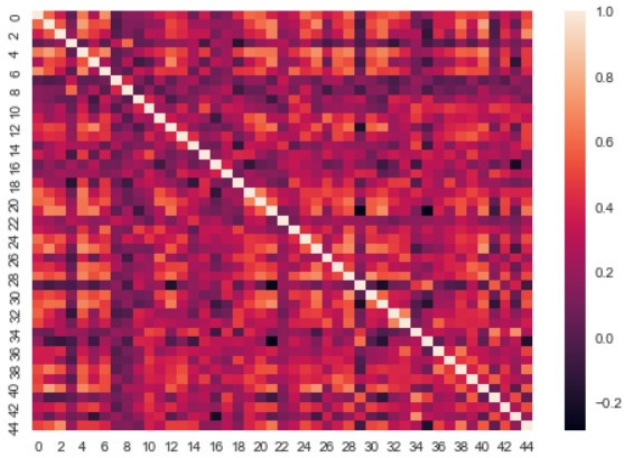

Figure 2: *Correlation heatmap of the samples*

Samples are quite neutral in terms of correlation, as only two of them (1 and 40) are correlated to 75% or more, and decorrelation doesn't go any further than -30%. It is not possible to say much about this. One reason could be that correlation is sensitive to encoding, so this lack of correlation might not be relevant for the cluster forming later.

4.2 Classical methods

First the data was scaled with a StandardScaler which brings all features to the same scale (same minimum of 0 and maximum of 1 for each feature) and normalized with the Z-score standardization $X_{norm} = \frac{X - \mu}{\sigma}$ in order to distribute it evenly on a Gaussian $\mathcal{N}(0, 1)$.

Then, in order to visualise the results, a Principal Component Analysis was performed on the data, bringing the number of features from 78 to 3 which are a linear combination of the 78. The choice was also made to perform PCA before the clustering as the clusters are more likely to be more visible because the important information has been retained. More precisely, the five most important features for the first principal components are features 45, 15, 46, 20 and 44, which correspond to CL3g, scenario 7b, CL3h, scenario 10a and CL3f. This component explains for 13,55% of the variance. The second principal component explains for 11,87% of the variance and the third one for 6,59%. Although 32,01% of the variance may not seem like a lot, the fourth component also explains for about 5% and the fifth one around 4% and it goes decreasingly, and they also cannot be visualised.

Some of the methods listed below need the number of clusters to form or the minimum number of samples to form a cluster as a hyperparameter. In order to wisely choose these, and because the labels are not available so it is impossible to test any accuracy of some sort, the elbow method (distortion score) and the silhouette score were applied to the data and algorithm before doing the actual clustering. The elbow method consists in plotting the explained variance against the number of clusters, and taking the number corresponding to the "elbow", where the function becomes linear-like. The silhouette score computes the mean intra-distance (between the samples belonging to the same cluster) defined as a, and the mean nearest-cluster distance (nearest cluster that the sample is not part of), defined as b. The silhouette score expresses

simply : $score = \dfrac{b-a}{\max(a,b)}$, and is again plotted again the number of clusters. Here the goal is to maximise this score, which can be comprised in $[-1, 1]$. A score close to 1 signals that the clusters are far away from each other but that samples within clusters are close to each other. On the other hand, if the score is close to -1, the clusters are very close and the samples within are far from each other, which may result in clusters overlapping or bad categorization. Both metrics were tested for $k \in [2, 30]$.

Because these two evaluation metrics mean very different things, when they returned a different number of clusters, both were taken into accounts and used (for K-Means).

Please note that if every figure and graph is reproducible with the code, it is possible that it doesn't return exactly the same results, depending on the random initialization of some methods.

All the following clustering methods are categorized in the unsupervised part of learning as the label of each data point in unknown for now.

4.2.1 K-Means

The first unsupervised algorithm that comes to mind is K-Means. The algorithms consists in randomly choosing the centroids of the clusters and minimizing the inertia function : $\sum_{i=0}^{n} \min_{\mu_j \in C}(\|x_i - \mu_j\|^2)$ where μ_j is the centroid of cluster C. Inertia can be understood as a measure of how internally coherent clusters are. As explained before, because K-Means requires the number of clusters as a hyperparameter, the elbow method and the silhouette score were performed on the data and method. The elbow method returned an optimal $k_{opt} = 8$.

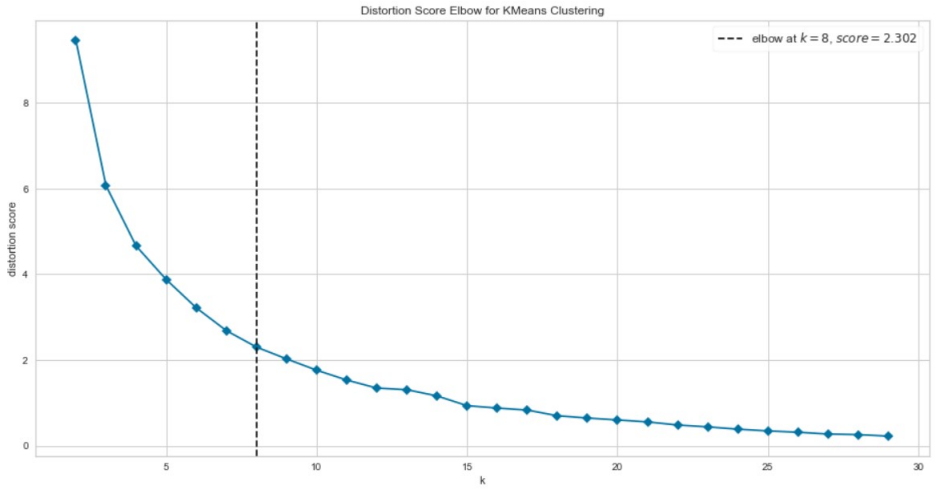

Figure 3: *Elbow method for K-Means*

And resulted in the following clusters :

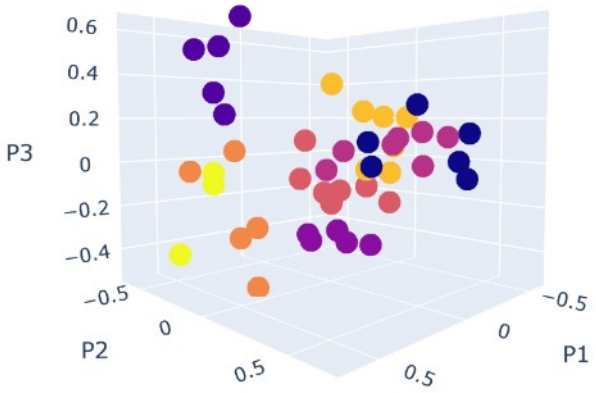

Figure 4: *K-Means clustering with* $k = 8$

Cluster n°	Color	Number of samples
0	Dark blue	6
1	Blue	5
2	Purple	5
3	Fushia	7
4	Salmon	7
5	Dark orange	5
6	Light orange	7
7	Yellow	3

Table 1: *K-Means clusters with* $k = 8$

The first observation that can be made is that the clusters are quite homogeneous by their sizes.

The silhouette score gave $k_{opt} = 4$:

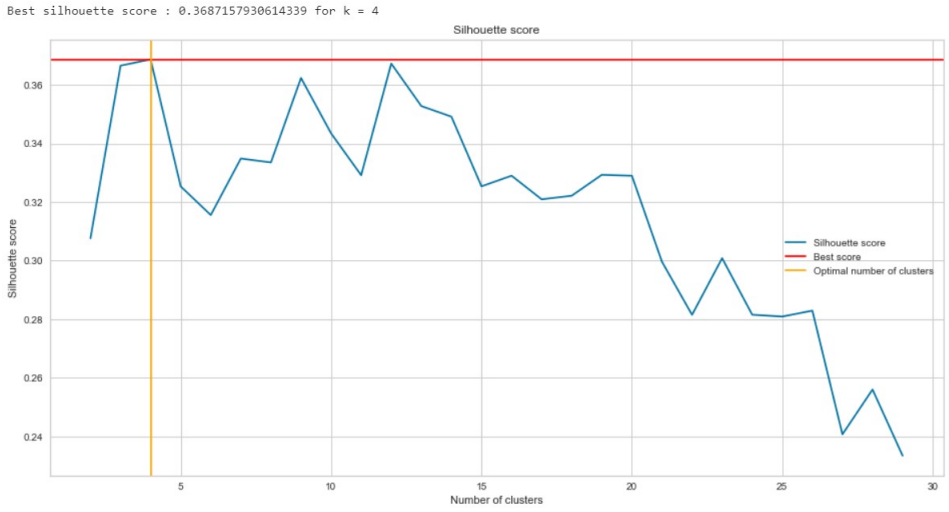

Figure 5: *Silhouette score for K-Means*

Which gave the following clusters :

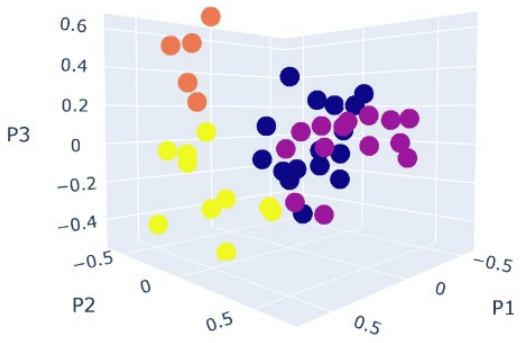

Figure 6: *K-Means clustering with* $k = 4$

Cluster n°	Color	Number of samples
0	Dark blue	16
1	Purple	14
2	Orange	5
3	Yellow	10

Table 2: *K-Means clusters with* $k = 4$

The clusters are less evenly distributed but it seems to make more sense when thinking about the actual data to have only four clusters instead of eight. More generally the problem with this sort of clustering is that the interpretability is quite poor. Indeed, it is necessary once the clusters are formed, to look "by hand" at the different answers and it is hard to learn anything from it.

4.2.2 Hierarchical clustering

Hierarchical clustering consists in recursively linking pairs or clusters of data using a linkage distance. Here the euclidean distance was used, and the linkage method is 'Ward', which aims to minimize variance of the clusters (or samples) being merged. Once again, the elbow method and the silhouette score were used to determine the optimal number of clusters, and both methods gave $k_{opt} = 8$.

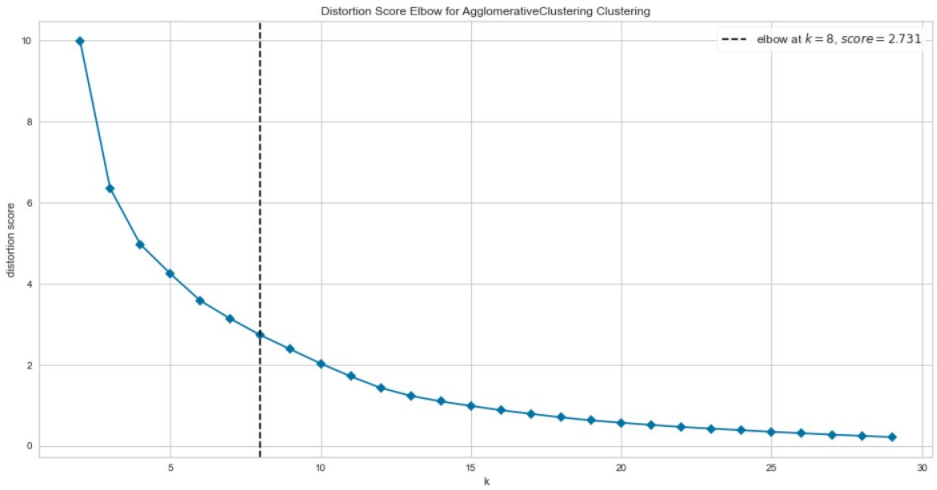

Figure 7: *Elbow method for hierarchical clustering*

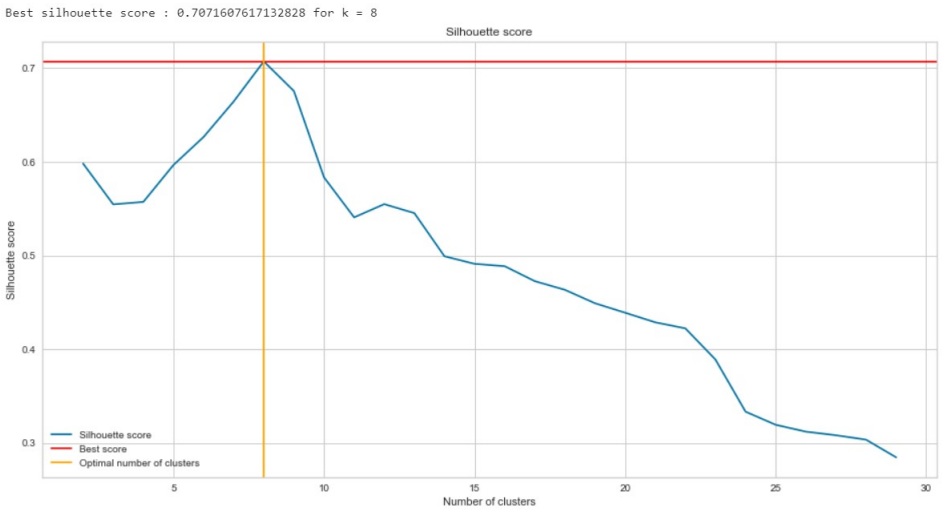

Figure 8: *Silhouette score for hierarchical clustering*

Which then gave the following clusters :

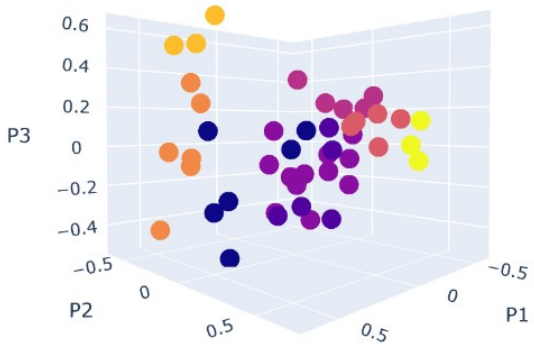

Figure 9: *Hierarchical clustering*

Cluster n°	Color	Number of samples
0	Dark blue	6
1	Blue	5
2	Purple	12
3	Fushia	5
4	Salmon	5
5	Dark orange	6
6	Light orange	3
7	Yellow	3

Table 3: *Hierarchical clustering clusters*

The samples are less evenly distributed in the clusters than with K-Means, and there are lots of groups, which complicates the analysis. Moreover, hierarchical clustering is very sensitive to initialization, so results may vary dependably.

Another way to visualize the clusters is with a dendrogram, reprensenting the clustered samples and euclidean distances. The orange line defines what is the distance between clusters for the ideal cluster number to be formed (the number of lines cut by the orange line is the number of clusters).

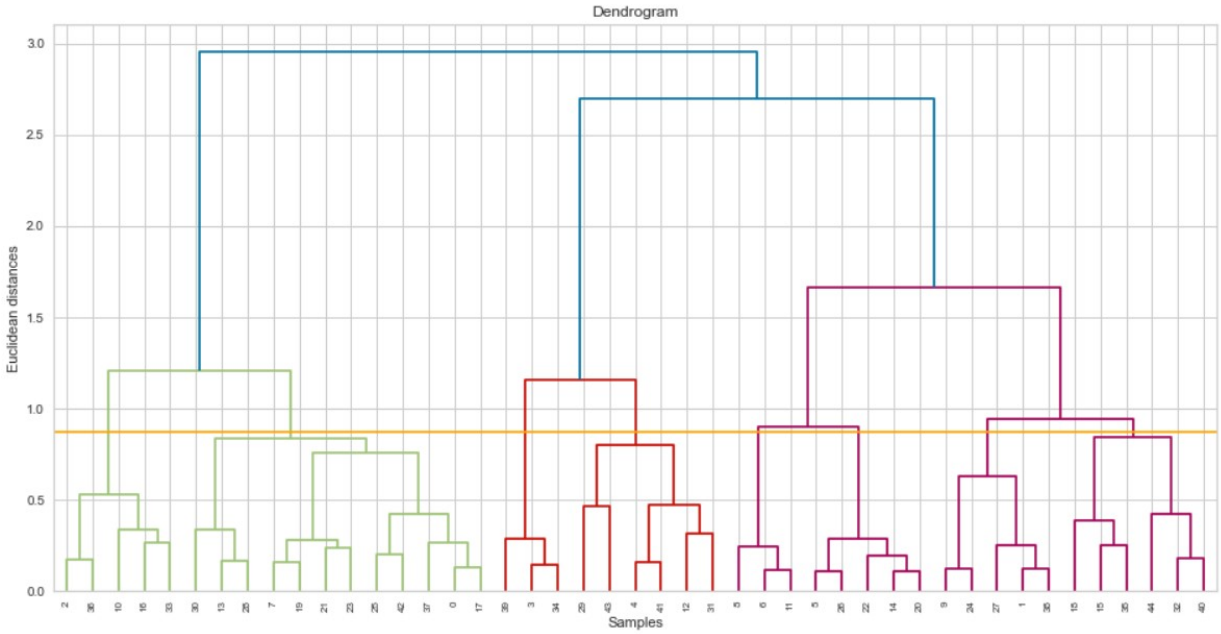

Figure 10: *Dendrogram*

4.2.3 DBSCAN

The Density-Based Spatial Clustering of Applications with Noise is an algorithm that finds core sample of high density, and expands clusters from them. Because the data looks quite dense but probably has some outliers, this could be a good method. The required arguments are the maximum distance ε between two samples for them to be considered as neighbors, and the minimum number of samples to be considered a cluster. The elbow method cannot be considered for this part because the number of clusters is not an argument, so only the silhouette score was used, looping on various ε and the minimum number of samples. It is to note that during the search for ε_{opt}, the scale was refined starting from 10^{-1} to 10^{-2}. The results are: $\varepsilon_{opt} = 0,40$ and $min_{samples,opt} = 5$.

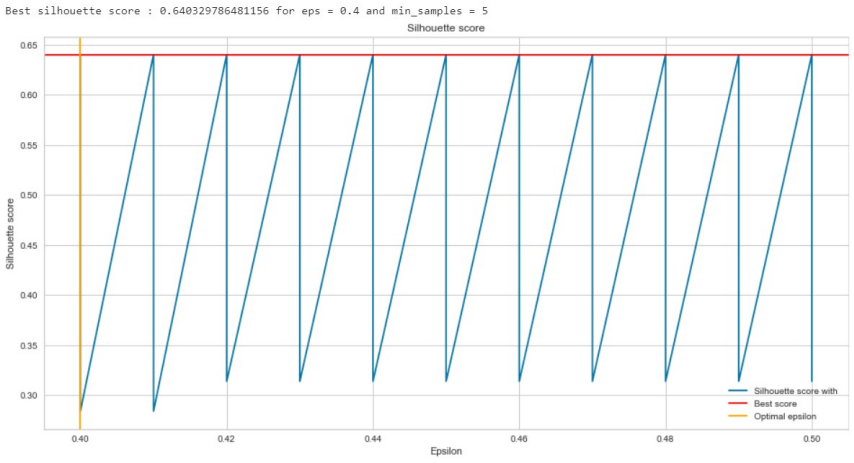

Figure 11: *Silhouette score for DBSCAN hyperparameters search*

With these parameters, the following clusters were formed :

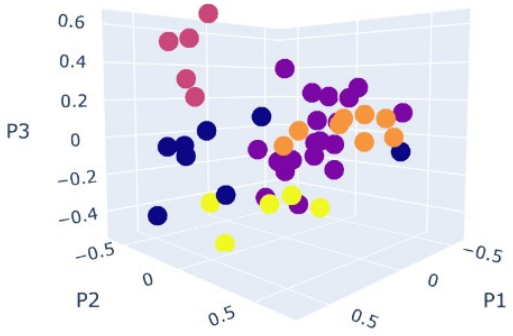

Figure 12: *DBSCAN clustering*

Cluster n°	Color	Number of samples
-1	Blue	8
0	Purple	19
1	Pink	5
2	Orange	8
3	Yellow	5

Table 4: *DBSCAN clusters*

It is important to note that samples classified as "-1" are considered outliers and not a real cluster. Besides, 4 clusters seems like a relatively sensed analysis, but the distribution is not even at all, with cluster 0 accounting for almost half of the samples, which allows to think that there might be a general way of thinking and some clusters with special ideologies.

4.2.4 Spectral clustering

Spectral clustering is a powerful technique for non-convex clusters, or when their center is not a good measure of the complete cluster (for example nested clusters in a 2D plane). The method projects the Laplacian and clusters afterwards. Both the elbow method for distortion and the silhouette score gave an optimal cluster number of $k = 9$:

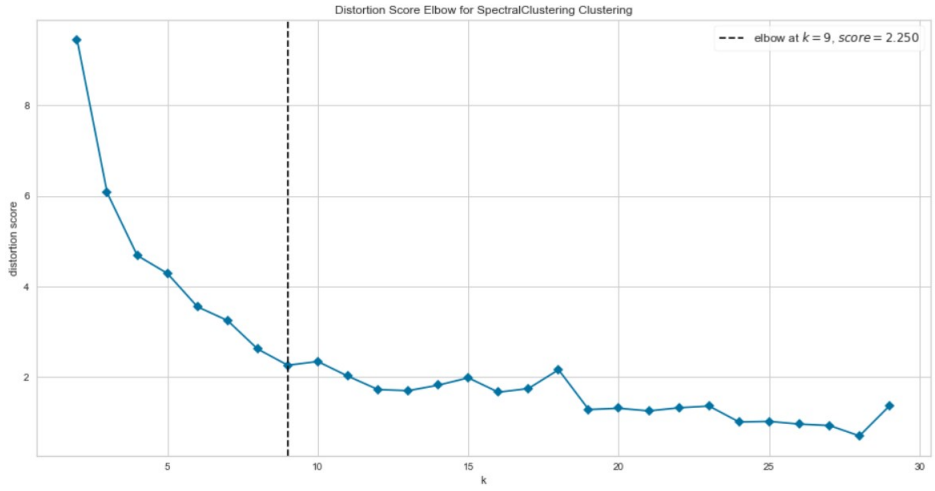

Figure 13: *Elbow method for spectral clustering*

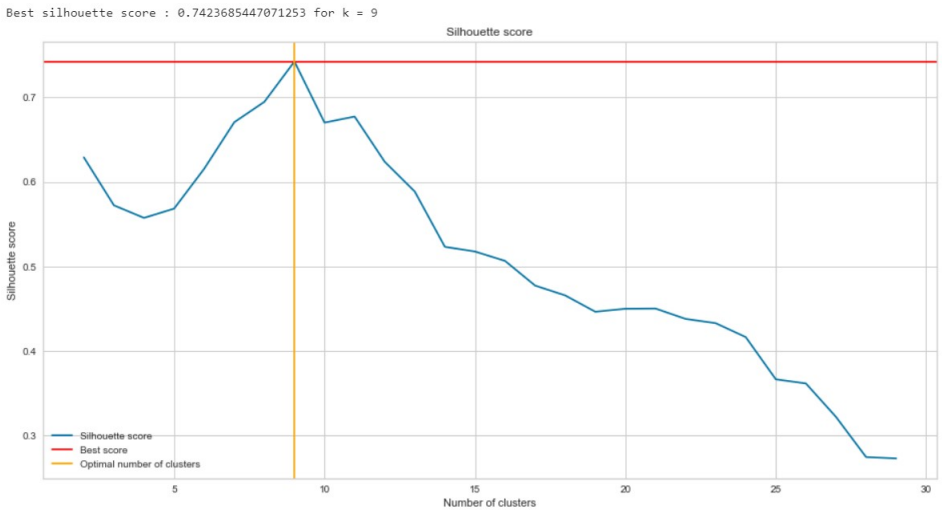

Figure 14: *Silhouette score for spectral clustering*

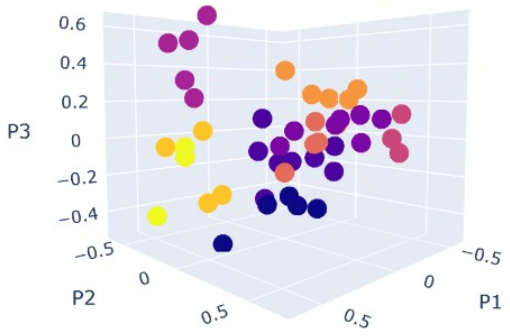

Figure 15: *Spectral clustering*

Cluster n°	Color	Number of samples
0	Dark blue	5
1	Blue	8
2	Purple	7
3	Fushia	5
4	Pink	3
5	Salmon	4
6	Dark orange	6
7	Light orange	4
8	Yellow	3

Table 5: *Spectral clustering clusters*

Again, the clusters are quite even, but their number make it quite hard to understand anything from it.

4.2.5 Gaussian Mixture Models

Gaussian Mixture models are probabilistic and assume that all data points are generated from a mix of a finite number of Gaussian distributions with unknown parameters. Because it is not a clustering algorithm per se, the distortion score is not available for this model, so only the silhouette score was used.

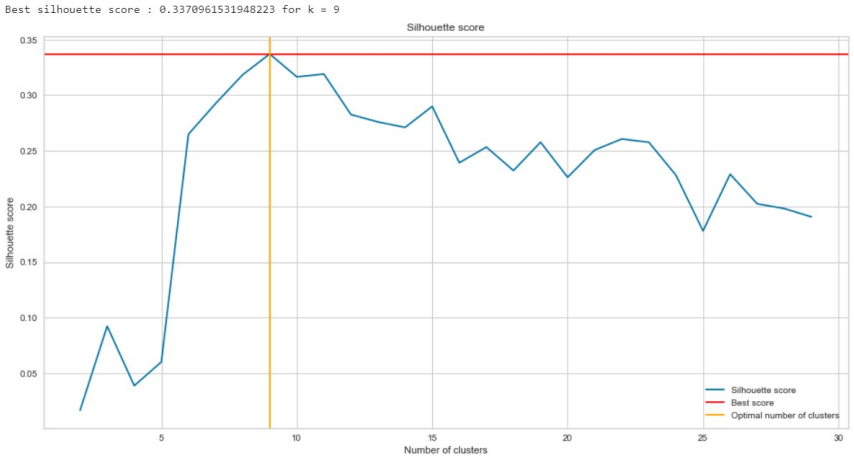

Figure 16: *Silhouette score for Gaussian Mixture Model*

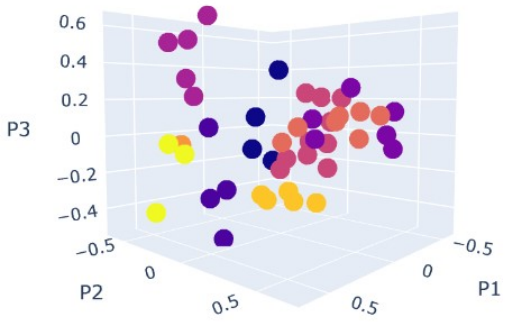

Figure 17: *Gaussian Mixture Model clustering*

Cluster n°	Color	Number of samples
0	Dark blue	4
1	Blue	4
2	Purple	6
3	Fushia	5
4	Pink	10
5	Salmon	7
6	Dark orange	1
7	Light orange	5
8	Yellow	3

Table 6: *Gaussian Mixture Model clusters*

We again have 9 clusters, that are quite hard to understand and also different from spectral clustering so it does not really help.

4.2.6 Mean shift

Mean shift is a non parametric feature-space, expectation-maximisation algorithm which locates the maxima of a density function. It is an iterative method that starts with an estimate x, and uses a kernel function $K(x_i - x)$ which computes the weight of the nearby points for (re)estimation of the mean. The weighted mean is then : $m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$, where $N(x)$ is the neighborhood of x, who verify $K(x_i - x) \neq 0$. The algorithm then performs the mean shift : $x_{new} = m(x)$ and repeats the estimation until $m(x)$ converges. Mean-shift is then defined by $m(x) - x$. Usual kernel functions are flat or gaussian.

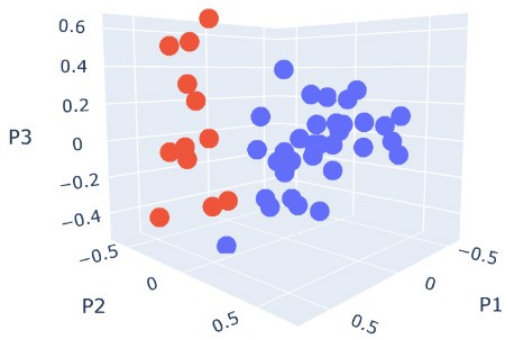

Figure 18: *Mean Shift clustering*

Cluster n°	Color	Number of samples
0	Blue	33
1	Red	12

Table 7: *Mean Shift clusters*

The results are quite different than from the other clustering methods. There are only two groups, disproportionate in size, do not seem to resemble to anything meaningful to the study.

4.3 New algorithms

All these methods are very fast and easy to implement, but because of the high dimensionality, the interpretabilty is quite low. It is hard to understand what the clusters mean only by looking at the graph and it is necessary to deep dive into the data and what it means. The following analysis was performed only on the answers to the questions contained in the five tables where it is asked to rank one situation versus another (ex : 1 person I know vs. 1 child). It was important to wonder how to analyse the data in a clever way so that it was easily interpretable and maybe easier to deduce the different ways of thinking.

4.3.1 Clustering by most common feature

The first idea was to look for the question that was answered in the most similar way by the majority of the people. To do that, a percentage of answers was calculated for each feature, and the highest was taken as "most similar feature". The feature was then removed from the dataset, and the people who answered in the most common way were put in a new table, as well as the few who thought different on that matter. It's almost a way to look for outliers, and those found near the beginning then have a rationally different way of thinking on questions that almost everybody agrees upon. This was done until every sample was singled out, for a tree depth of 30. The figure also indicates the sizes of the subsets : (samples x features).

Sadly, it is hard, and even with a dataset this small, to identify the "general way of thinking" - if it even exists - only by looking at the answers of the remaining samples in the dataset, but it seems good to find outliers.

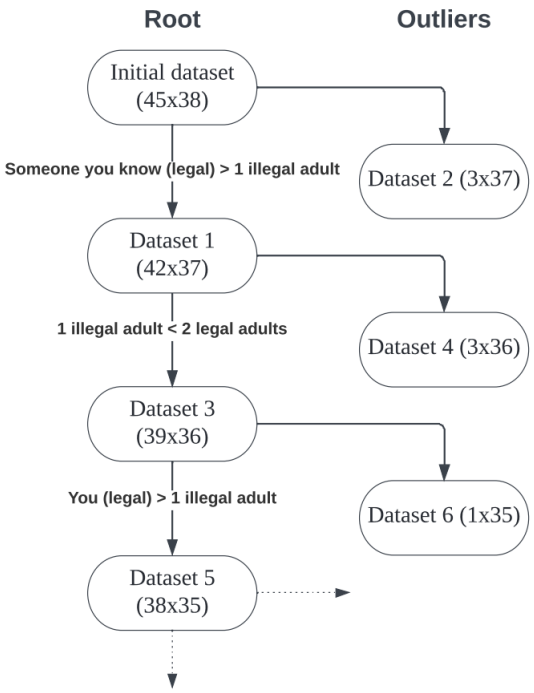

Figure 19: *Most common features*

4.3.2 Clustering by most different feature

Although the previous method is great to find outliers, it doesn't actually cluster people in different categories and does not give a general way of thinking either. By slightly adapting the functions used for the previous part, it was possible to find the feature that divided best the samples into equal quantities. The operation could then be repeated on the subsets, and so on, until once again singling out every sample.

The question that divided best the dataset was feature "$CL1g$" which corresponds to the question : when everyone is legal, how do you rank someone you know versus 2 adults ? (possible answers were $>$, $=$ or $<$). 51,11% answered that for them, someone they know is more valuable or more important than 2 adults they don't know. This split the data set into two new ones. One with all that answered like this (dataset 1), and on the other side all the other samples (dataset 2). It is important to note that among all the people who answered that for them someone they know is lesser or equal than two adults, exactly half of them said equal and half said lesser. Then the same principle was applied to the two new datasets.

For dataset 1, the new dividing feature was "$CL1e$" which asks, if everyone is legal once again, to rank yourself versus one child you don't know. 47,83% of this subset said that themselves were more important than a child and they were put in a new dataset (dataset 3). The other 52,17%, constituing dataset 4, were quite divided on the question, as 58,8% of them thought it was equal and 41,2% thought a child was more important than themselves.

For dataset 2, the dividing feature was "$CL1h$" which corresponds to ranking someone you know versus one child, in legal conditions for both. Exactly 50% think that "=" is the best answer, they'll belong to dataset 5. Among the other half(dataset 6), 72,73% think a child is more important than someone you know.

With only a tree of depth two, it is already possible to see the different ways of thinking :

- 1^{st} group : "someone I know $>$ two adults I don't know" & "myself $>$ one child"

- 2^{nd} group : "someone I know $>$ two adults I don't know" & "myself $\leq$ one child" (58,8% "$=$" & 41,2% "$<$")
- 3^{rd} group : "someone I know $\leq$ two adults I don't know" (72,7% "$=$", 27,6% "$<$") & "someone I know $=$ one child"
- 4^{th} group : "someone I know $\leq$ two adults I don't know" (72,7% "$<$", 27,6& "$=$") & "someone I know "$<$" or "$>$" one child" (72,7% "$<$", 27,6% "$>$")

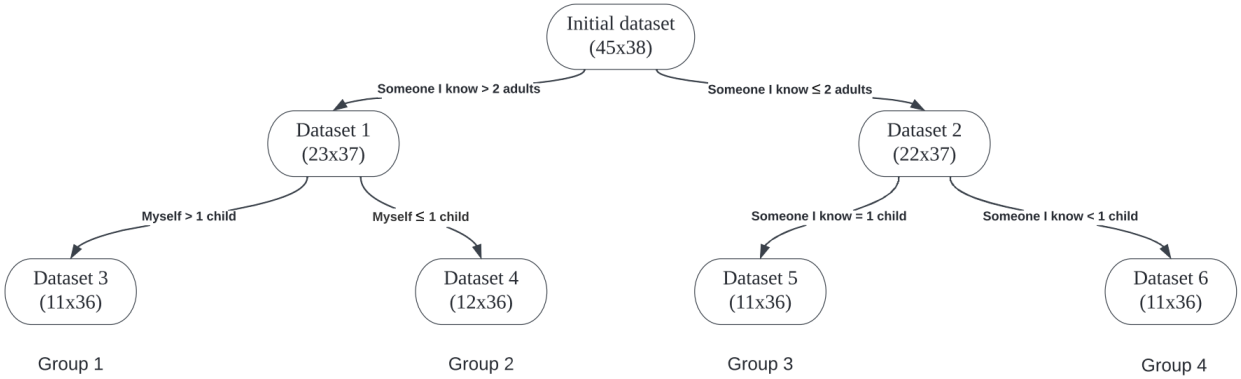

Figure 20: *Different ways of thinking*

Clearly, the first group prioritizes either themselves or people they know over number or age of other victims. The second group is more moderate and sensitive to children, even though they are more likely to save someone they know, disregarding the number of victims. The third group is also moderate in the sens that they will balance number with the fact that they know or not the victims, and feel like their life is as valuable as a child's. The fourth group will globaly always try to save the greatest number or children.

4.3.3 Merging the two ideas

The final idea is to try to remove outliers prior to separating the samples. Indeed, if a "general way of thinking" exists, the questions splitting the dataset should be very different and more into details rather than on highly quantitative or moral subjects. On the contrary, if the groups determined above are still valid, the questions may change but the thoughts of the groups should not.

First were removed the 6,7% (3 samples) that did not answer like the vast majority to $CL2f$. Almost everyone answered that when someone you know is legal, they should be prioritized over an illegal adult you don't know. The 3 outliers all said it was the same thing. Then, over the 42 remaining, again 3 were taken out concerning $CL3a$. 92,8% answered that one illegal adult had less importance than two legal adults. Again, all three said it was equal. The last one was someone who answered that when you are legal, you are equal to an illegal adult ($CL2c$), contrarily to the other 37 samples that said you should be prioritized. The removal of outliers was stopped arbitrarily here because the last split included only one sample.

Then, the same method as before was applied and gave the resulting tree (note that all determinant questions here come from table classifier n°1 in which everyone is legal):
Although the determinant questions changed, the ideology of the four groups is still very recognizable and doesn't change at all from before :

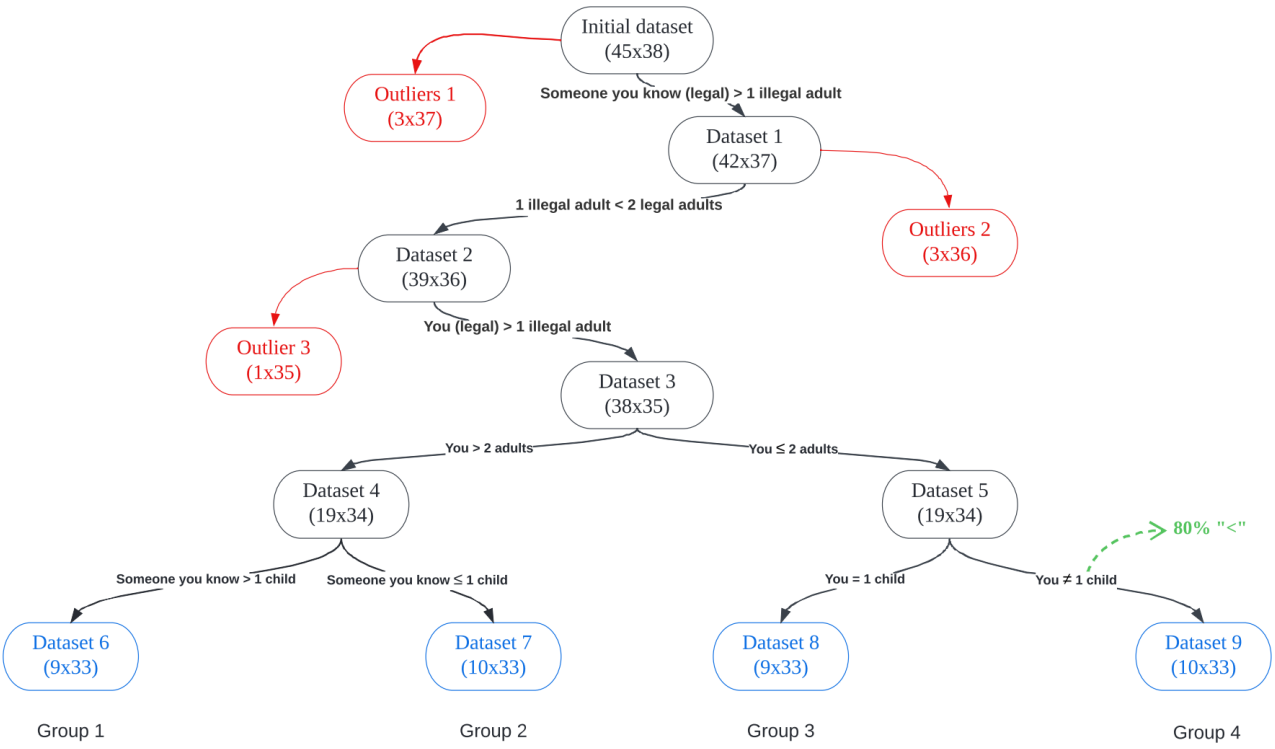

Figure 21: *Different ways of thinking after outliers removal*

- Group 1 answered that their selves should be prioritized over two adults, and someone they know over a child. This is the exact same analysis as before, these samples will favour themselves or someone they know over number or age of other victims.
- Group 2 also favoured their selves over 2 adults but did not think that someone they know should be prioritized over a child (70% said it was the same thing and 30% found a child should be more important). This again corresponds exactly to the child sensitive group 2 before.
- Group 3 disagrees with groups 1 & 2 and thinks in equal proportions that they are equal or inferior to two other adults. They also think that their life is equal to a child's. Here the ideology is slightly different than with the previous corresponding group because the questions changed, but it can still be considered as a moderate/progressive point of view.
- Group 4 also agrees with group 3 that their life is not more valuable than the ones of two adults, and they think by 80% that a child's life is more valuable than theirs. Again, the question is slightly different than for the previous part but the ideas are the same.

Removing outliers did not have a significant effect at all here, as the four clusters of different ideologies are still very present.

4.4 Results and discussion

Finally, it is very visible that usual clustering techniques do produce groups but they are very hard to interpret. Because of the complexity of the data and the difficulty to measure ethical values, handmade algorithms and less straight-forward techniques work better. However, five algorithms (K-Means with $k = 4$ and $k = 8$, DBSCAN, spectral clustering and gaussian mixture models) all identified a cluster of 5 samples (samples n°3, 7, 8, 16 and 22), which happen to be exactly the removed samples from the algorithm with the most common feature (although 32 and 43 were removed as well). This validates that the implemented new methods make

sense, but also that the classical clustering algorithms, despite being less interpretable, can still cluster the data in a sensitive way, and that the encoding also makes sense.

Moreover, an "empirical" accuracy was computed for each group. Within each group, the most common answer for each of the 13 validation scenarios was found, as well as the percentage of people within the group who chose it. This majoritarian answer was considered as the answer of the group for this scenario. Then, the majoritarian answer was compared to what the group "should" have answered considering their ideology determined just above. This is the part where this is empirical, as some scenarios can be hard to judge depending on the group. This gives a group accuracy over the 13 validation scenarios. The results are presented below:

	Group 1			Group 2			Group 3			Group 4		
Scenario ID	Prediction	Answer	% MAX	Prediction	Answer	% MAX	Prediction	Answer	% MAX	Prediction	Answer	% MAX
12	0	0	77.78	0	0	70	0	0	55.56	0	0	60
13	0	0	100	0	0	70	0	0	88.89	0	0	90
14	0	0	66.67	0	0	50	1	1	55.56	1	0	50
15	1	1	55.56	1	1	60	1	1	44.44	1	1	40
16	0	0	44.44	0	0	40	1	1	44.44	1	0	50
17	0	0	66.67	0	0	90	0	0	88.89	0	0	90
18	0	0	55.56	0	0	70	0	0	44.44	0	0	60
19	0	1 = 2	33.33	1	1	50	1	1	66.67	1	1	70
20	0	0	33.33	1	1	50	1	1	44.44	1	1	50
21	0	0	77.78	0	0	80	0	0	88.89	1	0	60
22	0	0	77.78	0	0	70	0	0	77.78	1	0	90
add1	0	3	44.44	0	1	40	0	0	55.56	0	0	70
add2	0	0	55.56	0	0	50	1	0	55.56	1	0	50
average of the MAX			60.68			60.77			62.39			63.85

Accuracy per group [%]	84.62	92.31	92.31	61.54
Accuracy per person in each group [%]	51.35	56.09	57.59	39.29

Legend:
0 : answer A
1 : answer B
2 : answer C
3 : answer D

Figure 22: *"Empirical accuracy"*

The way to interpret the table is as follows. Each group has a sub-table which contains 3 columns: the first is the predicted answer according to the group and thus the cluster, the second column contains the answer that was answered most often within the group and the third column indicates the percentage of people within this group who answered this most frequent answer.

The first thing is that the groups answered with about 80% accuracy **as a group** (with the majoritarian answer) to the scenarios. However, because samples within a group only have answers about 60% alike, this pulls down the individual accuracy to just over 50%. This is really a bad result but there are leads to improve it :

- Make more clusters (8 for example), in order to have "denser" clusters (interpretation of the clusters will get more difficult with more clusters).
- Fully describe and quantify the scenarios and use it as a part of learning to build a model more in adequation with the data rather than choosing "left or right". This seems like the best path to follow because it is the closest to reality. Indeed, the car will have all the scenario information just before making the decision, so it should be considered as an input.

5 Conclusion

This work has shown that it is possible to separate the population into clusters of thoughts and ethics. It would therefore be possible to adapt the decisions made by their own autonomous vehicle as far as possible. However, it would be preferable to develop questions less related to real cases. Indeed, it could be relevant to end up with questions that have no apparent connection with the topic (example given as an idea: "Do you prefer apples or pears?). Indeed, if the answers to these "off-topic" questions can be used to classify each person, it would then be possible to predict their ethics in the context of transport. This type of questions would allow the new user not to realise what he is influencing in his new car and therefore he would not have the possibility to adapt his answers according to who he wants to be and not who he really is. The answers to the questions would be quick, automatic and intuitive. Therefore, a lot of work and research still needs to be done. First of all, the results of this project need to be confirmed on a larger scale. Secondly, is it possible to reduce the number of questions to a minimum? And finally, how can these questions be transferred to questions that are seemingly unrelated to the topic of ethics in transport?

With regard to the work done in this project, the question of extreme cases remains open. What to do with these people? How to deal with them? Is it better to treat them separately so that they do not influence the results of the majority of people? Or on the contrary, is it necessary to treat them like the rest of the population? More time would have allowed us to think about how to quantify the scenarios in order to create a model that would allow the learning and testing parts to be used appropriately. It would then have been possible to determine whether the learning part is sufficient or not. A lack of information would have made it possible to add questions or learning scenarios. As for the surplus of information, it would have been possible to remove questions containing redundant information.

Although research is still in its infancy, autonomous vehicles are already operating in some parts of the world. For example, Waymo operates a fully autonomous robot taxi service in Arizona, where it has already made "tens of thousands" of trips without a human driver. In addition, Argo AI, a company backed by Volkswagen and Ford, among others, recently announced that it was starting to test driverless cars in Austin, Texas, and Miami, Florida. It is therefore clear that it is time to address the ethical challenges as soon as possible. This project focused on the appearance of the car and its occupants (of which the driver is the respondent). However, it might be interesting to look at the subject from another angle. Pedestrians are used to seeking eye contact with the driver of a vehicle when they are about to cross a pedestrian crossing, for example. So how can we take into account that this eye contact will be lost when the driver of the car is no longer human but artificial? Jaguar Land Rover has proposed a first solution by imagining a self-driving car with eyes that would allow the autonomous vehicle to interact and communicate with the pedestrians when it's safe to cross in front of them. Replacing the eyes of a human driver, the cars make eye contact with nearby human beings to acknowledge that they've seen them and that they will stop. The pedestrians can feel safe to cross the road.

The last question that imposes itself is the following : are people willing to use such softwares ? According to the data collection, 43% of them would rather use a generic one set by laws for example, if they were to ever use an autonomous vehicle at all.

6 References

- https://www.moralmachine.net/
- https://www.merriam-webster.com/words-at-play/trolley-problem-moral-philosophy-ethics
- https://www.nature.com/articles/s41586-018-0637-6
- http://web.mit.edu/francisc/www/24_131_Paper1.pdf
- https://siecledigital.fr/2022/06/06/cruise-san-francisco-permis/
- https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/MoralMachineMonster.pdf
- https://scikit-learn.org/stable/
- https://www.dezeen.com/2018/09/04/jaguar-land-rovers-prototype-driverless-car-makes-eye-contact-pedestrians-transport/