Bias and Morality Judgment AI

ENSE 480 Information Systems

Evan Geissler

200331033

Table of Contents

1. Introduction

1.1 Trolley Problem

The trolley problems are “series of hypothetical scenarios developed by British philosopher Philippa Foot in 1967. Each scenario presents an extreme environment that tests the subject's ethical prowess.” (Dean) Along with ethical prowess, these questions also challenge a person’s beliefs, morals, and idea of themselves i.e. if they were actually in this situation would the truly feel they would react in the way they said they would. These problems can help show a person’s decision-making and how they would react. It is important to note that different people could have very similar reactions, but very different thought processes. For example, a person may wonder if others will judge them, if it would affect who they are, if anyone would see, or if their actions are morally ok.

1.2 Why is it Relevant to AI?

The trolley problem is relevant to AI because it can help to assist in decision making of many AI systems. Since there are so many inputs of what could affect a decision, it is important to have some form of judgment and predefined constraints such as what is and is not morally ok. Practical applications of the problem and the information gathered can be applied to self-driving cars and robotics. Self-driving cars may come across instances where an accident will result regardless what the car does. By utilizing the trolley problem, the AI can make an informed decision on what would be considered the best result to lessen the affects of an accident. Robotics can use the problem by helping in their decisions and to be modeled after human behaviour. If future robots follow Isaac Asimov’s Three Laws of Robotics, then difficult situations and how the robot would react would need to be addressed. For example, a paradox can be created if Asimov’s first law, “a robot cannot injure a human being, or, through inaction, allow a human being to come to harm” (NASA) is presented in a situation where the robot has to choose to save a human, but the action harms another human and if it does nothing than that human is harmed.

1.3 Bias and Morality Judgment AI

The AI system that I have created uses a variation of the trolley problem that has different options and uses a self-driving car. Below is an example of a basic question that the AI uses:

> A Self-Driving Car is heading down a road when 1 person walks in front of the car. The car will either continue its path OR the car can swerve off the road hitting 2 people. Should the car:
> a. Continue its course OR
> b. Swerve off the road?

Each time the AI asks one of these questions the variables that define the continue option and/or swerve option are changed. The aim of the AI is to have the user always make the choice to swerve the car. This shows that the user made the decision to intervene and directly change the outcome of the accident.

To ask a question, the system gets information about the user (age, weight, and sex) and then asks the user initial questions to see if they have a possible bias against a list of traits. For example, an initial question that asks if a male user would continue to hit 1 person or swerve to hit 1 woman would increase the bias towards women and sex if the user picks to hit the woman. The biases that are asked in each question are used to continually update the list of biases and provide more detailed information about the user to the AI that will help decide which question will be asked next. Since the aim of the AI is to pick a question that will make the user choose option b, i.e. to swerve the car, the AI will try to find a question that the user has the most and least bias connection to. The system will also ask questions based on heuristics gained and learned from previous trails if available.

The AI will be successful if, after 10 questions asked, the user chooses b each time; semi-successful if b is picked 60% or more; and non-successful if b is picked less than 60%.

1.4 Scope

The scope of this project is to allow the AI to ask semi-specific questions to the user based on a limited set of traits and biases. This includes the ability to ask a single bias, such as male or female and also an amount such as 1 or 2 people. The system should also have some ability to learn from questions asked and from the results each time the program runs (each trial). This will involve saving to and reading in input from a text file so the AI can use heuristics and use machine learning. Finally, the AI should be able to ensure a user does not answer questions too fast and that questions are not repeated.

2. Techniques & Algorithms

2.1 Directive Management

Directive management is a set of algorithms that helps to ensure the program, user, and different aspects of the AI are kept on track while the system moves towards a goal or end state. By doing this, the manager can ensure the user is answering questions more effectively and to ensure that the system is not repeating questions. To do this, the manager checks to see if a user is answering questions too fast which would affect bias calculations and the seriousness of the user. The manager also checks to see if the AI picks a question that it has already asked the user. By repeating questions, the AI would add redundant information to the bias calculations.

To check if a user has answered a question too fast, a timer is started when a question is asked. After the user answers, the duration of the timer is checked against a minimum time limit and the question is asked again if the duration is lower than the limit. To ensure the AI has not repeated a question, all asked questions are stored in a 1D array. The manager checks through each set of values in the array against the new question that has just been created. If the question has been asked, then the AI is forced to pick the next option available. If the question has not been asked, then the AI will ask the user this question.

If the directive management was not included, the program becomes unsuccessful or flawed by gathering poor or incorrect information. Since the AI also learns from each trial of questions asked, future heuristics from the machine learning would have cascaded effects on future trials. Other techniques for management were not thought of.

2.2 Heuristics

Heuristics are used in the system to help ask more successful answers first in hopes that the user will choose b before more information about the user's biases are obtained. This is done by loading information from a text file about past trials into a 2D array; seeing if past questions have a minimum success rate and a minimum number of people asked; and finally asking all questions that have these two requirements met. The heuristic algorithms may not be used if the program is running for the first time or if past questions do not meet the minimum requirements. However, the questions that are asked have no guarantee to be successful, but the more information that is gathered the more likely the heuristics are correct in asking specific questions.

The text file that is read by the heuristic algorithm stores 4 values for each possible combination of questions. The 4 values, in order, are the continue choice; swerve choice; number of people asked; and the success rate from 0.0 to 1.0. The 2D array that is used to pick the next question also follows this order.

Another technique that was thought of was to have the biases in a graph. The AI would find the next question choices by searching through the graph by depth first, breadth first, or A* searching. Each edge would have a different weight to it that would be used to decide which question would be most successful. The paths to get to each choice would be saved and could be jumped to instead of using depth or breadth first searches. This option did not seem very efficient and would be more time consuming to run and implement than the current method used.

2.3 Machine Learning

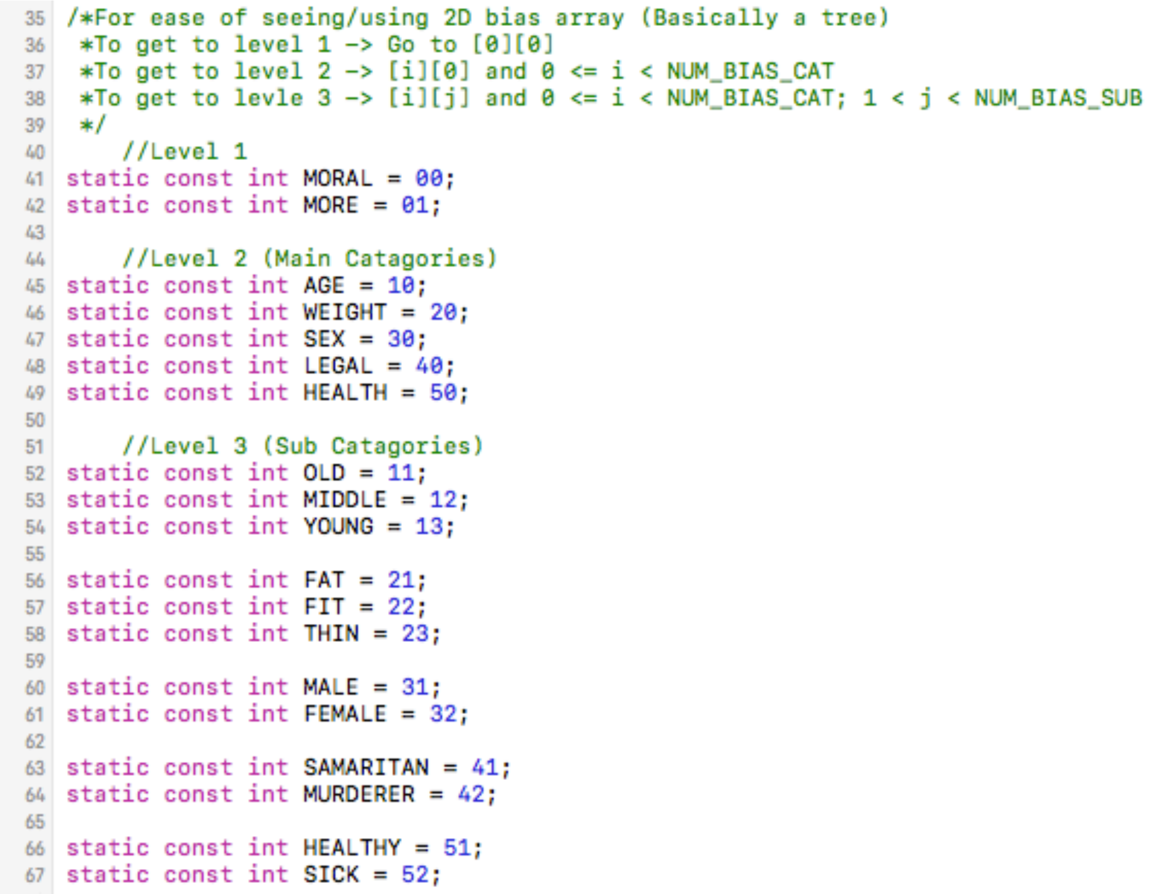
After a user answers a question, the AI saves the users answer and the question asked. At the end of the program, the AI goes through the values of the heuristic 2D array discussed above and the questions that were asked in the current trial. If a question matches a previously asked question then the number of people asked increases and a new success rate is calculated by checking the saved answer. If the current trial asks a question that has not been asked before, then the continue and swerve choices are added with 1 person asked and either a 1.0 or 0.0 success rate given. The information in the 2D array is then stored in the text file that the AI read in for the heuristics. This allows the heuristics to be constantly updated with new questions or better-tested questions with a more accurate success rate. Updating the heuristics helps to give the AI a better chance of success or, if the AI becomes unsuccessful can continue to be more successful.

The main alternatives I thought of for machine learning was to use log files or save directly into the .cpp or .h files. I decided against these options as I am unfamiliar with log files and I also did not want someone to access the program code to change or see the saved information. Saving in file would also require me to recompile every time a change was made to the values during testing, which would increase testing time.

3. Knowledge Representation & Data Structures

3.1 Bias Information

To represent biases I used constant integer codes relating to each one. The codes are made to make looking up values easier. Biases are broken into different groups that would let the code show if it was a category, subcategory, or the morality bias. The morality bias was represented as 00 and each group of biases would be shown as values starting at 10, 20, 30, etc. Category biases, such as age, are stored at the first value and each subcategory, such as old age, would start at the next number. This would allow uniform retrieval of specific biases.



```
/*For ease of seeing/using 2D bias array (Basically a tree)
 *To get to level 1 -> Go to [0][0]
 *To get to level 2 -> [i][0] and 0 <= i < NUM_BIAS_CAT
 *To get to levle 3 -> [i][j] and 0 <= i < NUM_BIAS_CAT; 1 < j < NUM_BIAS_SUB
 */
    //Level 1
static const int MORAL = 00;
static const int MORE = 01;

    //Level 2 (Main Catagories)
static const int AGE = 10;
static const int WEIGHT = 20;
static const int SEX = 30;
static const int LEGAL = 40;
static const int HEALTH = 50;

    //Level 3 (Sub Catagories)
static const int OLD = 11;
static const int MIDDLE = 12;
static const int YOUNG = 13;

static const int FAT = 21;
static const int FIT = 22;
static const int THIN = 23;

static const int MALE = 31;
static const int FEMALE = 32;

static const int SAMARITAN = 41;
static const int MURDERER = 42;

static const int HEALTHY = 51;
static const int SICK = 52;
```

Figure 1. Constant Bias Variable Codes

To store the actual value of each bias, a 2D array was used and was influenced by the original bias graph design. The graph was considered instead of arrays, but the arrays were easier to implement, easier to design, and made more sense for how the program was using and calling the information. To format the array, the first column was designed to be the main category bias. Each subsequent column of the array is the sub category biases and each row corresponds to an individual group of biases, such as age. If there are any places where a bias would not exist, a -1 is used as a placeholder. To find a specific value in the bias array, the bias codes are used. To find the row, the code is divided by 10 and to find the column, the code is mod divided by 10.

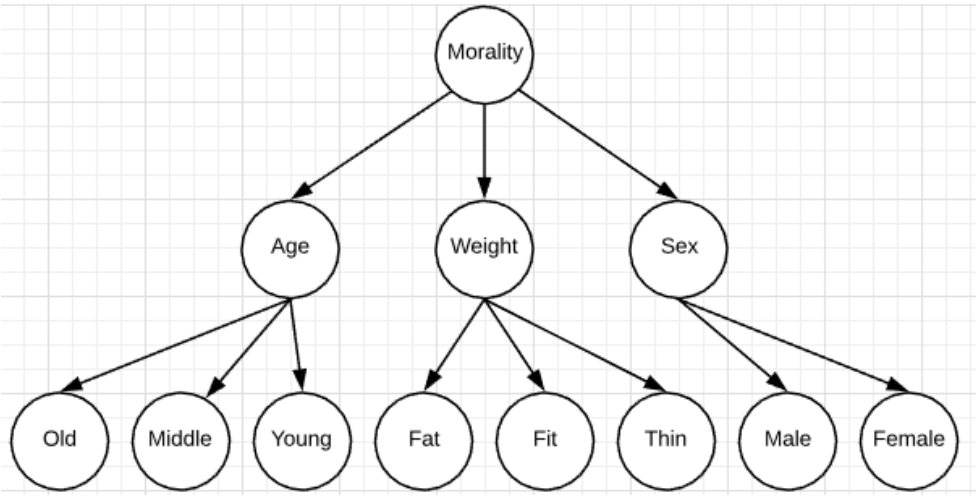
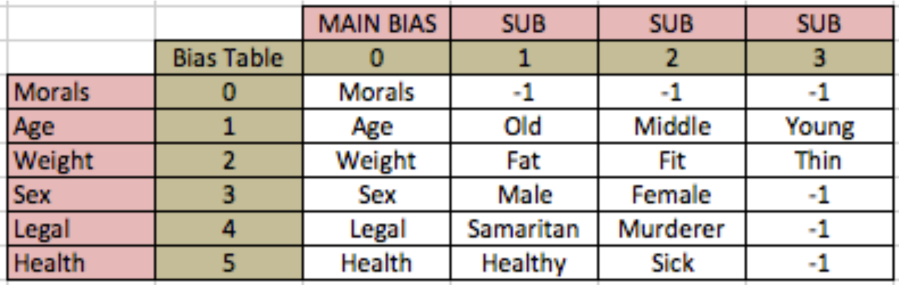
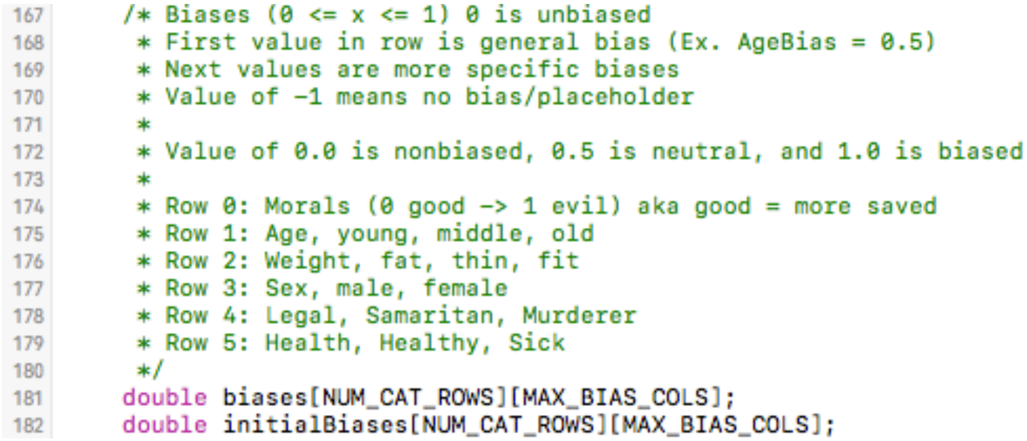


Figure 2. Graph Representation of Bias Categories



		MAIN BIAS	SUB	SUB	SUB
	Bias Table	0	1	2	3
Morals	0	Morals	-1	-1	-1
Age	1	Age	Old	Middle	Young
Weight	2	Weight	Fat	Fit	Thin
Sex	3	Sex	Male	Female	-1
Legal	4	Legal	Samaritan	Murderer	-1
Health	5	Health	Healthy	Sick	-1

Figure 3. Table Showing 2D Bias Value Array



```
    /* Biases (0 <= x <= 1) 0 is unbiased
     * First value in row is general bias (Ex. AgeBias = 0.5)
     * Next values are more specific biases
     * Value of -1 means no bias/placeholder
     *
     * Value of 0.0 is nonbiased, 0.5 is neutral, and 1.0 is biased
     *
     * Row 0: Morals (0 good -> 1 evil) aka good = more saved
     * Row 1: Age, young, middle, old
     * Row 2: Weight, fat, thin, fit
     * Row 3: Sex, male, female
     * Row 4: Legal, Samaritan, Murderer
     * Row 5: Health, Healthy, Sick
     */
    double biases[NUM_CAT_ROWS][MAX_BIAS_COLS];
    double initialBiases[NUM_CAT_ROWS][MAX_BIAS_COLS];
```

Figure 4. 2D Bias Array Code

3.2 Questions & Results

To represent an individual question I have each separated into 5 parts. This was needed to interchange the two choices being asked; to keep proper grammar structure; and to ensure the question sounded correct. The 5 parts of a question are:

1. The question intro
2. The continue choice
3. The middle of the question
4. The swerve choice
5. The remaining question

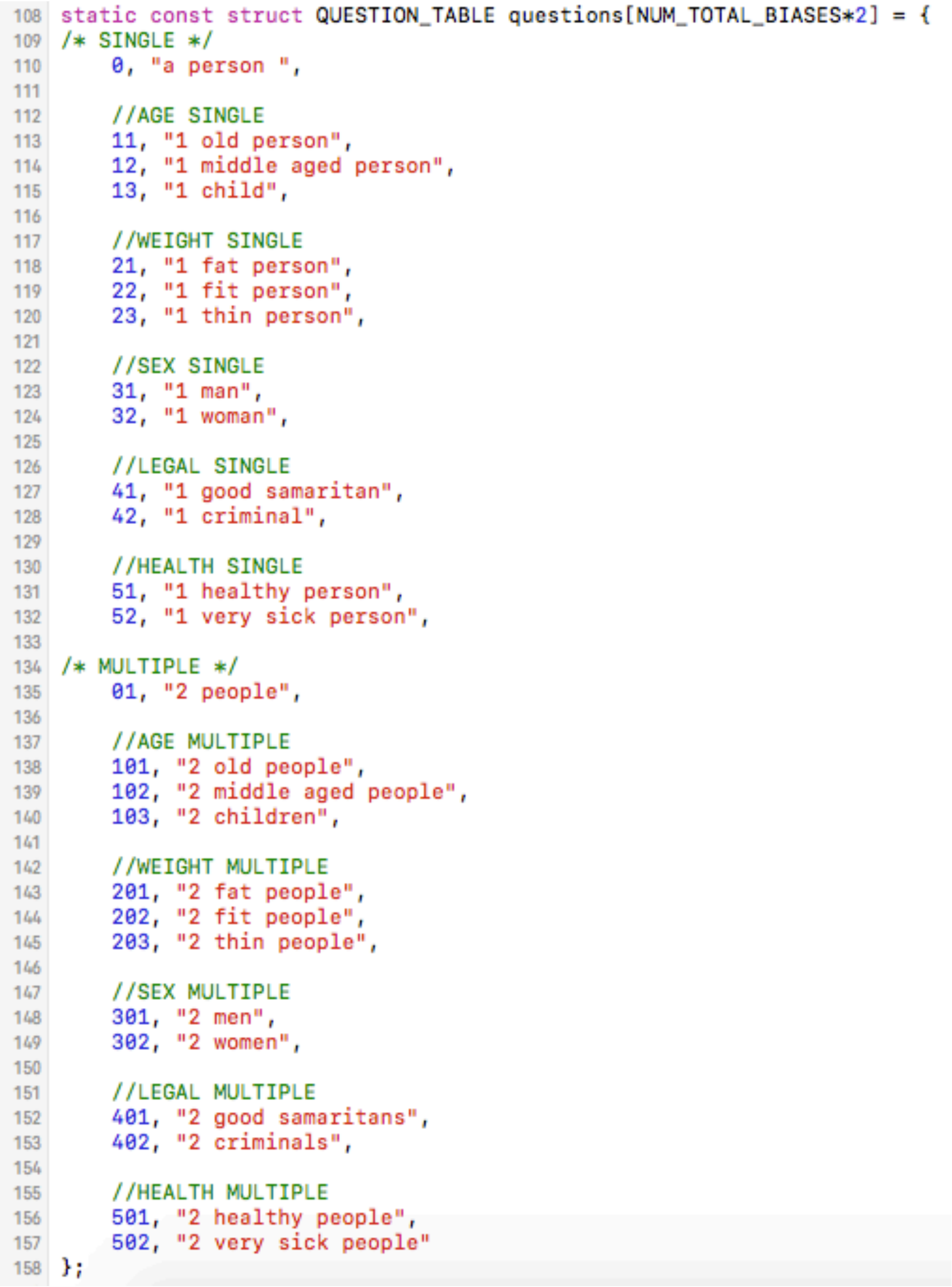
Parts 1, 3, and 5 do not change and are displayed by using print functions. Parts 2 and 4 require the AI to first decide what options it will ask, search the correct statement for each, and finally display it in order.

For example, in the question below part 1 is highlighted in blue, part 2 is highlighted in red, part 3 is highlighted in yellow, part 4 is highlighted in green, and part 5 is highlighted in grey.

> A Self-Driving Car is heading down a road when 1 person walks in front of the car. The car will either continue its path OR the car can swerve off the road hitting 2 people. Should the car:
> a. Continue its course OR
> b. Swerve off the road?

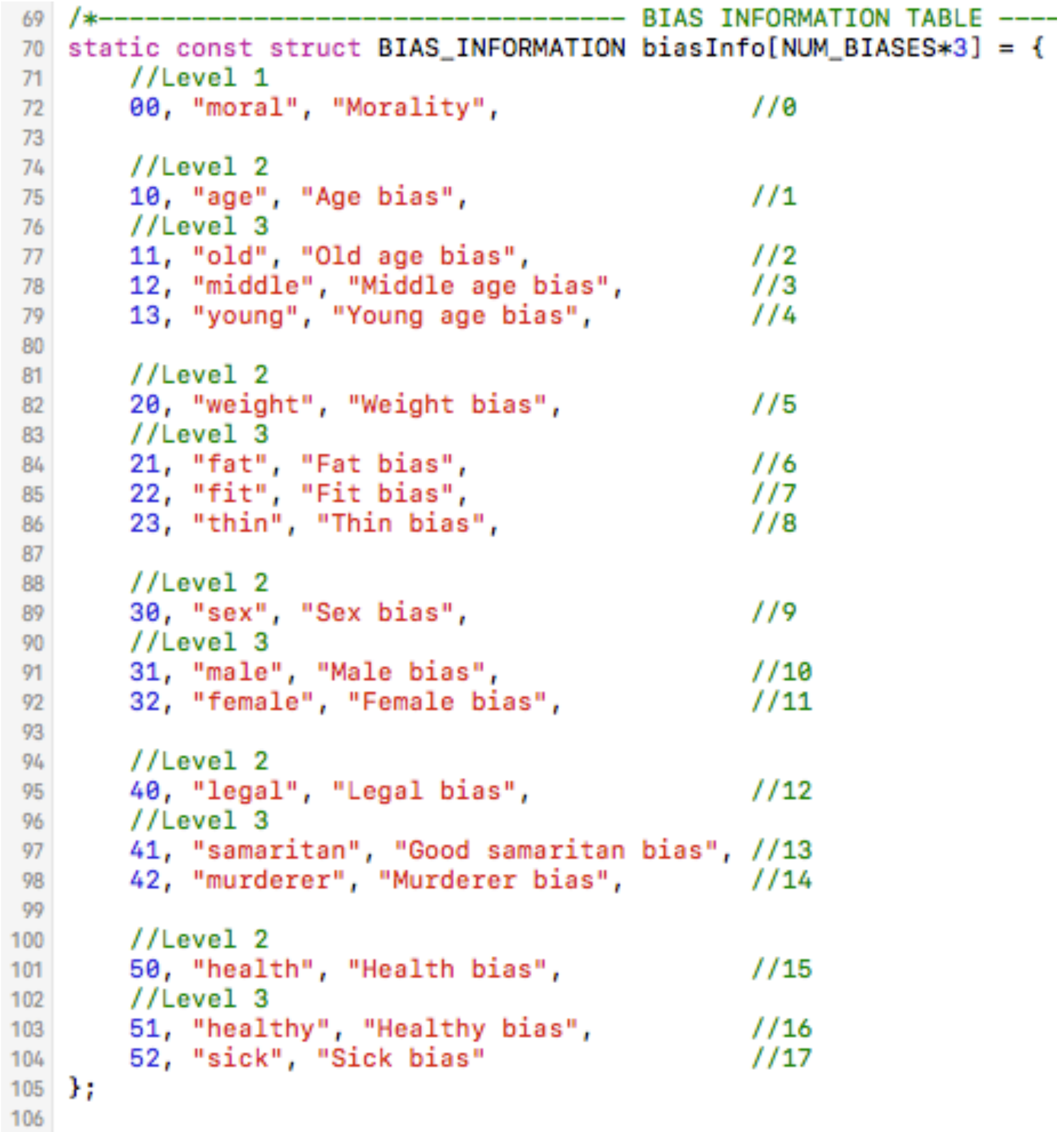
To properly ask parts 2 and 4, consideration for pluralisation and grammar was essential. To do this, I created a struct table of all possible options for both singular and plural questions. By using this table, I can quickly and easily see that each option follows the correct format to properly ask any question. To find any option, a look up code is used by the AI, which corresponds to each bias. Each option that is below 100 is for singular use and any option above is for plural options. To quickly look up values, the AI checks if the value is above or below 100. It then takes the code and can use value/10 or value/100 to get the category and value%10 or value%100 to get the sub category.

To represent the current and final results, another struct table is used with a look up code. By using this table, the AI can cycle through each bias in it, print the bias name, and then use the code to find the bias value in the bias value 2D array. The code works the same way to find a specific bias as it does in the question table. I used structs for both of these so I could easily see and search for information. Since I wanted to have an integer code and a string, I felt that using a struct was also beneficial as it can hold different variable types.



```
static const struct QUESTION_TABLE questions[NUM_TOTAL_BIASES*2] = {
/* SINGLE */
    0, "a person ",

    //AGE SINGLE
    11, "1 old person",
    12, "1 middle aged person",
    13, "1 child",

    //WEIGHT SINGLE
    21, "1 fat person",
    22, "1 fit person",
    23, "1 thin person",

    //SEX SINGLE
    31, "1 man",
    32, "1 woman",

    //LEGAL SINGLE
    41, "1 good samaritan",
    42, "1 criminal",

    //HEALTH SINGLE
    51, "1 healthy person",
    52, "1 very sick person",

/* MULTIPLE */
    01, "2 people",

    //AGE MULTIPLE
    101, "2 old people",
    102, "2 middle aged people",
    103, "2 children",

    //WEIGHT MULTIPLE
    201, "2 fat people",
    202, "2 fit people",
    203, "2 thin people",

    //SEX MULTIPLE
    301, "2 men",
    302, "2 women",

    //LEGAL MULTIPLE
    401, "2 good samaritans",
    402, "2 criminals",

    //HEALTH MULTIPLE
    501, "2 healthy people",
    502, "2 very sick people"
};
```

Figure 5. Question Table Code



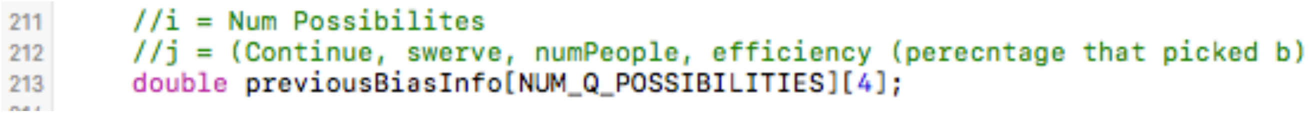
```
/*---------------------------------- BIAS INFORMATION TABLE ----
static const struct BIAS_INFORMATION biasInfo[NUM_BIASES*3] = {
    //Level 1
    00, "moral", "Morality",                //0

    //Level 2
    10, "age", "Age bias",                  //1
    //Level 3
    11, "old", "Old age bias",              //2
    12, "middle", "Middle age bias",        //3
    13, "young", "Young age bias",          //4

    //Level 2
    20, "weight", "Weight bias",            //5
    //Level 3
    21, "fat", "Fat bias",                  //6
    22, "fit", "Fit bias",                  //7
    23, "thin", "Thin bias",                //8

    //Level 2
    30, "sex", "Sex bias",                  //9
    //Level 3
    31, "male", "Male bias",                //10
    32, "female", "Female bias",            //11

    //Level 2
    40, "legal", "Legal bias",              //12
    //Level 3
    41, "samaritan", "Good samaritan bias", //13
    42, "murderer", "Murderer bias",        //14

    //Level 2
    50, "health", "Health bias",            //15
    //Level 3
    51, "healthy", "Healthy bias",          //16
    52, "sick", "Sick bias"                 //17
};
```

Figure 6. Bias Information Code for Output

3.3 Directive Management, Heuristics, & Machine Learning

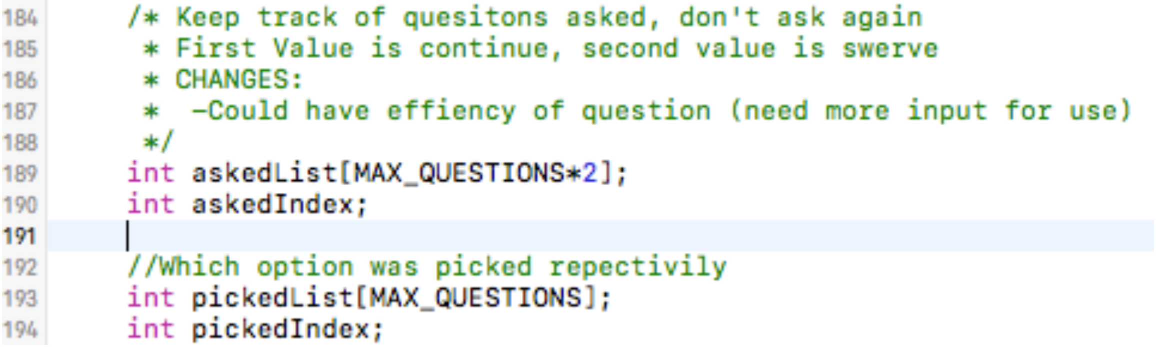
To represent the heuristics, a text file is used that holds hold of 4 values that, in order, are the continue choice; swerve choice; number of people asked; and the success rate from 0.0 to 1.0. However, these values are set to -1 show any questions not asked yet and to help shorten searching. In the file, there is a row for every possible combination of questions that can be asked. These values are then read into a 2D array that is used to check if heuristic questions can be asked. This array is called previousBias[][] and can store i = total number of possible questions and j = 4.

To use the directive management two arrays were used to compare against the current questions picked by the AI that may or may not be asked. These are a 1D array called askedList that holds every set of question choices already asked in the current trial. The first value is the continue choice and the next value is the swerve choice. The other array used is a 1D array called pickedList. This value corresponds to the answer the user picked relating to the choices in the askedList. To represent the machine learning, the previousBias array is used again and updated. This array is then placed back into the text file, overwriting old values.



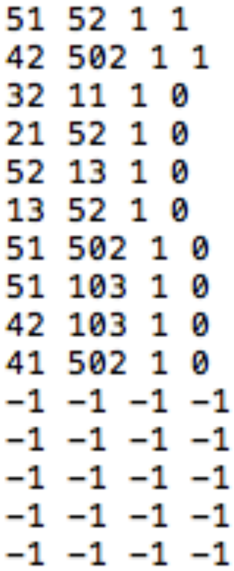
```
    //i = Num Possibilites
    //j = (Continue, swerve, numPeople, efficiency (perecntage that picked b)
    double previousBiasInfo[NUM_Q_POSSIBILITIES][4];
```

Figure 7. 2D Array of Previous Questions



```
    /* Keep track of quesitons asked, don't ask again
     * First Value is continue, second value is swerve
     * CHANGES:
     *  -Could have effiency of question (need more input for use)
     */
    int askedList[MAX_QUESTIONS*2];
    int askedIndex;
    
    //Which option was picked repectivily
    int pickedList[MAX_QUESTIONS];
    int pickedIndex;
```

Figure 8. 1D Asked and Picked Arrays



```
51 52 1 1
42 502 1 1
32 11 1 0
21 52 1 0
52 13 1 0
13 52 1 0
51 502 1 0
51 103 1 0
42 103 1 0
41 502 1 0
-1 -1 -1 -1
-1 -1 -1 -1
-1 -1 -1 -1
-1 -1 -1 -1
-1 -1 -1 -1
```

Figure 9. Example of Input Text File

4. Structure Diagrams

4.1 Class Diagrams

The class diagram in Figure 10 shows that the main file has both an AI and Player. The player in main is used to gather the user input and then copy into to the AI. The AI class is needed to run all the questions, bias calculations, etc. The AI also contains a player to get player information and use it for bias calculations and question asking.

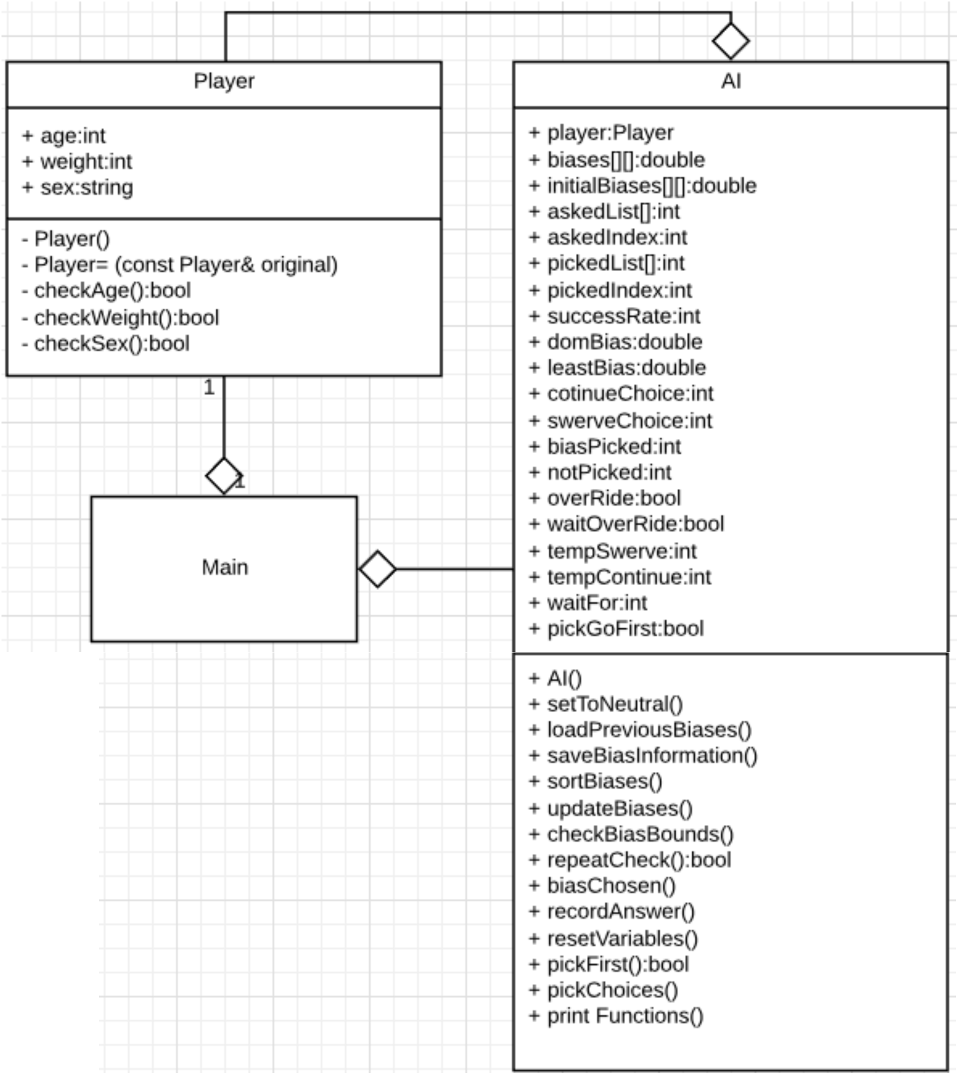


Figure 10. Class Diagram

4.2 Flow Chart of Program

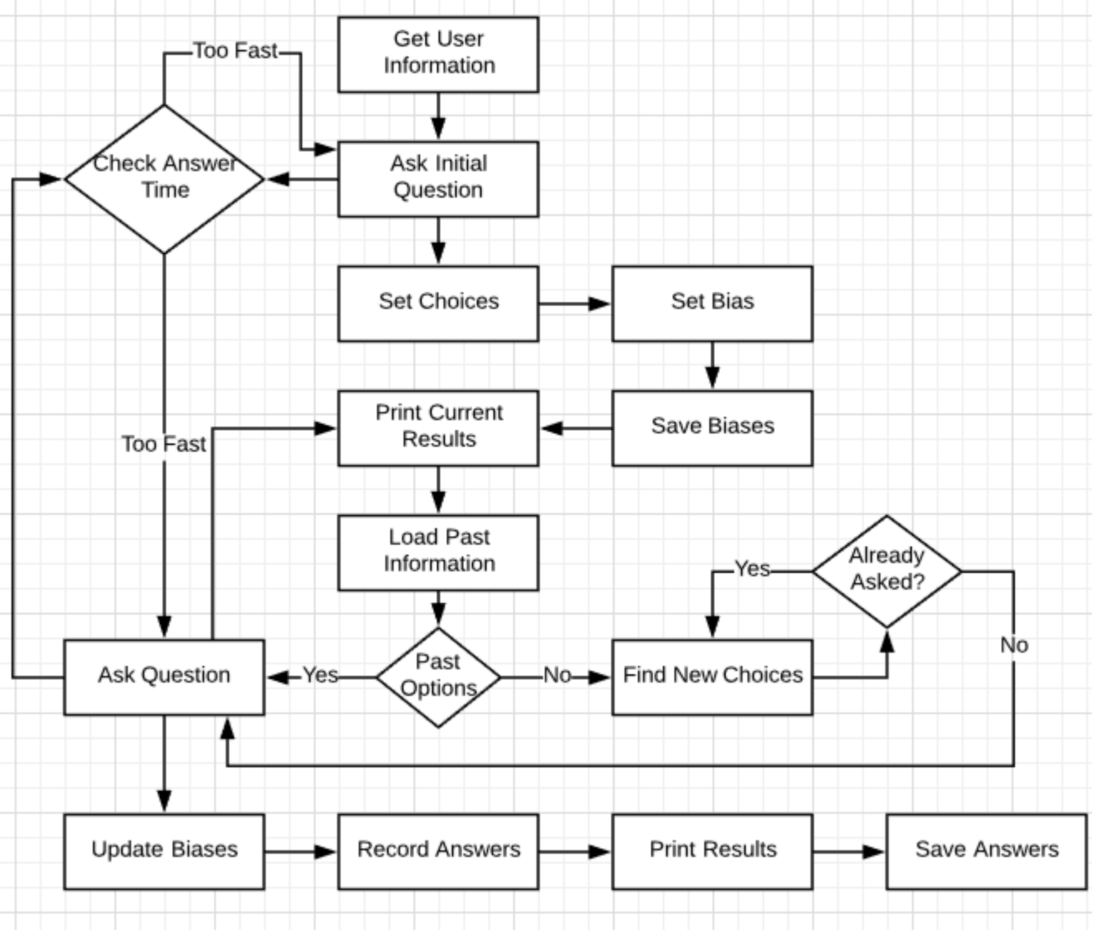


Figure 11. Sample Flow Chart of Program

5. How to Use

To use the Bias and Morality Judgment AI, the files need to be linked and compiled by either a command line interface (CLI) or program like Visual Studio. Since the project was tested and run through a CLI on a MacBook Pro, it will be outlined how to use the program through a terminal or other CLI. To run the program the user must:

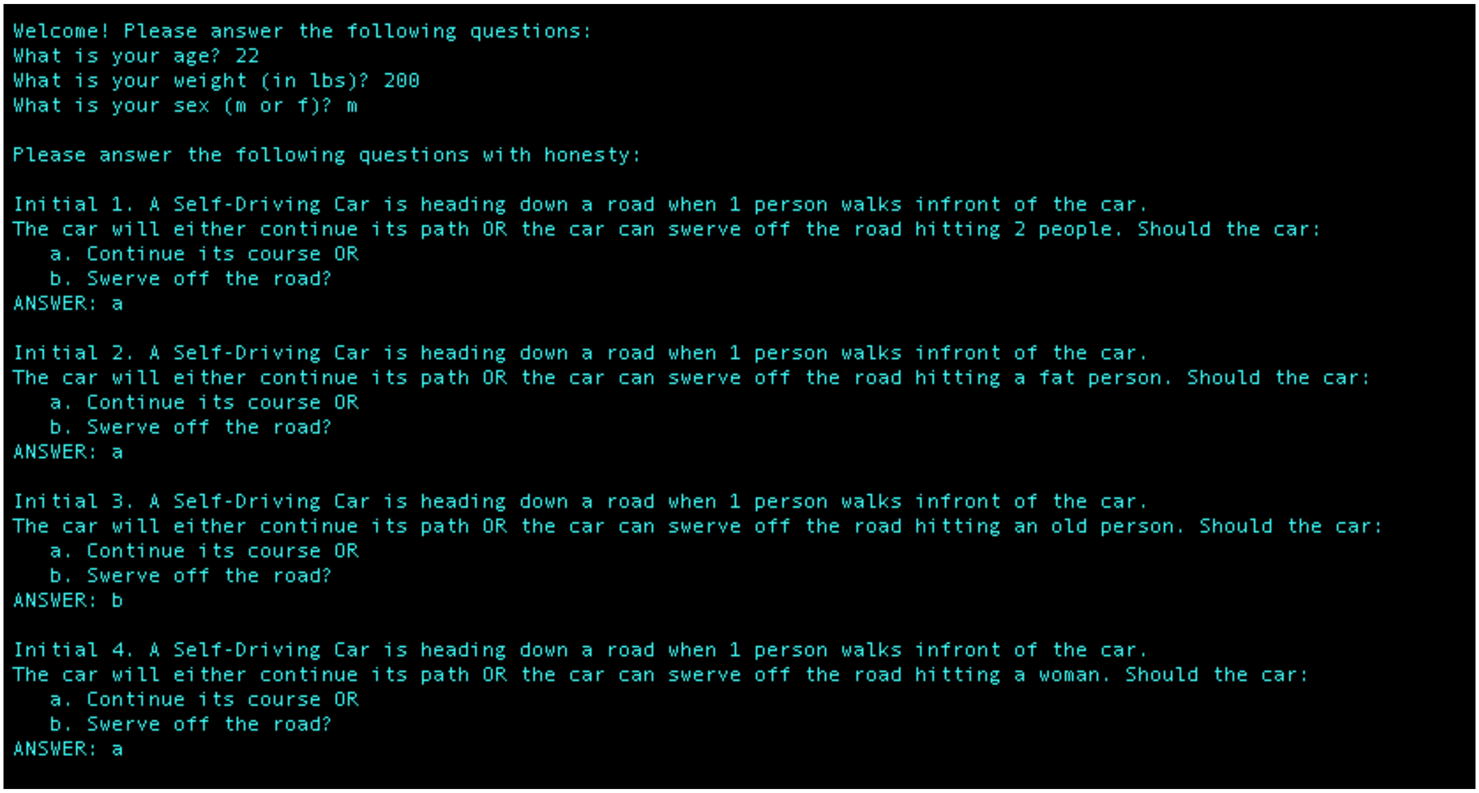
1. Open up the terminal and navigate to the folder containing all .cpp, .h, and .txt files used by the AI. This can be done by using *cd* followed by the folder path
2. Enter *g++ *.cpp –o main*
3. Enter *./main*

This will compile, link, and run the program. Text will then appear asking the user for information about their age, weight, and sex. After entering this information, questions will be asked and the user will type either *a* or *b* and press the enter key. The question and answer process will continue until the program has asked all possible questions or until the program has asked the max number allowed.

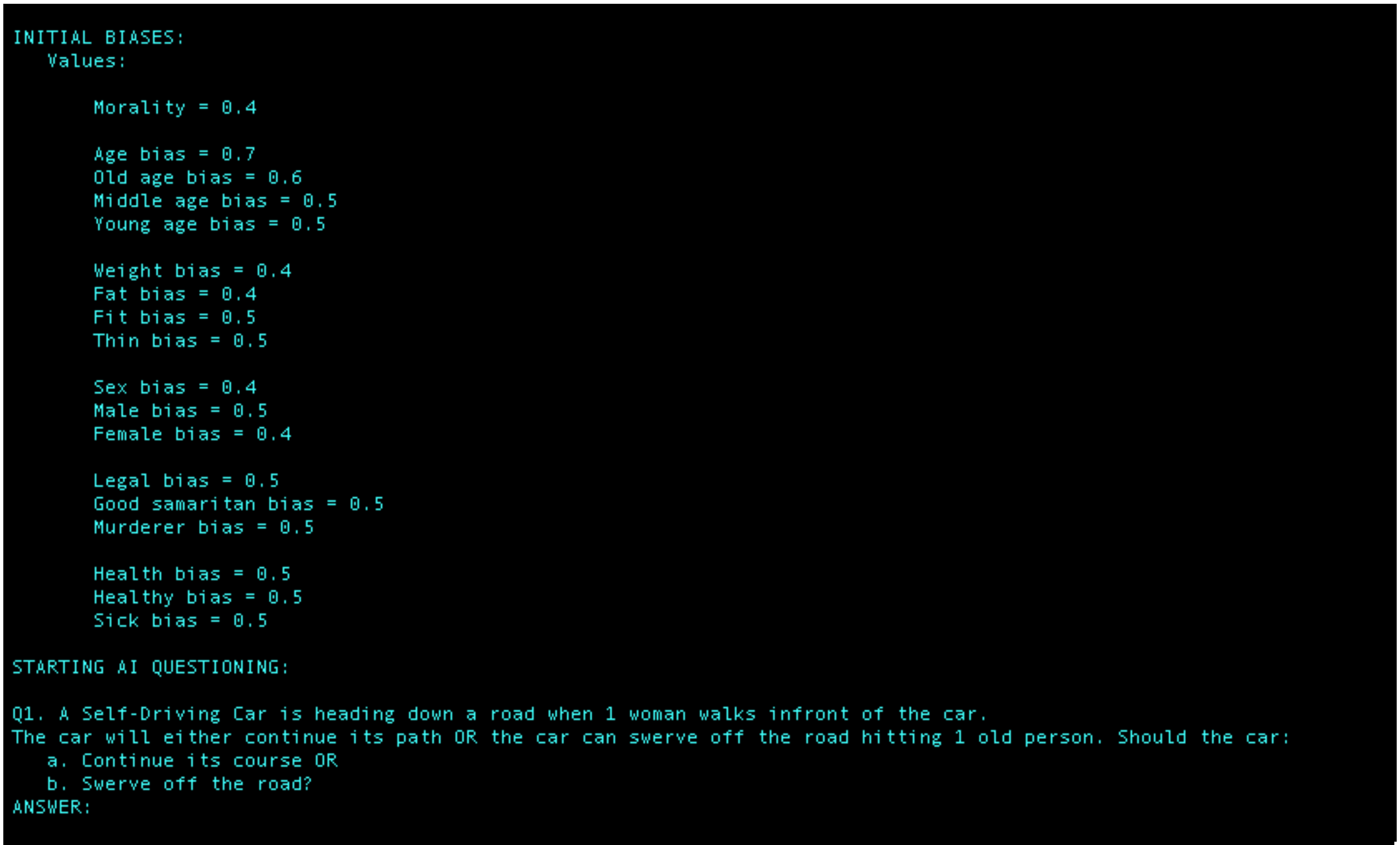
6. Sample Sessions

6.1 No Heuristics Sample

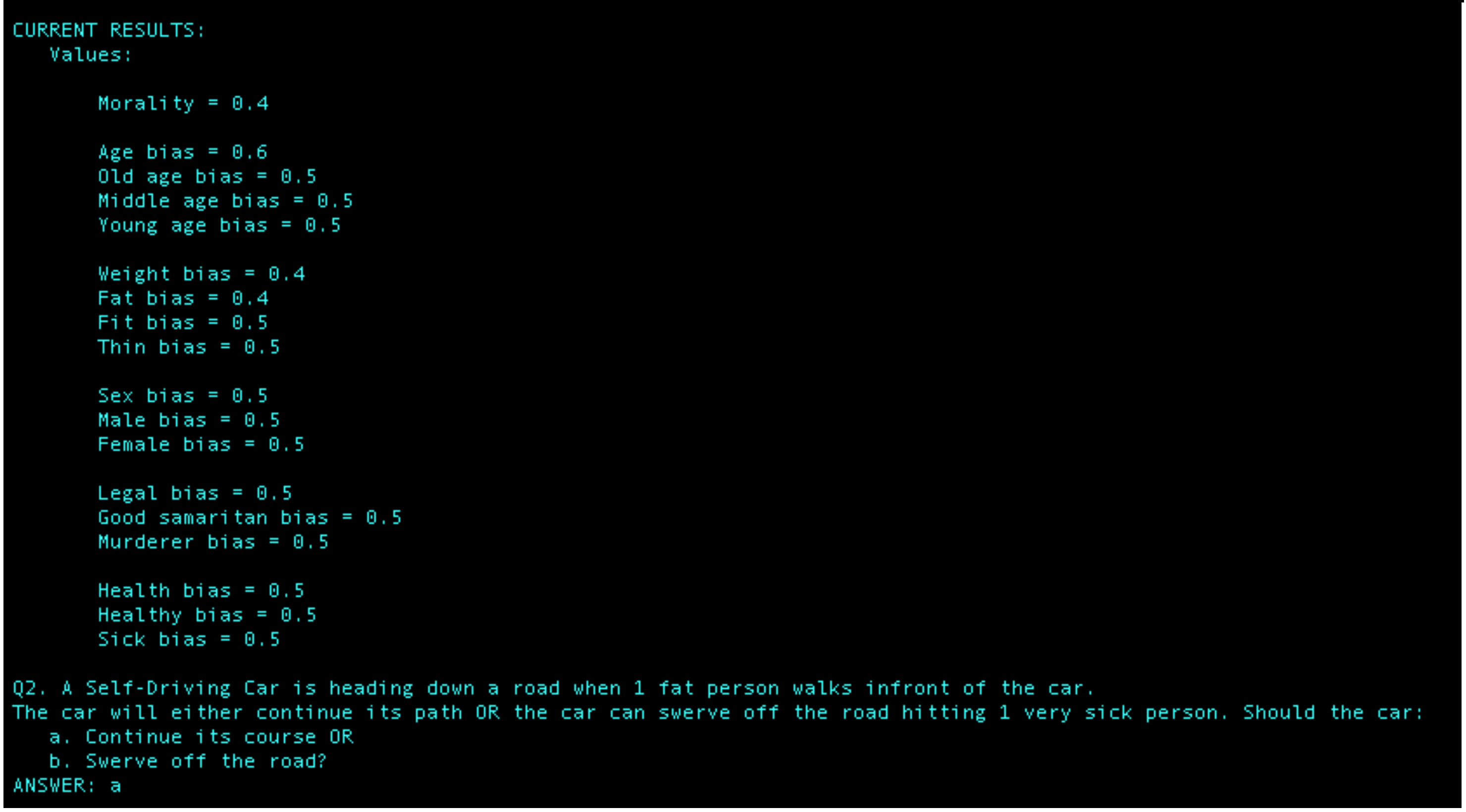
In the sample session below, a test with a male user that is 22 years old and 200lbs with no heuristics used. The initial questions are then answered to pick no bias against 2 people, fat people or women. The 3rd initial question selects an answer of b to increase the bias of age and old people. Below this, the initial biases are shown with age and old age increased and the others that were asked in the initial questions have decreased. From there, 10 questions are asked that choose the highest and lowest biases from the list and choose 1 or 2 people if morality is higher than or lower than neutral. Finally, after all questions have been asked, the initial biases and final biases are shown. Shown below this session is the .txt read in before and after. The first 15 values are shown of the .txt file.



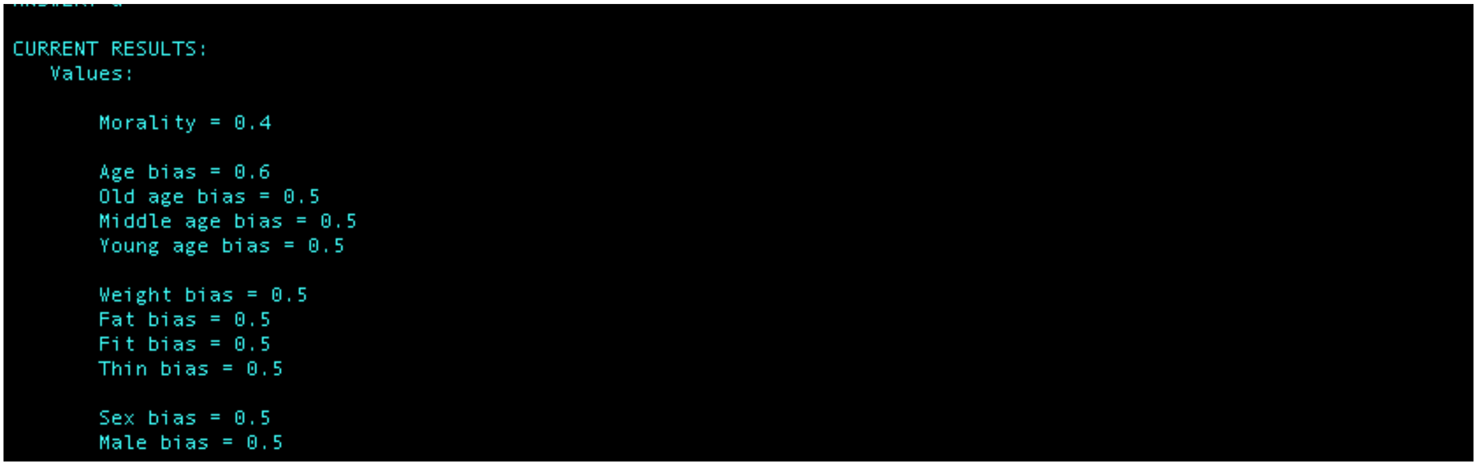
```
Welcome! Please answer the following questions:
What is your age? 22
What is your weight (in lbs)? 200
What is your sex (m or f)? m

Please answer the following questions with honesty:

Initial 1. A Self-Driving Car is heading down a road when 1 person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 2 people. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a

Initial 2. A Self-Driving Car is heading down a road when 1 person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting a fat person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a

Initial 3. A Self-Driving Car is heading down a road when 1 person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting an old person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: b

Initial 4. A Self-Driving Car is heading down a road when 1 person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting a woman. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a
```



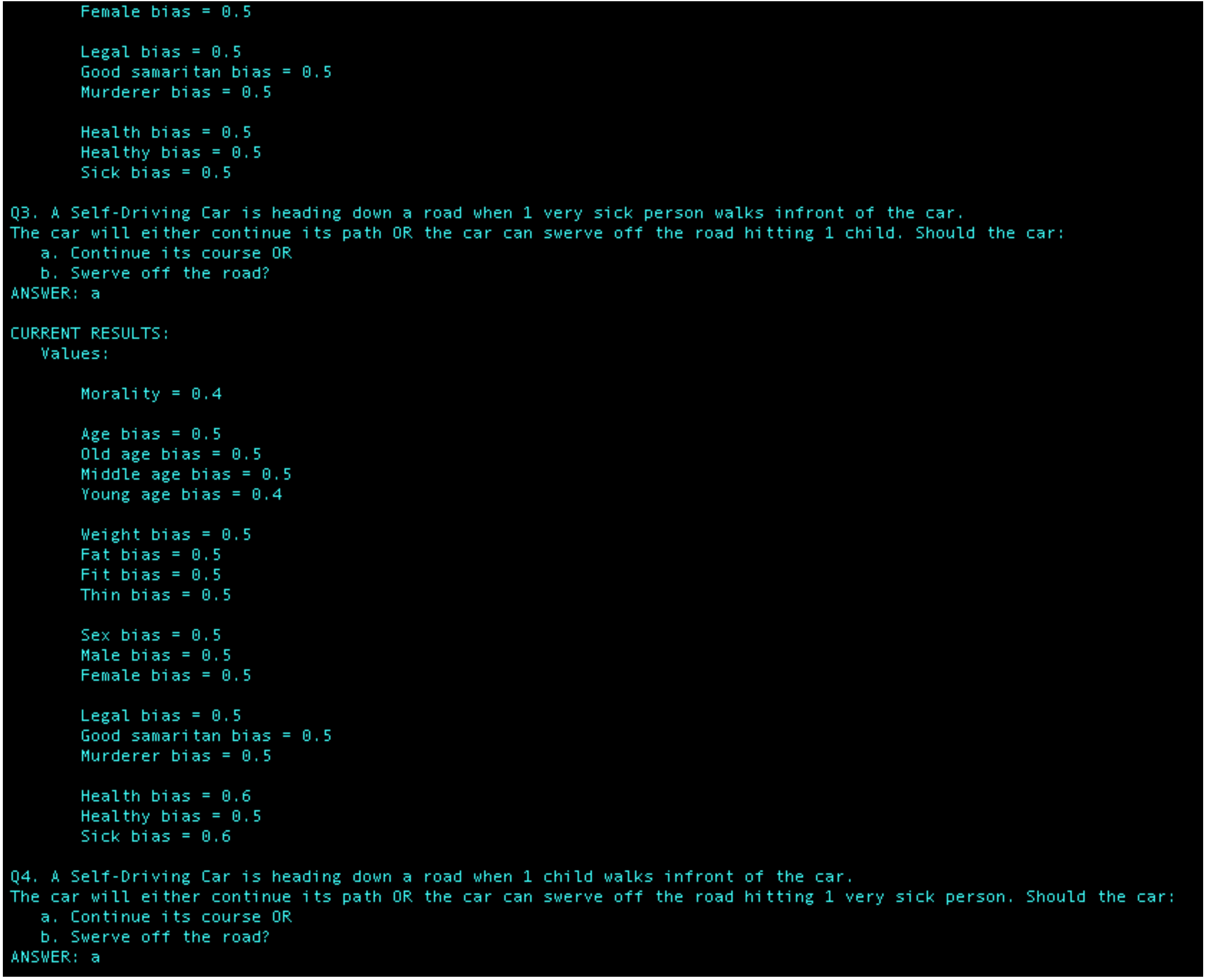
```
INITIAL BIASES:
   Values:

       Morality = 0.4

       Age bias = 0.7
       Old age bias = 0.6
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.4
       Fat bias = 0.4
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.4
       Male bias = 0.5
       Female bias = 0.4

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.5
       Healthy bias = 0.5
       Sick bias = 0.5

STARTING AI QUESTIONING:

Q1. A Self-Driving Car is heading down a road when 1 woman walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 1 old person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER:
```



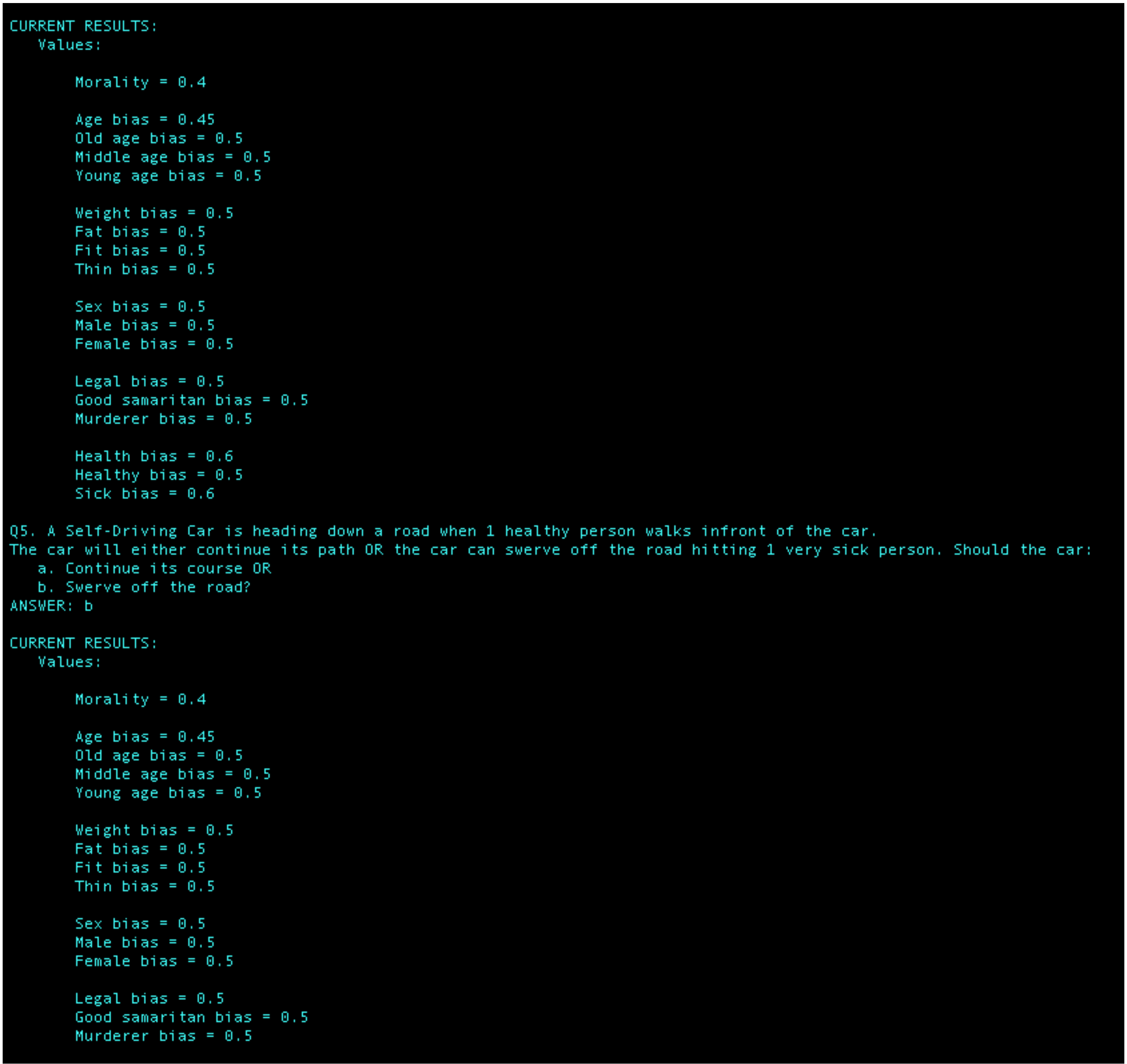
```
CURRENT RESULTS:
   Values:

       Morality = 0.4

       Age bias = 0.6
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.4
       Fat bias = 0.4
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.5
       Healthy bias = 0.5
       Sick bias = 0.5

Q2. A Self-Driving Car is heading down a road when 1 fat person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 1 very sick person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a
```



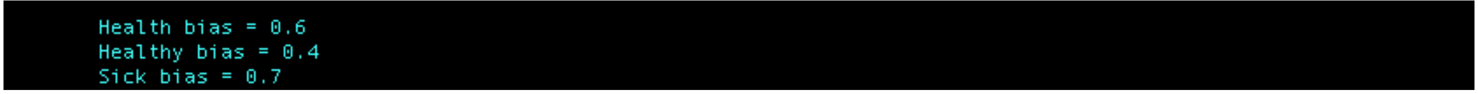
```
ANSWER: a

CURRENT RESULTS:
   Values:

       Morality = 0.4

       Age bias = 0.6
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
```



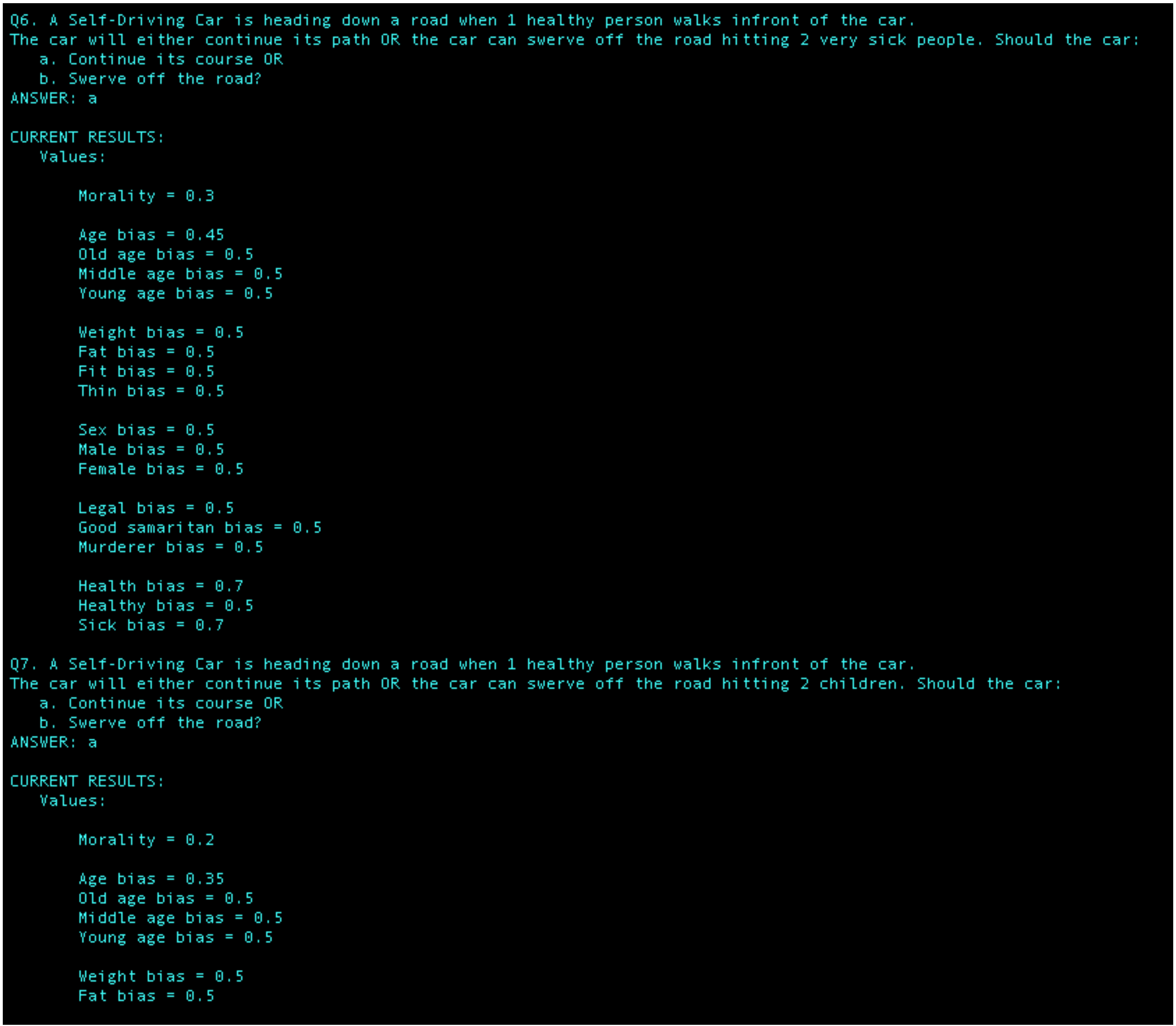
```
       Female bias = 0.5

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.5
       Healthy bias = 0.5
       Sick bias = 0.5

Q3. A Self-Driving Car is heading down a road when 1 very sick person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 1 child. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a

CURRENT RESULTS:
   Values:

       Morality = 0.4

       Age bias = 0.5
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.4

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.6
       Healthy bias = 0.5
       Sick bias = 0.6

Q4. A Self-Driving Car is heading down a road when 1 child walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 1 very sick person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a
```



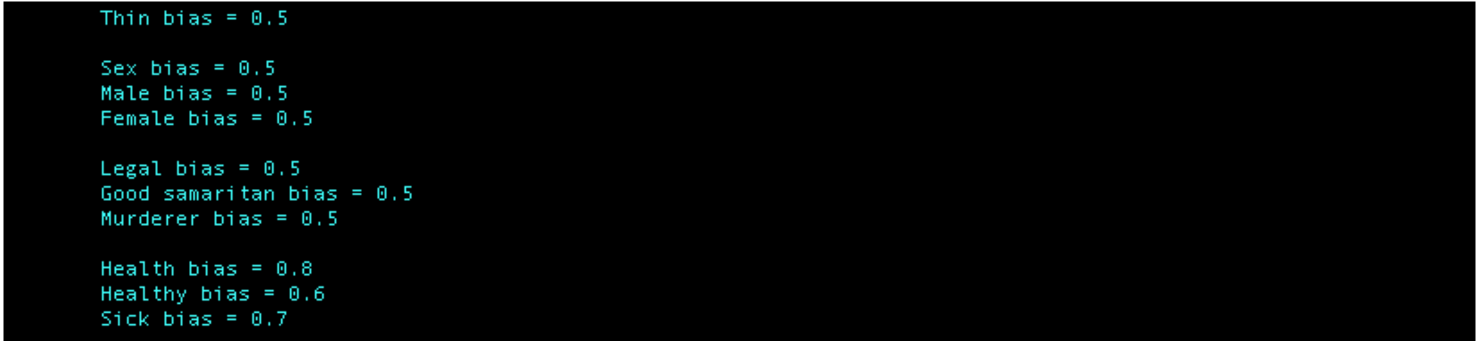
```
CURRENT RESULTS:
   Values:

       Morality = 0.4

       Age bias = 0.45
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.6
       Healthy bias = 0.5
       Sick bias = 0.6

Q5. A Self-Driving Car is heading down a road when 1 healthy person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 1 very sick person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: b

CURRENT RESULTS:
   Values:

       Morality = 0.4

       Age bias = 0.45
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5
```



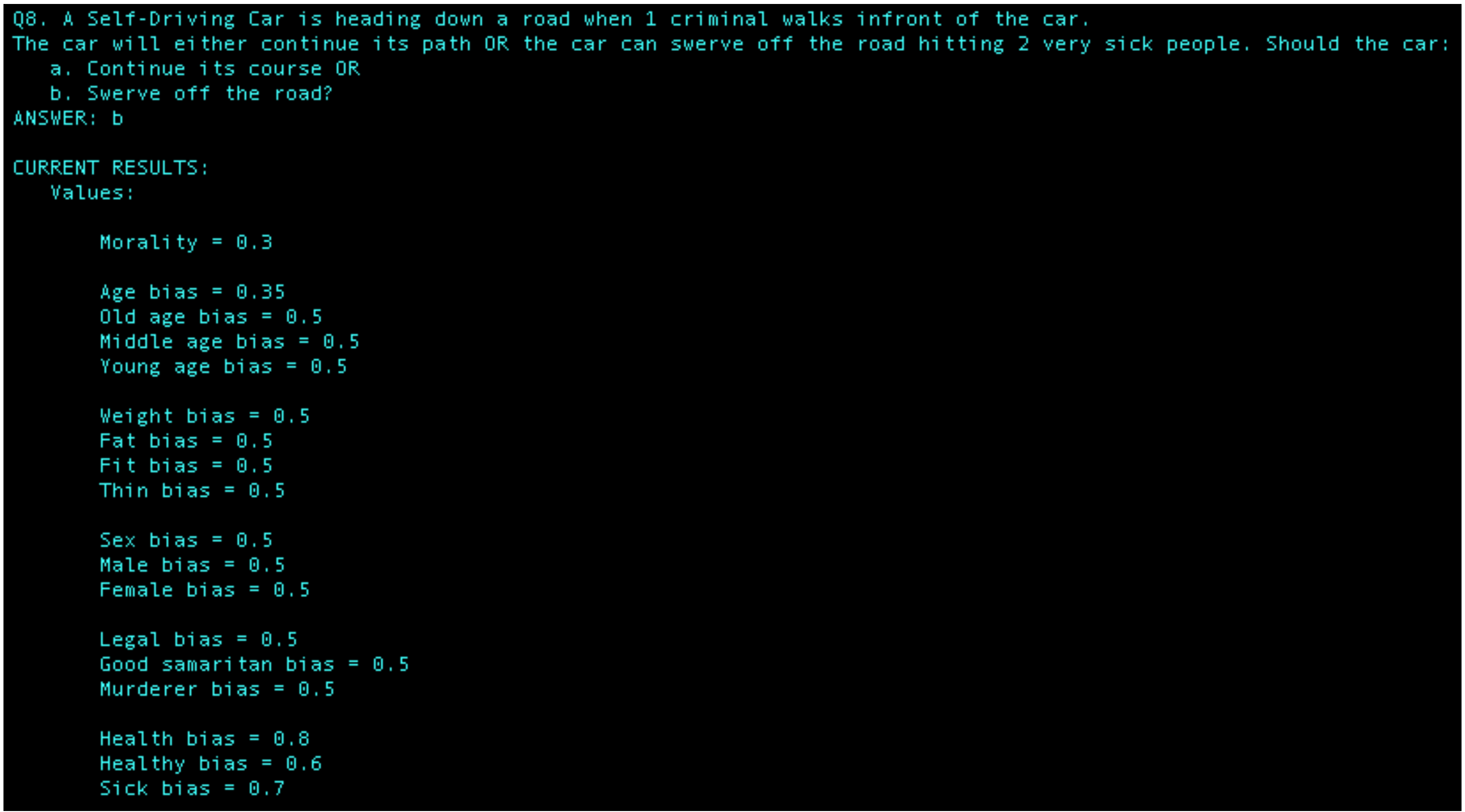
```
       Health bias = 0.6
       Healthy bias = 0.4
       Sick bias = 0.7
```



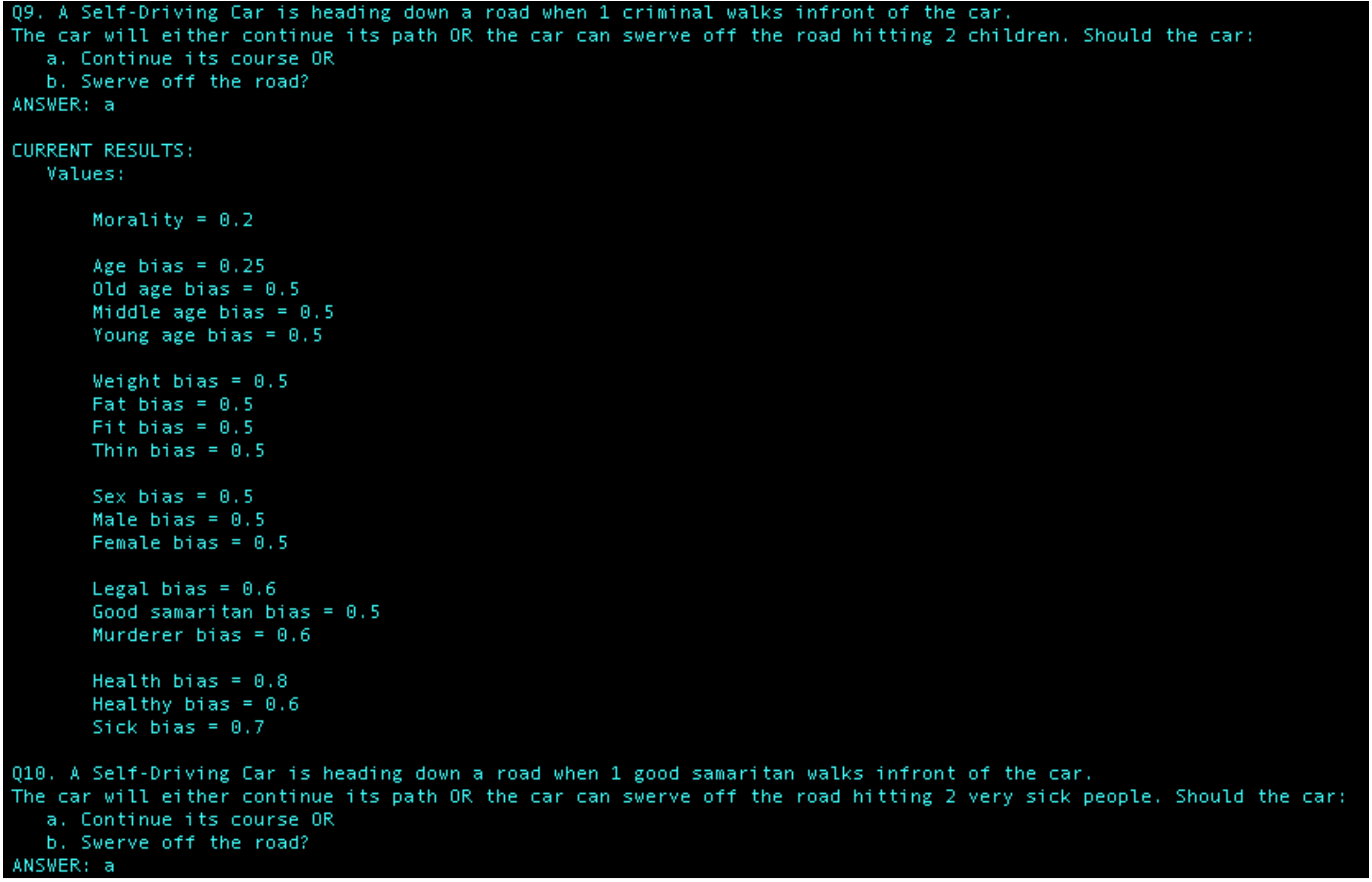
```
Q6. A Self-Driving Car is heading down a road when 1 healthy person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 2 very sick people. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a

CURRENT RESULTS:
   Values:

       Morality = 0.3

       Age bias = 0.45
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.7
       Healthy bias = 0.5
       Sick bias = 0.7

Q7. A Self-Driving Car is heading down a road when 1 healthy person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 2 children. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a

CURRENT RESULTS:
   Values:

       Morality = 0.2

       Age bias = 0.35
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
```



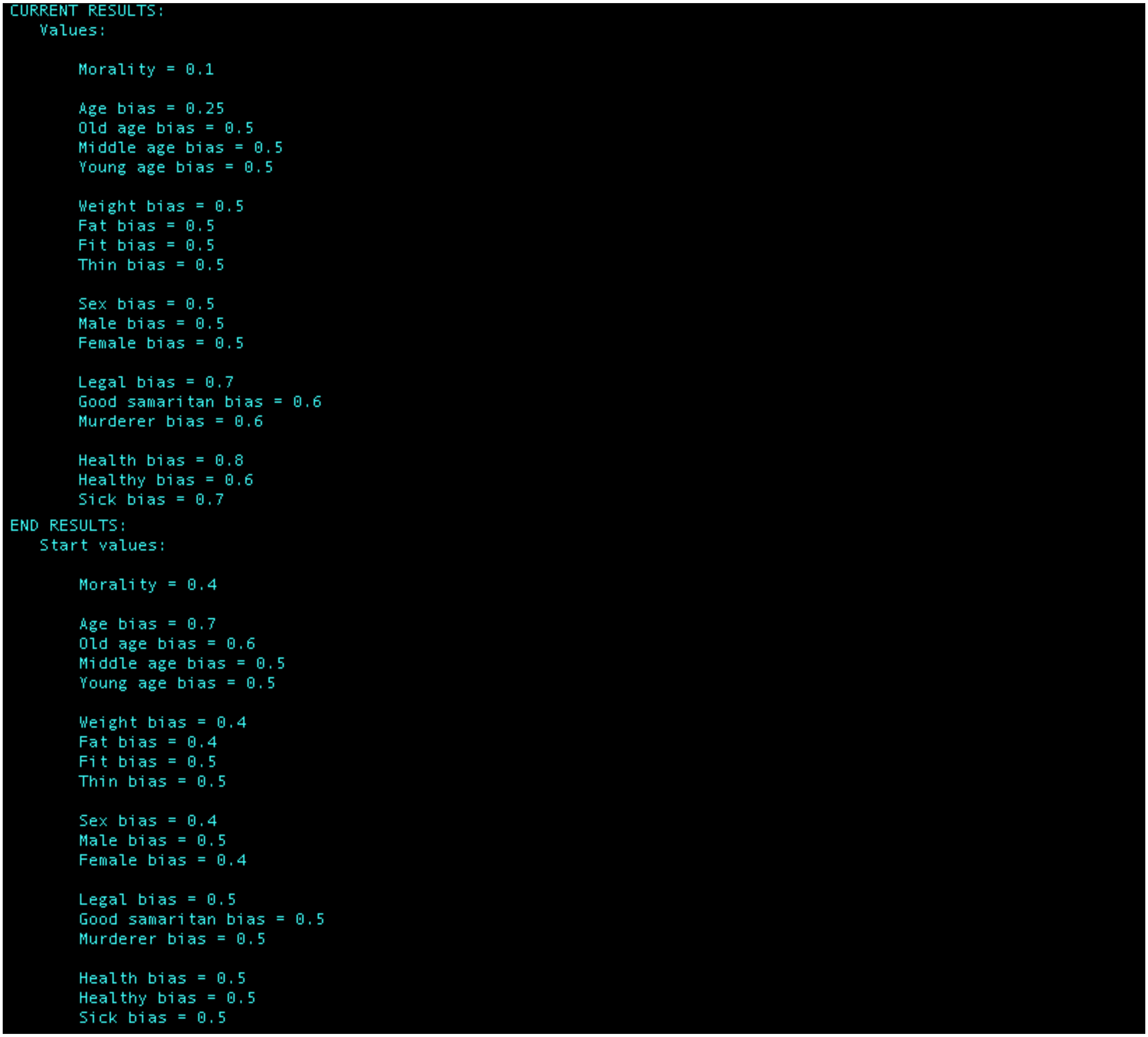
```
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.8
       Healthy bias = 0.6
       Sick bias = 0.7
```



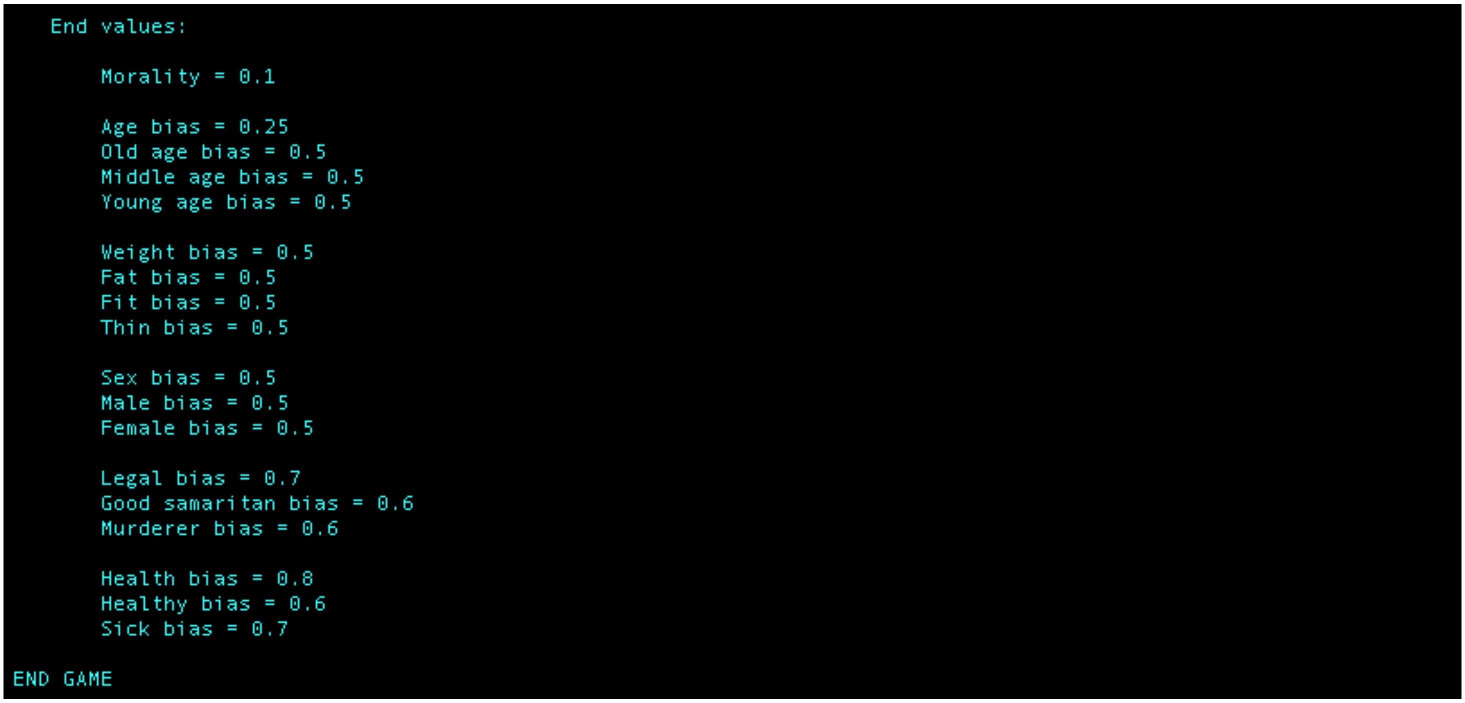
```
Q8. A Self-Driving Car is heading down a road when 1 criminal walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 2 very sick people. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: b

CURRENT RESULTS:
   Values:

       Morality = 0.3

       Age bias = 0.35
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.8
       Healthy bias = 0.6
       Sick bias = 0.7
```



```
Q9. A Self-Driving Car is heading down a road when 1 criminal walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 2 children. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a

CURRENT RESULTS:
   Values:

       Morality = 0.2

       Age bias = 0.25
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.6
       Good samaritan bias = 0.5
       Murderer bias = 0.6

       Health bias = 0.8
       Healthy bias = 0.6
       Sick bias = 0.7

Q10. A Self-Driving Car is heading down a road when 1 good samaritan walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 2 very sick people. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a
```

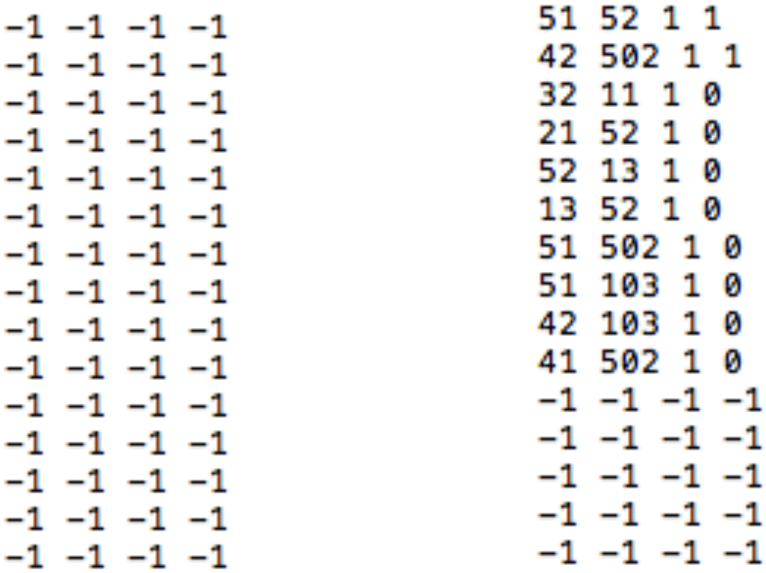


```
CURRENT RESULTS:
   Values:

       Morality = 0.1

       Age bias = 0.25
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.7
       Good samaritan bias = 0.6
       Murderer bias = 0.6

       Health bias = 0.8
       Healthy bias = 0.6
       Sick bias = 0.7
END RESULTS:
   Start values:

       Morality = 0.4

       Age bias = 0.7
       Old age bias = 0.6
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.4
       Fat bias = 0.4
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.4
       Male bias = 0.5
       Female bias = 0.4

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.5
       Healthy bias = 0.5
       Sick bias = 0.5
```



```
   End values:

       Morality = 0.1

       Age bias = 0.25
       Old age bias = 0.5
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.5
       Male bias = 0.5
       Female bias = 0.5

       Legal bias = 0.7
       Good samaritan bias = 0.6
       Murderer bias = 0.6

       Health bias = 0.8
       Healthy bias = 0.6
       Sick bias = 0.7

END GAME
```

Figure 12. Sample Run with No Heuristics

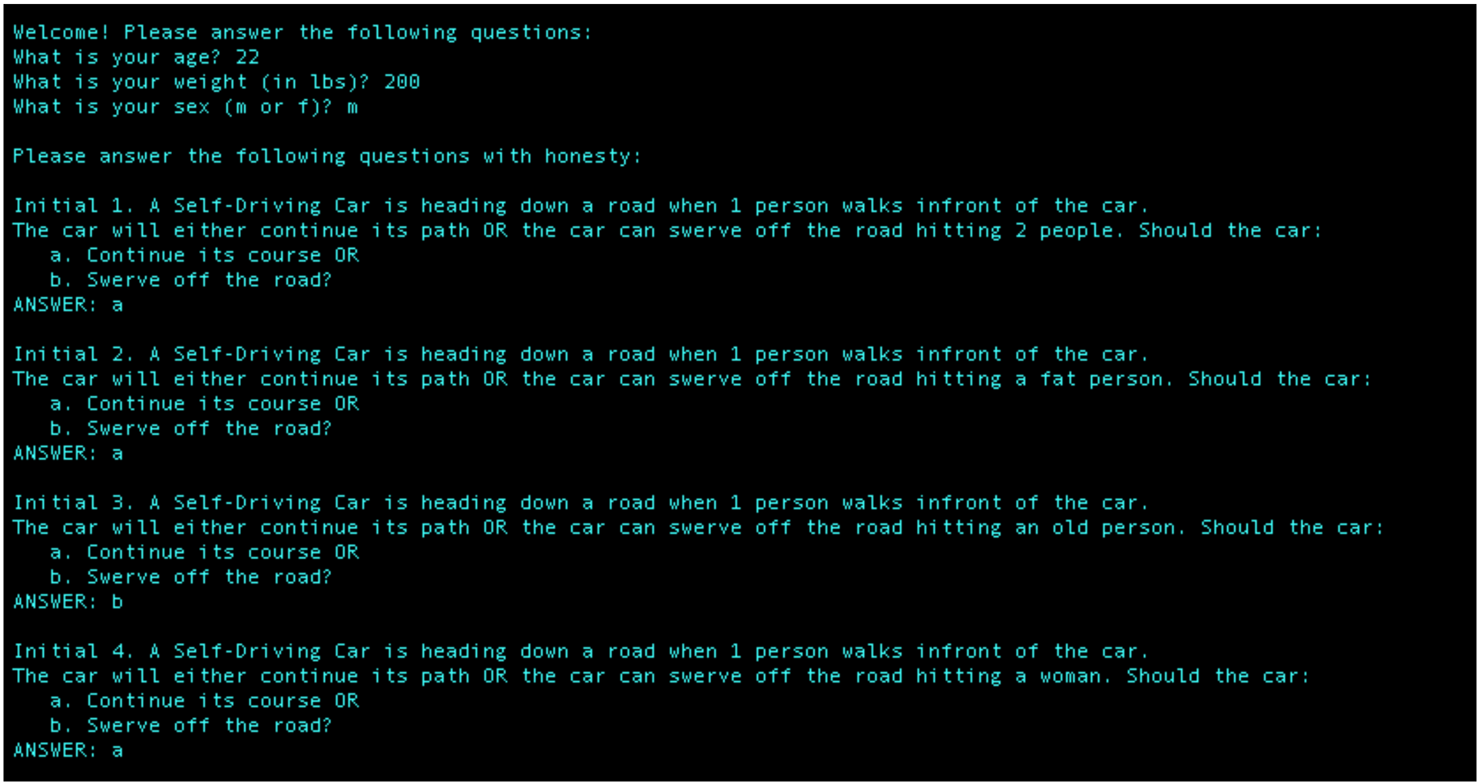


```
-1 -1 -1 -1          51 52 1 1
-1 -1 -1 -1          42 502 1 1
-1 -1 -1 -1          32 11 1 0
-1 -1 -1 -1          21 52 1 0
-1 -1 -1 -1          52 13 1 0
-1 -1 -1 -1          13 52 1 0
-1 -1 -1 -1          51 502 1 0
-1 -1 -1 -1          51 103 1 0
-1 -1 -1 -1          42 103 1 0
-1 -1 -1 -1          41 502 1 0
-1 -1 -1 -1          -1 -1 -1 -1
-1 -1 -1 -1          -1 -1 -1 -1
-1 -1 -1 -1          -1 -1 -1 -1
-1 -1 -1 -1          -1 -1 -1 -1
-1 -1 -1 -1          -1 -1 -1 -1
```

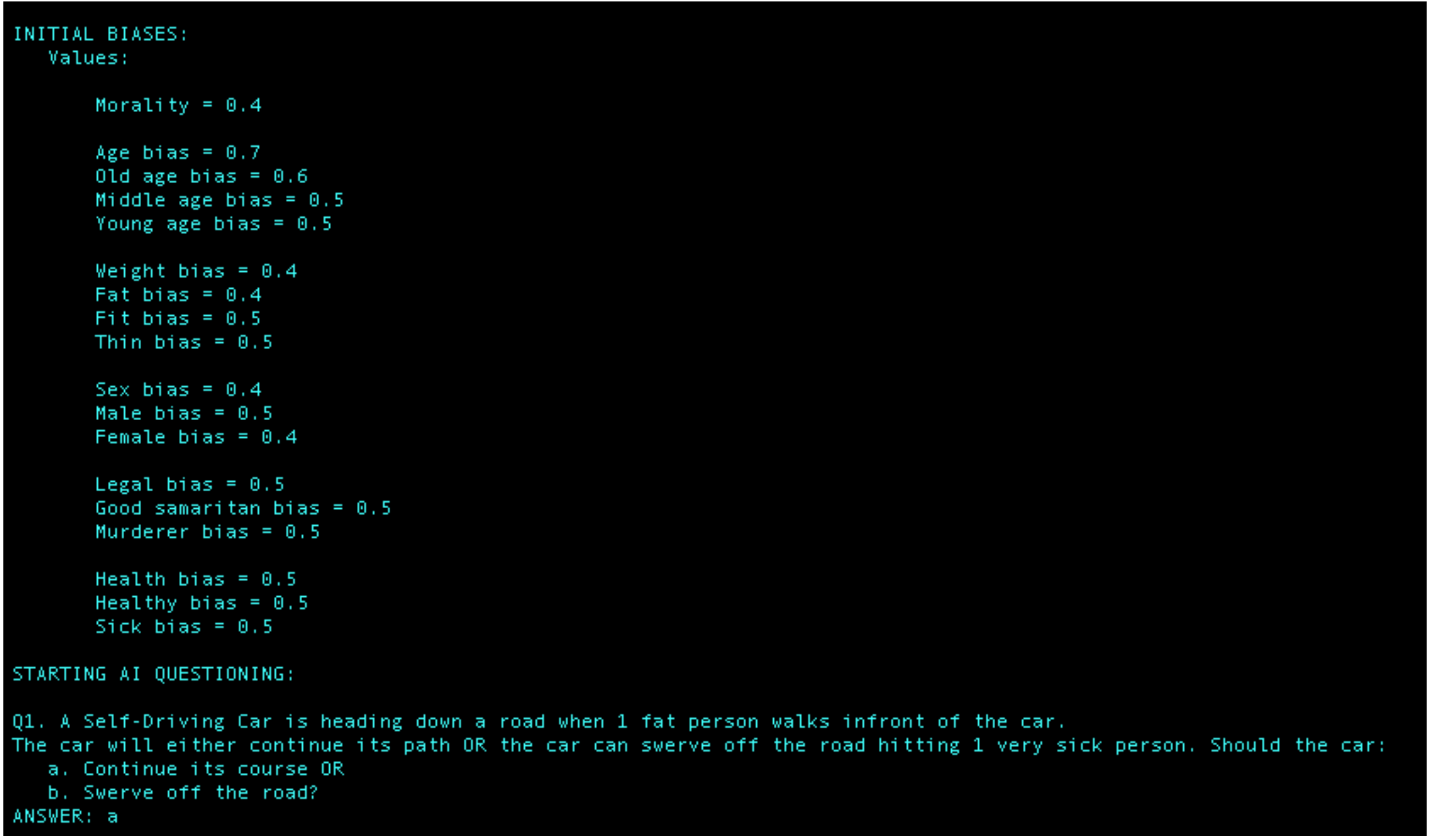
Figure 13. Text File Before and After Execution

6.2 Heuristics Sample

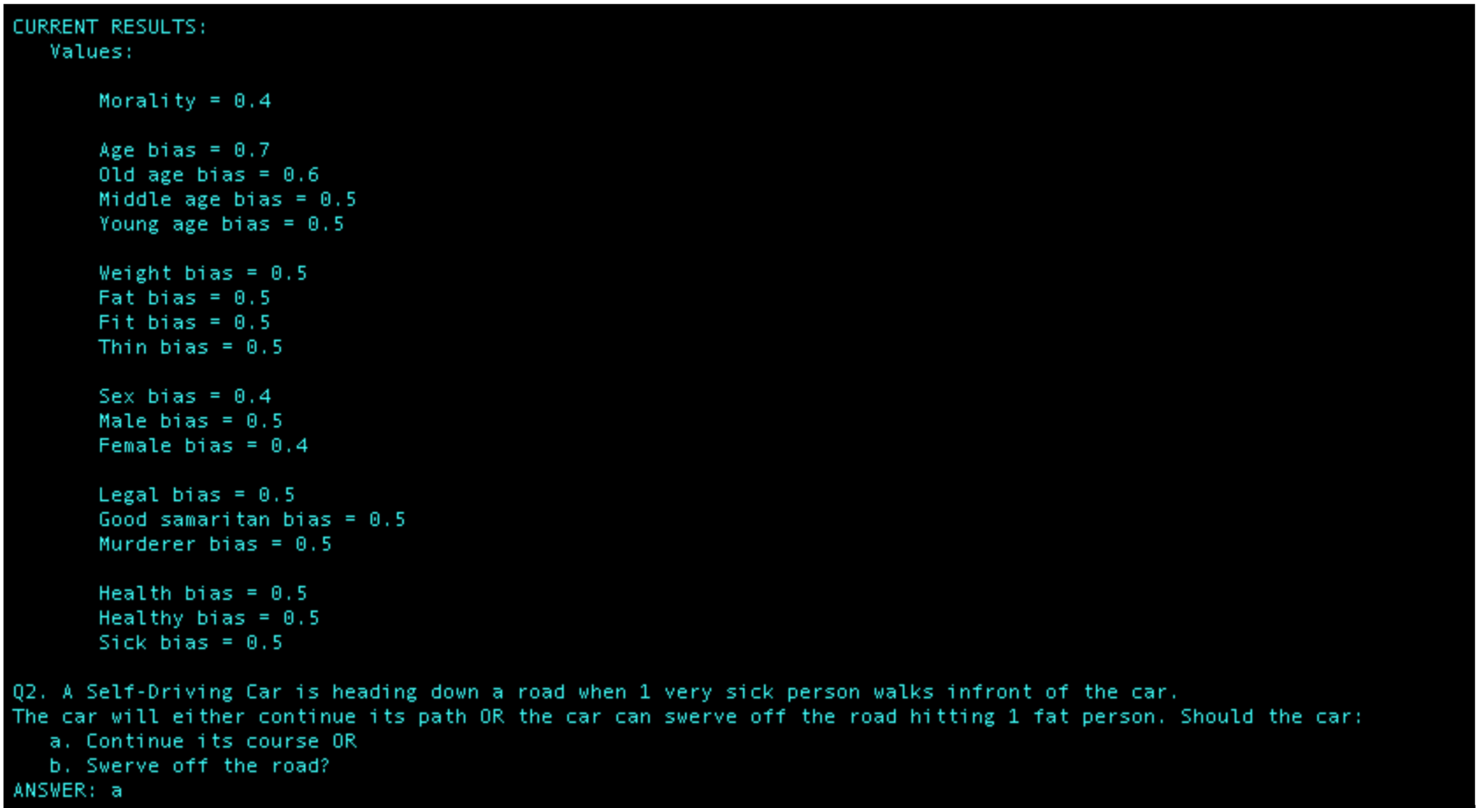
In this sample session, the user information is the same, but this time heuristics are used from a .txt file that will ask specific questions first. This session also asks fewer questions to show results sooner. It is important to note that the previous saved values allows the first and third entries to be asked because it has a success rate of 0.6 or greater and has asked 5 people or more. The last two questions that are asked are generated by the normal AI comparisons. Finally, it is shown at the end the first 17 values of the initial .txt inputs read in and the values saved after the program is finished. This is how the AI learns and uses better heuristics.



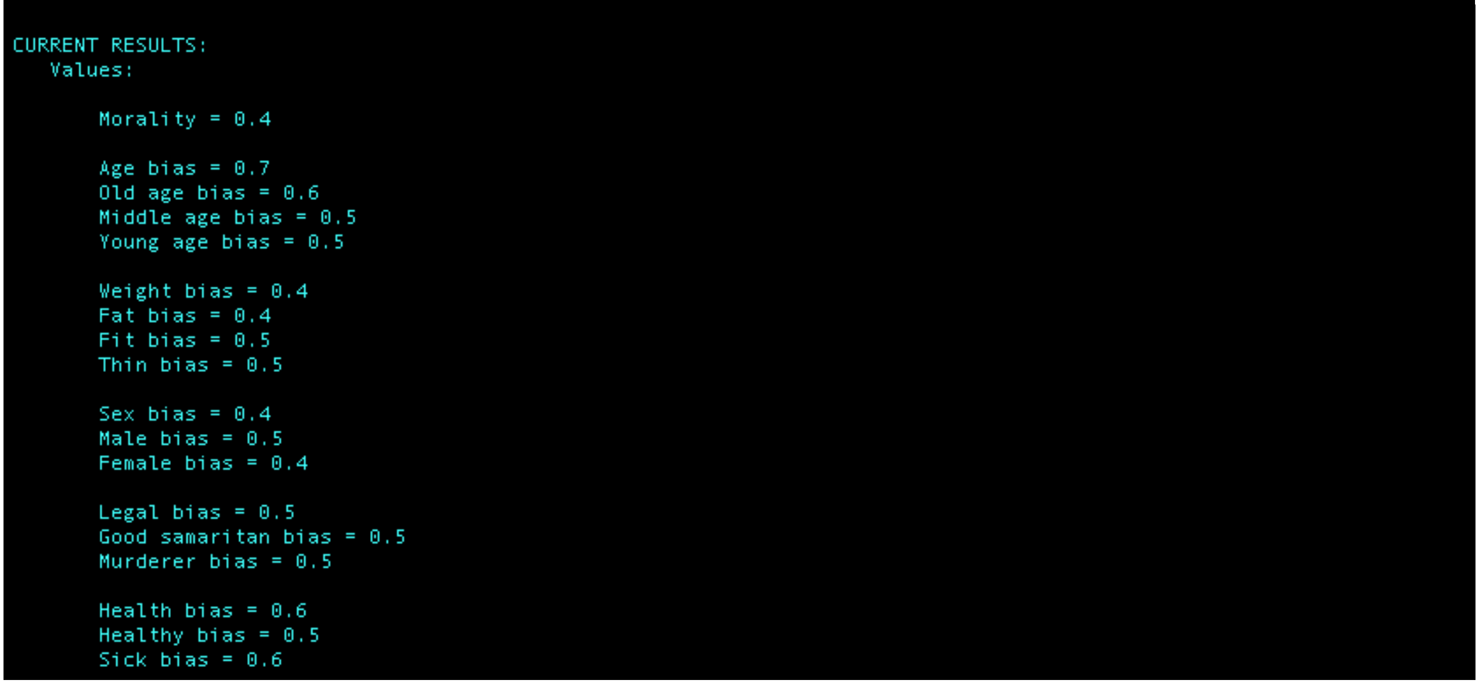
```
Welcome! Please answer the following questions:
What is your age? 22
What is your weight (in lbs)? 200
What is your sex (m or f)? m

Please answer the following questions with honesty:

Initial 1. A Self-Driving Car is heading down a road when 1 person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 2 people. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a

Initial 2. A Self-Driving Car is heading down a road when 1 person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting a fat person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a

Initial 3. A Self-Driving Car is heading down a road when 1 person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting an old person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: b

Initial 4. A Self-Driving Car is heading down a road when 1 person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting a woman. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a
```



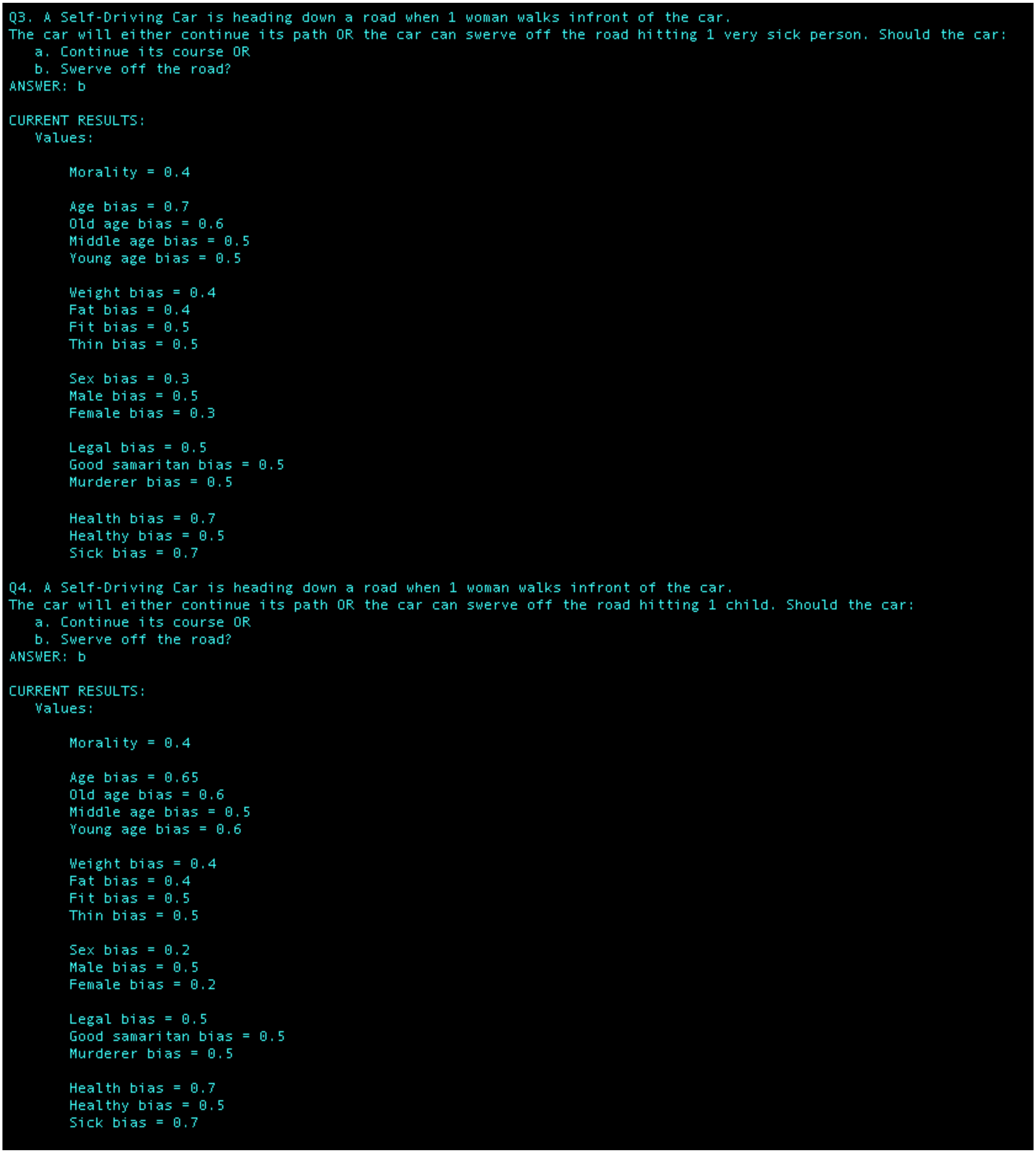
```
INITIAL BIASES:
   Values:

       Morality = 0.4

       Age bias = 0.7
       Old age bias = 0.6
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.4
       Fat bias = 0.4
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.4
       Male bias = 0.5
       Female bias = 0.4

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.5
       Healthy bias = 0.5
       Sick bias = 0.5

STARTING AI QUESTIONING:

Q1. A Self-Driving Car is heading down a road when 1 fat person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 1 very sick person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a
```



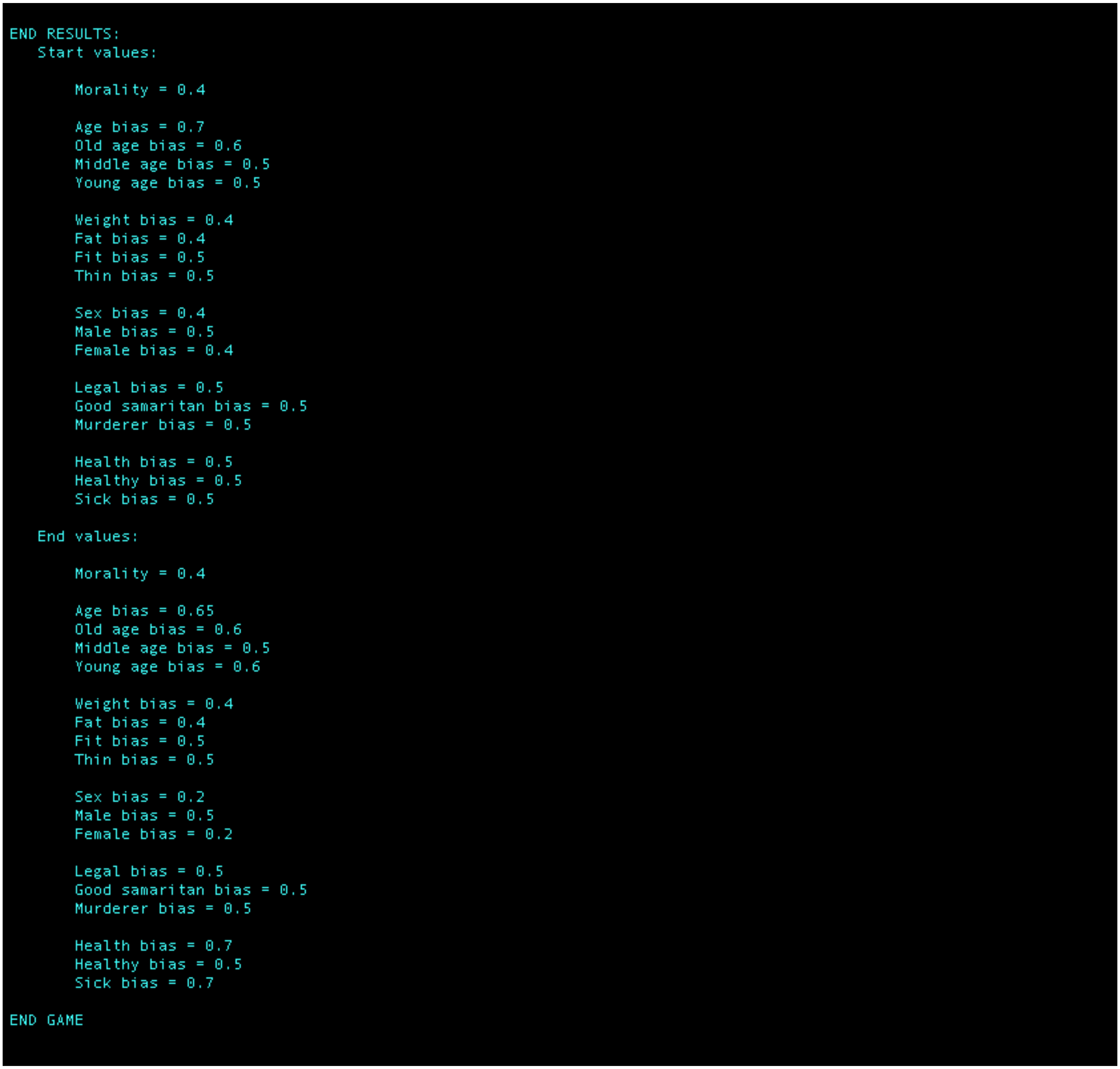
```
CURRENT RESULTS:
   Values:

       Morality = 0.4

       Age bias = 0.7
       Old age bias = 0.6
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.5
       Fat bias = 0.5
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.4
       Male bias = 0.5
       Female bias = 0.4

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.5
       Healthy bias = 0.5
       Sick bias = 0.5

Q2. A Self-Driving Car is heading down a road when 1 very sick person walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 1 fat person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: a
```



```
CURRENT RESULTS:
   Values:

       Morality = 0.4

       Age bias = 0.7
       Old age bias = 0.6
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.4
       Fat bias = 0.4
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.4
       Male bias = 0.5
       Female bias = 0.4

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.6
       Healthy bias = 0.5
       Sick bias = 0.6
```

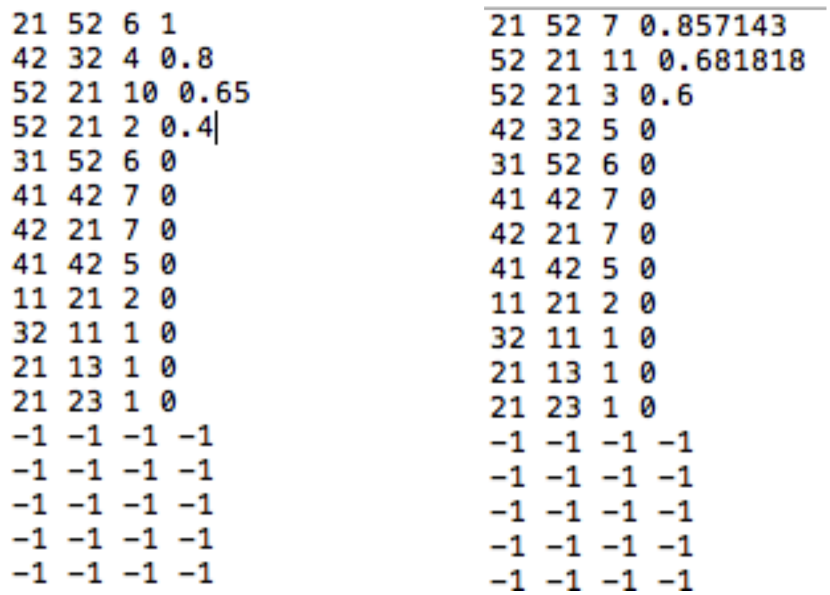


```
Q3. A Self-Driving Car is heading down a road when 1 woman walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 1 very sick person. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: b

CURRENT RESULTS:
   Values:

       Morality = 0.4

       Age bias = 0.7
       Old age bias = 0.6
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.4
       Fat bias = 0.4
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.3
       Male bias = 0.5
       Female bias = 0.3

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.7
       Healthy bias = 0.5
       Sick bias = 0.7

Q4. A Self-Driving Car is heading down a road when 1 woman walks infront of the car.
The car will either continue its path OR the car can swerve off the road hitting 1 child. Should the car:
   a. Continue its course OR
   b. Swerve off the road?
ANSWER: b

CURRENT RESULTS:
   Values:

       Morality = 0.4

       Age bias = 0.65
       Old age bias = 0.6
       Middle age bias = 0.5
       Young age bias = 0.6

       Weight bias = 0.4
       Fat bias = 0.4
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.2
       Male bias = 0.5
       Female bias = 0.2

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.7
       Healthy bias = 0.5
       Sick bias = 0.7
```



```
END RESULTS:
   Start values:

       Morality = 0.4

       Age bias = 0.7
       Old age bias = 0.6
       Middle age bias = 0.5
       Young age bias = 0.5

       Weight bias = 0.4
       Fat bias = 0.4
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.4
       Male bias = 0.5
       Female bias = 0.4

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.5
       Healthy bias = 0.5
       Sick bias = 0.5

   End values:

       Morality = 0.4

       Age bias = 0.65
       Old age bias = 0.6
       Middle age bias = 0.5
       Young age bias = 0.6

       Weight bias = 0.4
       Fat bias = 0.4
       Fit bias = 0.5
       Thin bias = 0.5

       Sex bias = 0.2
       Male bias = 0.5
       Female bias = 0.2

       Legal bias = 0.5
       Good samaritan bias = 0.5
       Murderer bias = 0.5

       Health bias = 0.7
       Healthy bias = 0.5
       Sick bias = 0.7

END GAME
```

Figure 14. Sample Run with Heuristics



```
21 52 6 1
42 32 4 0.8
52 21 10 0.65
52 21 2 0.4
31 52 6 0
41 42 7 0
42 21 7 0
41 42 5 0
11 21 2 0
32 11 1 0
21 13 1 0
21 23 1 0
-1 -1 -1 -1
-1 -1 -1 -1
-1 -1 -1 -1
-1 -1 -1 -1
-1 -1 -1 -1
```

```
21 52 7 0.857143
52 21 11 0.681818
52 21 3 0.6
42 32 5 0
31 52 6 0
41 42 7 0
42 21 7 0
41 42 5 0
11 21 2 0
32 11 1 0
21 13 1 0
21 23 1 0
-1 -1 -1 -1
-1 -1 -1 -1
-1 -1 -1 -1
-1 -1 -1 -1
-1 -1 -1 -1
```

Figure 15. Text File Before and After Execution with Heuristics

7. Discussion

7.1 Pros of Approaches & Methods

Pros of the approaches made it possible to talk to users about how they would have reacted differently to a better situation such as using their grandpa instead of an old person. This allowed me to improve the program a lot by adding better biases and questions. Another pro is that many users can try the program and the text file for the heuristics and machine learning takes in all the data asked from the user. This shortens the time needed to see what each user picked individually as it was more beneficial to see results from a larger group of people. The methods chosen also allow for flexible information to be added so improved heuristics or new biases can be quickly and easily added without changing the majority of the code.

7.2 Cons of Approaches & Methods

The biggest cons of my approaches, the methods, and the entire system itself are the assumptions that have been made. The other cons of the approaches are a lack of information gathered and the general questions asked. Assumptions that were made include assuming a person has any bias for or against the ones asked; assuming the person is answering truthfully; assuming the person cares enough to answer thoughtfully; and assuming that a person's answer reflects their stance and beliefs. These assumptions were originally creating 0% success rate because assumptions were made that biases would be found every time. However, this was not the case until questions became more specific or better biases were asked. For example, in the original set of testing only sex, weight, and age-based questions were asked. The users would always pick the option that involved fewer people because that's their main stance; the user was indifferent to the actual biases involved.

8. Conclusions

After a limited number of tests were performed with different users, the AI proved that it was semi-successful. When testing began, the program was completely unsuccessful with a success rate of 0% caused by very general questions and poor assumptions about the users. When the questions were improved and expanded, testing started to fall in the semi-successful range of about 60%. The better asked questions included new biases that were not as ethically confusing or challenging for users. These new biases included the legal and health biases.

I also found that like-minded people typically answered the same, but the thought processes behind their actions differed greatly. It would be beneficial to test more people with different backgrounds, views, beliefs, etc. that can put other biases into better use. For example, having someone who is sexist would let the AI utilize the age bias more effectively. By having larger, more diverse groups the heuristics improve, the prediction algorithms improve, and the success rate has shown to increase. The test showed that regardless of success of heuristics and the bias calculations of the users, a question has no guarantee if it will be successful due to the factors and thought processes a person has. Future work and updated results can help find the best heuristics, but I feel that a 100% success rate is not possible.

9. Future Work

9.1 Questioning & Reasoning

Since every person is different and reacts to different situations, it would be very beneficial to have a much larger variety of questions. To increase the variety the program can include more biases such as race, education, or income. By including more categories, the AI can try to find things that a specific person would react to easier. Along with more categories, more specific questions should be asked to have better comparisons. For example, asking if a person would choose between a male and female is not as effective as asking to choose between an old male murderer and a young educated female. People would react more thoughtful to questions with more information or questions with more weight to them. It would be beneficial to add a family and/or friend biases since many people would react to these biases over a very general bias such as sex. Finally, having different forms of choices would change a person's input such as saying killing is inherently wrong, so the user should not do anything and allow the car to continue.

9.2 Updated Bias Calculations

Currently, biases are calculated by differences in user information and the question asked. For example, the larger the weight difference is between the user and a bias, the more that bias will change in either direction by 0.05 to 0.1. However, some biases may be calculated better given different reactions from the public. A possible change would be making parabolic change weights for age. If the user shows a bias towards young or old people, but the user is middle aged then the bias is changed much more than a child would have against an old person. However, this child would have bias against middle age people increase or decrease more than it would for an old person, even if the difference of age is greater.

9.3 Directed Heuristics

To increase the effectiveness of the AI's machine learning and heuristics, more .txt files could be used for specific cases. For example, a specific heuristic file for a user who is a 300lb, 20-year-old male would be used, but a 100lb, 90-year-old woman would have a different heuristic file loaded. However, to do this would require many test cases for each set of inputs that would be used to load the file. If the current three input user (age, weight, sex) were used, there could be a total of 18 different heuristic files. The need for so many files may also create a need to redo how my heuristics and machine learning are performed.

9.4 User Interaction

Using a text based interface limits the full understanding of how people would react in these difficult situations, it would be beneficial to create more interactive experiences. For example, instead of text-based questions, it would be beneficial to have prerecorded videos of someone asking the user the questions. The user may react differently seeing someone ask the questions as they may begin to wonder if they will be judged, if what they are choosing is morally right, if anyone would see, etc. More advanced interactions could also be done by using a high graphic engine to make animations that the user can interact with in real time such as a video game. Taking it one step further, virtual reality simulated situations could be used where the user would be in the car and have to choose their first hand and with little time to think it through.

10. References

Bardram, J. E. (2015). Activity-Based Computing: Computational Management of Activities Reflecting Human Intention. *AI Magazine, 36*(2). Retrieved from https://www.aaai.org/ojs/index.php/aimagazine/issue/view/210

Clark, P. (2016). My Computer Is an Honor Student — but How Intelligent Is It? Standardized Tests as a Measure of AI. *AI Magazine, 37*(1). Retrieved from https://www.aaai.org/ojs/index.php/aimagazine/issue/view/213

Dean, Josh. "The Trolley Dilemma and How It Relates to Ethical Communication." *Trolley Dilemma | Homepage*, www.trolleydilemma.com/.

Kotthoff, L. (2014). Algorithm Selection for Combinatorial Search Problems: A Survey. *AI Magazine, 35*(3). Retrieved from https://www.aaai.org/ojs/index.php/aimagazine/issue/view/207

Lester, J. C. (2013). Serious Games Get Smart: Intelligent Game-Based Learning Environments. *AI Magazine, 34*(4). Retrieved from https://www.aaai.org/ojs/index.php/aimagazine/issue/view/204

Matt. "Robots and AI." NASA, NASA, robotics.nasa.gov/students/ai_robotics.php.

Riedl, M. O. (2013). Interactive Narrative: An Intelligent Systems Approach. *AI Magazine, 34*(1). Retrieved from https://www.aaai.org/ojs/index.php/aimagazine/issue/view/201

Rus, V. (2013). Recent Advances in Conversational Intelligent Tutoring Systems. *AI Magazine, 34*(3). Retrieved from https://www.aaai.org/ojs/index.php/aimagazine/issue/view/203