**Noname manuscript No.**
(will be inserted by the editor)

# Supplementary Information
# Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence

the date of receipt and acceptance should be inserted later

## 1 Choice Experiment Data

The data used in this research resulted from a choice experiment in which participants were required to express support for or opposition against a massive national transport infrastructure investment scheme. The choice task featured four attributes (vehicle ownership tax, travel time, non-fatal traffic injuries, and traffic fatalities) with two levels each (more/less than *status quo*) coded as [-1] for less than *status quo* and [1] for more than *status quo*. A set of 16 choice tasks as coded in the experiment is presented below [1].

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

**Table 1** Set of choice tasks in choice experiments

| Set | Tax | Time | NonFat | Fat |
|---|---|---|---|---|
| **1** | -1 | -1 | -1 | -1 |
| **2** | -1 | -1 | -1 | 1 |
| **3** | -1 | -1 | 1 | 1 |
| **4** | -1 | 1 | 1 | 1 |
| **5** | 1 | 1 | 1 | 1 |
| **6** | 1 | -1 | -1 | -1 |
| **7** | 1 | 1 | -1 | -1 |
| **8** | 1 | 1 | 1 | -1 |
| **9** | -1 | 1 | -1 | 1 |
| **10** | 1 | -1 | 1 | -1 |
| **11** | -1 | -1 | 1 | -1 |
| **12** | -1 | 1 | -1 | -1 |
| **13** | -1 | 1 | 1 | -1 |
| **14** | 1 | -1 | -1 | 1 |
| **15** | 1 | -1 | 1 | 1 |
| **16** | 1 | 1 | -1 | 1 |

## 2 Parameters in One-Class Model

The one-class model is used an improper benchmark to facilitate the comprehension of the values of the utilities in the three-class model. Here we provide the parameters of the one-class model which can be used to compare and contrast withe the parameters found in Table 5 for the three-class model.

**Table 2** Estimated parameters in the one-class model

| Name | Value | Std err | t-test | p-value |
|---|---|---|---|---|
| ACS Oppose | 0.927 | 0.0711 | 13.04 | 0 |
| BETA Fat | -0.792 | 0.0697 | -11.36 | 0 |
| BETA Inj | -1.11 | 0.0731 | -15.18 | 0 |
| BETA Tax | -0.978 | 0.0717 | -13.64 | 0 |
| BETA Time | -0.52 | 0.0671 | -7.74 | 0 |

## 3 Estimation of Utilities

### 3.1 One-Class Model

The baseline one-class model features only one class which means that computing the utility of actions in the model does not require the computation of the class membership probability. The formulation for estimating the utility of actions in this model is therefore: $\hat{V}(a_i) = \hat{V}_t(a_i)$. Given that $\hat{V}_t(a_i) = \beta_{it} + \sum_m \beta_{mt} \cdot x_{mi}$ for the one-class model we compute the following expression

$\hat{V}_t(a_i) = 0.927 + x_{mi}(-0.978) + x_{mi}(-0.52) + x_{mi}(-1.11) + x_{mi}(-0.792)$ where $x_{mi}$ is either [-1] or [1] as per Table 10 above. Below are the utilities for each action in the one-class model (Table 11).

**Table 3** Utility of actions in one-class baseline model

| $i$ | $\hat{V}(a_i)$ |
|---|---|
| **1** | 4.327 |
| **2** | 2.743 |
| **3** | 0.523 |
| **4** | -0.517 |
| **5** | -2.473 |
| **6** | 2.371 |
| **7** | 1.331 |
| **8** | -0.889 |
| **9** | 1.703 |
| **10** | 0.151 |
| **11** | 2.107 |
| **12** | 3.287 |
| **13** | -1.067 |
| **14** | 0.787 |
| **15** | -1.433 |
| **16** | -0.253 |

## 3.2 Three-Class Model

The formulation for estimating the utility of actions in the three-class model is $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$. To determine $\hat{V}_t(a_i) = \beta_{it} + \sum_m \beta_{mt} \cdot x_{mi}$ for the three-class model we compute the following expressions for the different classes:
For class 1: $\hat{V}_1(a_i) = -0.519 + x_{mi}(-2.56) + x_{mi}(-0.119) + x_{mi}(-0.209) + x_{mi}(-0.561)$ where $x_{mi}$ is either [-1] or [1] as per Table 10 above. For class 2: $\hat{V}_2(a_i) = 1.52 + x_{mi}(-0.967) + x_{mi}(-0.328) + x_{mi}(-1.92) + x_{mi}(-1.41)$ where $x_{mi}$ is either [-1] or [1] as per Table 10 above. And finally for class 3: $\hat{V}_3(a_i) = 1.24 + x_{mi}(-1.02) + x_{mi}(-1.72) + x_{mi}(-0.745) + x_{mi}(-0.36)$ where $x_{mi}$ is either [-1] or [1] as per Table 10 above.
Below are the utilities for each action as ascribed by each class in the three-class-model (Table 11).

**Table 4** Utilities of actions for each class in three-class model

| **Set** | $\hat{V}_1(a_i)$ | $\hat{V}_2(a_i)$ | $\hat{V}_3(a_i)$ |
|---|---|---|---|
| **1** | 2.93 | 6.145 | 5.085 |
| **2** | 1.808 | 3.325 | 4.365 |
| **3** | 1.39 | -0.515 | 2.875 |
| **4** | 1.152 | -1.171 | -0.565 |
| **5** | -3.968 | -3.105 | -2.605 |
| **6** | -2.19 | 4.211 | 3.045 |
| **7** | -2.428 | 3.555 | -0.395 |
| **8** | -2.846 | -0.285 | -1.885 |
| **9** | 1.57 | 2.669 | 0.925 |
| **10** | -2.608 | 0.371 | 1.555 |
| **11** | 2.512 | 2.305 | 3.595 |
| **12** | 2.692 | 5.489 | 1.645 |
| **13** | 2.274 | 1.649 | 0.155 |
| **14** | -3.312 | 1.391 | 2.325 |
| **15** | -3.73 | -2.449 | 0.835 |
| **16** | -3.55 | 0.735 | -1.115 |

To determine the class membership probabilities $\hat{P}(t)$ a logit function was used and the values are shown in Table 4. Recalling the formulation for estimating the utility in the three-class model $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$ we therefore proceed to estimate $\hat{V}(a_i) = \hat{P}(1) \cdot \hat{V}_1(a_i) + \hat{P}(2) \cdot \hat{V}_2(a_i) + ...\hat{P}(T) \cdot \hat{V}_T(a_i)$. Below are the utilities for each action in the three-class model (Table 13).

**Table 5** Utility of actions in three-class model

| $i$ | $\hat{V}(a_i)$ |
|---|---|
| **1** | 5.471330863 |
| **2** | 3.338994832 |
| **3** | 0.479295244 |
| **4** | -0.718695194 |
| **5** | -3.116784951 |
| **6** | 3.073241107 |
| **7** | 1.875250668 |
| **8** | -0.98444892 |
| **9** | 2.141004394 |
| **10** | 0.213541519 |
| **11** | 2.611631275 |
| **12** | 4.273340425 |
| **13** | 1.413640837 |
| **14** | 0.940905075 |
| **15** | -1.918794512 |
| **16** | -0.257085363 |

## References

1. Chorus, C., Mouter, N., Pudane, B.: A taboo trade off model for discrete choice analysis. In: International Choice Modelling Conference 2017 (2017)