Computer Says I Don't Know

An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence

Martinho, Andreia; Kroesen, Maarten; Chorus, Caspar

DOI
10.1007/s11023-021-09556-9

Publication date
2021

Document Version
Final published version

Published in
Minds and Machines

Minds and Machines
https://doi.org/10.1007/s11023-021-09556-9

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

Andreia Martinho[1] · **Maarten Kroesen**[1] · **Caspar Chorus**[1]

Received: 9 September 2019 / Accepted: 5 February 2021

Abstract
As AI Systems become increasingly autonomous, they are expected to engage in decision-making processes that have moral implications. In this research we integrate theoretical and empirical lines of thought to address the matters of moral reasoning and moral uncertainty in AI Systems. We reconceptualize the metanormative framework for decision-making under moral uncertainty and we operationalize it through a latent class choice model. The core idea being that moral heterogeneity in society can be codified in terms of a small number of classes with distinct moral preferences and that this codification can be used to express moral uncertainty of an AI. Choice analysis allows for the identification of classes and their moral preferences based on observed choice data. Our reformulation of the metanormative framework is theory-rooted and practical in the sense that it avoids runtime issues in real time applications. To illustrate our approach we conceptualize a society in which AI Systems are in charge of making policy choices. While one of the systems uses a baseline morally certain model, the other uses a morally uncertain model. We highlight cases in which the AI Systems disagree about the policy to be chosen, thus illustrating the need to capture moral uncertainty in AI systems.

Keywords Morality · Uncertainty · Metanormative Theory · Artificial Intelligence · Discrete Choice Analysis · Latent Class Choice Model

1 Introduction

The inner workings and innuendos of morality remain obscure yet the design of a moral compass for Artificial Intelligence (AI) Systems is pressing. The modern AI System benefits from robust computational power and sophisticated algorithms to feed on data for its own learning and adaptive processes thus becoming increasingly autonomous while, at the same time, engaging in complex decision-making

✉ Andreia Martinho
a.m.martinho@tudelft.nl

[1] Delft University of Technology, Delft, The Netherlands

Published online: 23 February 2021

processes. Such complexity is aggravated by the potential moral dimension of decisions and there are concerns about whether these systems will uphold moral values (Dignum 2017).

A number of cases where decisions made by AI Systems have morally problematic implications have been discussed in the literature. The archetypal case is the autonomous vehicle (AV) moral dilemma, a philosophical situation, modeled after the *trolley problem* thought experiment (Thomson 1984; Foot 1967), in which the AV is required to make a moral choice between actions in traffic that will result in different combinations of lives saved and sacrificed (Bonnefon et al. 2016; Awad et al. 2018; Goodall 2016; Lundgren 2020; Himmelreich 2018; Wolkenstein 2018; Keeling 2020). In the services industry it was reported that a machine learning algorithm used by Amazon, which was eventually dropped, penalized female applicants (Fritz et al. 2020; Maedche et al. 2019). Another well explored case is related to the use of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a machine learning based algorithm used in recidivism risk calculations for parole and bail in the U.S justice system, which was reported to be racially biased (Fritz et al. 2020; Rudin 2019; Flores et al. 2016; Feller et al. 2016; Wexler 2017).

Theoretical and empirical lines of thought have emerged in the literature to resolve issues associated with moral reasoning in and by AI systems. These researches shape Machine Ethics, a controversial ethical field, which aims at implementing ethical principles and moral decision-making faculties in machines to ensure that their behavior toward human users and other machines is ethically acceptable (Anderson and Anderson 2011; Allen et al. 2006; van Hartskamp et al. 2019; Hunyadi 2019; Poulsen et al. 2019).

In theoretical work, there appears to be a general consensus that conceptual agreement regarding the moral machinery of artificial agents should precede design endeavors. This has led to a renewed interest in normative ethics work that spans several domains of knowledge and research. Prospects were made for moral machines based on moral theories, such as deontology and consequentialism (Allen et al. 2000; Anderson et al. 2004; Powers 2006; Wallach et al. 2008; Wallach and Allen 2008; Tonkens 2009; Anderson and Anderson 2011) but the problem of moral disagreement between competing moral theories and conflicting moral judgments was never surmounted (Bogosian 2017). A solution that has been advanced in more recent literature is to design AI Systems to be fundamentally uncertain about morality (Bogosian 2017; Brundage 2014; Dobbe et al. 2019).

Decisions made by these systems within the realm of moral uncertainty would be based on the assumption that there is no certainty about which moral theory is correct. A particular theoretical framework for decision-making under moral uncertainty developed by William MacAskill (MacAskill 2014) has been outlined within the domain of AI morality by Kyle Bogosian (2017). It is based on the metanormative notion that moral uncertainty can be conceived as a voting problem among moral theories. The resulting moral preference of the AI is then a function of the credence in particular moral theories weighted by the moral acceptability ("choice-worthiness") of an action under each theory. This translates into an ordering of actions that maximizes overall choice-worthiness (MacAskill 2014, 2016; MacAskill et al. 2020; Bogosian 2017). Because this

framework treats moral uncertainty as a voting problem among moral theories, it overcomes common objections of incomparability and incommensurability of moral theories and moral judgments. By allowing AI systems to act in accordance with the diverse values endorsed by humans, the system accommodates the diversity of moral values in its moral decisions (Nissan-Rozen 2015; Żuradzki 2015; MacAskill 2014, 2016; Bogosian 2017).

Empirical attempts to address and embed moral reasoning in AI Systems rely on the idea that human morality should be reflected in AI moral reasoning. Human morality would first need to be empirically identified and subsequently embedded in the AI system (Bonnefon et al. 2016; Awad et al. 2018; Lin 2016; Zhao et al. 2016; Bergmann et al. 2018; Faulhaber et al. 2018). In the Moral Machine Experiment, a particularly impressive empirical research endeavor in which preference data of 1.3 million respondents from various regions of the world was compiled in the context of moral dilemmas, it was suggested that the relative agreement found is a positive indicator for consensual morality in AI (Awad et al. 2018). Although this notion of moral agreement is indeed attractive from a pragmatic viewpoint, it has also raised some criticism (Bigman and Gray 2020; Harris 2020). Other studies have made attempts to capture moral heterogeneity across individuals, inspired by the above mentioned line of theoretical argumentation. This has, however, proven to pose severe practical challenges in terms of empirical operationalization. Most pressingly, capturing every preference or vote gives rise to practical runtime problems (which could be particularly problematic in cases where the AI has to make split-second decisions), yet, averaging preferences or votes into one preference profile (Noothigattu et al. 2018) comes with the risk of failing to properly account for marginal preferences.

This paper contributes to the theoretical and empirical strands of literature, which focus on embedding moral values and judgments into AI systems, by providing a theory-rooted yet empirically practical approach to capture society's moral heterogeneity in a morally uncertain AI system. Our approach builds on the current theoretical understanding that moral uncertainty is paramount to the endeavor of implementing moral reasoning in AI Systems and it is practical by avoiding runtime issues and not requiring averaging efforts. We propose to generate such moral uncertainty by re-conceptualizing and operationalizing the metanormative framework for decision-making under moral uncertainty, briefly introduced earlier, as a utility-based latent class discrete choice model.

Moral heterogeneity is captured through a small set of latent classes, each with its own distinct moral preferences, which makes this theory-rooted approach for moral decision-making of AI systems practical in terms of runtime and interpretability. Without loss of generality we use a small-scale dataset that resulted from a choice experiment to provide an illustration in which an AI System makes policy choices on behalf of societies based on the conflicting moral preferences of latent classes in society.

The novelty of this work is the use of discrete choice analysis to codify human (moral) preferences and decision rules in order to embed these into a (moral) AI System and, moreover, an empirical illustration of moral uncertainty.

With this research we expect to contribute to the Machine Ethics and Artificial Moral Agents (AMA) literature (Anderson et al. 2004; Anderson and Anderson 2011; Allen et al. 2006; Floridi and Sanders 2004; Poulsen et al. 2019; van Wynsberghe and Robbins 2019; Cervantes et al. 2020), as well as the moral uncertainty literature, which has mainly explored theoretical case studies (Lockhart 2000; MacAskill et al. 2020), and also to broader lines of research emphasizing the need to embed values into AI (van de Poel 2020; Klenk 2020).

The remainder of the paper is organized as follows. Section two explains at a conceptual level how the recently proposed metanormative framework can be connected to Discrete Choice Analysis. Section three goes further by showing how an operational latent class discrete choice model can be used to codify moral uncertainty in AI. Sections four and five illustrate the approach in the context of a concrete example, where the latter section focuses on estimating the latent class model on choice data. Section six presents the proof of concept, by equipping AI systems with a moral framework under the assumptions of normative certainty versus normative uncertainty, building on the modeling efforts presented in preceding sections. Section seven draws conclusions and puts forward avenues for further research.

2 A Metanormative Framework and its connection with Discrete Choice Analysis

The metanormative framework, as construed by MacAskill and Bogosian, produces an ordering of actions in terms of their choice-worthiness in a particular decision-situation (Bogosian 2017; MacAskill 2014). The key elements in the decision-situation, i.e. a context in which an agent is required to make a decision, are the *decision-maker*, which in this research is an AI System defined as a regularly interacting or interdependent group of units which form an integrated structure that employs AI in any of its forms (learning, planning, reasoning, natural language processing, perception) separately or combined to perform a function while continuously interacting with the environment (Backlund 2000); a set of possible *actions* that the decision-maker has the power to bring about; the *normative theories* taken into account by the decision-maker; a *credence function* that represents the decision-maker's beliefs or trust in the various normative theories; and the *choice-worthiness*, which is the normative ordering of actions after all relevant considerations have been taken into account (MacAskill 2014; Bogosian 2017).

At the core of this metanormative framework for capturing moral uncertainty is the notion that the choice-worthiness of an action is determined by its choice-worthiness according to various competing normative or moral theories and the credence of the decision-maker in each of those theories. More precisely, the choice-worthiness of an action is the credence-weighted average of the choice-worthiness of the action in all of the individual theories. Using a slightly adapted notation, this core can be formalized as $W(a_i) = \sum_t^T [C(t) \cdot W_t(a_i)]$ where $W(a_i)$ denotes the total or over-all choice-worthiness of an action; a_i is an action from the exhaustive set A of mutually exclusive actions $\{a_1 ... a_i ... a_J\}$ where J is the cardinality of the choice set; $W_t(a_i)$ denotes the choice-worthiness of an action given a particular normative

theory t which is taken from the set T of available theories; and $C(t)$ denotes the credence of the theory[1].

The operationalization of this formulation entails two important challenges regarding the measure or inference of $W_t(a_i)$, i.e. the choice-worthiness of an action given a moral theory, and the measure or inference of $C(t)$, i.e. the credence of a moral theory. We present Discrete Choice Analysis as an intuitive method to make these inferences in an empirically rigorous way. A reconceptualization of the formulation introduced above is required so it can be re-positioned into the Discrete Choice Analysis domain. Firstly the choice-worthiness of an action given a moral theory is re-conceptualized as the utility of an action given a moral theory. Although this variation is in fact a matter of semantics, it facilitates the connection with the micro-econometric framework of Discrete Choice Analysis. Further details on the definition and operationalization of the concept of utility will be provided in the upcoming sections. A second and more relevant step is the re-conceptualization of the credence of a theory into the share of the population that adheres to that theory or, equivalently, the probability that a randomly sampled individual from the population adheres to the theory. It is therefore implicitly postulated that $C(t) \in [0, 1] \forall t$, and that $\sum_t^T C(t) = 1$ which is in fact congruent to the construction of credence in the metanormative framework as an "assignment of probabilities to various moral theories of being correct" (Bogosian 2017).

To avoid confusion, a new notation (V for utility of an action and P for the probability that a sampled individual adheres to a theory) is adopted thus leading to the following formulation for the utility of an action: $V(a_i) = \sum_t^T [P(t) \cdot V_t(a_i)]$. The challenge is to measure or infer $P(t)$ and $V_t(a_i)$. As elaborated below, the domain of Discrete Choice Analysis, and its sub-branch of latent class discrete choice modeling, offers a powerful approach to tackle this challenge. As an empirical base, we use experimental choice data (observed choices made by human decision-makers in a choice situation) to estimate the probability that an individual belongs to a specific class associated with a particular moral theory $P(t)$ as well as the utility of an action a_i given a moral theory $V_t(a_i)$.

3 Operationalization of the Metanormative Framework using Discrete Choice Analysis

As mentioned above, the model that will be used to operationalize the re-conceptualized metanormative framework is drawn from Discrete Choice Analysis. The key idea in this field is that a choice provides a signal of the latent utility of the choice options or alternatives (Samuelson 1938, 1948) and that the utility that the

[1] It is acknowledged that this formulation is a restricted version of the metanormative framework proposed by MacAskill and Bogosian in that their original framework features various important extensions to account for different types of moral theories. For now, however, the focus of this research is on what it is believed to be the core of the metanormative framework thus opening an avenue for further research that accommodates its various extensions.

decision-maker ascribes to each alternative in the choice set can be inferred by means of econometric methodology (McFadden et al. 1973; Train 2009; Ben-Akiva et al. 1985). For this purpose, a choice model has to be built that explicitly relates utilities to choices in such a way that utilities can be inferred from the choices in a statistically rigorous process.

In a broad sense, discrete choice theory generates mathematical models that formally describe and explain the decision-process of an agent or a group of agents that make a choice between two or more mutually exclusive discrete alternatives from a finite choice set. A choice model is defined in terms of the input variables it includes, their associated parameters, a decision rule, and an error structure. These models are probabilistic in the sense that they generate choice probabilities for each alternative in a choice set. This means that choice models reflect not only that decision-makers are to some extent inconsistent or random in their behavior but also that the model does not capture all information that may be relevant for every single choice, as well as the fact that preferences differ across individuals.

We present a brief notation of a choice model to elucidate the relation between latent class based discrete choice theory and the re-conceptualized metanormative framework. The model will be based on Random Utility Theory, which assumes that the utility of an action a_i is the sum of observable (deterministic or systematic utility) and unobservable (random utility) components of the total utilities: $U_{in} = V_{in} + \epsilon_{in}$ (Manski 1977; Walker and Ben-Akiva 2002; Azari et al. 2012).

The unobservable component in Random Utility Theory is an error term that captures noise. Although we will not elaborate on the intricacies of this disturbance term, in the remainder of this paper it is important to note that depending on the assumed distribution of the term, there are different formulations of the probability that a randomly sampled individual from the population chooses a_i from the set of actions A. The by far most used formulation is the so-called Logit function which assumes that errors are extreme value distributed type I with variance $\frac{\pi^2}{6}$. In this formalization of the error term, the probability that a_i is chosen is written as: $P(a_i) = \frac{exp(V(a_i))}{\sum_{j=1}^{J} exp(V(a_j))}$. In a process of maximum-likelihood based optimization the utilities are obtained for each alternative which together (through their implied choice probabilities) make the observed choices the most likely. The result of this process is an econometric estimate $\hat{V}(a_j)$ for every alternative j in A, including i.

Whereas conventional choice models, like the Logit model introduced above, implicitly assume that the utilities of the population can be represented in one single estimate for each alternative, the latent class approach alleviates this restrictive assumption by postulating that there may exist several latent classes in the population, with homogeneous utilities within each class, which are different from those in other classes. In other words, the latent class choice model is based on the assumption that a number of segments or classes exist within the population featuring different preferences albeit internally relatively homogeneous (Greene and Hensher 2003). These models provide insights into heterogeneity of preferences and decision rules of people while accounting for the fact that different segments of the population have different needs and values and, in consequence, may exhibit different choice preferences. Since it is not known *a priori* which decision-makers belong to

each class, the segments are treated as latent rather than predetermined by the analyst. This means that the choice model provides a solution to the problem of unobserved heterogeneity. It determines simultaneously the number of (latent) segments and the size of each segment and it also estimates a separate set of utility parameters for each segment (Magidson et al. 2003; Magidson and Vermunt 2004; Kroesen 2014; Araghi et al. 2016).

The choice probability given by a latent class choice model is written as: $P(a_i) = \sum_t^T \left[P(t) \cdot P_t(a_i)\right]$. This means that the probability that a_i is chosen is the probability that a randomly sampled individual belongs to a class t (this is called the class membership probability and is denoted by $P(t)$ for individuals in class $t \in T$ where T is the set of all mutually exclusive and commonly exhaustive classes of decision-makers[2]) multiplied by the probability that a_i is chosen by a decision-maker from a particular class t ($P_t(a_i)$), summed over all classes. Estimation of such a latent class choice model results in not only an estimate of the probability that a randomly sampled individual belongs to a class (i.e., the share of the population that belongs to that class) ($\hat{P}(t)$), but also an estimate of class-specific utility for each alternative j and for each class t ($\hat{V}_t(a_j)$).

Revisiting re-conceptualized metanormative formulation introduced above $V(a_i) = \sum_t^T \left[P(t) \cdot V_t(a_i)\right]$ it is now clear that through Discrete Choice Analysis econometric estimations of the two crucial elements in this formula can be obtained, leading to the following discrete choice analysis-based formulation of the metanormative framework: $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$, which gives the estimated utility of action i, taking into account that people in different classes ascribe different utilities to that action.

4 Implementation of the Discrete Choice Analysis-based Formulation of the Metanormative Framework

To guide the process of Discrete Choice Analysis for the implementation of the metanormative framework, we use a two-pronged approach with long pedigree in the applied choice literature. It comprises the specification of the utility function and the collection of data to allow for a statistically rigorous estimation of the utilities of alternatives. While the formulations presented in the previous section are general, we will now interpret them in light of the data we use for our empirical proof of concept. These data resulted from a choice experiment that took place in 2017 which entailed a public consultation for a massive national transport infrastructure scheme among car commuters in The Netherlands. The scheme was specified in terms of its consequences on a number of dimensions relative to the *status quo* (Chorus et al. 2018).

For the utility specification, we build on the seminal work in consumer theory by Lancaster, who postulated that utility is derived not from goods but rather from

[2] The class membership probability is computed by a logit function which ensures that $\sum_t^T P(t) = 1$.

properties or characteristics of the goods (Lancaster 1966), and we follow the utilitarianism view on ethics by considering the utility of an action a_i to be a linear function of the action's consequences: $V_t(a_i) = \beta_{it} + \sum_m \beta_{mt} \cdot x_{mi}$. Here, $V_t(a_i)$ is the utility ascribed by a member of class t to action a_i; x_{mi} is the m*th* consequence of action a_i; β_{mt} is the weight attached by a member of class t to that consequence; and β_{it} is the remaining – i.e., not associated with any observed consequence – utility of the action. Importantly, a vector of weights or β is considered to represent a multidimensional moral theory. Further research is to be explored using more sophisticated behavioral representations such as Random Regret Minimization or Taboo models instead of a linear utility function (Chorus et al. 2017; Chorus 2010), as such allowing for a richer representation of various moral preference structures and decision rules.

Concerning the data collection, it is noted that choice experiments are widely used in the applied choice domain (Louviere et al. 2000). The key point in these experiments is to systematically vary the consequences (attributes) of actions (alternatives) and construct choice tasks by combining different hypothetical actions. Choices between these, or between one action and an opt out (*status quo* action), then give maximum information about preferences which allow for statistically efficient estimation of the weights β_{mt} for each consequence β_{it}. The data which we use in our empirical proof of concept resulted from a full factorial choice experiment[3] in which 99 respondents, composing a roughly representative sample of the Dutch regular car commuters, were presented with a series of binary choice situations with a clear *status quo* (current situation) and an alternative (a proposed infrastructure investment scheme). The infrastructure investment scheme was specified in terms of its positive or negative consequences with respect to vehicle ownership tax, travel time for the average commute trip, as well as the number of seriously injured in traffic, and the number of traffic fatalities[4]. The final experiment resulted in $99 \times 16 = 1584$ choice observations (Table 1 and Table 2 in Supplementary Information) (Chorus et al. 2018).

The utility of a particular infrastructure transport policy is written as $V_j = \sum_m \beta_m \cdot x_{jm} = \beta_{tax} \cdot tax_j + B_{time} \cdot time_j + \beta_{inj} \cdot inj_j + \beta_{fat} \cdot fat_j$. The utility of the opt out (status quo) option is merely defined in terms of a so-called alternative specific constant (ASC) which represents a generic preferences for or against the status quo versus an infrastructure investment scheme. In the sampled population, different classes are found featuring a different vector of β, which implies a different weighing of the consequences, thus defining the different trade-offs the members of a class are willing to accept. Because such trade-offs involve the well-being of humans, we postulate that we can use the vector of β to infer the morality of the classes in this particular context.

[3] A description of the full set of actions $i = 1$ to $i = 16$ is found in the the Supplementary Information.

[4] The consequences were effect-coded as [−1] for a decrease in the level of attributes and [1] for an increase in the level of attributes (Table 10). For example [− 1] on vehicle ownership tax means a decrease of 300 euros in the vehicle ownership tax per year.

Table 1 Attributes in choice experiment

Attributes	Increase/decrease
Vehicle ownership tax (Euros)	300 per year
Travel time (Minutes)	20 per working day
Non-fatal traffic injuries	100 per year
Traffic fatalities	5 per year

Table 2 Example of a choice task in the choice experiment

	Proposed transport policy
Vehicle ownership tax (per year, for each car owner including yourself)	300 euro **less** tax
Travel time (per working day, for each car commuter including yourself)	20 minutes **less** travel time
Number of seriously injured in traffic (per year)	100 seriously injured **more**
Number of traffic fatalities (per year)	5 traffic fatalities **more**
Your choice	☐ I support the proposed policy ☐ I oppose the proposed policy

It is relevant to note that, while our operationalization is well aligned with a consequentialist view on ethics, in this research we refrain from relating this morality vector to particular moral theories, such as deontology or consequentialism and their ramifications (Hooker 2003; Shafer-Landau 2012). Rather, we explore the subtle differences in contextual moral preferences that characterize different classes and are captured in the vector of β that defines each class. The empirical work to implement the discrete choice analysis-based metanormative theory is described below, followed by a proof of concept that allows us to investigate whether the policy choices made by an AI System, based on the conflicting input of different sized moral classes, would differ from the same choices made by an AI System that overlooks such differences (Tables 1 and 2).

5 Empirical Analysis

We describe the empirical approach that is employed to capture and investigate the relevance of moral uncertainty in AI decision-making by first elaborating on the model estimation, followed by a brief characterization of the classes in the model, and an inspection of the utility of actions per class.

5.1 Model estimation

To decide on the optimal number of latent classes, consecutive models with one through four classes were estimated and compared on the dataset that resulted

Table 3 BIC and Log-Likelihood function models 1-4

MODEL	BIC	LOG-LIKELIHOOD FUNCTION
Model 1 class	1479.290	− 721.226
Model 2 classes	1400.273	− 659.614
Model 3 classes	1369.509	− 622.129
Model 4 classes	1360.966	− 595.754

Table 4 Class membership probability for classes in three-class model

CLASSES	CLASS MEMBERSHIP PROBABILITY
Class 1: Financially-driven	0.14
Class 2: Want-it-all's	0.65
Class 3: Efficient	0.22

from the choice experiment described above. In general, the decision to select a certain number of latent classes is a trade-off between model fit (in terms of the log-likelihood) and parsimony (in terms of the number of classes/parameters) and interpretability.

Typically, such a decision is therefore guided by an information criterion, which weighs both model fit and parsimonity. In the context of latent class modeling, the Bayesian Information Criterion (BIC) criterion has been shown to perform well (Nylund et al. 2007). The BIC is a fit criteria for model selection that measures the trade-off between model fit and complexity of the model (Neath and Cavanaugh 2012). The equation used to calculate this criterion is $BIC = (\ln N) \cdot k - 2(\ln L)$ where N is the number of recorded measurements (e.g. choices), k is the number of estimated parameters, and L is the maximum value of the likelihood function for the model. A lower BIC indicates a better model.

In the present application, this statistic indicated that the optimal solution is one with four or more classes, which would be too many to interpret meaningfully (Table 3). A straightforward and practical alternative to the BIC is to compute the percentage increase in the log-likelihood of each model compared to the baseline one-class model. This measure reveals that after three classes there is no substantial increase in the relative fit of the model (LL increase > 4%).

5.2 Size and Features of the Classes

The classes that compose the three-class model have different sizes and defining features. For the assessment of the size of the classes, we recall the discrete choice analysis-based formulation of the metanormative framework that was introduced above: $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$. The estimate of the probability $\hat{P}(t)$ that a randomly sampled individual belongs to class t equals the relative share of the population that belongs to that class (Table 4). A vector of β estimated in the

Table 5 Estimated parameters in the classes of three-class model

Name	Value	Std err	t-test	p-value
Class 1: Financially-driven				
ACS Oppose	− 0.519	0.359	− 1.44	0.15
BETA Fat	− 0.561	0.298	− 1.88	0.06
BETA Inj	− 0.209	0.288	− 0.73	0.47
BETA Tax	− 2.56	0.339	− 7.54	0
BETA Time	− 0.119	0.253	− 0.47	0.64
Class 2: Want-it-all's				
ACS Oppose	1.52	0.136	11.16	0
BETA Fat	− 1.41	0.14	− 10.09	0
BETA Inj	− 1.92	0.169	− 11.32	0
BETA Tax	− 0.967	0.117	− 8.25	0
BETA Time	− 0.328	0.111	− 2.97	0
Class 3: Efficient				
ACS Oppose	1.24	0.222	5.59	0
BETA Fat	− 0.36	0.189	− 1.9	0.06
BETA Inj	− 0.745	0.186	− 4	0
BETA Tax	− 1.02	0.189	− 5.38	0
BETA Time	− 1.72	0.264	− 6.52	0
s2	1.54	0.333	4.62	0
s3	0.442	0.425	1.04	0.3

empirical process is associated to each class allowing us to understand the key defining features and the subtle differences in moral preferences that characterize the different classes (Table 5). We provide below an interpretation of the features of each class, along with the full estimation and interpretation of the parameters, that define each class (Table 5).

Class 1: Financially-driven The first class is the smallest segment in the three-class model (≈ 14%). Its members care only about lowering vehicle ownership taxes. We therefore infer that members of this class are financially-driven. All parameters have negative signs but only vehicle ownership tax is statistically significant, which means that for the members of this class the utility of a policy decreases if it features an increase in the vehicle ownership tax.

Class 2: Want-it-all's The second class is the largest segment in the three-class model (≈ 65%). Its members show a generic disposition against policies and a preference for lower vehicle ownership taxes, lower travel time, lower number of seriously injured in traffic, and lower number of traffic fatalities; they are especially concerned with lowering the number of seriously injured in traffic. We infer that members of this class are maximizers that want it all and believe that changes should only occur if a substantial improvement is secured, specifically in terms of road safety. All parameters are statistically significant and, with the exception of the alternative specific constant (ASC) for the status quo option, have negative signs. This means that for the members of this class, the utility of a policy

Table 6 Utility of policies $\hat{V}(a_i)$ in one-class model, three-class model, and class-specific utility of policies $\hat{V}_t(a_i)$ in classes [1-3] of three-class model

i	One-class Model	Three-class Model	Class 1	Class 2	Class 3
1	4.327	5.471	2.93	6.145	5.085
2	2.743	3.339	1.808	3.325	4.365
3	0.523	0.479	1.39	− 0.515	2.875
4	− 0.517	− 0.719	1.152	− 1.171	− 0.565
5	− 2.473	− 3.117	− 3.968	− 3.105	− 2.605
6	2.371	3.073	− 2.19	4.211	3.045
7	1.331	1.875	− 2.428	3.555	− 0.395
8	− 0.889	− 0.984	− 2.846	− 0.285	− 1.885
9	1.703	2.141	1.57	2.669	0.925
10	0.151	0.214	− 2.608	0.371	1.555
11	2.107	2.612	2.512	2.305	3.595
12	3.287	4.273	2.692	5.489	1.645
13	1.067	1.414	2.274	1.649	0.155
14	0.787	0.941	− 3.312	1.391	2.325
15	− 1.433	− 1.919	− 3.73	− 2.449	0.835
16	− 0.253	− 0.257	− 3.55	0.735	− 1.115

decreases if it features an increase in the vehicle ownership tax, travel time, number of seriously injured in traffic, or number of traffic fatalities.

Class 3: Efficient The third class accounts for $\approx 22\%$ of the sampled population. Members of this class care mostly about low travel time and therefore we infer that they are (time-) efficient. With the exception of the ASC for the status quo, all parameters have negative signs and are all statistically significant, except for the traffic fatalities parameter which has low significance. This means that the members of this class show a disposition against policies and consider that the utility of a policy decreases if it features an increase in the vehicle ownership tax, travel time or number of seriously injured in traffic.

5.3 Utility of Actions

We have so far determined the number of latent classes in the model and provided a brief description of each class. Now we proceed to inspect the utility and rank of actions (i.e. policies) in the three-class model and in each of its classes. We recall once again the discrete choice analysis-based formulation of the metanormative framework that was introduced above $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$ to accentuate that the estimate of the utilities $\hat{V}_t(a_i)$ of policies in each class is given by $V_j = \sum_m \beta_m \cdot x_{jm} = \beta_{tax} \cdot tax_j + B_{time} \cdot time_j + \beta_{inj} \cdot inj_j + \beta_{fat} \cdot fat_j$. Such class-specific utilities are subsequently multiplied by the class membership probability of each class $\hat{P}(t)$ and summed over classes for the purpose of estimating the utility of policies $\hat{V}(a_i)$. To facilitate the comprehension about the utilities in the three-class

Table 7 Kendall Tau b-test rank correlation of policies between one-class model and three-class model; one-class model and classes [1–3] of three-class model; and three-class model and classes [1–3] of three-class model

	Three-class model	Class 1	Class 2	Class 3
One-class model	1.000	0.700	0.833	0.790
Three class model	(N/A)	0.688	0.969	0.764

Table 8 Kendall Tau b-test rank correlation of policies between classes [1-3] in the three-class model

	Class 1	Class 2	Class 3
Class 1	N/A	0.533	0.483
Class 2	0.533	N/A	0.450
Class 3	0.483	0.450	N/A

model, we compare it with the baseline one-class model (Table 6)[5]. This comparison will take a new meaning in Section 6, as we conceptualize a society in which AI Systems equipped with one-class and three-class rules make policy decisions on behalf of society.

Unsurprisingly, the policy with highest utility in both models is $i = 1$ which entails lower vehicle ownership taxes, lower travel time, lower number of injuries, and lower number of fatalities; followed by $i = 12$ which entails higher travel time but lower vehicle ownership taxes, lower number of injuries, and lower number of fatalities; and by $i = 2$ which entails more traffic fatalities but lower vehicle ownership tax, lower travel time, and lower number of injuries (Table 6 in Supplementary Information). It is clear that the utility of the policies in the three-class model is highly influenced by the preferences of the *Want-it-all's*, which make up the largest class in the model.

In order to measure the rank correlation of the utility of policies in the baseline one-class model, the three-class model, and the different classes within the three-class model a Kendall Tau test was used (Table 7). We observe that the baseline one-class model and the three-class model have the same ranking of policies (correlation = 1.0), which raises questions about the relevance of taking into account moral uncertainty for the purpose of ranking actions. This will be addressed later (Section 6) as we will randomly generate thousands of policies and evaluate the implications of moral uncertainty across simulated cases. We also used the Kendall Tau coefficient test the rank correlation of the utility of policies among the three classes in the three-class model. The ranking of the policies by the members of the class 1 (*Financially-driven*) and class 2 (*Want it all's*) show a low correlation (correlation

[5] The parameters for the one-class model can be found in the Supplementary Information (Table 11).A description of the full set of actions $i = 1$ to $i = 16$ effect-coded as [−1] for a decrease in the level of attributes and [1] for an increase in the level of attributes can also be found in the Supplementary Information (Table 10).

Table 9 Standard deviation of the ranks for each policy in classes [1-3] in three-model class

i	Standard deviation
1	0
2	1.41
3	3.40
4	2.62
5	0
6	2.62
7	3.40
8	1.41
9	1.41
10	1.41
11	1.89
12	2.36
13	2.87
14	2.87
15	2.36
16	1.89

= 0.533) similarly to the rankings of the policies by (*Financially-driven*) and class 3 (*Efficient*) (correlation = 0.483) (Table 8).

Looking specifically at the classes within the three-class model, we observe that, similarly to what was reported above concerning the baseline one-class model and the three-class model, the policy with highest utility in all classes is $i = 1$ which entails lower vehicle ownership taxes, lower travel time, lower number of injuries, and lower number of fatalities. On the other hand, the policy with lowest utility in all three classes is $i = 5$ which entails higher vehicle ownership taxes, higher travel time, higher number of injuries, and higher number of fatalities. Moreover, by computing the standard deviation of the ranks for each policy in the different classes of the three-class model, we determine the discrepancy in rankings among the classes (Table 9). Three policies registered high discrepancies in rankings: *i=3*, *i=7*, and *i=13*. We remark that policy *i=3*, which entails lower vehicle ownership tax and travel time, and higher injuries and traffic fatalities[6], ranks substantially higher among the *Efficient* and *Financially-driven* compared with the *Want it all's*. Policy *i=7* entails higher vehicle ownership tax and travel time, and lower injuries and traffic fatalities ranks much higher among *Want it all's* when compared to the *Efficient* and *Financially-driven*. And finally policy *i=13*, which entails lower vehicle ownership tax, higher travel time and also higher number of traffic injuries, and lower fatalities, ranks substantially higher among the *Financially-driven* when compared to the *Efficient* and in a lesser extent to the *Want it all's*.

[6] We refer once again to Table 10 in Supplementary Information.

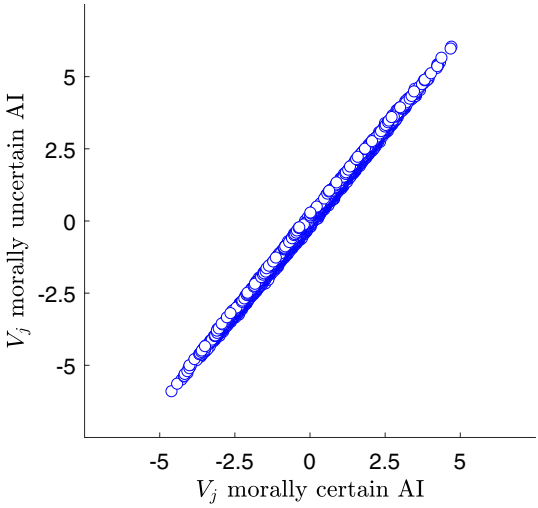

Fig. 1 Scatter plot of utilities of baseline and three-class morally uncertain model for 5000 randomly drawn policies

We have shown that there is variance in the utility assigned to particular policies by members of different classes. These discrepancies seem to arise when policies involve trade-offs as opposed to policies that merely have desirable or undesirable outcomes, which is relevant given that policies in general tend to involve trade-offs. Using the same dataset we will now provide a proof of concept to further explore the relevance of moral uncertainty in the context of AI.

6 Proof of Concept: AI Systems Make Policy Choices on Behalf of Societies

To illustrate the discrete choice analysis-based formulation of the metanormative framework, we capitalize on the fact that the AI field is traditionally lenient to remarkable thoughts. We therefore conceptualize AI Systems that make policy choices on behalf of a society. And we further investigate whether differences in policy choices arise as a result of accounting for moral uncertainty.

Building on the work outlined in previous sections, we consider an AI System equipped with a one-class rule and an AI System equipped with a three-class rule which factors in moral uncertainty. To make a larger action space available for the AI Systems, we randomly generated 5000 sets with 2 policies in each set by allowing the consequences of the policies, i.e. x_{jm} in the formulation of the utility of a infrastructure transport policy $V_j = \sum_m \beta_m \cdot x_{jm} = \beta_{tax} \cdot tax_j + B_{time} \cdot time_j + \beta_{inj} \cdot inj_j + \beta_{fat} \cdot fat_j$, to take on random values within the interval $[-1, 1]$ instead of taking only the extreme values as in the original dataset (Table 1 in Supplementary Information)[7].

[7] For example [-0.5] on vehicle ownership tax means a decrease of 150 euros in the vehicle ownership tax per year

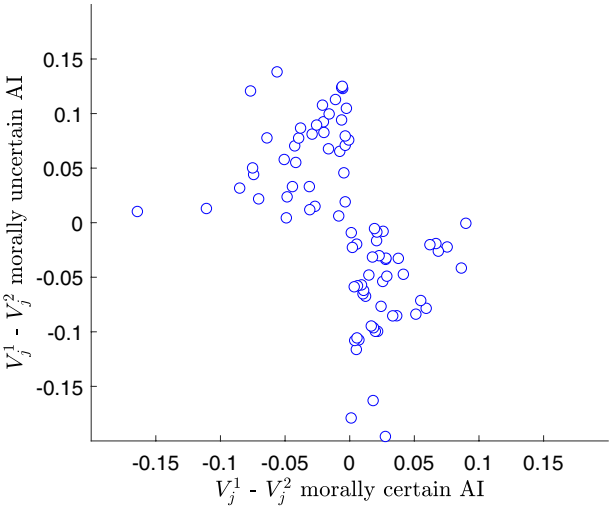

Fig. 2 Policies that caused disagreement among baseline and three-class morally uncertain model

The utility of each policy was estimated through the formulations introduced above for both the one-class baseline model (for the morally certain AI System) and for the three-class model (the morally uncertain AI System). The comparative value of the utilities of policies is interpreted as an indication of which policy is favored by each model and accordingly by the corresponding AI System. Although the utilities assigned to policies by the morally certain and the morally uncertain AI seem to be similar in most cases (Fig. 1), we observed 85 cases in which there was disagreement between the two models regarding the choice of the policy (Fig. 2).

In such instances of disagreement, the policy decisions of the morally uncertain AI system equipped with the three-class rule would contrast with the decisions of the morally certain AI equipped with the baseline one-class rule. Although the number of cases of disagreement is not large in this particular example it still allows us to hint at the potential relevance of capturing moral uncertainty. To visualize the contrasting policy decisions, we plotted the difference, for each of the 85 combinations of policies, between the utility of the first policy and that of the second policy for the morally certain AI System equipped with the one-class model (horizontal axis) and for the morally certain AI System equipped with the one-class model (vertical axis) (Fig. 2).

Out of the 85 cases mentioned above we selected three sets of policies that registered relatively high discrepancies between the utility values estimated for the baseline (morally certain) AI and those estimated for the morally uncertainty AI. The first set of policies that we selected featured a *time efficient and safe for injuries* policy favored by the base AI and a *safe but expensive and time inefficient* policy favored by the morally uncertain AI (Fig. 3).

Another set of policies featured a *time efficient but unsafe for fatalities* policy favored by the base model and a *safe but time inefficient* policy favored by the morally uncertain model (Fig. 4).

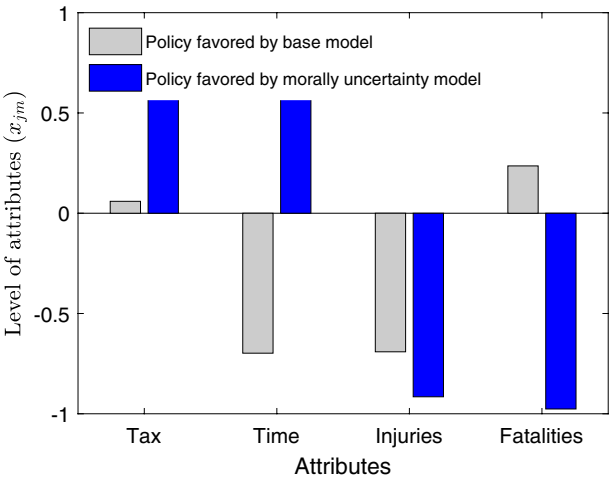

Fig. 3 Set of policies: *time efficient and safe for injuries* policy and *safe but expensive and time inefficient* policy

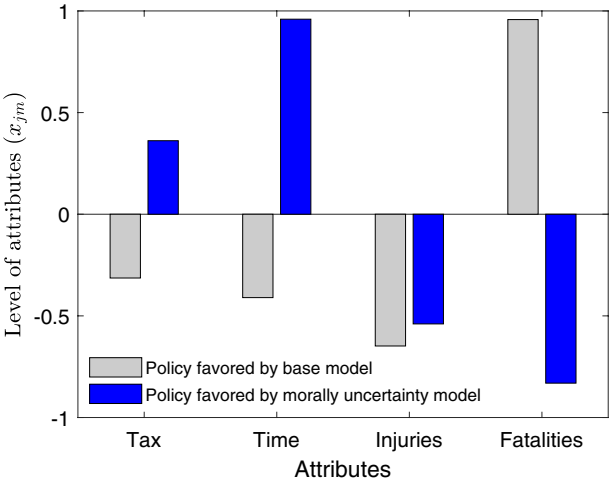

Fig. 4 Set of policies: *time efficient but unsafe for fatalities* policy and *safe but time inefficient* policy

Finally the third set of policies registering high discrepancies in utility value among the two competing models features a *time efficient but unsafe* policy favored by the base model and a *safe but expensive* policy favored by the morally uncertain model (Fig. 5).

These discrepancies in policies chosen by the morally certain AI and the morally uncertain AI emphasize the relevance of studying moral uncertainty and capturing it in AI systems.

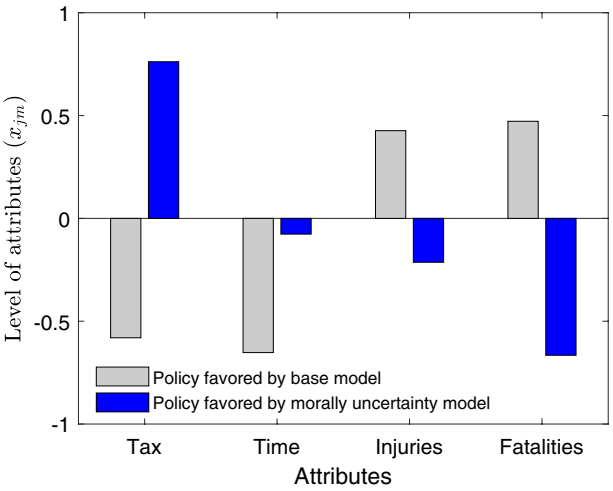

Fig. 5 Set of policies: *time efficient but unsafe* policy and *safe but expensive* policy

7 Conclusion and Discussion

As AI Systems become increasingly autonomous, they are expected to engage in complex moral decision-making processes. For the purpose of guidance of such processes, theoretical and empirical solutions within the controversial domain of Machine Ethics have been sought. In this research we integrate both theoretical and empirical lines of thought to address the matters of moral reasoning in AI Systems in a pragmatic yet statistically rigorous way that is firmly connected to theoretical considerations related to normative uncertainty in AI systems.

Our approach is built on the theoretical notion that moral uncertainty is paramount to the endeavor of implementing moral reasoning in AI Systems. More specifically, it employs the metanormative framework for decision-making under moral uncertainty, as construed by William MacAskill (2014) and Kyle Bogosian (2017), and re-conceptualizes it as a latent class discrete choice model. We assume that a number of classes featuring different preferences exist within a population where each class is internally relatively homogeneous in terms of its behaviors and preferences (Greene and Hensher 2003). By codifying the moral preferences and decision rules of different classes, and aggregating them across the population, we are able to obtain a moral representation for the AI System: its resulting normative uncertainty is embedded in the form of an empirical model of the moral heterogeneity of the society it represents.

In the empirical installment of our approach we specify a multi-dimensional utility function which represents a moral theory or set of moral preferences (i.e., weights attached to different criteria), and we allow this vector of weights to vary across classes. The final ranking of the actions available to the AI System is provided by computing the class membership weighted average of the utility of each action in each class. Importantly, our approach does not involve the analyst a priori

selecting classes and class sizes in the population, they rather emerge - just like the class-specific weights assigned to each criteria - in the process of estimating the choice model from observed choice data.

The discrete choice analysis-based formulation of the metanormative framework is theory-rooted and practical, as it captures moral uncertainty through a small set of latent classes, thus avoiding runtime issues which are common in applications that aim to capture the full level of individual-to-individual heterogeneity in the population.

For the purpose of illustrating our approach we conceptualize a society in which AI Systems are in charge of making policy choices. In the proof of concept two AI systems make policy choices on behalf of a society, but while one of the systems uses a baseline morally certain model the other uses a morally uncertain model. Specifically, we used our approach in a dataset that resulted from a choice experiment that took place in 2017 which entailed a consideration for a massive national transport infrastructure scheme among car commuters in The Netherlands, having implications on morally salient dimensions such as the number of road fatalities. It was observed that there are cases in which the two AI Systems disagree about the policy to be chosen, which we believe is an indication about the relevance of moral uncertainty.

We are aware that our finding that a morally uncertain AI might in some cases decide differently than a morally certain AI not only validates the notion that moral uncertainty is a topic worthy of further investigation by the AI community, but that it also generates another question: in cases where the two AIs would make different decisions, which AI should prevail? This question is not one with a clear-cut answer, but the following observations and considerations could help address this matter.

First, it is important to revisit the starting point of this research: our aim was to present an approach to capture moral uncertainty in AI that a) has a firm theoretical foundation, and b) is empirically and practically feasible. To achieve both aims simultaneously, we proposed a latent class approach, which is a compromise between not acknowledging moral uncertainty at all (i.e., estimating one set of moral preferences to represent an entire society's morality) and taking into account subtle moral differences between each and every individual in society. By allowing a small number of latent classes with distinct morality to emerge in a process of econometric model estimation, we connect to the theoretical framework of normative uncertainty, and avoid run-time issues which plague the estimation and application of individual-level models. Building on this argument, we believe that, in general terms, the decisions made by a morally uncertain AI (equipped with a latent class choice model of morality) should be preferred to the decisions made by an AI that is morally certain.

Whether or not the decisions made by a morally uncertain AI equipped with a limited number of latent moral classes are to be preferred over those made by an AI that tries to embed the morality of each individual in society (supposing that this would be feasible), is another matter.

Here, we propose to use the notion of Occam's razor, which puts a premium on the most simple explanation behind empirical observations. In statistics, this generic scientific notion is operationalized in metrics such as the adjusted rho-square, the Bayesian Information Criterion and others, which penalize a model for the number

of parameters it uses to explain data. In the field of machine learning, this relates to the notion of regularization, which prevents artificial neural networks from overfitting training data-sets. Such statistical tools and metrics offer a formal, theory-rooted approach to select one model (or: AI system) over another. For example, in our case, allowing for a small number of latent moral classes clearly led to a better explanation for the observed choice data than a model that attempts to embed all society's preferences in one utility function, also after correcting for the increased number of parameters.

However, further increasing the number of latent classes beyond a handful inevitably leads to increases in the number of parameters that no longer are offset by increases in explanatory power. The same holds for morally uncertain AI Systems that aim to capture the differences in morality between each and every individual: the resulting model, besides being difficult to handle in real time decision contexts, will most likely be statistically inferior to more parsimonious models that attempt to cluster individuals that are relatively speaking like-minded in terms of their morality. In sum, by employing statistical model selection techniques that appropriately penalize for the number of parameters used, helps the designer of the AI choose the optimal level of heterogeneity (uncertainty) to embed in the AI.

The novelties in this research are the idea that discrete choice analysis can be used to codify human morality and, as such, provides a tool to embed morality in AI systems; moral uncertainty can be operationalized by re-conceptualizing the metanormative framework of normative uncertainty through latent class choice models; and also the empirical illustration of the concept of moral uncertainty. We acknowledge that our re-conceptualization fails to take into account the richness and subtleties of the work developed originally by MacAskill (2014, 2016) yet opening an avenue for further research that accommodates its various extensions. Additionally, instead of using a linear utility function, as it was the case in this research, other utility functions such as Random Regret Minimization (Chorus 2010), taboo trade off aversion, or lexicographic choice may be explored, as these indirectly refer to different moral theories. Finally, through the proof of concept this research also opens avenues for further research on the meaning and practical implications of moral uncertainty in artificial decision-making.

Supplementary Information The online version of this article (https://doi.org/10.1007/s11023-021-09556-9) contains supplementary material, which is available to authorized users.

Acknowledgements The authors acknowledge the European Research Council for financial support of this research (ERC Consolidator grant BEHAVE—724431).

Funding European Research Council Consolidator grant BEHAVE – 724431.

References

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, *21*(4), 12–17.

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge: Cambridge University Press.

Anderson, M., Anderson, S.L., & Armen, C. (2004). Towards machine ethics. In AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA.

Araghi, Y., Kroesen, M., Molin, E., & Van Wee, B. (2016). Revealing heterogeneity in air travelers' responses to passenger-oriented environmental policies: A discrete-choice latent class model. *International Journal of Sustainable Transportation*, *10*(9), 765–772.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Azari, H., Parks, D., & Xia, L. (2012). Random utility theory for social choice. In Advances in Neural Information Processing Systems (pp. 126–134).

Backlund, A. (2000). The definition of system. *Kybernetes*, *29*(4), 444–451.

Ben-Akiva, M. E., Lerman, S. R., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand* (Vol. 9). Cambridge: MIT press.

Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., et al. (2018). Autonomous vehicles require socio-political acceptance–an empirical and philosophical perspective on the problem of moral decision making. *Frontiers in Behavioral Neuroscience*, *12*, 31.

Bigman, Y. E., & Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature*, *579*(7797), E1–E2.

Bogosian, K. (2017). Implementation of moral uncertainty in intelligent machines. *Minds and Machines*, *27*(4), 591–608.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654. URL https://science.sciencemag.org/content/352/6293/1573

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 355–372.

Cervantes, J. A., López, S., Rodríguez, L. F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. Science and Engineering Ethics (pp. 1–32).

Chorus, C. G. (2010). A new model of random regret minimization. *European Journal of Transport and Infrastructure Research, 10*(2).

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

Chorus, C. G., Pudāne, B., Mouter, N., & Campbell, D. (2018). Taboo trade-off aversion: A discrete choice model and empirical analysis. *Journal of Choice Modelling*, *27*, 37–49.

Dignum, V. (2017). Responsible artificial intelligence: Designing ai for human values. *Discoveries*, *1*, 1–8.

Dobbe, R., Gilbert, T. K., & Mintz, Y. (2019). Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments. arXiv preprint arXiv:1911.09005.

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G., & König, P. (2018). Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. Science and engineering ethics (pp. 1–20).

Feller, A., Pierson, E., Corbett-Davies, S., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. The Washington Post.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, *80*, 38.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379.
Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.
Fritz, A., Brandt, W., Gimpel, H., & Bayer, S. (2020). Moral agency without responsibility? analysis of three ethical models of human-computer interaction in times of artificial intelligence (ai). *De Ethica*, *6*(1), 3–22.
Goodall, N. J. (2016). Can you program ethics into a self-driving car? *IEEE Spectrum*, *53*(6), 28–58.
Greene, W. H., & Hensher, D. A. (2003). A latent class model for discrete choice analysis: Contrasts with mixed logit. *Transportation Research Part B: Methodological*, *37*(8), 681–698.
Harris, J. (2020). The immoral machine. *Cambridge Quarterly of Healthcare Ethics*, *29*(1), 71–79. https://doi.org/10.1017/S096318011900080X.
Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. Ethical Theory and Moral Practice (pp. 1–16).
Hooker, B. (2003). Rule consequentialism.
Hunyadi, M. (2019). Artificial moral agents. really? In Wording Robotics (pp. 59–69). Springer.
Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, *26*(1), 293–307.
Klenk, M. (2020). How do technological artefacts embody moral values? Philosophy & Technology (pp. 1–20).
Kroesen, M. (2014). Modeling the behavioral determinants of travel behavior: An application of latent transition analysis. *Transportation Research Part A: Policy and Practice*, *65*, 56–67.
Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, *74*(2), 132–157.
Lin, P. (2016). Why ethics matters for autonomous cars. In Autonomous driving (pp. 69–85). Springer, Berlin, Heidelberg.
Lockhart, T. (2000). *Moral uncertainty and its consequences*. Oxford: Oxford University Press.
Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: Analysis and applications*. Cambridge: Cambridge University Press.
Lundgren, B. (2020). Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles. AI & SOCIETY (pp. 1–11).
MacAskill, W. (2014). Normative uncertainty. Ph.D. thesis, University of Oxford.
MacAskill, W. (2016). Normative uncertainty as a voting problem. *Mind*, *125*(500), 967–1004.
MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral Uncertainty*. Oxford: Oxford University Press.
Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). Ai-based digital assistants. Business & Information Systems Engineering (pp. 1–10).
Magidson, J., Eagle, T., & Vermunt, J. K. (2003). New developments in latent class choice models. In Sawtooth Software Conference Proceedings (pp. 89–112).
Magidson, J., & Vermunt, J. K. (2004). Latent class models. The Sage handbook of quantitative methodology for the social sciences (pp. 175–198).
Manski, C. F. (1977). The structure of random utility models. *Theory and Decision*, *8*(3), 229–254.
McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.
Neath, A. A., & Cavanaugh, J. E. (2012). The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199–203.
Nissan-Rozen, I. (2015). Against moral hedging. *Economics & Philosophy*, *31*(3), 349–369.
Noothigattu, R., Gaikwad, S. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. D. (2018). A voting-based system for ethical decision making. In Thirty-Second AAAI Conference on Artificial Intelligence.
Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569.
Poulsen, A., Anderson, M., Anderson, S. L., Byford, B., Fossa, F., Neely, E. L., Rosas, A., & Winfield, A. (2019). Responses to a critique of artificial moral agents. CoRR **abs/1903.07021**. URL http://arxiv.org/abs/1903.07021
Powers, T. M. (2006). Prospects for a kantian machine. *IEEE Intelligent Systems*, *21*(4), 46–51.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.
Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica*, *15*(60), 243–253.
Shafer-Landau, R. (2012). *Ethical theory: an anthology* (Vol. 13). New York: Wiley.
Thomson, J. J. (1984). The trolley problem. *Yale LJ*, *94*, 1395.
Tonkens, R. (2009). A challenge for machine ethics. *Minds and Machines*, *19*(3), 421.
Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge: Cambridge University Press.
van Hartskamp, M., Consoli, S., Verhaegh, W., Petkovic, M., & van de Stolpe, A. (2019). Artificial intelligence in clinical health care applications. *Interactive Journal of Medical Research*, *8*(2), e12100.
van de Poel, I. (2020). Embedding values in artificial intelligence (ai) systems. Minds and Machines (pp. 1–25).
van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, *25*(3), 719–735.
Walker, J., & Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, *43*(3), 303–343.
Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.
Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society*, *22*(4), 565–582.
Wexler, R. (2017). When a computer program keeps you in jail: How computers are harming criminal justice. *New York Times, 13*.
Wolkenstein, A. (2018). What has the trolley dilemma ever done for us (and what will it do in the future)? on some recent debates about the ethics of self-driving cars. Ethics and Information Technology (pp. 1–11).
Zhao, H., Dimovitz, K., Staveland, B., & Medsker, L. (2016). Responding to challenges in the design of moral autonomous vehicles. In The 2016 AAAI Fall Symposium Series: Cognitive Assistance in Government and Public Sector Applications, Technical Report FS-16-02 (pp. 169–173).
Żuradzki, T. (2015). Meta-reasoning in making moral decisions under normative uncertainty. In 2016). Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon, vol. 2 (pp. 1093–1104).