Computer says YES!

Towards a human inspired moral compass for AI

Caspar Chorus

Professor of choice behavior modeling

Head of department: Engineering Systems & Services

Entrepreneur (disclaimer…)

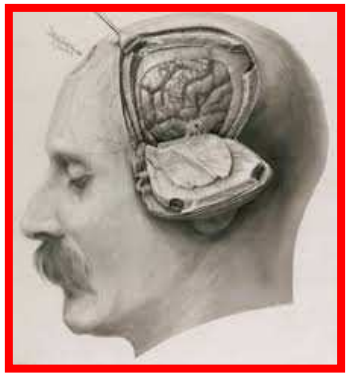
What to expect?

A quick introduction into the ERC-BEHAVE program: models of moral decision making of humans an AI

An empirical and model-based study of how humans make taboo-trade-offs

How to create a human-inspired moral compass for a morally uncertain AI.

Choice analysis on one slide



Core idea: your choices offer a window into your brain

(as long as I have a good choice model)

1. Build a choice model (based on behavioural science)

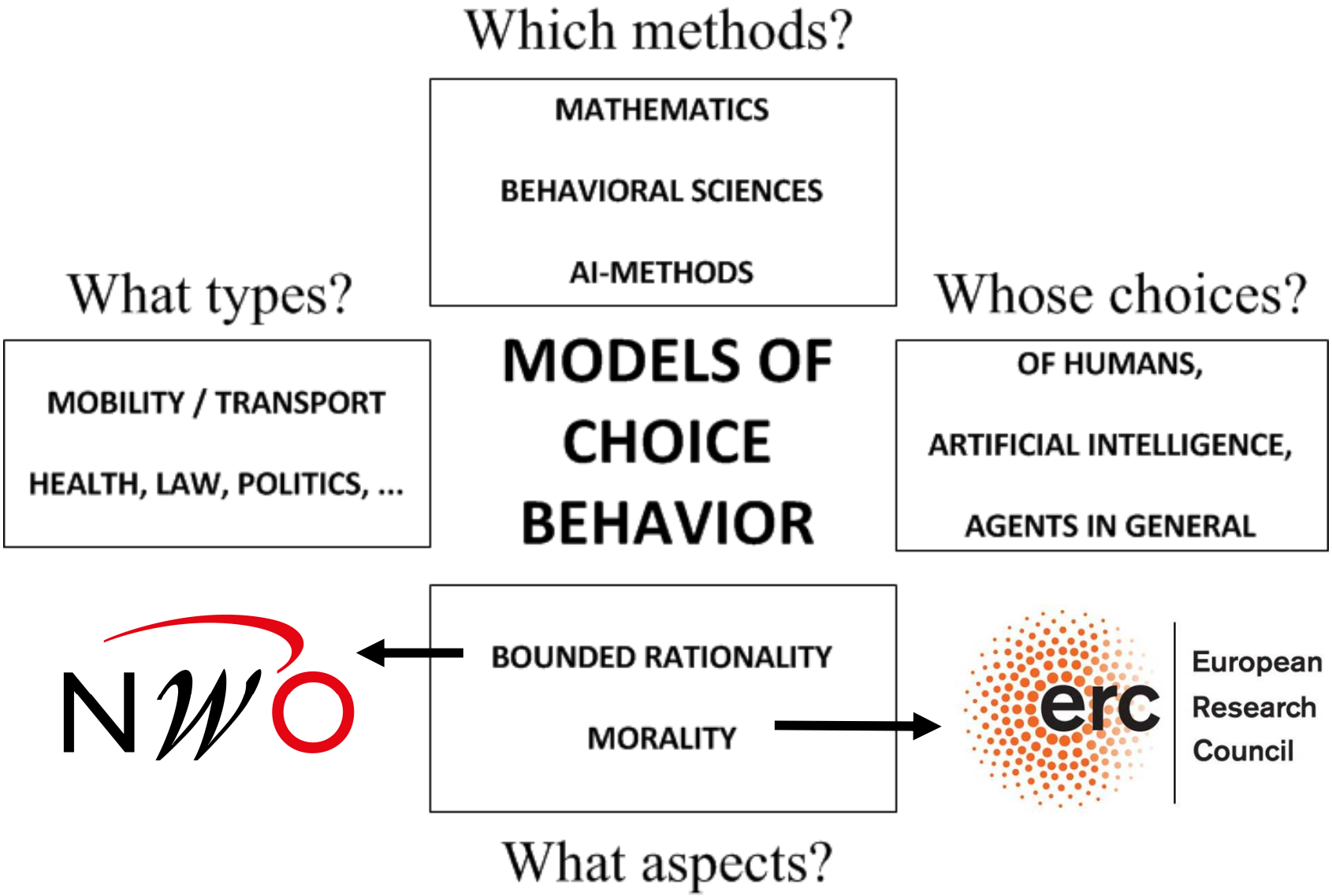
2. Observe people's choices (in real life or experiments)

3. Estimate/validate model, infer preferences, trade-offs, decision rules

4. Based on these inferences, predict future choices

Widely used throughout Social Sciences; Nobel for McFadden (2000)

My (team's) research on one slide



Mind the gap!
The Morality of Choice

Non-moral, 'consumer' choices

- **Optimal** decisions
- Budget constraints, trade-offs

Moral Choices

- **Right** versus wrong
- Heuristics, norms,…

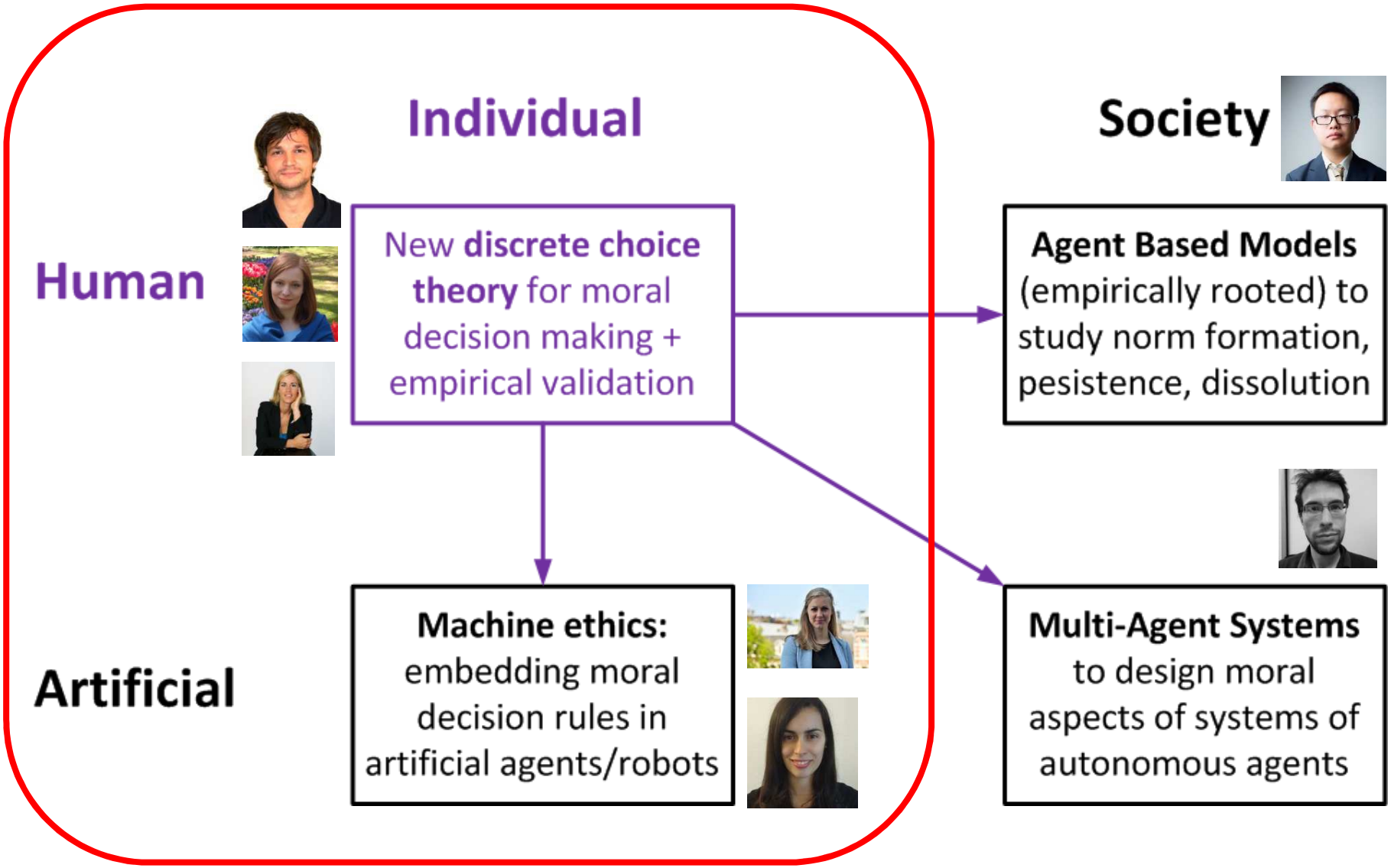
The missing piece of the puzzle

CONTRIBUTION	Economics, Decision Science	Behavioral Sciences
Consumer choices	*'Consumer' Choice models* ←	*Consumer Psychology*
Moral choices		*Moral Psychology*

The missing piece of the puzzle

CONTRIBUTION	Economics, Decision Science	Behavioral Sciences
Consumer choices	*'Consumer' Choice models* ←	*Consumer Psychology*
Moral choices	*'Moral' Choice models* ←	*Moral Psychology*

Not just *human* agents...



Taboo trade-off aversion: A choice model and empirical analysis

Caspar Chorus, N. Mouter, B. Pudane, D. Campbell

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

European Research Council

What is a taboo trade-off?

Willing to sacrifice an hour of travel time to meet a friend, inform how he is doing.

My Value of Time = €20 / hour

NOT willing to pay him €20 to come over to me instead...

'paying' in terms of time, attention: OK. In terms of money: taboo.

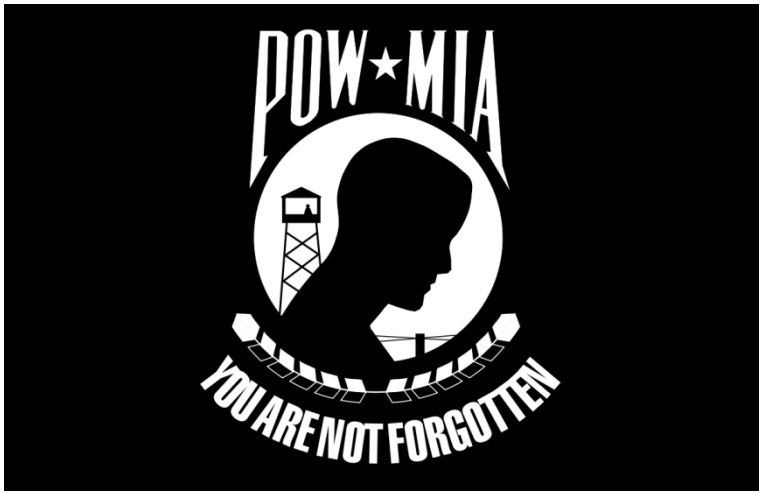
Why?

- Time, friendship belong to the same sphere (social relations)
- Money belongs to a different sphere (economic transactions)

What is a taboo trade-off? (II)

(The Economist, 17 March 2017)

± 700,000 USD per identified and repatriated remains of a single US soldier (MIA).



"You cannot associate a dollar value with this national imperative," says General Spindler.

The mere idea of trading off the anguish of left-behind families against budget constraints, is awkward and politically dangerous.

What is a taboo trade-off? (IV)

Key concept in Moral Psychology (**Tetlock**), Economic Law (Radin)

People hesitate, refuse to trade off 'sacred' values with non-sacred ones (usually money):

- Love *versus money*
- Health of one's child *versus money*
- Loyalty to one's country *versus money*

Since Lancaster (1966), Keeney & Raiffa (1976), trade-offs at the core of decision theory, microeconomic consumer theory.

Our contribution: tractable model of decision-making that allows for taboo trade off aversion + an empirical test.

Empirical context

Support or oppose comprehensive national infrastructure plan.

Effects in terms of **in**crease or **de**crease in:

• Vehicle ownership tax (€)	300	p. year	TAX
• Travel time (min.)	20	p. working day	TIME
• Non-fatal traffic injuries	100	p. year	INJ
• Traffic fatalities	5	p. year	FAT

Some examples of trade-offs

TAX ↓ & TIME ↑ : Secular trade-off

TAX ↓ & FAT ↑ : **Taboo** trade-off

INJ ↓ & FAT ↑ : Tragic trade-off

Data

Specifically designed Stated Choice survey (see earlier slide)

Experimental design: full factorial (every combination occurs)

Ensures (theoretical) identification of taboo-penalties and tastes

9 out of 16 tasks contained (1, 2, 3 or 4) taboo trade-offs

Sample of 99 representative regular car commuters, 16 choice tasks

First: pilot study (20 people), interviews with respondents.

Final data collected February 2017, random sample Dutch >18.

Example choice task

	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

A conventional linear model

- Policy variant j constitutes change w.r.t. Status Quo
  (V_{SQ} = utility of Status Quo, i.e. of opposing the policy)

$$V_j = \sum_m \beta_m \cdot x_{jm} = \beta_{tax} \cdot tax_j + \beta_{time} \cdot time_j + \beta_{fat} \cdot fat_j + \beta_{inj} \cdot inj_j$$

$$P(j) = \frac{\exp(V_j)}{\exp(V_j)+\exp(V_{SQ})} = \frac{\exp(\sum_m \beta_m x_{jm})}{\exp(\sum_m \beta_m x_{jm})+\exp(V_{SQ})}$$

- m and n denote attributes, x attribute-values, β attribute weights
- Linear utility function, implies fully compensatory decision making.
- Weights (β_m) found by means of Maximum Likelihood Estimation

Modeling taboo trade-off aversion

- The following, generic specification is adopted:

$$V_j^{TTOA} = \sum_m \beta_m \cdot x_{jm} + \tau_G \cdot \max_{(m,n)\in T} I_{m\to n}$$

- T represents the set of ordered pairs (m, n) where m is a 'sacred' attribute and n is a 'secular' attribute
- I indicates taboo trade-off: a worse value is accepted for m to obtain a better value for n
- τ_G is generic taboo-penalty associated with having **one or more** taboo-trade offs embedded in the policy alternative

Results – Taboo trade-off aversion

Mixed Logit (Panel), 4000 draws
(all parameters $\sim N$.)

Null log-likelihood: -1098
Final log-likelihood: -589

Name	Value	Rob.SE	Rob.t	Rob. p
V_SQ	1.48	0.354	4.19	0.00
BETA_Fat	-1.52	0.234	-6.50	0.00
BETA_Inj	-2.19	0.310	-7.07	0.00
BETA_Tax	-2.27	0.330	-6.87	0.00
BETA_Time	-1.25	0.227	-5.50	0.00
SIGMA_OPPOSE	1.36	0.336	3.70	0.00
SIGMA_Fat	1.03	0.249	4.12	0.00
SIGMA_NonFat	1.75	0.384	4.57	0.00
SIGMA_Tax	1.58	0.253	6.23	0.00
SIGMA_Time	1.31	0.272	4.82	0.00
BETA_Taboo	**-1.02**	**0.473**	**-2.16**	**0.03**
SIGMA_Taboo	**2.14**	**0.499**	**4.29**	**0.00**

Effects on parameters, choice probs.

Parameters

- Relative to Taboo-model, linear RUM **overestimates** importance of traffic fatality, injury parameters (both 19% inflated)
- Correlation found between weights of injuries and fatalities, but **not** between these weights and taboo penalty!
- Much heterogeneity: *deontologists, utilitarians, "I don't-care-ans"*

Choice probabilities

- Relative to linear model, Taboo model assigns lower support for policies which contains taboo trade-off(s)
- On our data, Taboo model predictions much closer to observed support-levels

Computer says “I don’t know”

How to build a morally uncertain AI using Latent Class choice models

Andreia Martinho
Maarten Kroesen
Caspar Chorus

TU Delft Delft University of Technology
Challenge the future

erc European Research Council

Context: Evolution of AI – PAST

Analysis of BigData ("pattern recognition", "classification")

Radiologist/Surgeon:	is this a tumour / what kind of?
Stock-trader:	what is the risk profile of this investment?
Autonomous drone:	is this friend or foe (civilian or not)?
Immigration office:	is this a migrant or refugee (radicalized)?
HR department:	what kind of CV is this?

Context: Evolution of AI – PRESENT

AI-powered autonomous systems ("decision-making")

Radiologist/Surgeon: is this a tumour / what kind of?
Operate or not / which kind of treatment?

Stock-trader: what is the risk profile of this investment?
To invest or not, how much and when to divest?

Autonomous drone: is this friend or foe (civilian or not)?
Shoot or not? First fire warning shot?

Immigration office: is this a migrant or refugee (radicalized)?
Admit to the country? With which status?

HR department: what kind of CV is this?
Invite for interview, offer job?

Challenge

Great and justified societal anxiety; fear of losing:

Meaningful Human Control over autonomous artificial agents (AAA)

Important conditions for MHC:

- We need to understand fully **why** AAA decided to choose 'A or B'
- The AAA's motivations, preferences, values need to **align** w/ ours

So that a human can always be called to answer for AAA's choices

In practice, devilish trade-off between unleashing the full capacity of AI (e.g. deep learning) and retaining MHC (e.g. rule-based).

Solution:

Discrete Choice Analysis for AI

[Use Machine Learning to analyse data; black box, highly flexible, surpasses human capabilities]

Use **Econometrics** (Discrete Choice Analysis) to help the AAA make explainable **decisions** that align with human (moral) values.

DATA: Discrete Choice Experiments

- Carefully crafted and statistically efficient choice tasks
- Participants: professionals / domain experts

MODEL: Discrete Choice Theory

- Use observed choices to estimate weights for criteria, trade-offs
- And derive decision rules implicit in people's behaviour

Example
(far-fetched, just for illustration)

Wish to develop an AI-system that makes transport policy choices.

Based on Deep Learning, it can predict any future transport policy's effects on taxes, time-gains, injuries and fatalities.

But we do not want the AI to weigh those aspects and choose a policy, based on opaque neural networks.

Hence: use choice experiment and choice analysis to design human (citizen) inspired moral compass for the AI.

But: how to deal with **heterogeneity among citizens**??

A morally diverse society

The utility V of an action a_i:

$V(a_i) = \sum_{t \in T}[P(t) \cdot V_t(a_i)]$

where $V_t(a_i)$ denotes the utility of a_i given a particular normative theory t taken from the set T of available theories;

and $P(t)$ is the **share of the population that adheres to the theory** (as implicitly underlying the choices they make).

"moral heterogeneity within society"

Latent Class Choice Model: weights estimated, and sizes of classes

A morally uncertain AI

The choice-worthiness W of an action a_i:

$$W(a_i) = \sum_{t \in T} [C(t) \cdot W_t(a_i)]$$

where $W_t(a_i)$ denotes the choice-worthiness of a_i given a particular normative theory t taken from the set T of available theories;

and $C(t)$ denotes the credence of the theory.

"moral conflict within the AI"

Bogosian (2017): AIs **should be** morally uncertain
(building on MacAskill 2014)

Empirical context

Support or oppose comprehensive national infrastructure plan.

Effects in terms of *increase* or *decrease* in:

• Vehicle ownership tax (€)	300	p. year	TAX
• Travel time (min.)	20	p. working day	TIME
• Non-fatal traffic injuries	100	p. year	INJ
• Traffic fatalities	5	p. year	FAT

$$V_j = \sum_m \beta_m \cdot x_{jm} = \beta_{tax} \cdot tax_j + \beta_{time} \cdot time_j + \beta_{fat} \cdot fat_j + \beta_{inj} \cdot inj_j$$

- Try and find classes of morally 'like-minded' people
- E.g. large weights for safety or for tax-breaks
- For the moment, ignore taboo trade off aversion

Name	Value	Std err
Class_1_ASC_Oppose	-0.519	0.359
Class_1_BETA_Fat	-0.561	0.298
Class_1_BETA_NonFat	-0.209	0.288
Class_1_BETA_Tax	**-2.56**	**0.339**
Class_1_BETA_Time	-0.119	0.253
Class_2_ASC_Oppose	1.52	0.136
Class_2_BETA_Fat	**-1.41**	**0.14**
Class_2_BETA_NonFat	**-1.92**	**0.169**
Class_2_BETA_Tax	-0.967	0.117
Class_2_BETA_Time	-0.328	0.111
Class_3_ASC_Oppose	1.24	0.222
Class_3_BETA_Fat	-0.36	0.189
Class_3_BETA_NonFat	-0.745	0.186
Class_3_BETA_Tax	**-1.02**	**0.189**
Class_3_BETA_Time	**-1.72**	**0.264**

14% tax-avoiders

65% safety-deliberators

21% tax- and time-conscious

Does this matter?

Q: Does a morally uncertain AI make different choices than an AI calibrated on 'average Joe'?

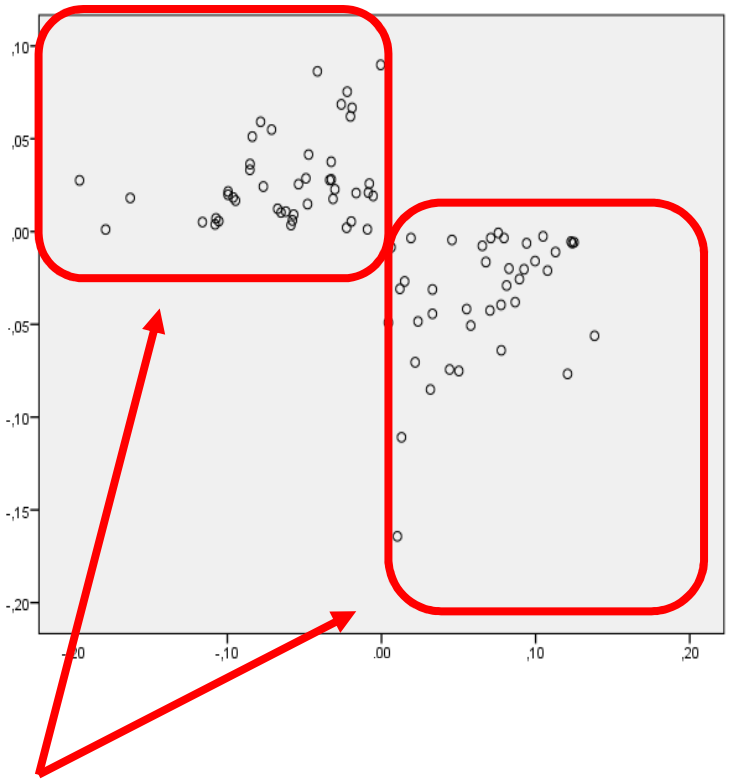
A: Sometimes…

Method:

We randomly created 10,000 policies within the boundaries of the choice experiment.

Had the two AI's decide to support or oppose.

In 1.5% of case, they disagree



Take-aways

A quick introduction into the ERC-BEHAVE program: models of moral decision making of humans an AI

An empirical and model-based study of how humans make taboo-trade-offs

How to create a human-inspired moral compass for a morally uncertain AI.

Thank you!

European Research Council