A RACE TO CONSTRAINT

EMBEDDING ETHICAL AND CULTURAL PREFERENCES IN AUTONOMOUS VEHICLE PLANNING

HTTPS://GITHUB.COM/RASHKOVAN/MORALMACHINE

Sonya Rashkovan / Julia Zdzilowska

Why taking on this project

Love cars

AVs are a technology that is converging the physical world and daily life around the world with sophisticated technology.

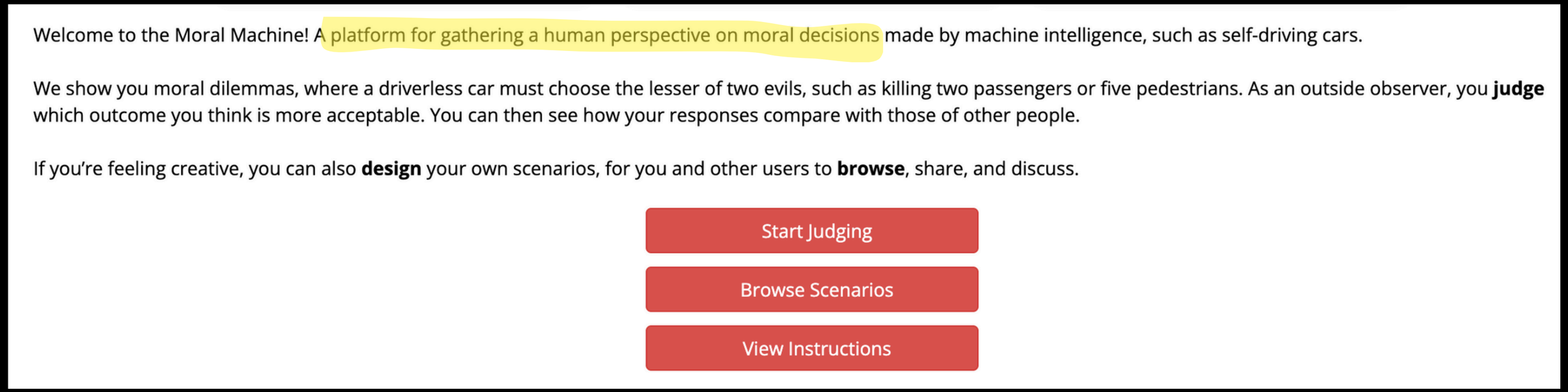
Curiosity about embedding cultural preferences

There are clearly different driving cultures and styles, while tech solutions often attempt to be universal.

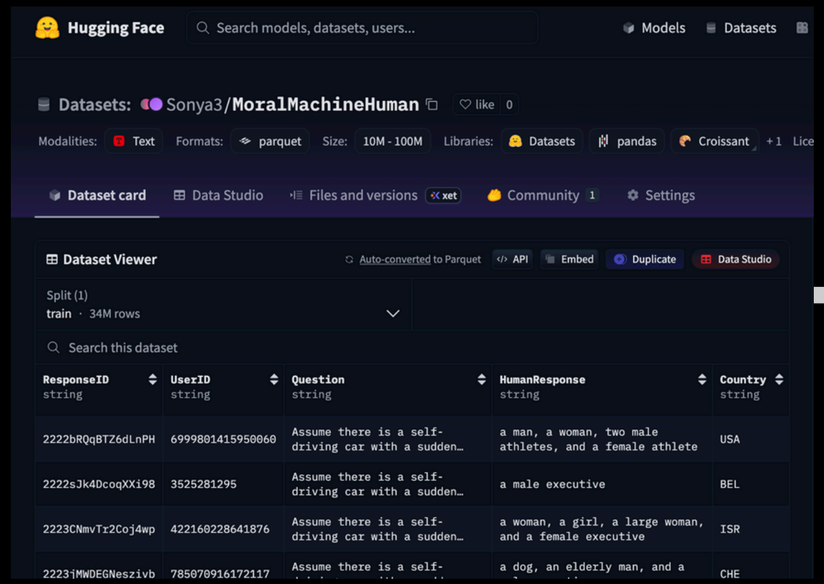
Interest in debate and social deliberation

AVs are one of the technologies that countries/states have been hesitant to implement, so there’s been much public discourse on the topic.

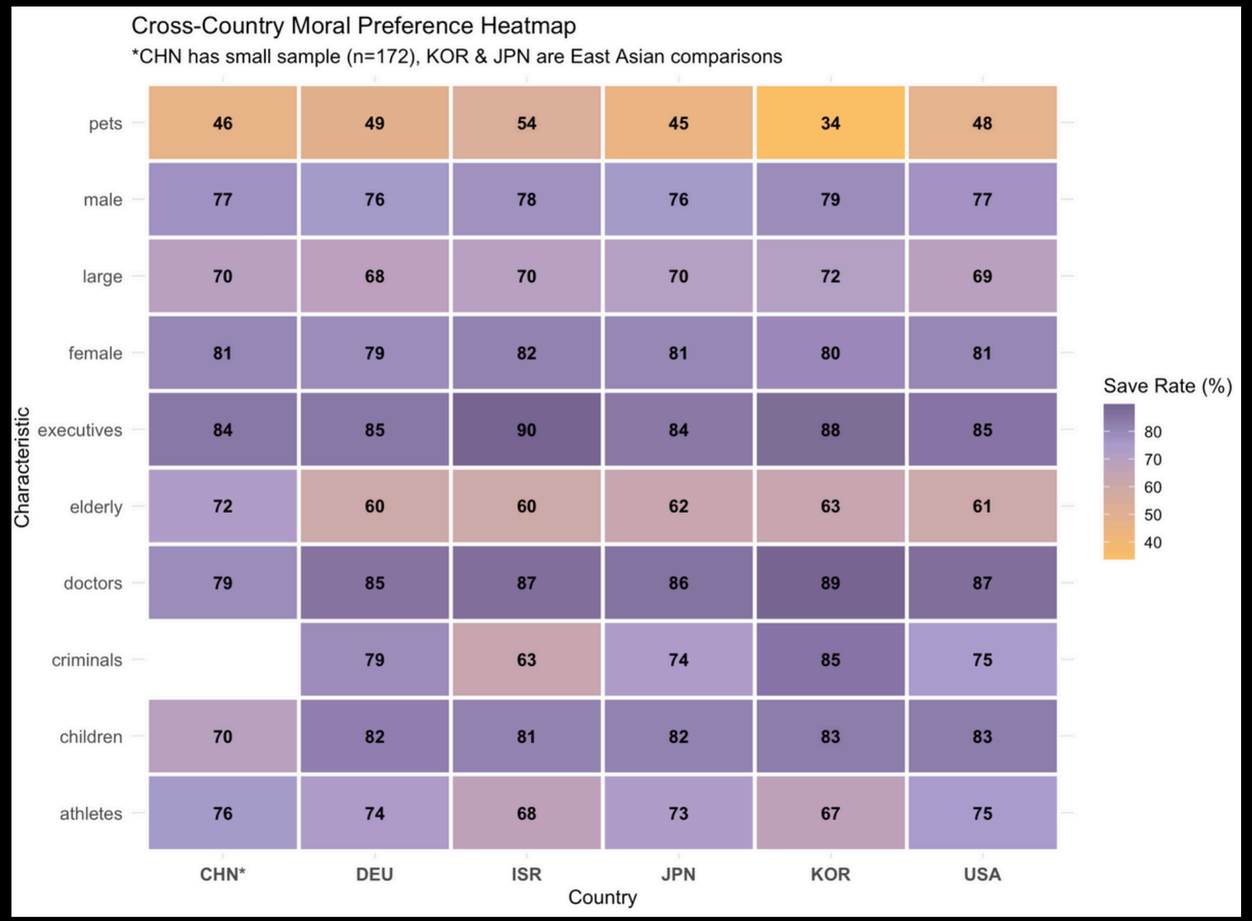
Moral Machine was an attempt to address these questions



Focusing on our select countries, we ran our own MM



CHN	DEU	ISR	JPN	KOR	USA
172	5475	597	1893	392	21471



After we ran our own "perfect" AVs protocol for each country

```
CHNAV PROTOCOL (n=172)
--------------------------------------------------
  1. executives    :  84.4% save rate [High Priority]
  2. female        :  80.8% save rate [High Priority]
  3. doctors       :  79.5% save rate [High Priority]
  4. male          :  77.5% save rate [Medium Priority]
  5. athletes      :  75.8% save rate [Medium Priority]
  6. elderly       :  72.3% save rate [Medium Priority]
  7. children      :  70.0% save rate [Low Priority]
  8. large         :  69.8% save rate [Low Priority]
  9. pets          :  46.2% save rate [Low Priority]
  10. criminals    :     NA% save rate [Low Priority]
```

```
ISRAV PROTOCOL (n=597)
--------------------------------------------------
  1. executives    :  89.8% save rate [High Priority]
  2. doctors       :  86.7% save rate [High Priority]
  3. female        :  81.9% save rate [High Priority]
  4. children      :  81.4% save rate [Medium Priority]
  5. male          :  77.8% save rate [Medium Priority]
  6. large         :  70.2% save rate [Medium Priority]
  7. athletes      :  67.6% save rate [Low Priority]
  8. criminals     :  63.5% save rate [Low Priority]
  9. elderly       :  59.8% save rate [Low Priority]
  10. pets         :  53.6% save rate [Low Priority]
```

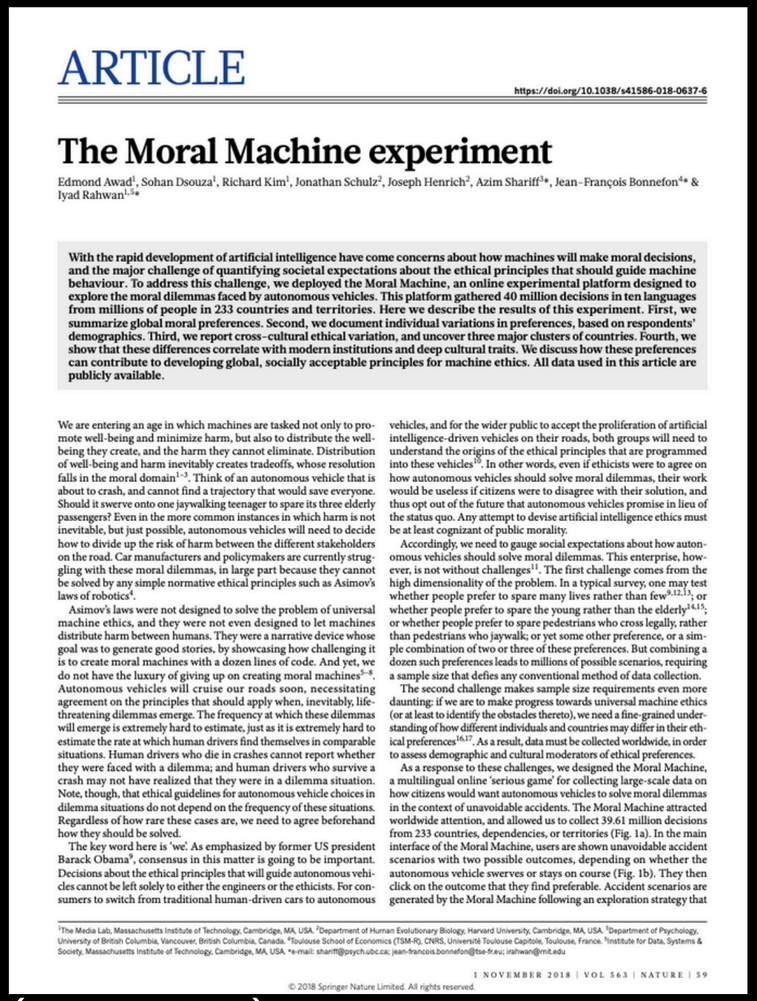
```
DEUAV PROTOCOL (n=5475)
--------------------------------------------------
  1. doctors       :  85.1% save rate [High Priority]
  2. executives    :  84.5% save rate [High Priority]
  3. children      :  82.2% save rate [High Priority]
  4. female        :  79.0% save rate [Medium Priority]
  5. criminals     :  78.8% save rate [Medium Priority]
  6. male          :  76.0% save rate [Medium Priority]
  7. athletes      :  73.8% save rate [Low Priority]
  8. large         :  68.0% save rate [Low Priority]
  9. elderly       :  59.9% save rate [Low Priority]
  10. pets         :  49.3% save rate [Low Priority]
```

```
USAAV PROTOCOL (n=21471)
--------------------------------------------------
  1. doctors       :  86.6% save rate [High Priority]
  2. executives    :  85.0% save rate [High Priority]
  3. children      :  83.1% save rate [High Priority]
  4. female        :  80.7% save rate [Medium Priority]
  5. male          :  77.3% save rate [Medium Priority]
  6. athletes      :  75.0% save rate [Medium Priority]
  7. criminals     :  74.8% save rate [Low Priority]
  8. large         :  68.7% save rate [Low Priority]
  9. elderly       :  60.8% save rate [Low Priority]
  10. pets         :  48.0% save rate [Low Priority]
```

Through decision trees, we trained the AVs model for each country and presented it with "challenging" scenarios

scenario_id	scenario	CHN	DEU	ISR	JPN
1	3 elderly people vs 2 children	A	A	A	A
2	1 elderly doctor vs 1 young criminal	A	A	A	A
3	2 elderly executives vs 2 young athletes	A	A	A	A
4	2 males vs 3 females	B	B	B	B
5	1 male doctor vs 1 female athlete	A	A	A	A
6	2 male executives vs 2 female large people	A	A	A	A
7	1 doctor vs 2 executives	B	B	B	B
8	1 executive vs 2 criminals	A	B	B	B
9	3 homeless vs 1 doctor	B	B	B	B
10	3 large people vs 2 athletes	A	A	A	A
11	1 large doctor vs 1 athletic criminal	A	A	A	A
12	2 large elderly vs 2 athletic children	B	B	B	B
13	1 elderly female doctor vs 2 young male criminals	A	B	B	B
14	1 child with pet vs 1 elderly person with pet	B	A	A	A
15	2 male athletes vs 2 female executives	B	B	B	B
16	3 large elderly people vs 1 athletic child	A	A	A	A
17	1 female criminal vs 1 male homeless person	A	A	A	A
18	5 criminals vs 1 doctor	B	A	A	A
19	4 large people vs 1 child	A	A	A	A
20	6 elderly vs 2 children with 1 pet	A	A	A	A

Yeah... right?



ARTICLE

https://doi.org/10.1038/s41586-018-0637-6

The Moral Machine experiment

Edmond Awad[1], Sohan Dsouza[1], Richard Kim[1], Jonathan Schulz[2], Joseph Henrich[2], Azim Shariff[3]*, Jean-François Bonnefon[4]* & Iyad Rahwan[1,5]*

With the rapid development of artificial intelligence have come concerns about how machines will make moral decisions, and the major challenge of quantifying societal expectations about the ethical principles that should guide machine behaviour. To address this challenge, we deployed the Moral Machine, an online experimental platform designed to explore the moral dilemmas faced by autonomous vehicles. This platform gathered 40 million decisions in ten languages from millions of people in 233 countries and territories. Here we describe the results of this experiment. First, we summarize global moral preferences. Second, we document individual variations in preferences, based on respondents' demographics. Third, we report cross-cultural ethical variation, and uncover three major clusters of countries. Fourth, we show that these differences correlate with modern institutions and deep cultural traits. We discuss how these preferences can contribute to developing global, socially acceptable principles for machine ethics. All data used in this article are publicly available.

We are entering an age in which machines are tasked not only to promote well-being and minimize harm, but also to distribute the well-being they create, and the harm they cannot eliminate. Distribution of well-being and harm inevitably creates tradeoffs, whose resolution falls in the moral domain[1–3]. Think of an autonomous vehicle that is about to crash, and cannot find a trajectory that would save everyone. Should it swerve onto one jaywalking teenager to spare its three elderly passengers? Even in the more common instances in which harm is not inevitable, but just possible, autonomous vehicles will need to decide how to divide up the risk of harm between the different stakeholders on the road. Car manufacturers and policymakers are currently struggling with these moral dilemmas, in large part because they cannot be solved by any simple normative ethical principles such as Asimov's laws of robotics[4].

Asimov's laws were not designed to solve the problem of universal machine ethics, and they were not even designed to let machines distribute harm between humans. They were a narrative device whose goal was to generate good stories, by showcasing how challenging it is to create moral machines with a dozen lines of code. And yet, we do not have the luxury of giving up on creating moral machines[5–8]. Autonomous vehicles will cruise our roads soon, necessitating agreement on the principles that should apply when, inevitably, life-threatening dilemmas emerge. The frequency at which these dilemmas will emerge is extremely hard to estimate, just as it is extremely hard to estimate the rate at which human drivers find themselves in comparable situations. Human drivers who die in crashes cannot report whether they were faced with a dilemma; and human drivers who survive a crash may not have realized that they were in a dilemma situation. Note, though, that ethical guidelines for autonomous vehicle choices in dilemma situations do not depend on the frequency of these situations. Regardless of how rare these cases are, we need to agree beforehand how they should be solved.

The key word here is 'we'. As emphasized by former US president Barack Obama[9], consensus in this matter is going to be important. Decisions about the ethical principles that will guide autonomous vehicles cannot be left solely to either the engineers or the ethicists. For consumers to switch from traditional human-driven cars to autonomous vehicles, and for the wider public to accept the proliferation of artificial intelligence-driven vehicles on their roads, both groups will need to understand the origins of the ethical principles that are programmed into these vehicles[10]. In other words, even if ethicists were to agree on how autonomous vehicles should solve moral dilemmas, their work would be useless if citizens were to disagree with their solution, and thus opt out of the future that autonomous vehicles promise in lieu of the status quo. Any attempt to devise artificial intelligence ethics must be at least cognizant of public morality.

Accordingly, we need to gauge social expectations about how autonomous vehicles should solve moral dilemmas. This enterprise, however, is not without challenges[11]. The first challenge comes from the high dimensionality of the problem. In a typical survey, one may test whether people prefer to spare many lives rather than few[9,12,13]; or whether people prefer to spare the young rather than the elderly[14,15]; or whether people prefer to spare pedestrians who cross legally, rather than pedestrians who jaywalk; or yet some other preference, or a simple combination of two or three of these preferences. But combining a dozen such preferences leads to millions of possible scenarios, requiring a sample size that defies any conventional method of data collection.

The second challenge makes sample size requirements even more daunting: if we are to make progress towards universal machine ethics (or at least to identify the obstacles thereto), we need a fine-grained understanding of how different individuals and countries may differ in their ethical preferences[16,17]. As a result, data must be collected worldwide, in order to assess demographic and cultural moderators of ethical preferences.

As a response to these challenges, we designed the Moral Machine, a multilingual online 'serious game' for collecting large-scale data on how citizens would want autonomous vehicles to solve moral dilemmas in the context of unavoidable accidents. The Moral Machine attracted worldwide attention, and allowed us to collect 39.61 million decisions from 233 countries, dependencies, or territories (Fig. 1a). In the main interface of the Moral Machine, users are shown unavoidable accident scenarios with two possible outcomes, depending on whether the autonomous vehicle swerves or stays on course (Fig. 1b). They then click on the outcome that they find preferable. Accident scenarios are generated by the Moral Machine following an exploration strategy that

[1]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [3]Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada. [4]Toulouse School of Economics (TSM-R), CNRS, Université Toulouse Capitole, Toulouse, France. [5]Institute for Data, Systems & Society, Massachusetts Institute of Technology, Cambridge, MA, USA. *e-mail: shariff@psych.ubc.ca; jean-francois.bonnefon@tse-fr.eu; irahwan@mit.edu

(Awad 2018)

New Technology > Infrastructure & Transportation

Out of Two Million People, Most Prefer That a Self-Driving Car Kill the Elderly

Questions towards the MM data set

Data provenance

"The sample is self-selected... policymakers should not embrace this data as the final word on societal preferences" (Etienne 2021).
Also no way of checking the country provenance

High-stake issues in low-stake environment

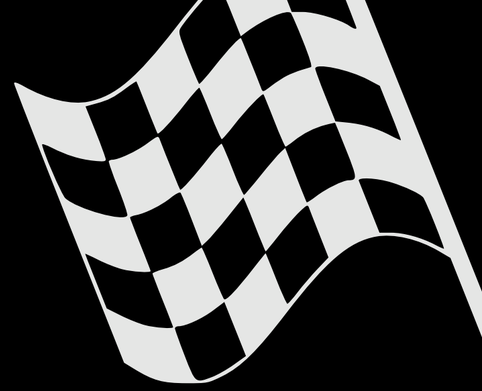
- Binary choices with "no information, no deliberation, and no consequences" (Etienne 2021)
- captures stereotypes about categories of people, not reasoned principles (Luft 2020).

Hypotheticals aren't real policy-making

AV ethics "is precisely the wrong question" when framed as an individual's split-second choice (Jaques 2025)

Computational ethics fail us

(Awad 2022)



Descriptive click-data as normative moral truths

The fallacy of measuring attitudes as building moral rules (Etienne 2021).

Normalizes discriminatory logics under the guise of “optimization”

By forcing choices, MM operationalizes harmful social stereotypes (Luft 2020).

Narrows the entire ethics of AVs to a single artificial dilemma, ignoring real risks

AV ethics is about infrastructures, institutions, and power, not trolley hypotheticals (Pink 2022).

This is just not how conversations happen)

Law and social sanctions express and renegotiate shared moral norms as social differentiation increases (Durkheim 1893).

AV Stack Overview

01: Perception

Detects and classifies objects (pedestrians, vehicles, signs); Estimates positions, velocities, and uncertainties.

Example: Camera + LiDAR identify a pedestrian partially occluded by a parked van

02: Prediction

Forecasts future trajectories of road users; models uncertainty and multi-modal behaviors.

Example: Predicting whether a cyclist will continue straight, merge, or yield

03: Planning

Chooses the safest and most appropriate maneuver; integrates cultural/ethical preferences.

Example: Deciding whether to brake aggressively or yield gradually depending on local norms

04: Control

Executes steering, throttle, braking to follow the chosen path; ensures smooth, stable actuation.

Example: Applying precise braking force to avoid wheel lock or passenger discomfort

AV Stack Overview

Ethics mainly enter in planning/decision-making via:
1) **Reward function design** (soft preferences, tuned weights)
2) **Ethical/safety constraint module** (hard rules, veto layer)

01: Perception

Detects and classifies objects (pedestrians, vehicles, signs); Estimates positions, velocities, and uncertainties.
Example: Camera + LiDAR identify a pedestrian partially occluded by a parked van

02: Prediction

Forecasts future trajectories of road users; models uncertainty and multi-modal behaviors.
Example: Predicting whether a cyclist will continue straight, merge, or yield

03: Planning

Chooses the safest and most appropriate maneuver; integrates cultural/ethical preferences.
Example: Deciding whether to brake aggressively or yield gradually depending on local norms

04: Control

Executes steering, throttle, braking to follow the chosen path; ensures smooth, stable actuation.
Example: Applying precise braking force to avoid wheel lock or passenger discomfort

Soft Ethics - Reward Function Structure

$$R(s,a) = -\Big(C_{\text{human}} N_{\text{human harm}} + C_{\text{animal}} N_{\text{animal harm}} + C_{\text{property}} D_{\text{property}} + C_{\text{law}} I_{\text{law viol}} + C_{\text{comfort}} A_{\text{jerk}}\Big),$$

Sources of Cultural & Legal Variation:

1. Safety vs. legality vs. efficiency weighting: risk-averse vs. assertive driving cultures
2. Per-category treatment of humans: some regions allow distinctions (young/old, many/few); some (e.g., Germany) require uniform treatment
3. Hard constraints vs. soft penalties: non-discrimination, no “one for many” sacrifice → hard rule

Learning Cultural Preferences

- Handcrafted weights are brittle, subject to designer bias
- Cultural norms differ across regions and evolve over time
- **IRL:** infer reward from demonstrations
- **Preference learning:** infer utilities from pairwise judgments, e.g.

$$U_\theta(\cdot) \text{ such that } U_\theta(x) > U_\theta(y)$$

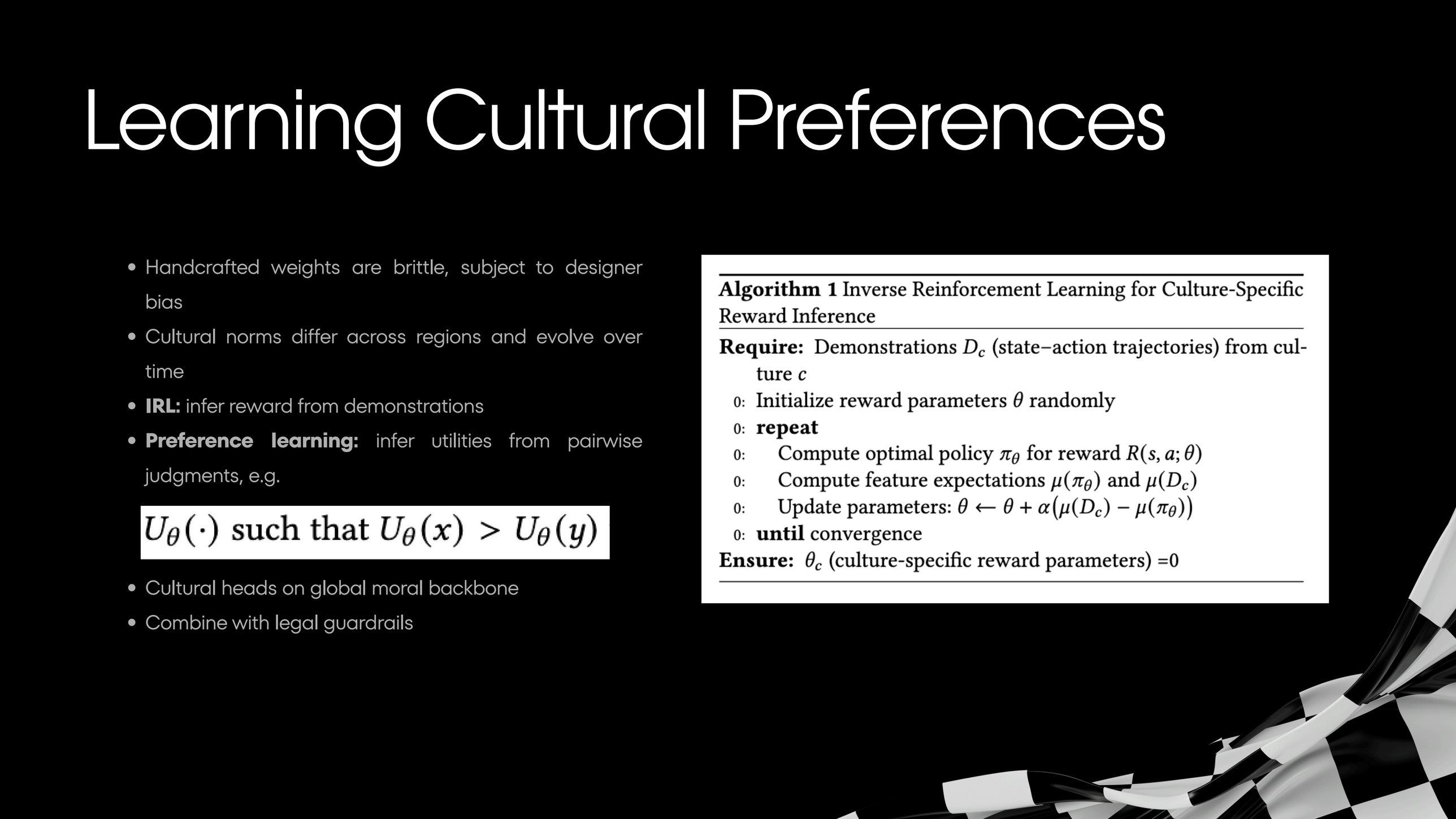
- Cultural heads on global moral backbone
- Combine with legal guardrails

Algorithm 1 Inverse Reinforcement Learning for Culture-Specific Reward Inference

Require: Demonstrations D_c (state–action trajectories) from culture c

Initialize reward parameters θ randomly
repeat
  Compute optimal policy π_θ for reward $R(s, a; \theta)$
  Compute feature expectations $\mu(\pi_\theta)$ and $\mu(D_c)$
  Update parameters: $\theta \leftarrow \theta + \alpha\big(\mu(D_c) - \mu(\pi_\theta)\big)$
until convergence

Ensure: θ_c (culture-specific reward parameters) =0



Case Studies

Legal Principles ('17 Ethics Commission, Autonomous Driving Act)

- Human dignity is inviolable → no discrimination by personal characteristics
- No numerical trade-offs → can't sacrifice one to save many
- Risk minimization only → braking, deceleration, evasive maneuvers are allowed; choosing whom to hit is not

Hard constraints + infinite penalties for human harm

Verification via attribute-swapping tests

Regulatory sandbox (C4IR)

Process-oriented, not rule-prescriptive

Iterative policy adjustment (AV ethics shaped through observed behavior rather than fixed in advance)

Focus on aligning AV behavior with local traffic norms and public expectations

Co-development with industry and public stakeholders (Mobileye, GM, Cruise, Waymo)

Real-Time Decision-Making Challenges & Uncertainties

1. Computational Limits

Planner runs at 50–100 ms per cycle;
Rich ethical optimization too slow; (Monte Carlo: 10–50 ms per rollout)
Runtime uses precomputed policies + 5–10 sampled trajectories;

$$\mathbf{a}^* = \arg \max_{\mathbf{a} \in \mathcal{A}_{\text{sample}}} V_\theta(\mathbf{s}_t, \mathbf{a})$$

Region-specific policies loaded offline (Germany vs. Israel)

2. Layered Architecture

Nominal planner: 10–20 Hz, evaluates candidate trajectories
Safety layer: 100 Hz, monitors Time-to-Collision (TTC)

$$\text{TTC}_i = \frac{\text{distance}_i(\mathbf{s}_t)}{\max\left(0, \text{relative\_velocity}_i(\mathbf{s}_t)\right)}.$$

If TTC < 0.5–1.0 s → emergency braking override
Hard constraints must hold across both layers

3. Perception Uncertainty

Errors: misclassification, localization noise, occlusions, overconfidence
Planner uses probabilistic outputs, not hard labels
RS-NNs → belief masses over classes

```
TRIGGER_FALLBACK ← false
for each detected agent i do
  if P(human_i | z_t) > p_threshold and TTC_i < t_threshold then
    TRIGGER_FALLBACK ← true
    break
  end if
end for=0
```

Fallback triggered if P(human) > p_threshold AND TTC < t_threshold
Thresholds reflect cultural risk tolerance

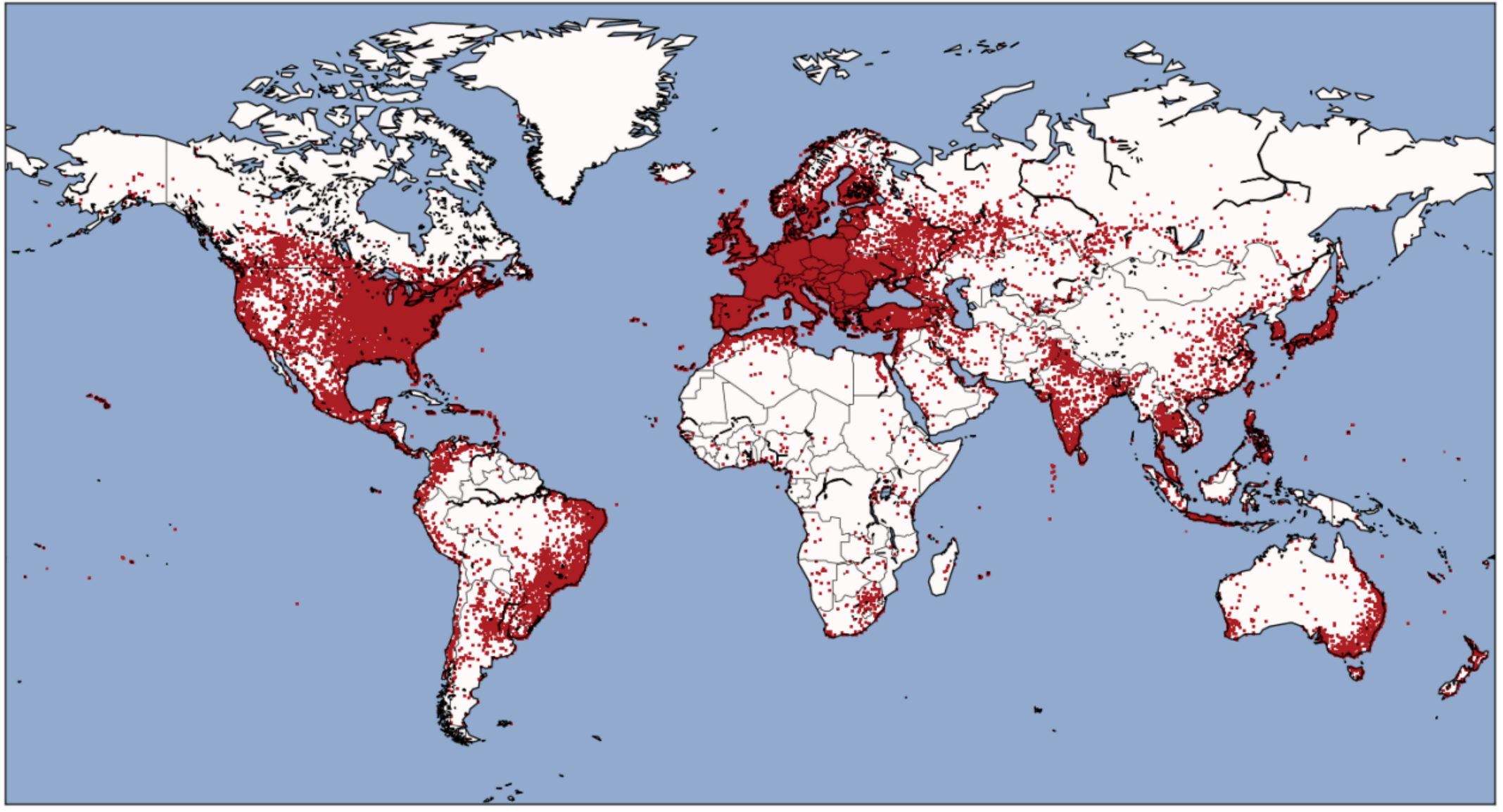
A RACE TO CONSTRAINT

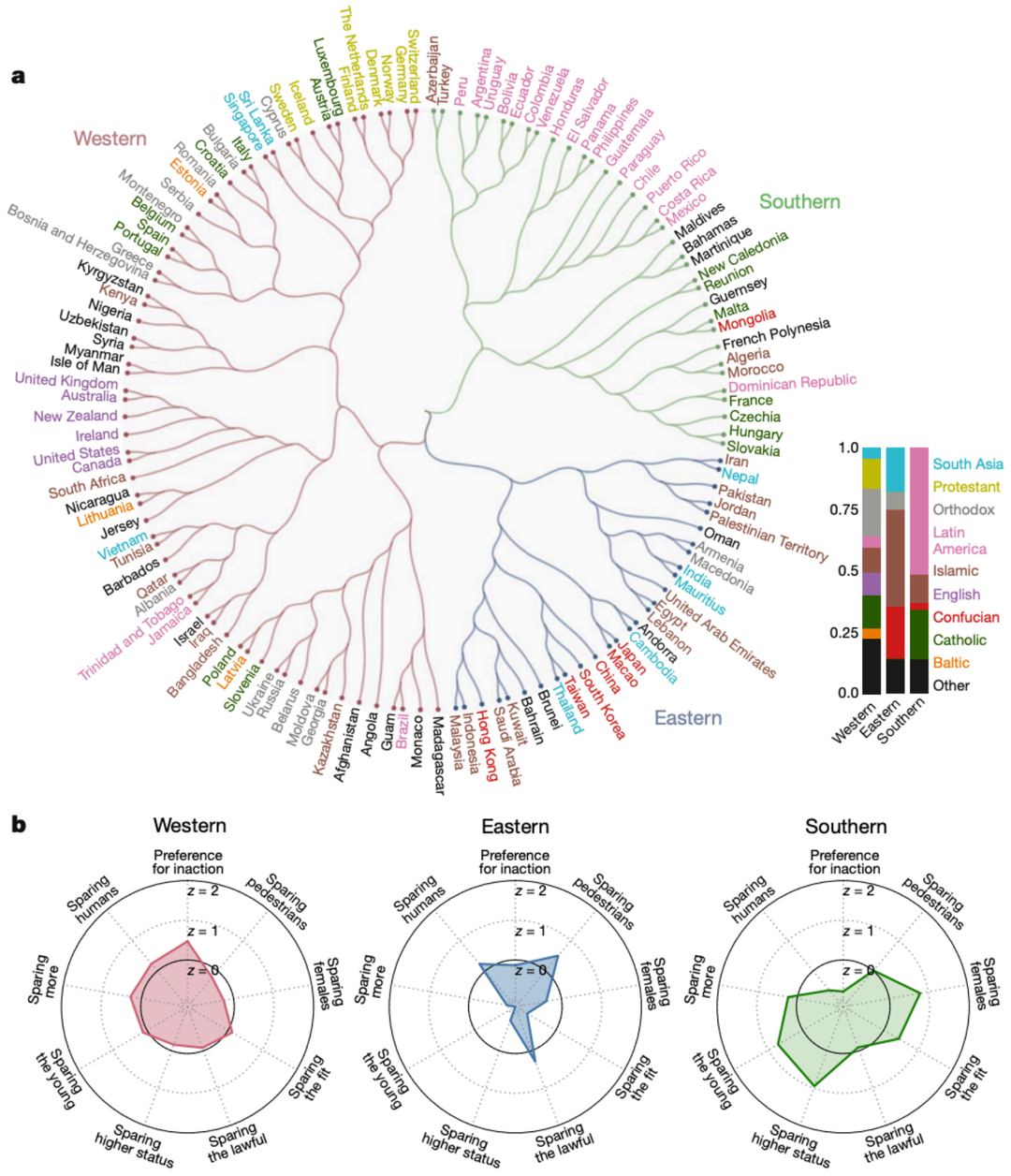
EMBEDDING ETHICAL AND CULTURAL PREFERENCES IN AUTONOMOUS VEHICLE PLANNING

Sonya Rashkovan / Julia Zdzilowska

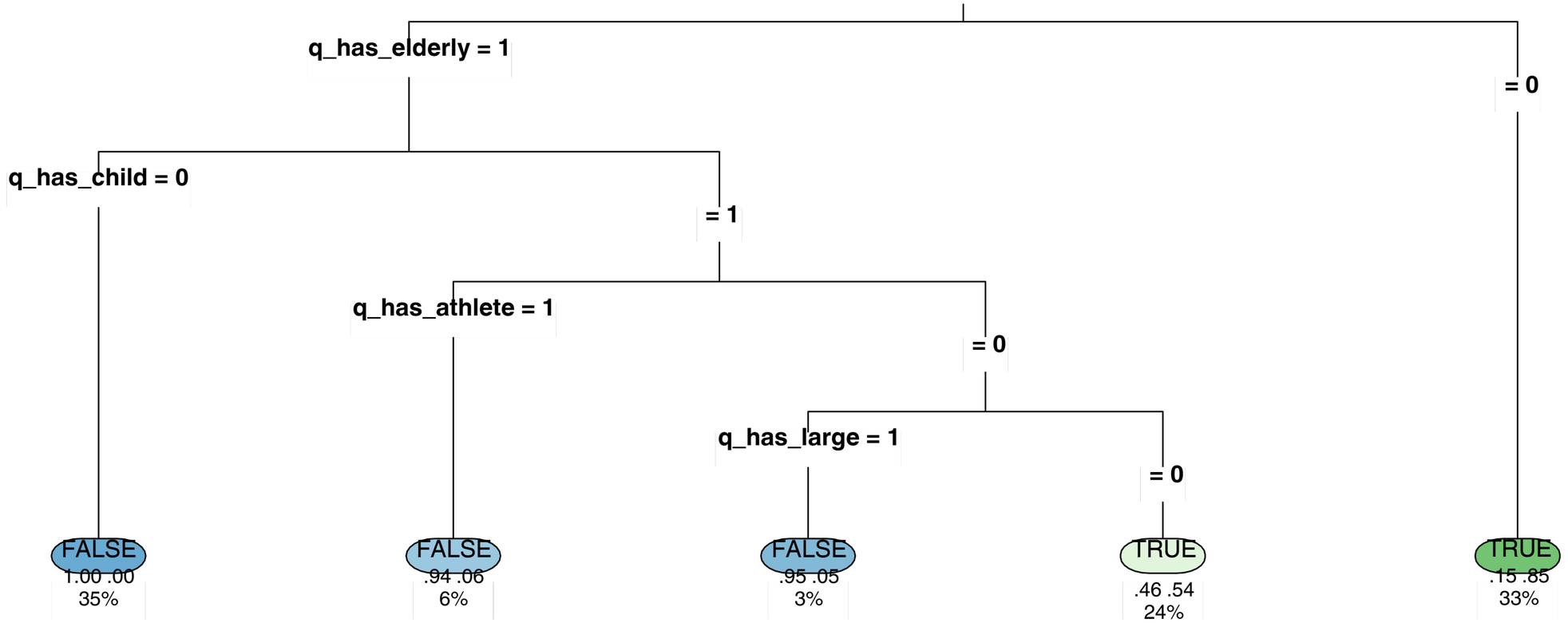
APPENDIX

a

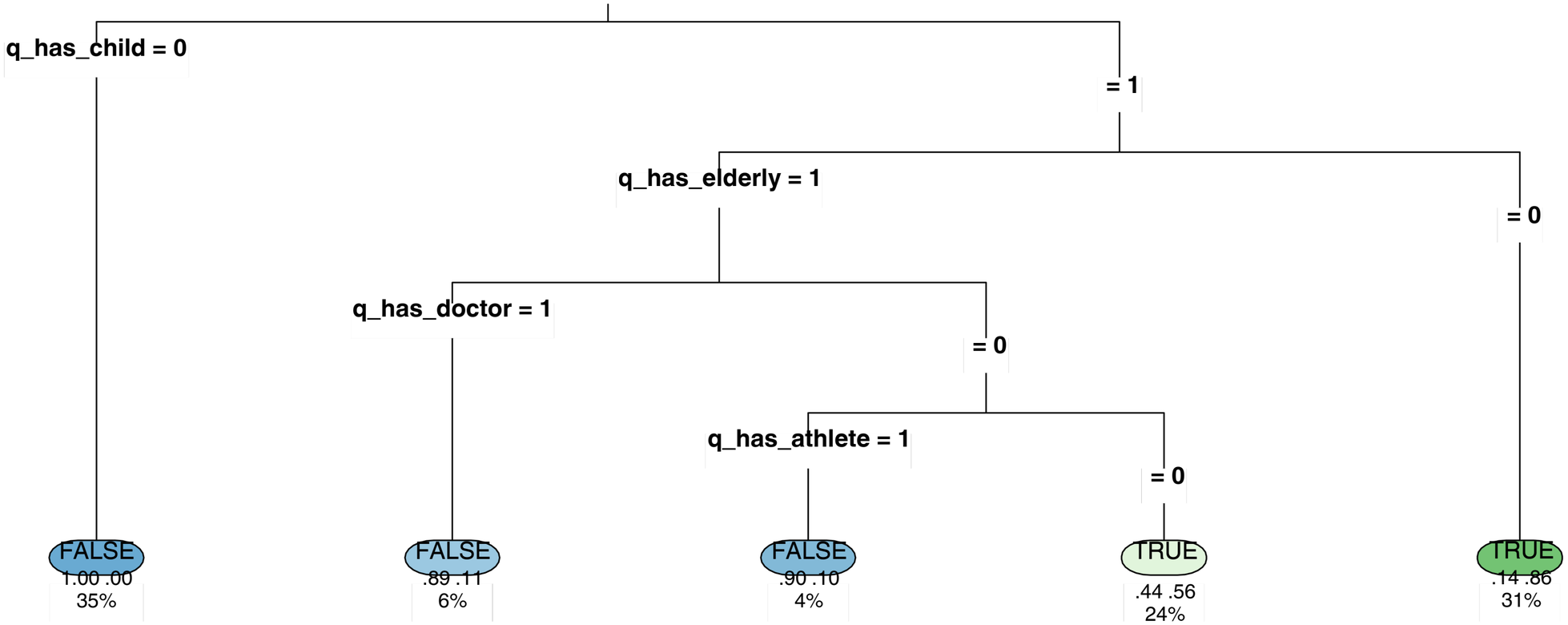



a
Western
Southern
Eastern
Switzerland
Azerbaijan
Turkey
Peru
Argentina
Uruguay
Bolivia
Ecuador
Colombia
Venezuela
Honduras
El Salvador
Panama
Philippines
Guatemala
Paraguay
Chile
Puerto Rico
Costa Rica
Mexico
Maldives
Bahamas
Martinique
New Caledonia
Reunion
Guernsey
Malta
Mongolia
French Polynesia
Algeria
Morocco
Dominican Republic
France
Czechia
Hungary
Slovakia
Iran
Nepal
Pakistan
Jordan
Palestinian Territory
Oman
Armenia
Macedonia
India
Mauritius
United Arab Emirates
Egypt
Lebanon
Andorra
Cambodia
Japan
Macao
China
South Korea
Taiwan
Thailand
Brunei
Bahrain
Kuwait
Saudi Arabia
Hong Kong
Indonesia
Malaysia
Madagascar
Monaco
Brazil
Guam
Angola
Afghanistan
Kazakhstan
Georgia
Moldova
Belarus
Russia
Ukraine
Slovenia
Latvia
Poland
Bangladesh
Iraq
Israel
Jamaica
Trinidad and Tobago
Albania
Qatar
Barbados
Tunisia
Vietnam
Jersey
Lithuania
Nicaragua
South Africa
Canada
United States
Ireland
New Zealand
Australia
United Kingdom
Isle of Man
Myanmar
Syria
Uzbekistan
Nigeria
Kenya
Kyrgyzstan
Greece
Bosnia and Herzegovina
Portugal
Spain
Belgium
Montenegro
Serbia
Estonia
Romania
Croatia
Bulgaria
Italy
Singapore
Sri Lanka
Cyprus
Sweden
Iceland
Austria
Luxembourg
Finland
Denmark
Norway
Germany
The Netherlands
1.0
0.75
0.5
0.25
0.0
Western
Eastern
Southern
South Asia
Protestant
Orthodox
Latin America
Islamic
English
Confucian
Catholic
Baltic
Other
b
Western
Eastern
Southern
Preference for inaction
Sparing pedestrians
Sparing females
Sparing the fit
Sparing the lawful
Sparing higher status
Sparing the young
Sparing more
Sparing humans
z = 2
z = 1
z = 0

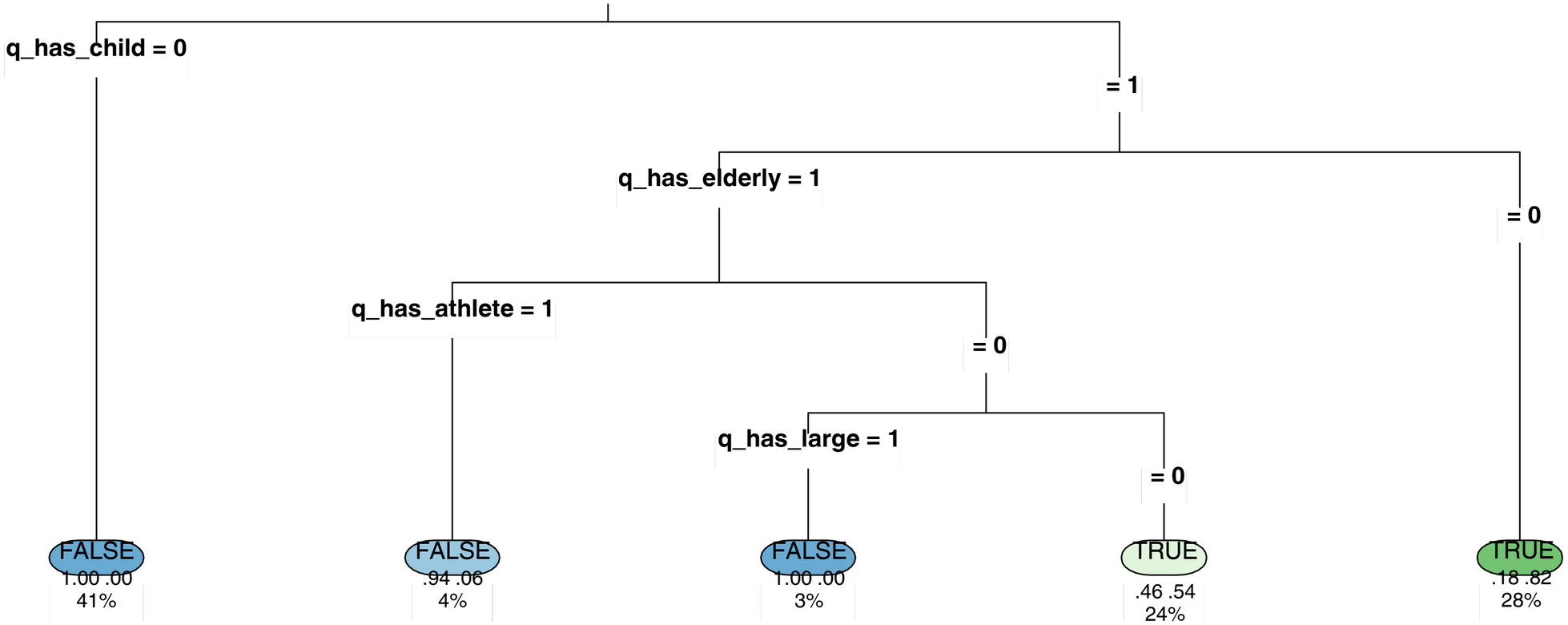
DEU AV Decision Tree: Choose Youth vs Elderly



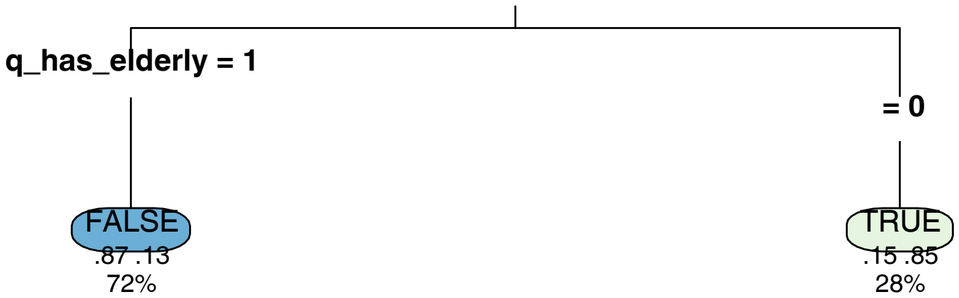
USA AV Decision Tree: Choose Youth vs Elderly

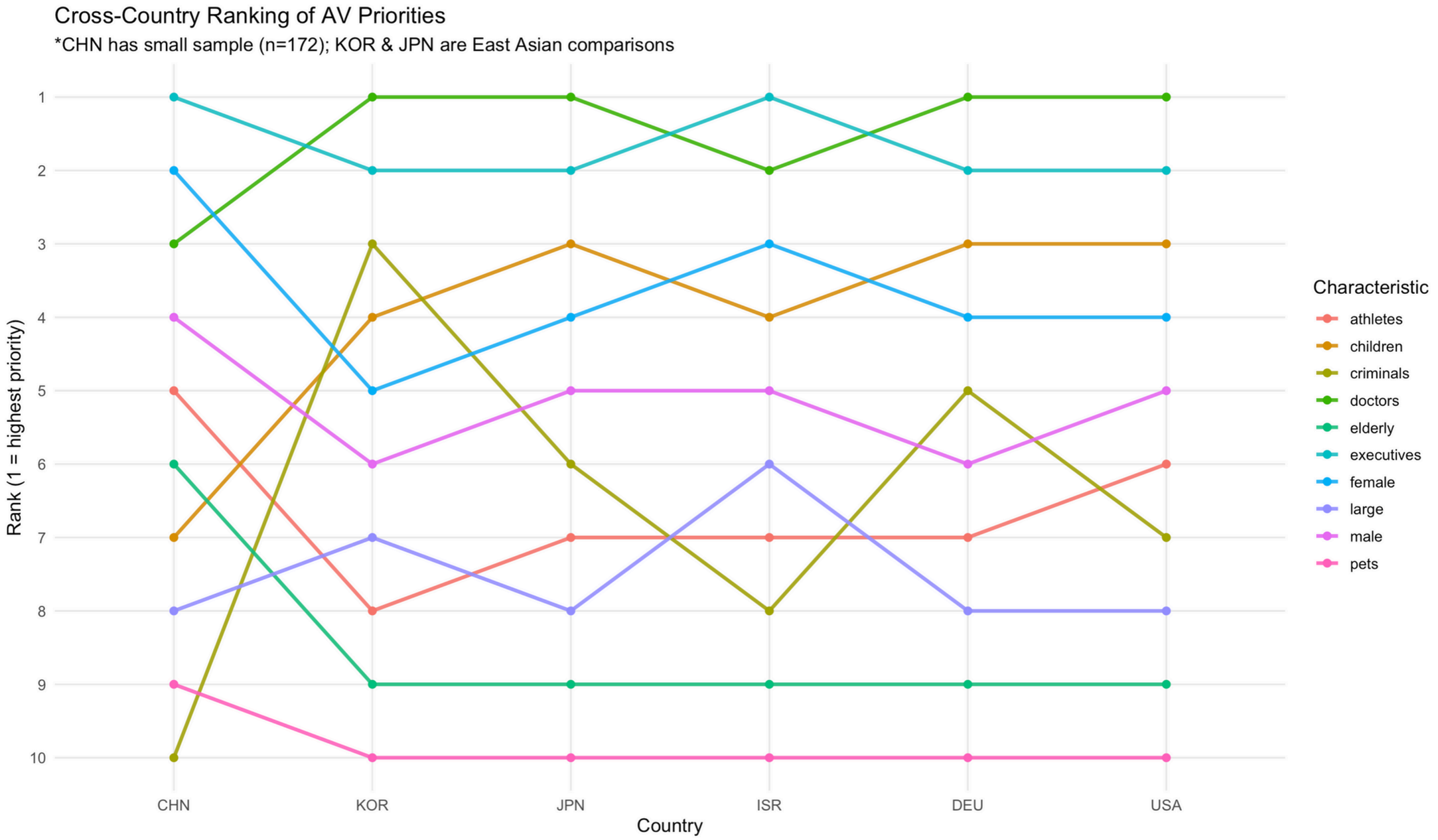


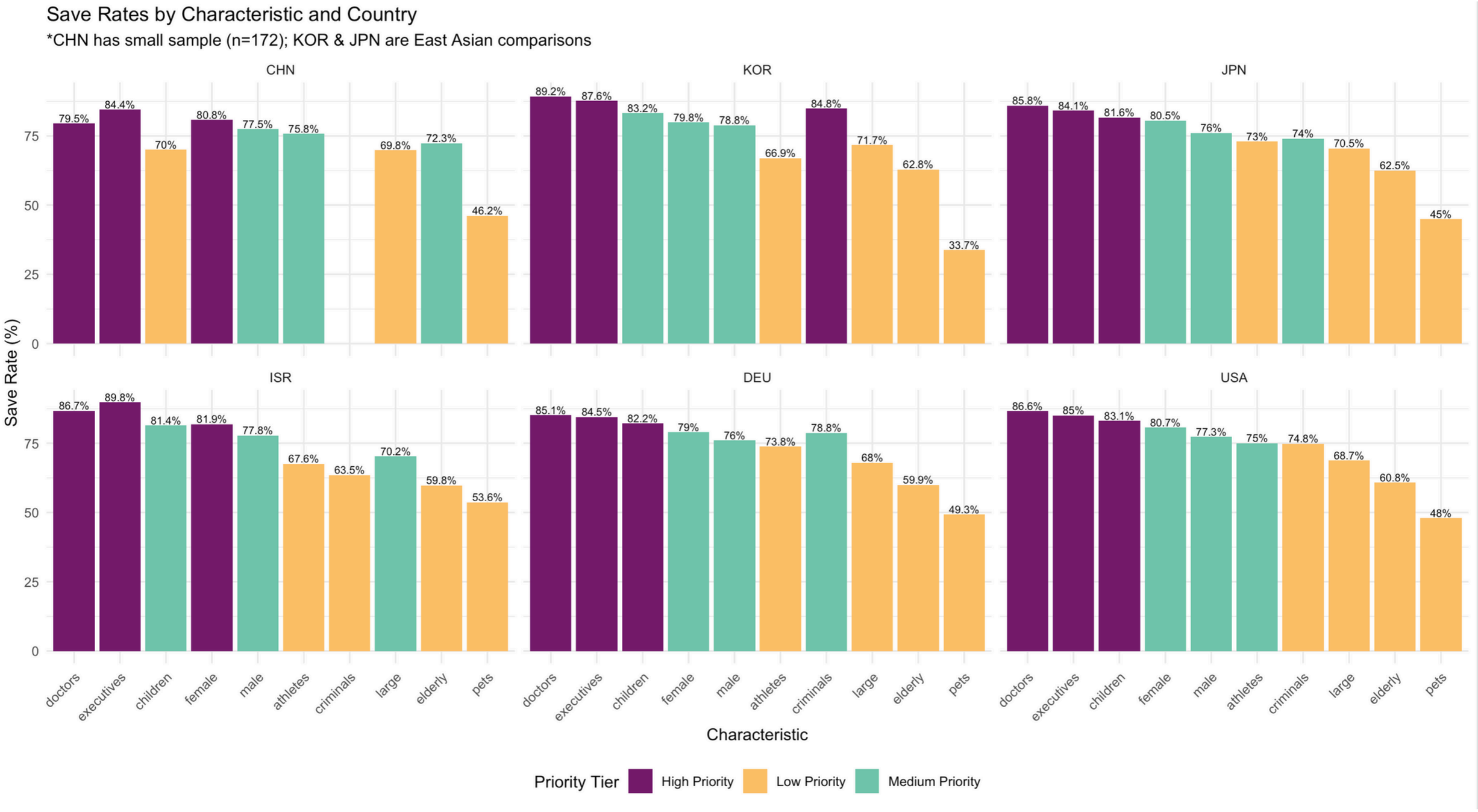
ISR AV Decision Tree: Choose Youth vs Elderly



CHN AV Decision Tree: Choose Youth vs Elderly




Cross-Country Ranking of AV Priorities
*CHN has small sample (n=172); KOR & JPN are East Asian comparisons
Rank (1 = highest priority)
1
2
3
4
5
6
7
8
9
10
CHN
KOR
JPN
ISR
DEU
USA
Country
Characteristic
athletes
children
criminals
doctors
elderly
executives
female
large
male
pets


Save Rates by Characteristic and Country
*CHN has small sample (n=172); KOR & JPN are East Asian comparisons
CHN
KOR
JPN
ISR
DEU
USA
Save Rate (%)
0
25
50
75
79.5%
84.4%
70%
80.8%
77.5%
75.8%
69.8%
72.3%
46.2%
89.2%
87.6%
83.2%
79.8%
78.8%
66.9%
84.8%
71.7%
62.8%
33.7%
85.8%
84.1%
81.6%
80.5%
76%
73%
74%
70.5%
62.5%
45%
86.7%
89.8%
81.4%
81.9%
77.8%
67.6%
63.5%
70.2%
59.8%
53.6%
85.1%
84.5%
82.2%
79%
76%
73.8%
78.8%
68%
59.9%
49.3%
86.6%
85%
83.1%
80.7%
77.3%
75%
74.8%
68.7%
60.8%
48%
doctors
executives
children
female
male
athletes
criminals
large
elderly
pets
Characteristic
Priority Tier
High Priority
Low Priority
Medium Priority