Ethical Decision Making in Artificial Intelligence

Pradeep Ravikumar
Machine Learning Department
School of Computer Science
Carnegie Mellon University

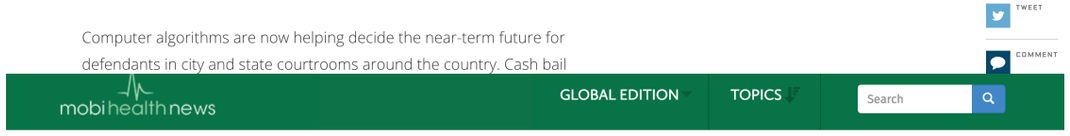
AI making societally important decisions

- Artificial Intelligence systems are being used to make societally important decisions
- Bail decisions
- Healthcare
- Autonomous cars

— FIRST LOOK —

Artificial intelligence plays budding role in courtroom bail decisions

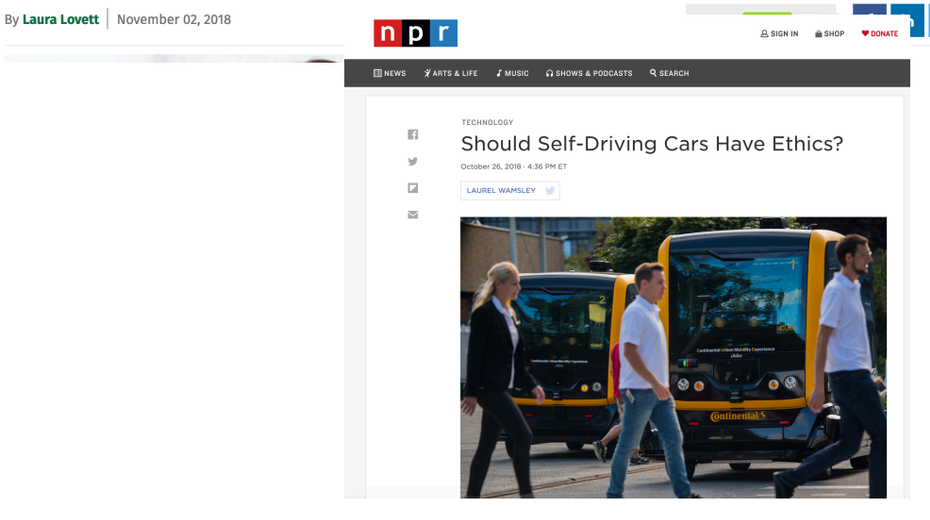
Computer algorithms are now helping decide the near-term future for defendants in city and state courtrooms around the country. Cash bail



AI triage chatbots trekking toward a standard of care despite criticism

Savvy hospitals like Boston Children's and NHS facilities are working with chatbot startups to create new ways to interface with patients seeking care.

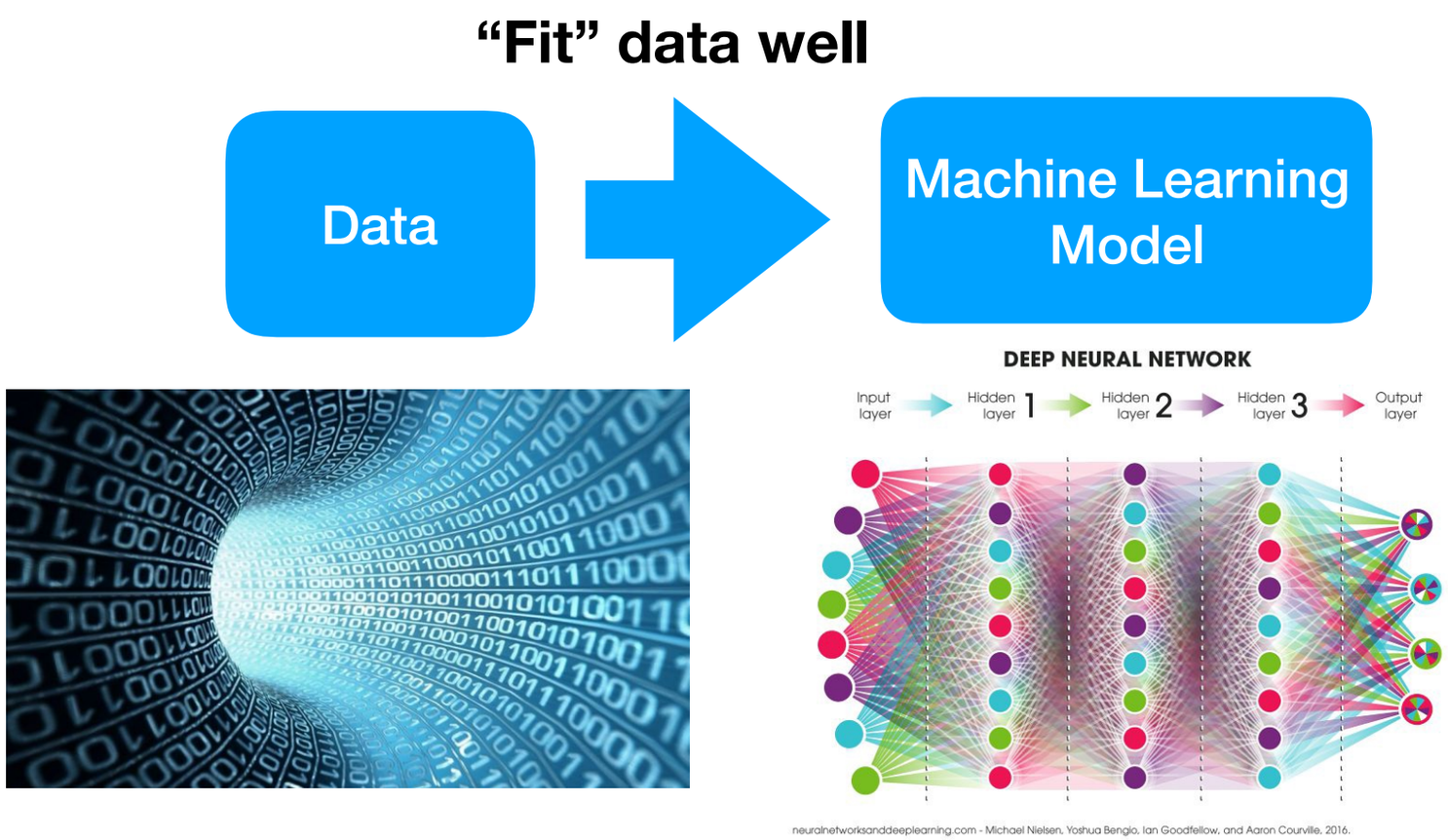
By **Laura Lovett** | November 02, 2018



AI and ethics

- Can AI systems behave ethically?

- Typical ML pipeline:



- How do we incorporate “ethical” thinking?

Ethics / Moral Philosophy

- What is right and what is wrong?
- How to make decisions that are right?
- Questions studied by philosophers for 1000s of years with no consensus formal framework

Three Main Ethical Frameworks

- **Deontological:** take action according to a specified set of rules
- **Virtue Ethics:** multiple “values” or “virtues”; take action that follows these values or virtues
- **Consequentialism:** take an action that has the most desirable future consequences
    - **Utilitarianism:** assign a utility to world states, and take action that leads to highest utility

Deontological Ethics

- Rules based (e.g. Ten Commandments)
    - Requires a priori specification of ethical rules
- Given a set of rules, or constraints, we can ensure that AI actions follow these rules
- Caveat: rules not always available, and when available, too broad to be applicable to specific situation
    - e.g. just the ten commandments not helpful for self-driving car ethics

Virtue Ethics

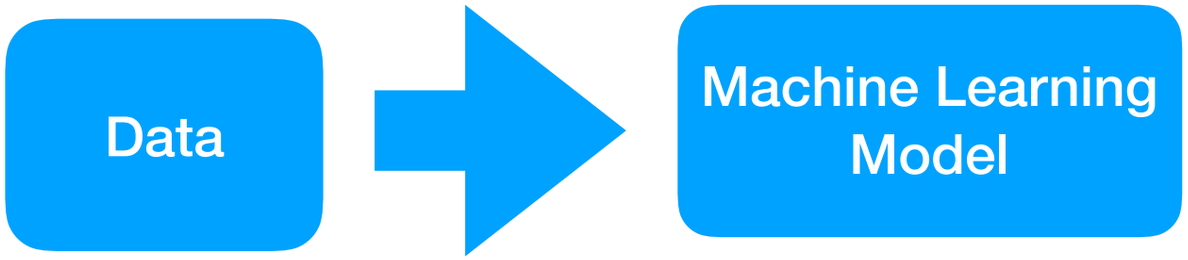
- Take actions based on values/virtues … Aristotle (and others)

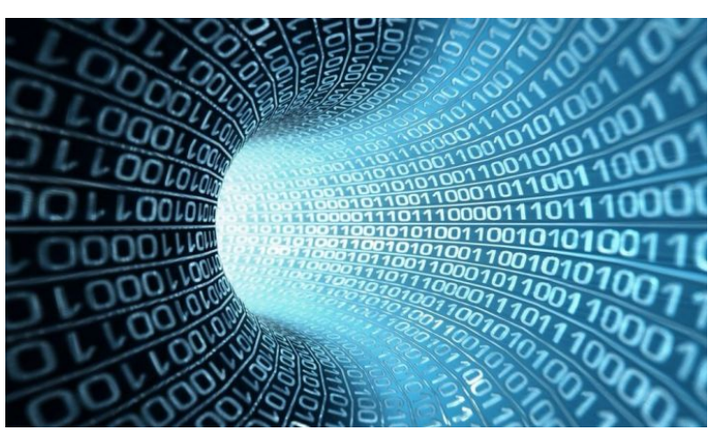
- Similar caveats to deontological ethics
  - values not always available, and when available (e.g. be honest) not always applicable to specific situation
  - differing values could conflict (e.g. equality, and freedom)
  - Ongoing work: virtue ethics driven decision making

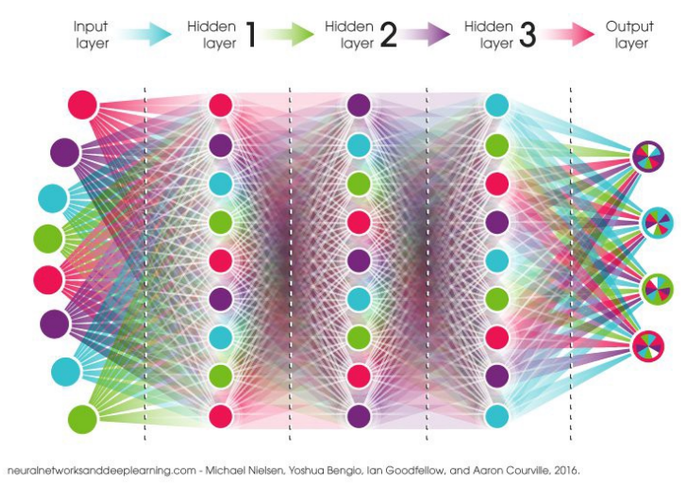
Utilitarian Ethics

- Decision theoretic foundations of machine learning based largely on utilitarianism

“Fit” data well: involves a loss/utility function





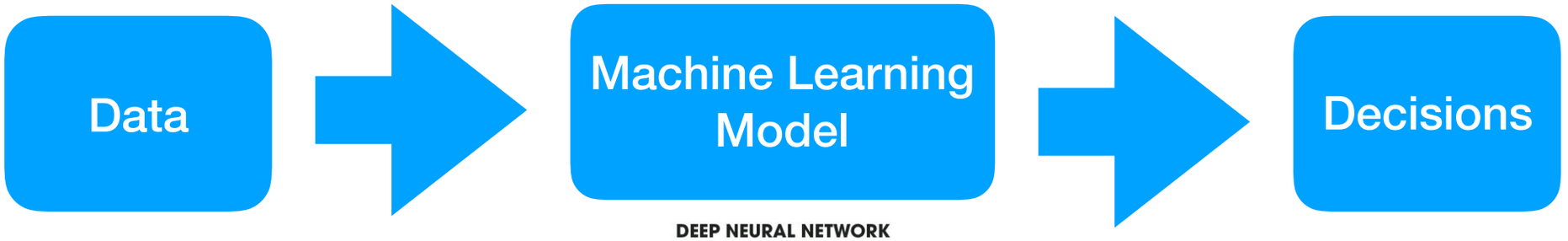


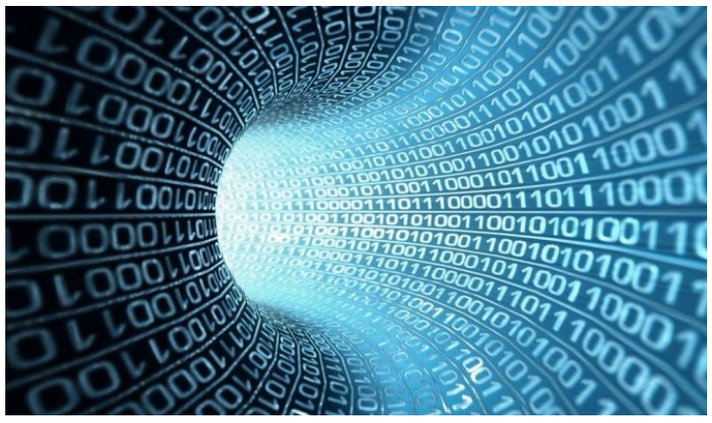
$\ell_{\text{model}}(\theta, D)$: loss i.e. negative utility associated with model parameter θ, and data D

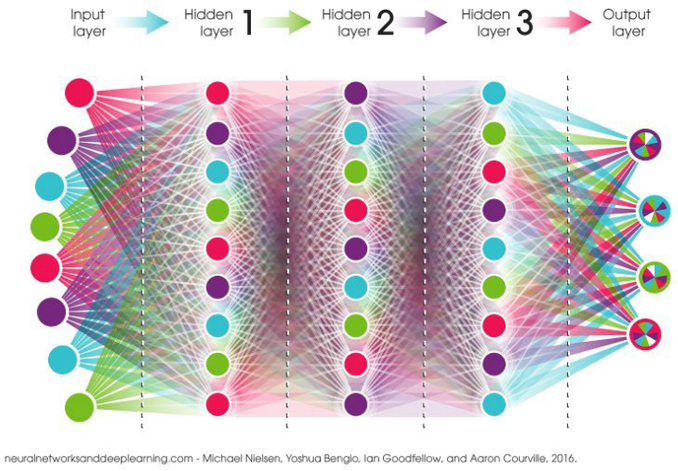
Utilitarian Ethics

- Decision theoretic foundations of machine learning based largely on utilitarianism

"Fit" data well **"optimal action" involves a loss function**







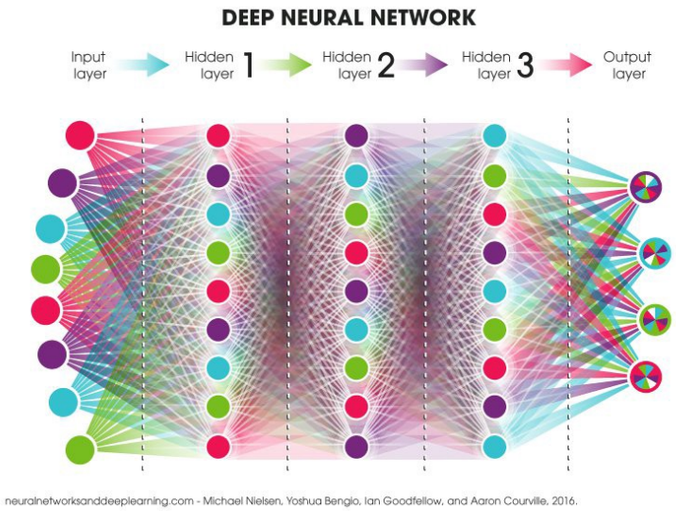
$\ell_{\text{model}}(\theta, D)$: loss i.e. negative utility associated with model parameter θ, and data D

$\ell_{\text{action}}(\theta, a)$: loss i.e. negative utility associated with model parameter θ, and action a

Example: Finance

"Fit" data well **Minimize loss function**

Data → Machine Learning Model → Decisions



neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

Data: past stock prices, **Model:** for predicting future stock price movements

Loss function: error in predicting future stock price movements

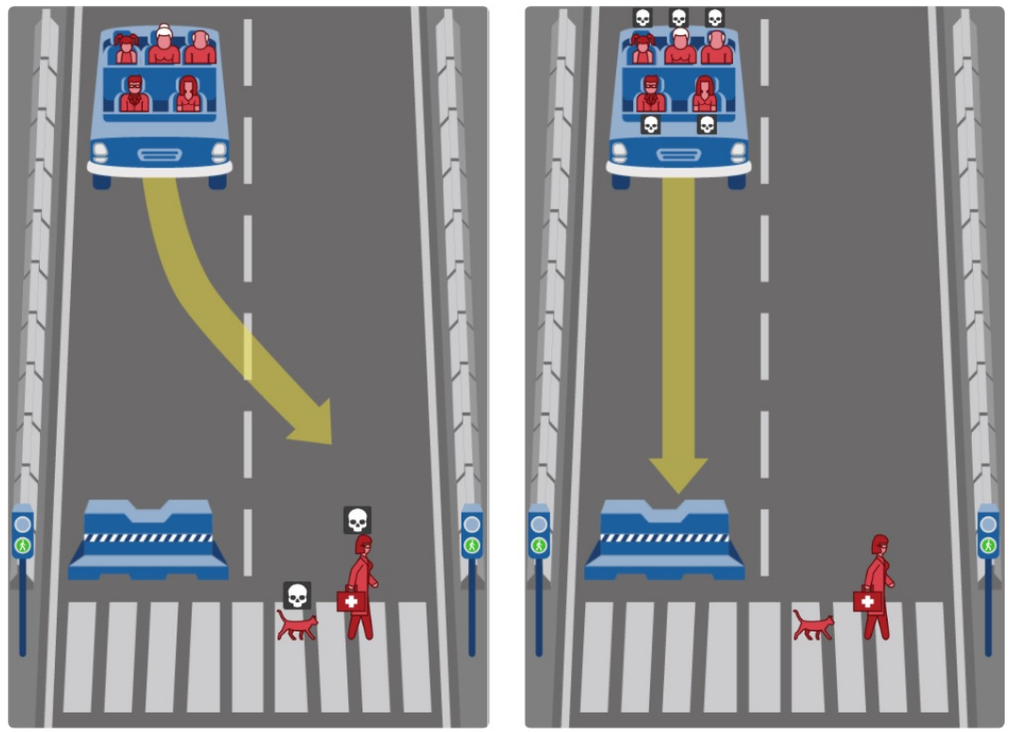
Actions: buy/sell x amount of y stock

Loss function: risk-adjusted return of action given stock market movements

Loss functions

- But where do we get these loss functions?
- Typically specified apriori, via domain knowledge
- But what would **ethical loss functions** look like?
    - 1000s years of moral philosophy provide a qualitative rather than quantitative picture of ethical loss functions
    - e.g. if airline has to decide who to not board due to overbooking, how do they decide if pregnant woman with two kids is not to be bumped over say a college student?
- We address this in two ways:
    - we **learn** ethical loss functions from data
    - we allow for the fact that there need not be a consensus single ethical loss function, and hence learn multiple ethical loss functions and **aggregate** them in a social-choice theoretically optimal way

Trolley problems, ethical dilemmas, self-driving cars



- Brakes of self-driving car have failed
- Should it swerve and hit a doctor and a cat?
- Or should it crash into a concrete barrier that will kill all five passengers?

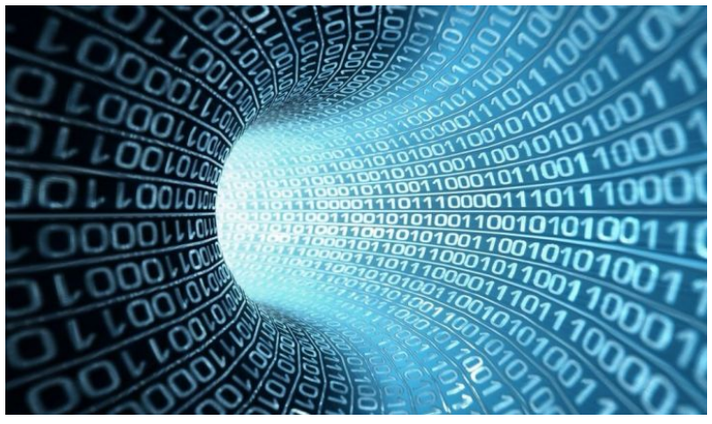
Trolley problems

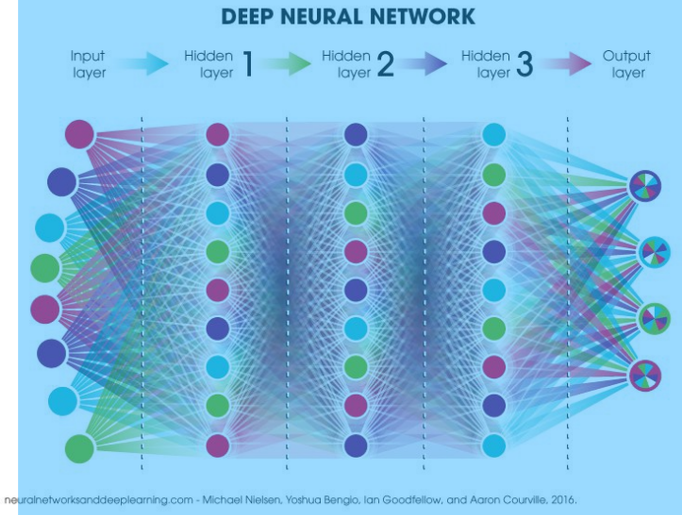
- Variant of the classical trolley problem (Thomson 1985)
- Different people, especially from different cultures and backgrounds, differ with respect to the optimal ethical action
- Moral Machine: dataset collected by collaborators at MIT
  - website where individuals could provide their optimal ethical action for varying self-driving car trolley dilemmas
  - each dilemma has two alternatives, characterized by 22 features (passengers or pedestrians, legality, differing character types (man/woman/child/cat/…), with varying characteristics (age/gender/…)
  - dataset of responses from 1,303,778 individuals, from multiple countries, each with around 14 responses

Utilitarian Ethics

"Fit" data well **"optimal action" involves a loss function**

Data → Machine Learning Model → Decisions





$\ell_{\text{action}}(\theta, a)$: loss i.e. negative utility associated with model parameter θ, and action a

Individual Utility Model

“optimal action” involves a loss function

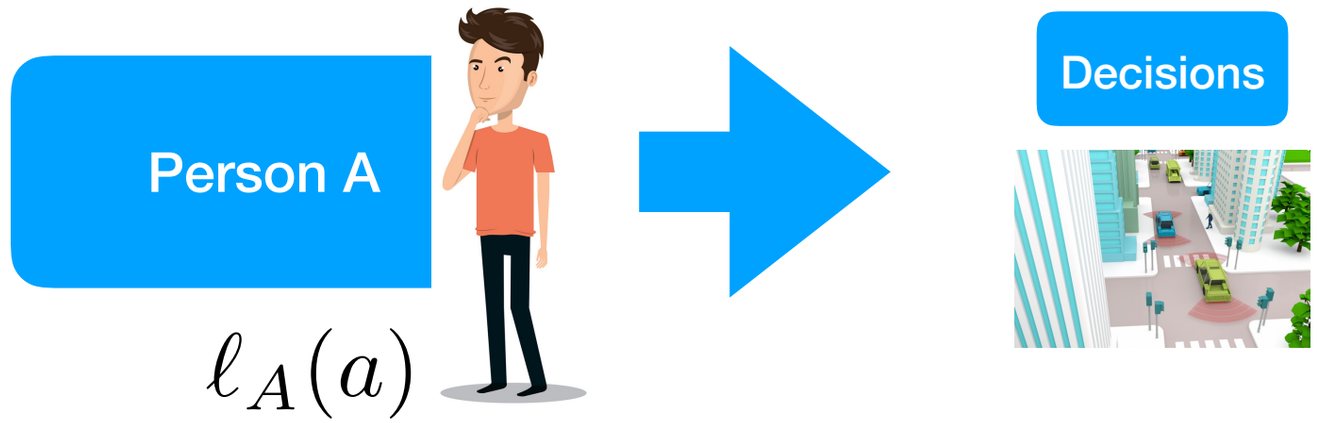
Person A

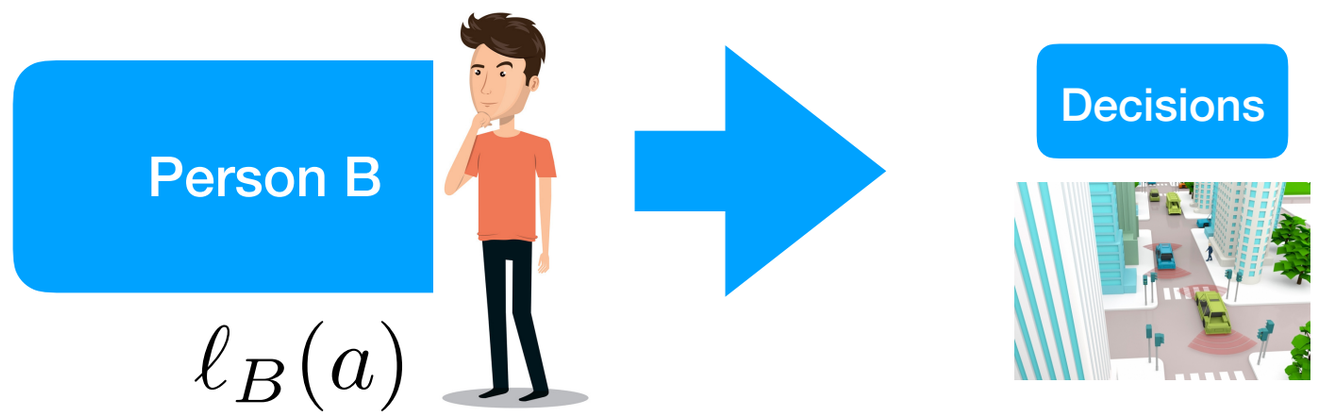
$\ell_A(a)$

Decisions

Varied Individual Utility Models

“optimal action” involves a loss function

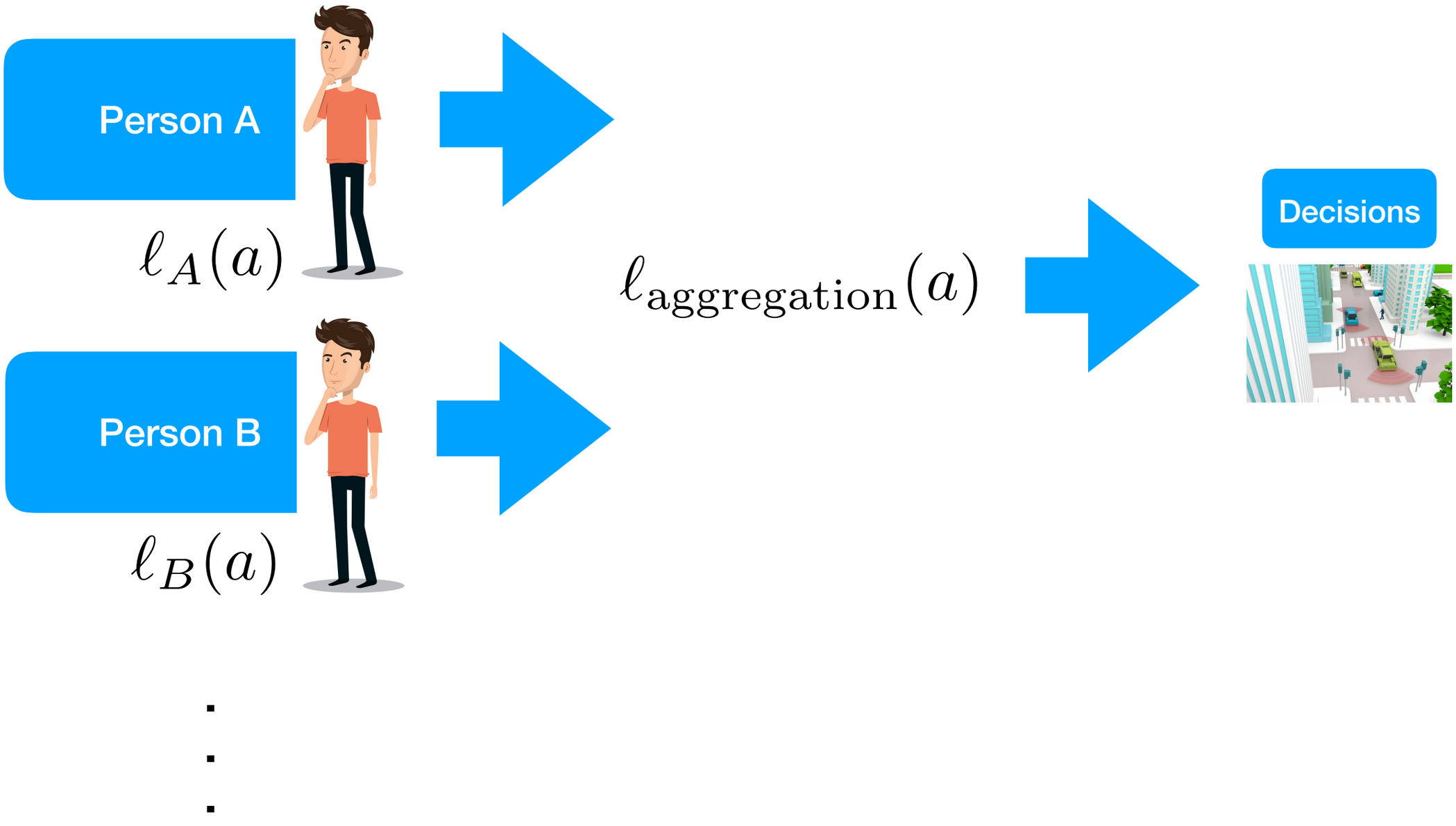




.
.
.

.
.
.

Aggregation of Utility Models



Ethical AI via aggregation of learned utility models

- Task I: Learn individual utility (or loss) models

- Task II: Aggregate individual utility models to create a “consensus” utility model

Learning individual utility models

- Random Utility Models (RUM): Given a set A of actions/ alternatives, a random utility model **U** is a stochastic process where **U(a),** for any alternative **a in A,** denotes the random utility (negative loss) associated with alternative **a**

- **Thurstone-Mosteller (TM) RUM:**

$$U(a) \sim \mathcal{N}(\mu_a, \sigma^2), \text{ where } \mu_a \text{ is mean utility for alternative } a$$

- **Plackket-Luce (PL) RUM:**

$$U(a) \sim \text{Gumbel}(\mu_a, \gamma), \text{ where } \mu_a \text{ is mean utility for alternative } a$$

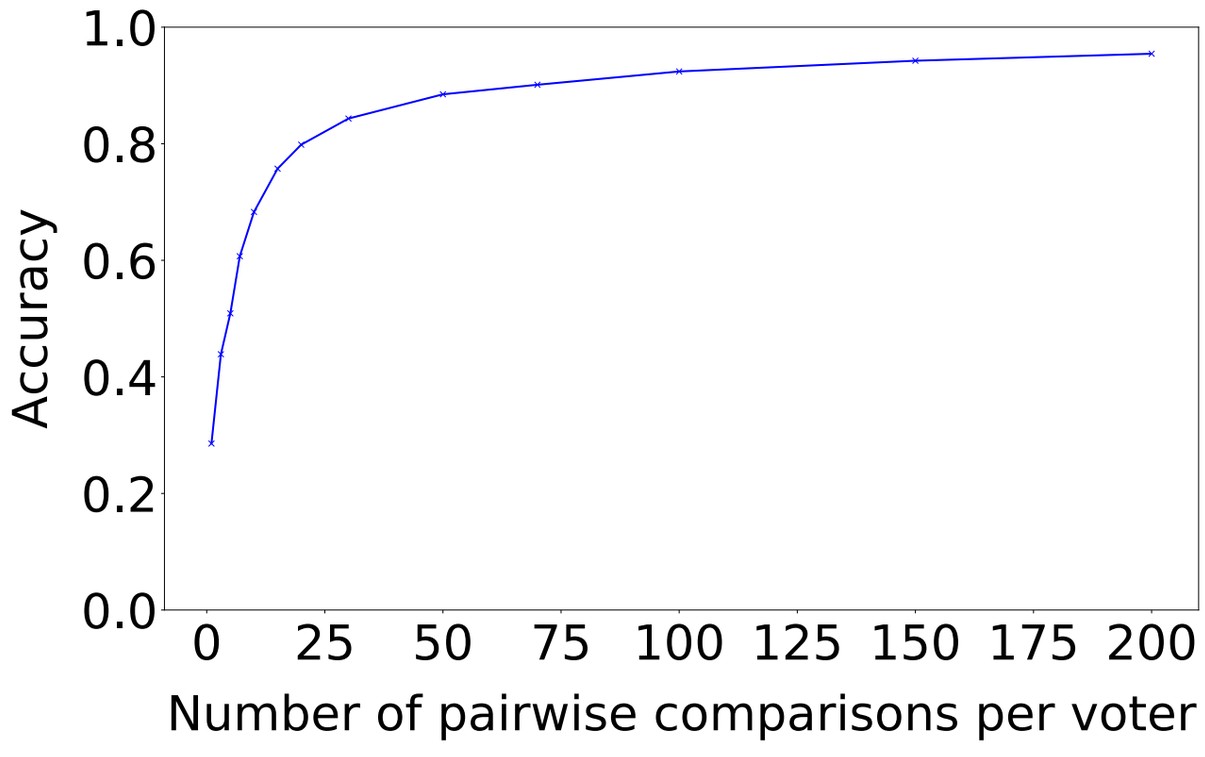
  - Parameterized by mean utility parameters $\{\mu_a\}_{a \in A}$

Learning a TM RUM

- Data: pairwise comparisons $\{a_i \succ b_i\}_{i=1}^n$
  - e.g. {5 passengers > cat + doctor}
- Linear parameterization:
  - $U(a) \sim \mathcal{N}(\langle \beta, a \rangle, 1/2)$
- $\mathbb{P}_\beta(a_i \succ b_i) = \mathbb{P}(U_\beta(a_i) > U_\beta(b_i))$
- Estimator: $\widehat{\beta} \in \arg\sup_\beta \left\{ \prod_{i=1}^n \mathbb{P}_\beta(a_i \succ b_i) \right\}$

Learning a TM RUM

- Estimator: $\widehat{\beta} \in \arg\sup_{\beta} \left\{ \prod_{i=1}^{n} \mathbb{P}_{\beta}(a_i \succ b_i) \right\}$

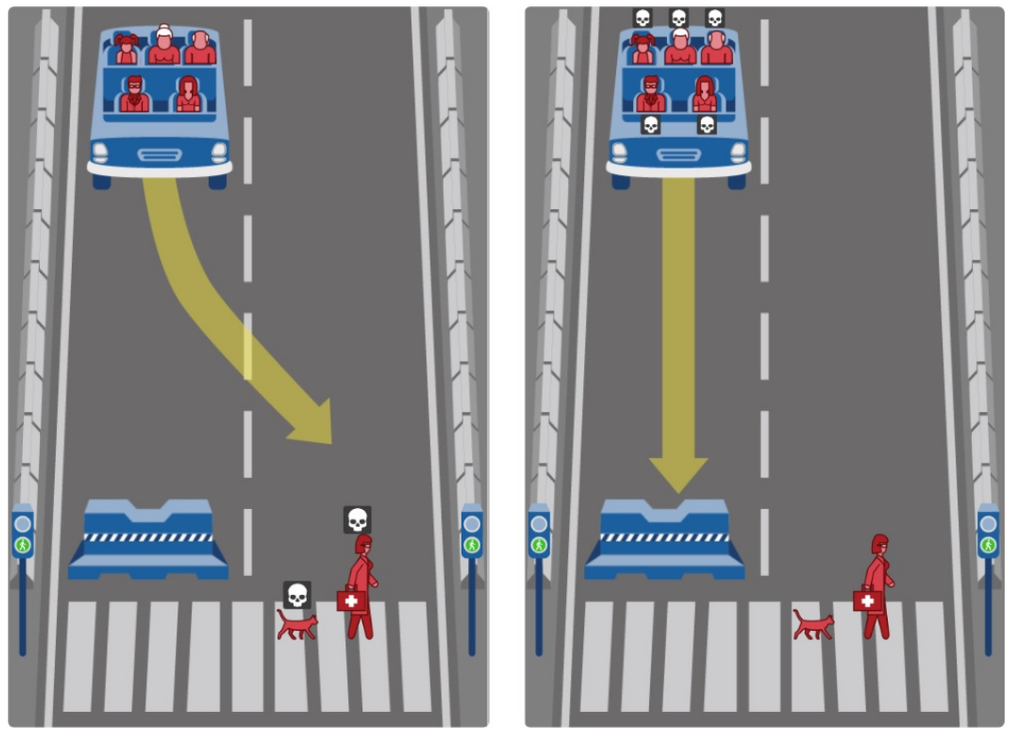


need very few comparisons per voter to learn their preferences

Aggregating TM RUMs

- Suppose N individuals give their ethical opinions, and for each of them, we learn a separate TM RUM
- How do we aggregate these RUMs $\{U_{\beta_i}(\cdot)\}_{i=1}^{N}$
- A reasonable estimator:
  - $\widehat{\beta}_{\mathrm{AGG}} \in \arg\inf_{\beta} \mathrm{KL}\left(\frac{1}{N}\sum_{i=1}^{N} U_{\beta_i} \| U_{\beta}\right)$
  - Finds a TM RUM that is closest to average utility (giving one vote to each person)
  - **Theorem:** $\widehat{\beta}_{\mathrm{AGG}} = \frac{1}{N}\sum_{i=1}^{n} \beta_i$

Ethical Decisions via Aggregate TM RUM

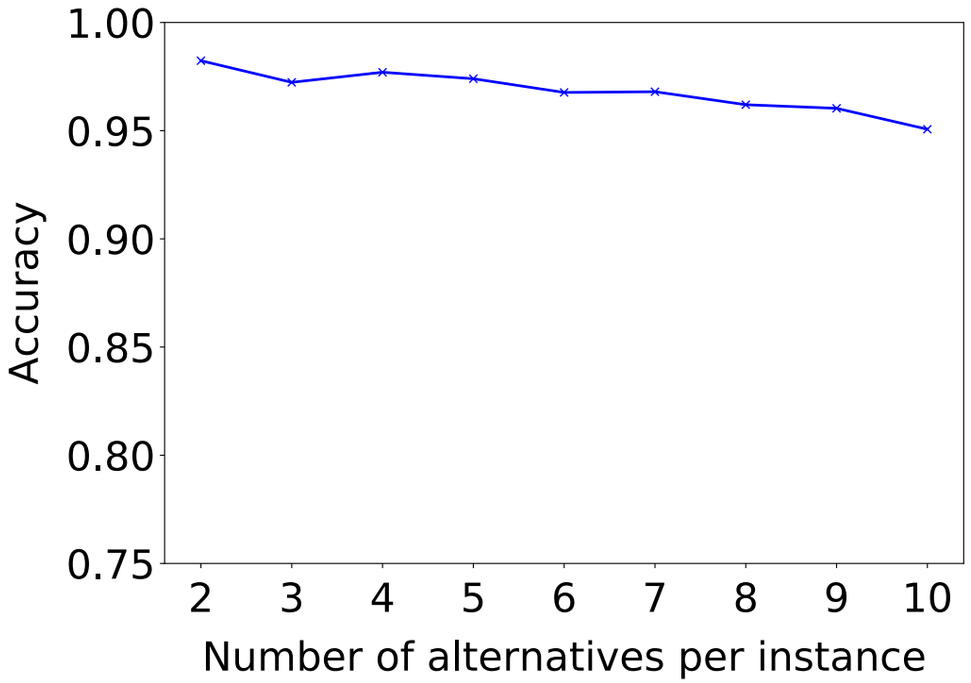


Given alternatives $\{a_1, \ldots, a_m\}$, pick the alternative:

$$a \in \arg \max_{\{a_1,\ldots,a_m\}} \mathbb{E}U_{\beta_{\text{AGG}}}(a)$$

$$\equiv a \in \arg \max_{\{a_1,\ldots,a_m\}} \beta_{\text{AGG}}^T a$$

Validating Aggregate TM RUM



- Suppose we could conduct a real-time election: faced with a **fresh** set of alternatives, we ask each one of the millions of the voters, get their preferences, and then aggregate to get a consensus winning action
    - impractical, computationally expensive
    - decision made by aggregate TM RUM mimicked social-choice theoretically optimal aggregation of (large sample of) all the voter preferences

Validating Aggregate TM RUM

- **Theorem (Stability):** If our system picks action **a** as the most ethical action when presented with a set **A** of alternatives, then it will again pick **a** as the most ethical action when presented with a set **B** of alternatives that is a subset of **A**, if it includes **a**.

- if the system prefers to save a dog over a cat or a mouse, then it should prefer to save a dog over a cat.

Validating Aggregate TM RUM

- **Theorem (Swap Efficient):** If our system picks action **a** as the most ethical action when presented with a set **A** of alternatives, then if there are two preferences which are identical except for swapped preferences between items **a** and **b**, then more people would have voted for the preference order where **a** is preferred to **b**.

Summary: Ethical AI

- Machine Learning has a utilitarian foundation
    - loss functions (or utilities) for (a) fitting model, (b) making decisions
- We learn per person utilities (loss functions), and aggregate them to form a consensus utility
- When doing so in the context of ethical decisions (for trolley dilemmas for self-driving cars), this results in an automated system that can make ethical decisions that represents the “ethical consensus” of millions of individuals
    - computationally practical, satisfies strong social choice theoretic properties
- In ongoing work, we are developing ethical AI systems built on virtue ethics, and deontological ethics
    - and learning more complex human utility models e.g. for suicidal behaviors