In the lab, simple objectives are good...
... but in reality, simple objectives have unintended side effects

On March 21, Navajo activist and social worker Amanda Blackhorse learned her Facebook account had been suspended. The social media service suspected her of using a fake last name.

This halt was more than an inconvenience. It meant she could no longer use the network to reach out to young Native Americans who indicated they might commit suicide.

Many other Native Americans with traditional surnames were swept up by Facebook’s stringent names policy, which is meant to authenticate user identity but has led to the suspension of accounts held by those in the Native American, drag and trans communities.

Uber drew criticism on Sunday by London users accusing the cab-hailing app of charging surge prices around the London Bridge area during the moments after the horrific terror attack there.

On Saturday night, some 7 people were killed and dozens injured when three terrorists mowed a white van over pedestrians and attacked people in the Borough Market area with knives. Police killed the attackers within eight minutes of the first call reporting the attack.

Furious Twitter users accused the app of profiting from the attack with surge prices. Amber Clemente claimed that the surge price was more than two times the normal amount.
AAAI /ACM Conference on
Artificial Intelligence, Ethics, and Society
Honolulu, Hawaii, USA
January 27-28, 2019
CALL FOR PAPERS
Moral Decision Making Frameworks for Artificial Intelligence

[AAAI’17]

with:

Walter Sinnott-Armstrong
Jana Schaich Borg
Yuan Deng
Max Kramer
Two main approaches

Extend **game theory** to directly incorporate moral reasoning

Cf. top-down vs. bottom-up distinction [Wallach and Allen 2008]

Generate data sets of human judgments, apply machine learning
Scenarios

• You see a woman throwing a stapler at her colleague who is snoring during her talk. How morally wrong is the action depicted in this scenario?
  • Not at all wrong (1)
  • Slightly wrong (2)
  • Somewhat wrong (3)
  • Very wrong (4)
  • Extremely wrong (5)

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in:
- The deaths of a female doctor, a female executive, a girl, a woman and an elderly woman.

Note that the affected pedestrians are flouting the law by crossing on the red signal.


Noothigattu et al., “A Voting-Based System for Ethical Decision Making”, AAAI’18
In this case, the self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in:

- The deaths of 3 cats.

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in:

- The deaths of 3 pregnant women.

Note that the affected pedestrians are abiding by the law by crossing on the green signal.
The Merging Problem
[Sadigh, Sastry, Seshia, and Dragan, RSS 2016]

(thanks to Anca Dragan for the image)
Concerns with the ML approach

• What if we predict people will disagree?
  • Social-choice theoretic questions [see also Rossi 2016, and Noothigattu et al. 2018 for moral machine data]

• This will *at best* result in current human-level moral decision making [raised by, e.g., Chaudhuri and Vardi 2014]
  • ... though might perform better than any *individual* person because individual’s errors are voted out

• Feedback to people about how the AI assesses their decisions can change how they make decisions!! [Chan, Doyle, McElfresh, C., Dickerson, Schaich Borg, and Sinnott-Armstrong AIES 2020]

• How to generalize appropriately? Representation?
Adapting a Kidney Exchange Algorithm to Align with Human Values

[AAAI’18]

with:

Rachel Freedman
Jana Schaich Borg
Walter Sinnott-Armstrong
John P. Dickerson
Kidney exchange [Roth, Sönmez, and Ünver 2004]

- Kidney exchanges allow patients with willing but incompatible live donors to swap donors

![Diagram of kidney exchange](image)

**Figure 1:** A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parentheses.

- Algorithms developed in the AI community are used to find optimal matchings (starting with Abraham, Blum, and Sandholm [2007])
Another example

Figure 2: A compatibility graph with four patient-donor pairs and two maximal solutions. Donor and patient blood types are given in parentheses.
Different profiles for our study

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Alternative 0</th>
<th>Alternative 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>30 years old (Young)</td>
<td>70 years old (Old)</td>
</tr>
<tr>
<td>Health -</td>
<td>1 alcoholic drink per month (Rare)</td>
<td>5 alcoholic drinks per day (Frequent)</td>
</tr>
<tr>
<td>Behavioral</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health -</td>
<td>no other major health problems (Healthy)</td>
<td>skin cancer in remission (Cancer)</td>
</tr>
<tr>
<td>General</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: The two alternatives selected for each attribute. The alternative in each pair that we expected to be preferable was labeled “0”, and the other was labeled “1”.
### MTurkers’ judgments

<table>
<thead>
<tr>
<th>Profile</th>
<th>Age</th>
<th>Drinking</th>
<th>Cancer</th>
<th>Preferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (YRH)</td>
<td>30</td>
<td>rare</td>
<td>healthy</td>
<td>94.0%</td>
</tr>
<tr>
<td>3 (YRC)</td>
<td>30</td>
<td>rare</td>
<td>cancer</td>
<td>76.8%</td>
</tr>
<tr>
<td>2 (YFH)</td>
<td>30</td>
<td>frequently</td>
<td>healthy</td>
<td>63.2%</td>
</tr>
<tr>
<td>5 (ORH)</td>
<td>70</td>
<td>rare</td>
<td>healthy</td>
<td>56.1%</td>
</tr>
<tr>
<td>4 (YFC)</td>
<td>30</td>
<td>frequently</td>
<td>cancer</td>
<td>43.5%</td>
</tr>
<tr>
<td>7 (ORC)</td>
<td>70</td>
<td>rare</td>
<td>cancer</td>
<td>36.3%</td>
</tr>
<tr>
<td>6 (OFH)</td>
<td>70</td>
<td>frequently</td>
<td>healthy</td>
<td>23.6%</td>
</tr>
<tr>
<td>8 (OFC)</td>
<td>70</td>
<td>frequently</td>
<td>cancer</td>
<td>6.4%</td>
</tr>
</tbody>
</table>

Table 2: Profile ranking according to Kidney Allocation Survey responses. The “Preferred” column describes the percentage of time the indicated profile was chosen among all the times it appeared in a comparison.
Bradley-Terry model scores

<table>
<thead>
<tr>
<th>Profile</th>
<th>Direct</th>
<th>Attribute-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (YRH)</td>
<td>1.0000000000</td>
<td>1.0000000000</td>
</tr>
<tr>
<td>3 (YRC)</td>
<td>0.236280167</td>
<td>0.13183083</td>
</tr>
<tr>
<td>2 (YFH)</td>
<td>0.103243396</td>
<td>0.29106507</td>
</tr>
<tr>
<td>5 (ORH)</td>
<td>0.070045054</td>
<td>0.03837135</td>
</tr>
<tr>
<td>4 (YFC)</td>
<td>0.035722844</td>
<td>0.08900390</td>
</tr>
<tr>
<td>7 (ORC)</td>
<td>0.024072427</td>
<td>0.01173346</td>
</tr>
<tr>
<td>6 (OFH)</td>
<td>0.011349772</td>
<td>0.02590593</td>
</tr>
<tr>
<td>8 (OFC)</td>
<td>0.002769801</td>
<td>0.00341520</td>
</tr>
</tbody>
</table>

Table 3: The patient profile scores estimated using the Bradley-Terry Model. The “Direct” scores correspond to allowing a separate parameter for each profile (we use these in our simulations below), and the “Attribute-based” scores are based on the attributes via the linear model.
Effect of tiebreaking by profiles

Figure 3: The proportions of pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box. The numbers are the scores assigned (for tiebreaking) to each profile by each algorithm type. Because the STANDARD algorithm treats all profiles equally, it assigns each profile a score of 1. In this figure and later figures, each box represents the interquartile range (middle 50%), with the inner line denoting the median. The whiskers extend to the furthest data points within 1.5 × the interquartile range of the median, and the small circles denote outliers beyond this range.
Monotone transformations of the weights seem to make little difference.
Classes of pairs of blood types
[Ashlagi and Roth 2014; Toulis and Parkes 2015]

• When generating sufficiently large random markets, patient-donor pairs’ situations can be categorized according to their blood types
  • *Underdemanded* pairs contain a patient with blood type O, a donor with blood type AB, or both
  • *Overdemanded* pairs contain a patient with blood type AB, a donor with blood type O, or both
  • *Self-demanded* pairs contain a patient and donor with the same blood type
  • *Reciprocally demanded* pairs contain one person with blood type A, and one person with blood type B
Most of the effect is felt by underdemanded pairs.

Figure 4: The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box.
Crowdsourcing:
Societal Tradeoffs

(AAMAS’15 blue sky paper; AAAI’16; ongoing work.)

with Rupert Freeman, Markus Brill, Yuqian Li
Example Decision Scenario

• Benevolent government would like to get old inefficient cars off the road

• But disposing of a car and building a new car has its own energy (and other) costs

• Which cars should the government aim to get off the road?
  • even energy costs are not directly comparable (e.g., perhaps gasoline contributes to energy dependence, coal does not)
The basic version of our problem is as bad as producing 1 bag of landfill trash is as bad as using $x$ gallons of gasoline. How to determine $x$?
One Approach: Let’s Vote!

• What should the outcome be...?
  • Average? Median?

• Assuming that preferences are single-peaked, selecting the median is strategy-proof and has other desirable social choice-theoretic properties

\[ 1 \text{gal} = x \]
Consistency of tradeoffs

Consistency:

\[ z = xy \]
A paradox

Just taking medians pairwise results in inconsistency
A first attempt at a rule satisfying consistency

- Let $t_{a,b,i}$ be voter $i$’s tradeoff between $a$ and $b$
- Aggregate tradeoff $t$ has score $\sum_i \sum_{a,b} |t_{a,b} - t_{a,b,i}|$

\[
\begin{array}{c|c|c}
\text{forest} & 100 & 200 \\
\text{trash} & 2 & \\
\text{gasoline} & 100 & 200 \\
\end{array}
\begin{array}{c|c|c}
\text{forest} & 300 & 300 \\
\text{trash} & 1 & \\
\text{gasoline} & 200 & 600 \\
\end{array}
\begin{array}{c|c|c}
\text{forest} & 300 & \\
\text{trash} & 3 & \\
\text{gasoline} & 200 & 300 \\
\end{array}
\]

Distance:
- Forest to $v_1$: 100
- Forest to $v_2$: 100
- Forest to $v_3$: 300
- Trash to $v_1$: 100
- Trash to $v_2$: 1
- Trash to $v_3$: 3/2

Total distance: 602.5 (minimum)
A nice property

- This rule agrees with the median when there are only two activities!

- $x$ should be 2
- $x$ should be 4
- $x$ should be 10

- Distance:
  - $2 + 8 = 10$
  - $2 + 6 = 8$
  - $8 + 6 = 14$
Not all is rosy, part 1

- What if we change units? Say forest from $m^2$ to $cm^2$ (divide by 10,000)

![Diagram showing the change in calculations after changing units. The distances between nodes are marked as distances: (negligible) or different from before! fails independence of other activities’ units.]
Not all is rosy, part 2

• Back to original units, but let’s change some edges’ direction

![Diagram](image-url)

- Distance: (negligible)
- Different from before! Fails independence of other edges’ directions
Summarizing

• Let $t_{a,b,i}$ be voter $i$’s tradeoff between $a$ and $b$

• Aggregate tradeoff $t$ has score

  $\sum_i \sum_{a,b} |t_{a,b} - t_{a,b,i}|$

• Upsides:
  • Coincides with median for 2 activities

• Downsides:
  • Dependence on choice of units:

    $|t_{a,b} - t_{a,b,i}| \neq |2t_{a,b} - 2t_{a,b,i}|$

  • Dependence on direction of edges:

    $|t_{a,b} - t_{a,b,i}| \neq |1/t_{a,b} - 1/t_{a,b,i}|$

  • We don’t have a general algorithm
A generalization

• Let $t_{a,b,i}$ be voter $i$’s tradeoff between $a$ and $b$
• Let $f$ be a monotone increasing function – say, $f(x) = x^2$
• Aggregate tradeoff $t$ has score
  \[
  \sum_i \sum_{a,b} |f(t_{a,b}) - f(t_{a,b,i})|
  \]
• Still coincides with median for 2 activities!
• **Theorem:** These are the only rules satisfying this property, agent separability, and edge separability

<table>
<thead>
<tr>
<th>$t_{a,b}$</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(t_{a,b})$</td>
<td>1</td>
<td>4</td>
<td>9</td>
</tr>
</tbody>
</table>
So what’s a good $f$?

- **Intuition**: Is the difference between tradeoffs of 1 and 2 the same as between 1000 and 1001, or as between 1000 and 2000?

- So how about $f(x) = \log(x)$?
  - (Say, base e – remember $\log_a(x) = \log_b(x)/\log_b(a)$)

<table>
<thead>
<tr>
<th>$t_{a,b}$</th>
<th>1</th>
<th>2</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln(t_{a,b})$</td>
<td>$\ln(1)$</td>
<td>$\ln(2)$</td>
<td>$\ln(1000)$</td>
<td>$\ln(2000)$</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.69</td>
<td>6.91</td>
<td>7.60</td>
</tr>
</tbody>
</table>
On our example
Properties

• Independence of units
  \[ | \log(1) - \log(2) | = | \log(1/2) | = \]
  \[ | \log(1000/2000) | = | \log(1000) - \log(2000) | \]
  More generally:
  \[ | \log(ax) - \log(ay) | = | \log(x) - \log(y) | \]

• Independence of edge direction
  \[ | \log(x) - \log(y) | = | \log(1/y) - \log(1/x) | = \]
  \[ | \log(1/x) - \log(1/y) | \]

• **Theorem.** The logarithmic distance based rule is unique in satisfying independence of units.*
  
  * Depending on the exact definition of independence of units, may need another minor condition about the function locally having bounded derivative.
Consistency constraint becomes additive

\[ xy = z \]

is equivalent to

\[ \log(xy) = \log(z) \]

is equivalent to

\[ \log(x) + \log(y) = \log(z) \]
An additive variant

- “I think basketball is 5 units more fun than football, which in turn is 10 units more fun than baseball”
Natural objective:

\[
\text{minimize } \sum_i \sum_{a,b} d_{a,b,i} \text{ where } d_{a,b,i} = |t_{a,b} - t_{a,b,i}| \text{ is the distance between the aggregate difference } t_{a,b} \text{ and the subjective difference } t_{a,b,i}
\]

objective value 70 (optimal)
A linear program for the additive variant

$q_a$: aggregate assessment of quality of activity $a$ (we’re really interested in $q_a - q_b = t_{a,b}$)

d$_{a,b,i}$: how far is $i$’s preferred difference $t_{a,b,i}$ from aggregate $q_a - q_b$, i.e., $d_{a,b,i} = |q_a - q_b - t_{a,b,i}|$

minimize $\sum_i \sum_{a,b} d_{a,b,i}$

subject to

for all $a,b,i$: $d_{a,b,i} \geq q_a - q_b - t_{a,b,i}$

for all $a,b,i$: $d_{a,b,i} \geq t_{a,b,i} - q_a + q_b$

(Can arbitrarily set one of the $q$ variables to 0)
Applying this to the logarithmic rule in the multiplicative variant

Just take logarithms on the edges, solve the additive variant, and exponentiate back
A simpler algorithm (hill climbing / greedy)

- Initialize qualities $q_a$ arbitrarily
- If some $q_a$ can be individually changed to improve the objective, do so
  - WLOG, set $q_a$ to the median of the $(#voters)*(#activities-1)$ implied votes on it
- Continue until convergence (possibly to local optimum)
Flow-based exact algorithm [AAAI’19]

with:

Hanrui Zhang

Yu Cheng
Deomposition

• Idea: Break down activities to relevant attributes

- gasoline use
  - contributes a units to
  - contributes b units to
  - contributes c units to
- global warming
- energy dependence
- ...
Another Paradox

activity A  
(gasoline)

activity B  
(trash)

Agent 1
Agent 2
Agent 3

attribute 1  
(global warming)

attribute 2  
(energy dependence)

aggregation on attribute level ≠ aggregation on activity level
Other Issues

• **Objective vs. subjective** tradeoffs
  • separate process?
  • who determines which is which?

• **Who gets to vote?**
  • how to bring **expert knowledge** to bear?
  • incentives to **participate**

• **Global vs. local** tradeoffs
  • different entities (e.g., countries) may wish to reach their tradeoffs **independently**
  • only care about opinions of **neighbors in my social network**

• ...
Why Do We Care?

• Inconsistent tradeoffs can result in inefficiency
  • Agents optimizing their utility functions individually leads to solutions that are Pareto inefficient

• Pigovian taxes: pay the cost your activity imposes on society (the externality of your activity)
  • If we decided using 1 gallon of gasoline came at a cost of $x to society, we could charge a tax of $x on each gallon
  • But where would we get x?
Inconsistent tradeoffs can result in inefficiency

- Agent 1: 1 gallon = 3 bags = -1 util
  - I.e., agent 1 feels she should be willing to sacrifice up to 1 util to reduce trash by 3, but no more
- Agent 2: 1.5 gallons = 1.5 bags = -1 util
- Agent 3: 3 gallons = 1 bag = -1 util
- Cost of reducing gasoline by $x$ is $x^2$ utils for each agent
- Cost of reducing trash by $y$ is $y^2$ for each agent
- Optimal solutions for the individual agents:
  - Agent 1 will reduce by 1/2 and 1/6
  - Agent 2 will reduce by 1/3 and 1/3
  - Agent 3 will reduce by 1/6 and 1/2
- But if agents 1 and 3 each reduce everything by 1/3, the total reductions are the same, and their costs are 2/9 rather than 1/4 + 1/36 which is clearly higher.
  - Could then reduce slightly more to make everyone happier.
Single-peaked preferences

• **Definition:** Let agent $a$’s most-preferred value be $p_a$.

Let $p$ and $p'$ satisfy:
- $p' \leq p \leq p_a$, or $p_a \leq p \leq p'$

• The agent’s preferences are **single-peaked** if the agent always weakly prefers $p$ to $p'$
Perhaps more reasonable...

- $x$ should be between 0 and 4
- $x$ should be between 2 and 6
- $x$ should be between 9 and 11

- E.g., due to missing information or plain uncertainty

- How to aggregate these interval votes? [Farfel & Conitzer 2011]
Median interval mechanism

• Construct a consensus interval from the median lower bound and the median upper bound

• Strategy-proof if preferences are single-peaked over intervals
Single-peaked preferences over intervals

• **Definition:** Let agent $a$’s most-preferred value interval be $P_a = [l_a, u_a]$.

Let $S = [l, u]$ and $S' = [l', u']$ be any two value intervals satisfying the following constraints:
- Either $l' \leq l \leq l_a$, or $l_a \leq l \leq l'$
- Either $u' \leq u \leq u_a$, or $u_a \leq u \leq u'$

• The agent’s preferences over intervals are **single-peaked** if the agent always weakly prefers $S$ to $S'$