Heuristic-Based Weak Learning for Automated Decision-Making

Ryan Steed $^{1,2}$  Benjamin Williams $^2$

$^1$Carnegie Mellon University
$^2$George Washington University

July 17, 2020
PAML @ ICML 2020
Motivation

Question

How can we lower the barrier to collective participation in algorithmic policy?

- One solution: elicit & aggregate user preferences.
  - Self-driving cars (Noothigattu et al., 2018; Kim et al., 2018)
  - Kidney exchange (Freedman et al., 2018)
  - Food donation allocation (Kahng et al., 2019; Lee et al., 2019)
Motivation

Question

How can we lower the barrier to collective participation in algorithmic policy?

- One solution: elicit & aggregate user preferences.
  - Self-driving cars (Noothigattu et al., 2018; Kim et al., 2018)
  - Kidney exchange (Freedman et al., 2018)
  - Food donation allocation (Kahng et al., 2019; Lee et al., 2019)

- Usually relies on many hand-labeled pairwise comparisons.
  - Costly labor from stakeholders or a crowd
  - May be less trustworthy than explicit rules (Lee et al., 2019)
Weak Supervision

**Idea**

Improve preference elicitation with decision-making heuristics.
Weak Supervision

Idea

Improve preference elicitation with decision-making heuristics.

**Heuristic:** a practical rule for decision-making.

```python
@labeling_function()
def utilitarian(x):
    """Save the most human lives.""
    saved_by_int = x['intervention']['Human']
    saved_by_no_int = x['no_intervention']['Human']
    return argmax([saved_by_int, saved_by_no_int])
```

Figure 1: A simple utilitarian heuristic in Python using the open-source Snorkel labeling function interface (snorkel.org).
Figure 2: A heuristic-based, weak supervision pipeline for automating decision-making.
What should the self-driving car do?

- Over 1.5 million decisions from around 50,000 respondents - mostly white male college graduates from U.S. & Europe (Awad et al., 2018)
- We wrote 15 heuristics based on estimated global preferences
Figure 3: Mean accuracy (rate of agreement with respondents’ pairwise decisions) across 50 trials with 95% confidence interval (shaded).

Benchmark: Kim et al. (2018) approach 75% accuracy as the number of respondents increases.
Who gets the kidney?

**Patient W.A.**
30 years old
Had 1 alcoholic drink per month
No major health problems

**Patient R.F.**
70 years old
Had 5 alcoholic drinks per day
Skin cancer in remission

*Figure 4: Freedman et al. (2018)* asked 289 Mechanical Turk users to allocate a kidney between two patients in 28 pairwise comparisons like the one shown here.
Who gets the kidney?

**Patient W.A.**
30 years old
Had 1 alcoholic drink per month
No major health problems

**Patient R.F.**
70 years old
Had 5 alcoholic drinks per day
Skin cancer in remission

Figure 4: Freedman et al. (2018) asked 289 Mechanical Turk users to allocate a kidney between two patients in 28 pairwise comparisons like the one shown here.

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Avg. Borda Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choose younger patient</td>
<td>3.42</td>
</tr>
<tr>
<td>Choose patient who drinks less</td>
<td>2.71</td>
</tr>
<tr>
<td>Choose patient with no other health issues</td>
<td>2.10</td>
</tr>
<tr>
<td>Choose patient with other health issues</td>
<td>0.19</td>
</tr>
<tr>
<td>Choose older patient</td>
<td>0.11</td>
</tr>
<tr>
<td>Choose patient who drinks more</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 1: Reported heuristics for the kidney exchange, ranked by popularity (Borda counts calculated from manual ranked choice coding of text responses).
Figure 5: Mean accuracy (rate of agreement with respondents’ pairwise decisions) across 50 trials with 95% confidence interval (shaded).

Benchmark: Freedman et al. (2018) agree with respondents 85.8% of the time.
Summary

- Why heuristics for collective participation?
  - For participants, an alternative means to express complex preferences
  - Empirically comparable performance, especially when heuristics are ranked
Summary

Why heuristics for collective participation?
- For participants, an alternative means to express complex preferences
- Empirically comparable performance, especially when heuristics are ranked

Future work:
- Are heuristic-based models more trustworthy?
- Performance in domains requiring rare expertise or more numerous/complex features?
- Heuristics for allocation, matching (not just classification)?
Questions?
ryansteed@cmu.edu

Code and data can be accessed at rbsteed.com/heuristic-moral-machine.
Slides can be accessed at rbsteed.com/paml-2020.

Acknowledgements
Special thanks to Rahul Simha, Brian Wright, and Rachel Riedner for their helpful comments.
References I


References II

Figure 6: Rate of agreement with Moral Machine respondents (accuracy) vs. rate of non-abstention (coverage). Heuristics are sized by strength of preference, as measured by Awad et al. (2018).