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.

```
@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).

Method

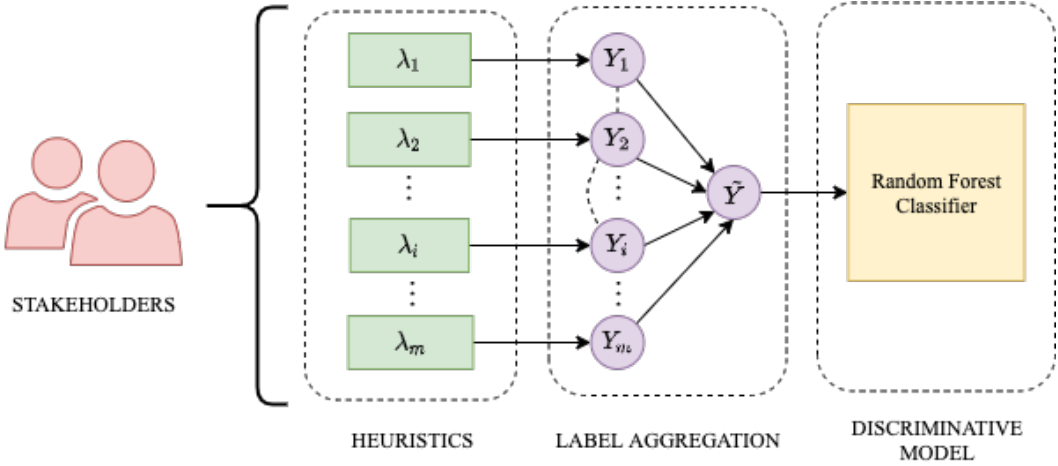
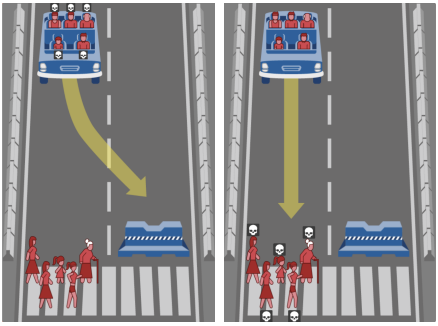


Figure 2: A heuristic-based, weak supervision pipeline for automating decision-making.

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 ▸ Appendix

Discriminative Accuracy

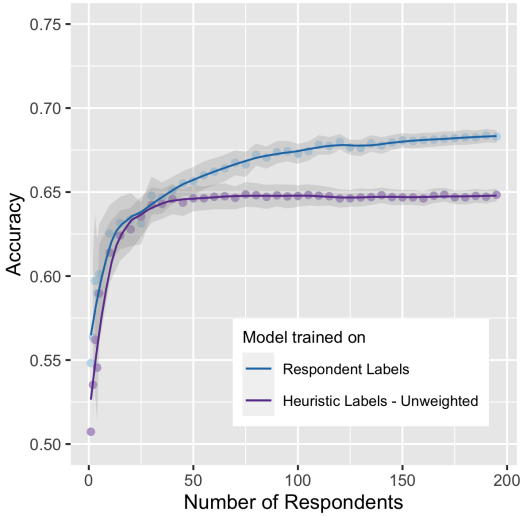


Figure 3: Mean accuracy (rate of agreement with respondents' pairwise decisions) across 50 trials with 95% confidence interval (shaded).

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.

Heuristic	Avg. Borda Count
Choose younger patient	3.42
Choose patient who drinks less	2.71
Choose patient with no other health issues	2.10
Choose patient with other health issues	0.19
Choose older patient	0.11
Choose patient who drinks more	0.04

Table 1: Reported heuristics for the kidney exchange, ranked by popularity (Borda counts calculated from manual ranked choice coding of text responses).

Discriminative Accuracy

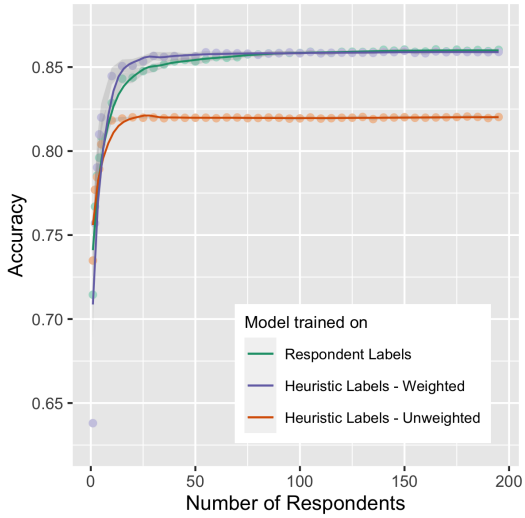


Figure 5: Mean accuracy (rate of agreement with respondents' pairwise decisions) across 50 trials with 95% confidence interval (shaded).

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

Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J. F. Bonnefon, and I. Rahwan (2018, 11). The Moral Machine experiment. *Nature 563*(7729), 59–64.

Freedman, R., J. Schaich Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer (2018). Adapting a Kidney Exchange Algorithm to Align with Human Values. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, New York, New York, USA, pp. 115–115. ACM Press.

Kahng, A., M. K. Lee, R. Noothigattu, A. Procaccia, and C.-A. Psomas (2019). Statistical Foundations of Virtual Democracy. In *International Conference on Machine Learning*, pp. 3173–3182.

Kim, R., M. Kleiman-Weiner, A. Abeliuk, E. Awad, S. Dsouza, J. B. Tenenbaum, and I. Rahwan (2018, 12). A Computational Model of Commonsense Moral Decision Making. In *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 197–203. Association for Computing Machinery, Inc.

Lee, M. K., A. Kahng, J. T. Kim, X. Yuan, A. Chan, S. Lee, A. D. Procaccia, D. Kusbit, D. See, R. Nooth-Igattu, and A. Psomas (2019). WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact 3*.

References II

Noothigattu, R., S. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia (2018). A voting-based system for ethical decision making. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Appendix

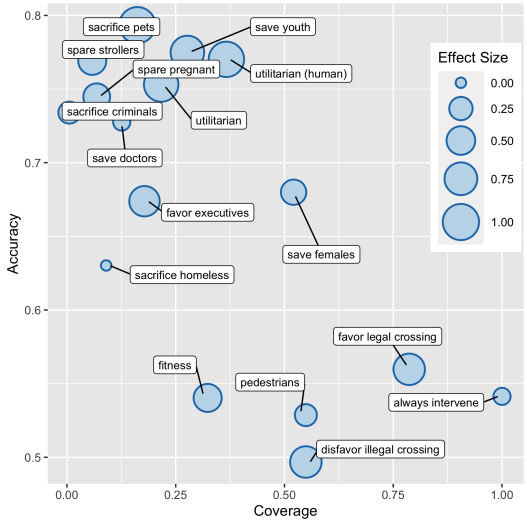


Figure 6: Rate of agreement with Moral Machine respondents (accuracy) vs. rate of non-abstention (coverage). Heuristics are sized by strength of preference, as measured by Awad et al. (2018).