Fair Machine: A Crowdsourcing Platform for Human Conceptions of Algorithmic Fairness

Levi Lian
Stanford University
levilian@stanford.edu

Alan Davoust
University of Edinburgh
adavoust@inf.ed.ac.uk

Michael Rovatsos
University of Edinburgh
mrovatso@inf.ed.ac.uk

Abstract

In the current spate of events and ensuing concerns around fairness of artificial intelligence (AI), the demand for "fair" data-driven algorithms is increasing. However, what is fair is both context-dependent and varies across individuals and cultures. So to program fairness into the algorithms is to first articulate what is fairest in a given context. The crucial step therefore is to formalize how individuals and societies perceive fairness and how they balance trade-offs. This paper extends the idea of crowdsourcing to identify human conceptions of fairness in different contexts. While empirical validation is still work to be done, the paper demonstrates the promise and perils of deriving societal preferences for different fairness criteria in different contexts through large-scale on-line experiments.

1 Introduction

AI has advanced rapidly in both its theoretical underpinnings and application in the past decade. Deep Neural Network architectures, Reinforcement Learning and economic reasoning in multi-agent systems have brought about unprecedented societal and individual benefits [1, 2, 3]. Among the prominent examples are voice recognition technologies that help hunt down criminals and simplify contactless payments, and image recognition advancements that expedite medical diagnosis and drug discovery.

However, the AI advances have raised serious concerns around issues of fairness, following a series of events. Certain groups have been unfairly denied loans, insurance, or employment opportunities while others suffer from existing biases in predictive policing or outrageous discrimination on search engines and personalization.

Therefore, the whole realm of system fairness, algorithmic fairness and data fairness is thrown into question. Furthermore, the black-box algorithms tend to refuse algorithmic governance or auditing. Even if oversight becomes possible, the difficulty with tracing upstream the responsible agents would be an insurmountable task. While attempts have been made to detect and mitigate unfairness in existing algorithms, designing fair procedures should not be delayed [4].

However, measures of fairness differ greatly and, as it is proven mathematically and reasoned intuitively, trade-offs must arise, particularly with algorithmic decision-making [5]. Is fairness through blindness necessary for due process and procedural justice? Should fairness consist of ensuring everyone has an equal probability of obtaining some benefit, or should we aim instead to minimize the harms to the least advantaged?

AA93: Building Trust in Autonomy, Summer 2018, Stanford University, CA

One question which immediately arises in such an endeavor is the need for formalization, i.e., how should algorithms decide what is fair in different contexts? To answer this question is to have a sense of what society as a whole would consider the fairest given the information we have. Some have argued for a “society in the loop” AI governance framework, where societal values would be embedded into algorithmic decision making [6]. This paper follows in the footstep of such design principles.

In this paper, we instantiate a crowdcomputing model proposed by the MIT Moral Machine team, tapping into the scale and reach of the Internet to answer the key question: How do people perceive different kinds of fairness with regard to AI algorithms? We are particularly interested in the case of resource allocation, mainly because the parent project of this research, "UnBias: Emancipating Users Against Algorithmic Biases for a Trusted Digital Economy," strives to design "fair" on-line platforms, which is essentially a resource allocation problem.

As a simple example, we evaluate what kinds of fairness criteria people consider when they are shown the outcomes of a course allocation problem with limited resources and differing individual preferences. We first analyze the results of a small-scale experiment, the limitations of which with eliciting weights for fairness criteria lead to our proposed methodology. Then we discuss representative prior work on crowdcomputing in AI to showcase the promise of a similar method for algorithmic fairness. Moving forward, we show a functional prototype, version 0.2, of the web application that satisfies all the minimum product criteria. Finally, we illustrate the issues our method may run into despite a lack of empirical analysis (note the research started less than a month ago), followed by additional discussion, final conclusions and suggested future work. While the Moral Machine approach provides a promising formalization of fairness criteria, there exists important limitations that will delay if not forestall the quest for an answer.

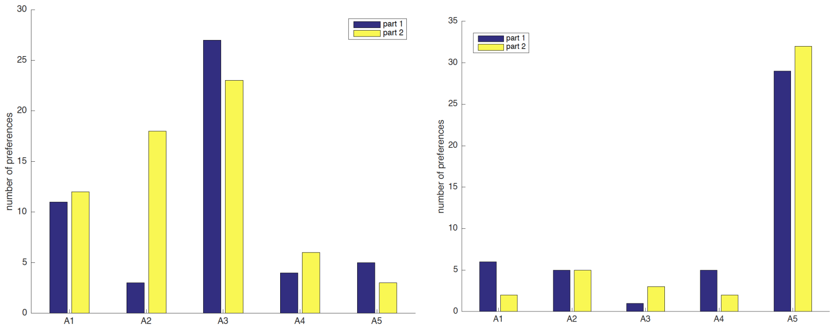
2 UnBias Course Allocation Experiment

The course allocation experiment of the UnBias project explores the different fairness criteria encoded in algorithms that people consider in a specific context and how transparency of the process would affect perceptions of fairness. The course allocation problem is that there are 34 spots for 7 courses to 34 students in a university, and each student has a ranking of the courses they want to take. 39 participants of the experiment recruited from both professional fields and UK universities were presented with 1) the aforementioned course allocation problem, 2) five algorithms denoted from $A1$ to $A5$ to solve it, and 3) the tables and graphs showing different utility values obtained by the students for each algorithm, as well as the mean of students' individual utilities, the total utility and the total distance between utilities [7].

The five algorithms used either i) maximize the sum of students' individual utilities (total utility), ii) maximize the lowest utility of any of the students for the allocation (focusing on limiting the “damage” to the student who is least well off given an overall allocation), or iii) minimize the sum of differences between the different students' utilities (aiming to reduce the total “distance” among all students' individual outcomes). Additional algorithms were obtained by combining several of these criteria, i.e. optimizing for one while guaranteeing a certain level of another. Participants were then asked to choose the most preferred and least preferred algorithms from the five. The experiment was then repeated but this time disclosed the details of the algorithms, a sample of which can be found in the appendix. Finally, their preferences for algorithms in two different cases were compared in Figures 1a and 1b.

The results for the experiment are mainly qualitative. Specific to the course allocation case, participants preferred multi-objective optimizing algorithms better than the single-objective ones, and have been influenced by the revelations of the algorithms in use. Moreover, the differing and incompatible fairness criteria within a group of 39 suggests the difficulty with arriving at one communally derived fair algorithm. While these conclusions point future research to the right direction, they lack the empirical insights needed to answer the question: what fairness criteria people prefer more in a specific context and the weights they have subconsciously assigned to them.

In addition, the cost of conducting similar research at a scale of merely 39 people is still prohibitive, with no flexibility left for changing key variables and analyzing the impact and interactions of them on the outcome. There are far more areas that the UnBias researchers hope to investigate with the limited resource and time we have, which is also a resource allocation problem in and of itself. All



(a) Figure 1a: Most preferred algorithms in Parts 1 and 2 of the questionnaire.

(b) Figure 1b: Least preferred algorithms, in Parts 1 and 2 of the questionnaire.

Figure 1: The influence of transparency of algorithms on participants' perceptions of fairness

the aforementioned limitations force us to look at the possibility of crowdcomputing. At the first sight, the flexibility of the experimental design, extreme reach to people across the world, relatively simple aggregation and analysis of data, possibility of tracking certain features of users (time to make a decision, cursor movement, etc.) all make the future of Fair Machine promising.

3 MIT's Moral Machine

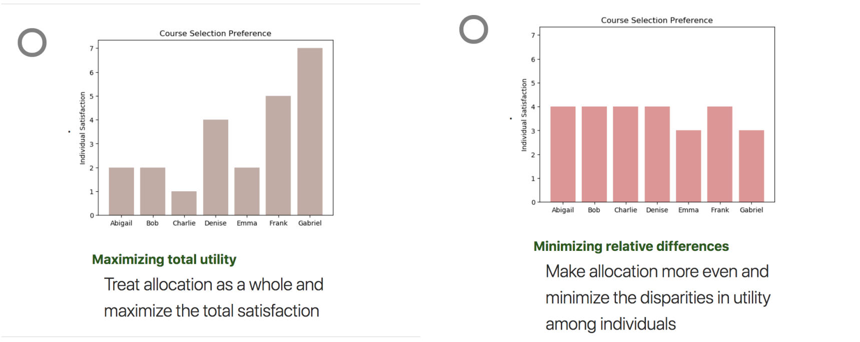
A case in point of the crowdcomputing experiment in AI is Moral Machine. Moral Machine is a web application built to collect and analyze human perceptions of moral dilemmas involving autonomous vehicles. As of June 2018, the application has collected over 50 million responses from over 5 million unique respondents from over 180 countries around the world [8].

Because features such as age, profession, and the law-abiding nature of the pedestrian affect moral values, each scenario can be easily represented as a vector. Two mathematical formulations have attempted to derive the weights people assign to different features when facing a moral dilemma. The first one uses permutations processes to aggregate individuals preferences such that the decision reached after the aggregation ensures global utility maximization [4]. The other one uses hierarchical Bayesian inference and utility function to represent different features [8]

4 Method: Large-scale On-line Experiment

Employing large-scale on-line experiments to collect, derive and aggregate individual choices shoulders the promise of answering how individuals and societies perceive fairness and balance trade-offs. The general framework dates back to 2016 when [9] suggest computational social choice as the tool for ethical decision making with autonomous driving. Instead, here we extend the notion to designing fair algorithmic systems.

We use the first two of the four-step approach by [4] to execute the experimental design, data collection and mathematical formulation. First is data collection. We have designed a functional prototype, a web application supported by Django, Bootstrap and Heroku, to gather user preferences over pairs of alternative algorithms and outcomes in the course allocation scenario. When users land on the main page and click "Start fairness evaluations," they will be presented with two options: "Course Allocation A" and "Course Allocation B." These two options represent the allocation scenarios with and without revealing the details of algorithms respectively, as described in section "UnBias Course Allocation Experiment". After the user has chosen either one of the surveys, they will be shown a description of the aforementioned course allocation problem and the relevant instructions. The instructions are simply to choose the one way of allocation to a sample of 7 students that looks fairer to the user and click "Submit" button to continue. All the 10 pairings per survey are generated randomly according to the algorithm in the back-end. Therefore, users should expect new scenarios to compare with even if they repeatedly doing the survey multiple times.


Course Selection Preference
Individual Satisfaction
0
1
2
3
4
5
6
7
Abigail
Bob
Charlie
Denise
Emma
Frank
Gabriel
Maximizing total utility
Treat allocation as a whole and maximize the total satisfaction
Course Selection Preference
Individual Satisfaction
0
1
2
3
4
5
6
7
Abigail
Bob
Charlie
Denise
Emma
Frank
Gabriel
Minimizing relative differences
Make allocation more even and minimize the disparities in utility among individuals

(a) Figure 2a: One of the two options presented to the user with algorithm in use stated explicitly.

(b) Figure 2b: The other of the two options presented to the user with algorithm in use stated explicitly.

Figure 2: Demo of one of ten pairs of alternatives in course allocation A, with two outcomes generated randomly.

Notably, the pairwise comparisons should be tied to each individual to take into account cultural, sex, age and educational differences. Users therefore need to complete a post-experiment survey to fill out demographic information. This allows us to create a useful database that not only stores the results and scenarios of ten comparisons, but also establishes their relationships with the user making the choice.

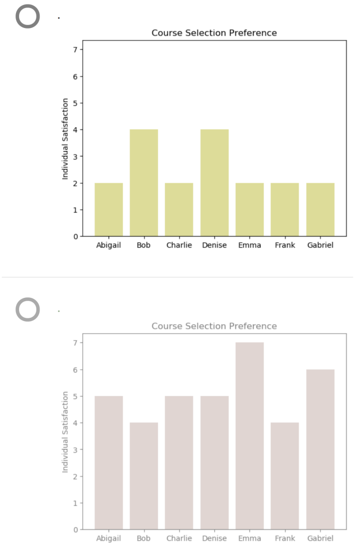


Figure 3: Demo of one of ten pairs of alternatives in course allocation B, with two outcomes generated randomly.

Second, we can analyze how users arrives at their decision based on the values they ascribe to the abstract dimensions of fairness criteria. For instance, if a user chooses an allocation that corresponds to all 7 students receiving equal utility, it suggests that the user places high value on absolute equality of outcome. Therefore, we need to count different fairness concerns, and a mapping of input vector composed of individual utilities to output vector composed of abstract fairness features can be established [8].

However, when we think about the three dimensional case where the outcome can be of either high stake or low stake to the user, e.g., losing a job, losing a housing contract, or simply failing to make a fortune of 5,000 dollars. Such complexity is to the context, not a fundamental feature of the input. Another dimension is whether the algorithm in use is stated explicitly. Because in most cases the allocation outcome does not tell the whole story and for events that have an average normal distribution, each sample case can differ by a significant amount. Therefore, the information is important to due process or fairness of the algorithm [10].

5 Conclusion and Future Work

The ideas proposed in this article draw upon an amalgam of important prior work in the field of AI and ethics. The Fair Machine is also a salute to the Moral Machine approach that stays at the forefront of crowdcomputing experiments. I have attempted to show the motivation for answering the question of how individuals and societies perceive fairness and balance trade-offs, expose the advantages and limitations with small-scale experiments. I present the minimum viable product that serves as a prototype for the online large-scale experimentation, and reflect on the nature of fairness to assess the possibility of correctly quantifying fairness concerns both for individuals and societies.

In the future, researchers could substantiate the prototype and reality-test its promises on real datasets. Such a task would not be difficult as the problem is not new, but would be time-consuming because of the iterative design process and foreseeable cycles of usability testing. Besides, building an online fairness scenario community would tap into the widom of the crowd when thinking about different measures of fairness. This would in turn help enumerate the full spectrum of cases where fairness concerns would diverge. While Fair Machine provides a promising formalization of fairness criteria, there does exist important limitations that will delay if not forestall the quest for an answer.

6 Acknowledgement

I am grateful to Michael Rovatsos for being my advisor as a visiting researcher at the University of Edinburgh, guiding me through the whole process of question framing to experimentation. I am indebted to Alan Davoust for conversations and comments that helped shape this article.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, p. 436, 2015.

[2] M. L. Littman, “Reinforcement learning improves behaviour from evaluative feedback,” *Nature*, vol. 521, no. 7553, p. 445, 2015.

[3] D. C. Parkes and M. P. Wellman, “Economic reasoning and artificial intelligence,” *Science*, vol. 349, no. 6245, pp. 267–272, 2015.

[4] R. Noothigattu, S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia, “A voting-based system for ethical decision making,” *arXiv preprint arXiv:1709.06692*, 2017.

[5] R. Freedman, J. S. Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer, “Adapting a kidney exchange algorithm to align with human values,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] I. Rahwan, “Society-in-the-loop: programming the algorithmic social contract,” *Ethics and Information Technology*, vol. 20, no. 1, pp. 5–14, 2018.

[7] A. Koene, E. Perez, S. Ceppi, M. Rovatsos, H. Webb, M. Patel, M. Jirotka, and G. Lane, “Algorithmic fairness in online information mediating systems,” in *Proceedings of the 2017 ACM on Web Science Conference*, WebSci ’17, (New York, NY, USA), pp. 391–392, ACM, 2017.

[8] R. Kim, M. Kleiman-Weiner, A. Abeliuk, E. Awad, S. Dsouza, J. Tenenbaum, and I. Rahwan, “A computational model of commonsense moral decision making,” *arXiv preprint arXiv:1801.04346*, 2018.

[9] F. Kamiran, I. Žliobaitė, and T. Calders, “Quantifying explainable discrimination and removing illegal discrimination in automated decision making,” *Knowledge and information systems*, vol. 35, no. 3, pp. 613–644, 2013.

[10] J. Broome, “Uncertainty and fairness,” *The Economic Journal*, vol. 94, no. 375, pp. 624–632, 1984.