Axioms Should Explain Solutions

Ariel D. Procaccia

Abstract In normative economics, axiomatic properties of mechanisms are often formulated by taking the viewpoint of the designer on what is desirable, rather than that of the participants. By contrast, I argue that, in real-world applications, the central role of axioms should be to help explain the mechanism's outcomes to participants. I specifically draw on my practical experience in two areas: fair division, which I view as a success story for the axiomatic approach; and voting, where this approach currently falls short.

The *axiomatic approach* advocates the design of mechanisms that provide rigorous guarantees, and, indeed, compares mechanisms based on the properties that they do, or do not, satisfy. This approach has long been a staple of economic theory, and, more recently, has guided the analysis of social computing systems (Tennenholtz and Zohar 2016).

The axioms in question typically attempt to formally capture notions of justice, fairness, efficiency, or just plain reasonableness. My view is that they are typically formulated with the *designer* in mind, that is, they are meant to convince an authority to adopt a specific mechanism. In this essay I wish to examine the idea that axioms should be formulated to help *explain* the mechanism's choices to *participants*. To demonstrate the appeal, and real-world implications, of this idea, I start by presenting a positive case study—fair division. I then discuss the practical shortcomings of the axiomatic approach to voting, and point to a way forward.

In fair division, an especially intuitive setting is that of rent division, where the goal is to fairly assign the n rooms of an apartment to n players, and divide the predetermined total rent between them. The preferences of the players are assumed to be *quasi-linear*, that is, the utility of a player for receiving a room that he values at x for the price of y is $x - y$.

From the axiomatic viewpoint, the primary question is what one means by "fairly". The gold standard of fairness axioms is (arguably) *envy-freeness*, which, in the rent division context, requires that the utility of each player for his room at its price be

A. D. Procaccia (✉)
Carnegie Mellon University, Pittsburgh, USA
e-mail: arielpro@cs.cmu.edu

J.-F. Laslier et al. (eds.), *The Future of Economic Design*, Studies in Economic Design, https://doi.org/10.1007/978-3-030-18050-8_27

at least as high as his utility for any other room at the price of that room. The true power of this axiom stems from the fact that an envy-free rent division always exists, under mild assumptions (Svensson 1983; Su 1999).

By contrast, when allocating indivisible goods *without* money, envy-freeness clearly cannot be guaranteed. For example, imagine a situation with two players and a single good: whichever player receives the good will be envied by the other. This observation underlies a significant body of work on fairness axioms that *can* be guaranteed in this setting, and the design of algorithms that achieve them. My current favorite is *envy-freeness up to one good*: one player may strictly prefer the bundle of another player to his own, but it is always possible to remove a single good from the bundle of the latter player to eliminate the former player's envy. Lipton et al. (2004) have shown that such an allocation always exists, even under general (monotonic) combinatorial valuations.

The foregoing axioms play a central role in the design of the not-for-profit fair division website *Spliddit* (Goldman and Procaccia 2014), which currently provides algorithms for the division of rent, goods, credit, tasks, and fare. The website's tagline—and, indeed, its central principle—is *provably fair solutions*, that is, each of the algorithms deployed on Spliddit guarantees axiomatic notions of fairness.

This brings us to the main point of this essay: The reason I view these provable fairness guarantees as crucial is that they allow us to explain to users why a given solution is fair. Indeed, the website contains an accessible description of the relevant axioms on the page devoted to each application. More importantly, when displaying a solution to a specific problem instance, users are also shown a *personalized* explanation of why their solution satisfies the promised fairness guarantees.

For example, for rent division, Spliddit uses a polynomial-time implementation of the *maximin solution* (Alkan et al. 1991; Gal et al. 2016), which is envy free, and might display the following text:

> Why is my assignment envy free? You were assigned the room called 'Master Bedroom' for $290.00. Since you valued the room at $535.00, you gained $245.00. You valued the room called 'Basement' at $64.00. Since this room costs $554.00, you would have lost $490.00. You valued the room called '2nd Floor' at $401.00. Since this room costs $156.00, you would have also gained $245.00.

Similarly, for goods division, Spliddit employs a highly optimized implementation of the Maximum Nash Welfare solution, which guarantees an allocation that is envy free up to one good (Caragiannis et al. 2016), and displays a personalized fairness explanation to that effect.

I firmly believe that these fairness explanations make the solutions more appealing to users, and, therefore, make it more likely that users will accept them. And, while I only have anecdotal evidence to support this belief, the fact that Spliddit has had more than 124,000 users since its launch in November 2014 (as of April 2018) suggests that the approach is quite effective.

Putting my artificial intelligence (AI) hat on for a moment, the notion of *explainable AI* is all the rage now, driven by the popularity of machine learning and the opacity of most machine learning algorithms. I am often asked why machine learning cannot be used to find solutions that people perceive as fair, thereby replacing the

axiomatic approach; my answer is twofold. First, there is insufficient data. Second, more pertinently, explainability is crucial, and, even if we had enough data, current machine learning algorithms would not be able to explain why a solution is fair.

Now for the bad news (although the next few paragraphs just set the stage for it). Since the publication of the book by Arrow (1951), social choice theory—which deals with aggregating individual preferences or opinions towards collective decisions—has relied heavily on the axiomatic approach. In the context of voting, specifically, several impossibility results are perhaps the most famous, but axioms are also routinely used to make the case for particular voting rules.

More concretely, the standard model of voting involves a set of voters, whose preferences over a set of alternatives are expressed as *rankings* of the alternatives. A *voting rule* (also known as a *social choice function*) takes as input a *preference profile* (the reported rankings), and returns the winning alternative. Axioms, in this context, are desirable properties of voting rules. Here are a few representative examples:

1. *Condorcet consistency:* We say that alternative x beats y in a pairwise comparison if a majority of voters rank x above y. An alternative is a *Condorcet winner* if it beats every other alternative in a pairwise comparison. A voting rule is *Condorcet consistent* if it selects a Condorcet winner whenever one exists in a given preference profile.
2. *Monotonicity:* If x is the winning alternative under a certain preference profile, and another profile is obtained by pushing x upwards in the votes while keeping the order of all other alternatives fixed, then x is also the winner under the new profile.
3. *Consistency*: If alternative x is the winner under two different preference profiles, x also wins under the union of the two profiles.

In terms of explainability, there is a significant difference between these axioms and, say, envy-freeness in rent division. Monotonicity and consistency tie together multiple preference profiles; explaining the outcome on any particular profile is difficult, because this would require voters to reason counterfactually (thinking about what the outcome would be if the profile was different).[1] And Condorcet consistency is very effective in explaining the outcome on a given profile—but only when a Condorcet winner exists. Consequently, while axioms do play a central role in presenting the voting systems that are deployed on several voting websites—including Pnyx, Whale, and CIVS—they are not used to explain outcomes, to the best of my knowledge.[2]

[1] Cailloux and Endriss (2016) develop an algorithm that automatically derives a justification for any outcome of the Borda rule, by applying both 'single-profile' and 'multi-profile' axioms to a sequence of hypothetical profiles. Their approach nicely formalizes the idea of explaining an outcome, but, in its current form, may not produce explanations that people would be able to follow.

[2] Two caveats are in order. First, the website Whale does visualize outcomes. For example, for Condorcet-based methods, the website displays the pairwise majority graph. These visualizations are useful insofar as they explain how the voting rule works, but, in my view, they do not explain its outcomes. Second, for the case of *multi-winner* elections, there are some examples of axioms that directly give rise to explainable outcomes. Notably, Aziz et al. (2017) recently developed the notion of *justified representation* for approval-based multi-winner elections, which, roughly speaking,

The shortcomings of the axiomatic approach (or, at least, the perception thereof) have motivated my work, and the work of others, on *optimization-based* approaches to voting. To be clear, implementations of rules derived from the axiomatic approach also frequently rely heavily on optimization, but here I am referring to voting rules that choose an alternative that optimizes an objective function. At the risk of sounding cryptic, let me mention that, in the context of the aggregation of subjective preferences, I am especially interested in optimizing utilitarian social welfare with respect to implicit utility functions (Boutilier 2015; Caragiannis et al. 2017). And when aggregating objective opinions, we have looked at optimizing the worst-case or expected position of the selected alternative in a set of feasible ground truth rankings (Procaccia et al. 2016; Benade et al. 2017).

These optimization-based rules are deployed on *RoboVote*, a not-for-profit voting website, which we launched in November 2016. RoboVote attempts to communicate to users the reasoning behind its calculations, but, in my view, its explanations are not nearly as compelling as their counterparts on Spliddit. For example:

> Based on the medium level of disagreement in the reported votes, we estimate that voters ordered at most 2.4 pairs of alternatives incorrectly, on average. Given that, the winning alternative returned by Robovote is guaranteed to compare favorably with all but at most 1 other alternative according to the objectively correct order. No other alternative can give a better guarantee.

In fact, some RoboVote users—especially those who have used the website to make relatively high-stakes group decisions, such as selecting papers for awards at a prestigious conference—have reached out to us and asked for additional explanation. In some cases, we were lucky that the alternative selected by RoboVote happened to be a Condorcet winner. In other cases, where a Condorcet winner did not exist, we could point to an agreement between RoboVote and several simple voting rules. But no general methodology emerged from this exercise. In summary, optimization-based approaches to voting currently do *not* give rise to satisfying explanations.

Through deployment via online platforms, I believe that research in voting has the potential to change the way group decisions are made worldwide—starting from (many) small groups, and, in a utopian world, ultimately shaping global institutions. But I view the problem of explaining outcomes as a significant barrier to (at least partially) realizing this ideal. The case study of fair division suggests that axiomatically explaining solutions is an effective approach; it remains an open question whether an equivalent axiomatic framework can be developed for voting.

Acknowledgements I thank Felix Brandt, Umberto Grandi, Domink Peters, Marcus Pivato, Nisarg Shah, and Bill Zwicker for insightful feedback. This work was partially supported by NSF grants IIS-1350598, IIS-1714140, CCF-1525932, and CCF-1733556; by ONR grants N00014-16-1-3075 and N00014-17-1-2428; as well as a Sloan Research Fellowship and a Guggenheim Fellowship.

---

requires that if a sufficiently large group of voters approve the same alternative, then the winning subset must contain at least one alternative approved by some member of the group. This at least allows addressing complaints by large groups that are not represented in the outcome, by arguing that the group members themselves cannot even agree on a single alternative.

References

Alkan, A., Demange, G., & Gale, D. (1991). Fair allocation of indivisible goods and criteria of justice. *Econometrica*, *59*(4), 1023–1039.

Arrow, K. (1951). *Social choice and individual values*. Hoboken: Wiley.

Aziz, H., Brill, M., Elkind, E., Freeman, R., & Walsh, T. (2017). Justified representation in approval-based committee voting. *Social Choice and Welfare*, *42*(2), 461–485.

Benade, G., Kahng, A., & Procaccia, A. D. (2017). Making right decisions based on wrong opinions. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC)* (pp. 267–284).

Boutilier, C., Caragiannis, I., Haber, S., Lu, T., Procaccia, A. D., & Sheffet, O. (2015). Optimal social choice functions: A utilitarian view. *Artificial Intelligence*, *227*, 190–213.

Cailloux, O., & Endriss, U. (2016). Arguing about voting rules. In *Proceedings of the 15th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)* (pp. 287–295).

Caragiannis, I., Kurokawa, D., Moulin, H., Procaccia, A. D., Shah, N., & Wang, J. (2016). The unreasonable fairness of maximum Nash welfare. In *Proceedings of the 17th ACM Conference on Economics and Computation (EC)* (pp. 305–322).

Caragiannis, I., Nath, S., Procaccia, A. D., & Shah, N. (2017). Subset selection via implicit utilitarian voting. *Journal of Artificial Intelligence Research*, *58*, 123–152.

Gal, Y., Mash, M., Procaccia, A. D., & Zick, Y. (2016). Which is the fairest (rent division) of them all? In *Proceedings of the 17th ACM Conference on Economics and Computation (EC)* (pp. 67–84).

Goldman, J., & Procaccia, A. D. (2014). Spliddit: Unleashing fair division algorithms. *SIGecom Exchanges*, *13*(2), 41–46.

Lipton, R. J., Markakis, E., Mossel, E., & Saberi, A. (2004). On approximately fair allocations of indivisible goods. In *Proceedings of the 6th ACM Conference on Economics and Computation (EC)* (pp. 125–131).

Procaccia, A. D., Shah, N., & Zick, Y. (2016). Voting rules as error-correcting codes. *Artificial Intelligence*, *231*, 1–16.

Su, F. E. (1999). Rental harmony: Sperner's lemma in fair division. *American Mathematical Monthly*, *106*(10), 930–942.

Svensson, L.-G. (1983). Large indivisibles: An analysis with respect to price equilibrium and fairness. *Econometrica*, *51*(4), 939–954.

Tennenholtz, M., & Zohar, A. (2016). The axiomatic approach and the Internet. In F. Brandt, V. Conitzer, U. Endress, J. Lang, & A. D. Procaccia (Eds.), *Handbook of computational social choice*, Chap. 18. Cambridge: Cambridge University Press.