Explainability in Social Choice

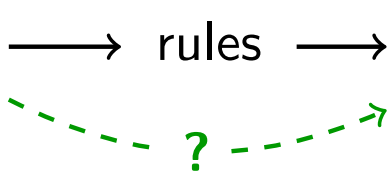
Ulle Endriss
Institute for Logic, Language and Computation
University of Amsterdam

Workshop on Computational and Topological Social Choice
16th Meeting of the Society for Social Choice and Welfare

Explainability in Social Choice

How do you explain why a given collective decision is the right one?

The axiomatic method of social choice theory seems relevant, given that axioms can justify voting rules, which produce decisions. But:



Plan for this talk:

- Recap: some basic voting theory
- Idea: justification = normative basis + step-by-step explanation
- Realisation: algorithmic thinking required to make it work

Three Voting Rules

Suppose several *voters* need to choose from a set of m *alternatives* by stating their preferences as *strict linear orders* over these alternatives.

Here are three *voting rules:*

- *Plurality:* elect the alternative ranked first most often
- *Plurality with runoff:* run a plurality election and retain the two front-runners; then run a majority contest between them
- *Borda:* each voter gives $m-1$ points to her top alternative, $m-2$ points to the alternative she ranks second, etc.

Example: Choosing a Beverage for Lunch

Consider this election, with nine *voters* having to choose from three *alternatives* (namely what beverage to order for a common lunch):

2 Germans	Beer	$\succ$	Wine	$\succ$ Milk
3 French people	Wine	$\succ$	Beer	$\succ$ Milk
4 Dutch people	Milk	$\succ$	Beer	$\succ$ Wine

Exercise: *Which beverage wins the election for*

- *the plurality rule?*
- *plurality with runoff?*
- *the Borda rule?*

The Model

Fix a finite set $X = \{a, b, c, \ldots\}$ of *alternatives*, with $|X| = m \geqslant 2$.
Let $\mathcal{L}(X)$ be the set of linear orders on X. Used to model *preferences*.

When the members of a *subelectorate* N drawn from an *electorate* $N^\star$ of voters express their preferences, we obtain a *profile* $R : N \to \mathcal{L}(X)$.

Given such a profile R, we need to decide on an *outcome* by choosing one or more "best" alternatives (a nonempty subset of X).

A *voting rule* is a function that returns an outcome for every profile:

$$F : \mathcal{L}(X)^{N \subseteq N^\star} \to 2^X \setminus \{\emptyset\}$$

Thus, voting rules here are *irresolute* and would have to be paired with a *tie-breaking rule* if we require single winners for all profiles.

Axioms

What's a good voting rule? One approach is to look for rules that satisfy *axioms* (basic normative principles) we care about. Examples:

- *Anonymity:* Treat all voters the same.
- *Neutrality:* Treat all alternatives the same.
- *Responsiveness:* If a voter raises a (possibly tied) winner $x^\star$ in her own ballot, then $x^\star$ should become the sole winner.
- *Reinforcement:* If $\mathrm{Dom}(R) \cap \mathrm{Dom}(R') = \emptyset$ and $F(R) \cap F(R') \neq \emptyset$, then we should have $F(R \oplus R') = F(R) \cap F(R')$.
- *Pareto:* There should be no alternative that every voter strictly prefers to an alternative selected by the voting rule.
- *Condorcet:* If some alternative is preferred to every other alternative by a majority of voters, it should be the sole winner.
- *Cancellation:* If all majority contests tie, everyone should win.

Characterisation Results

Important seminal results in SCT allow us to characterise certain voting rules in terms of the axioms they satisfy. Examples:

Theorem 1 (May, 1952) *For $m = 2$, a rule satisfies anonymity, neutrality, and responsiveness iff it is the simple majority rule.*

Theorem 2 (Young, 1975) *A rule satisfies neutrality, reinforcement, faithfulness, and cancellation iff it is the Borda rule.*

Note: The axiom of *faithfulness* says that when there is just one voter that voter's top alternative should be the sole winner.

Significance: Such results provide justifications for using certain rules. Arguing for one axiom at a time is much easier than arguing for a rule.

K.O. May. A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decisions. *Econometrica*, 1952.

H.P. Young. Social Choice Scoring Functions. *SIAM Journal Appl. Math.*, 1975.

Discussion

Can the axiomatic method help us *explain* / *justify* why some outcome might be *the right outcome* for a given profile? Maybe:

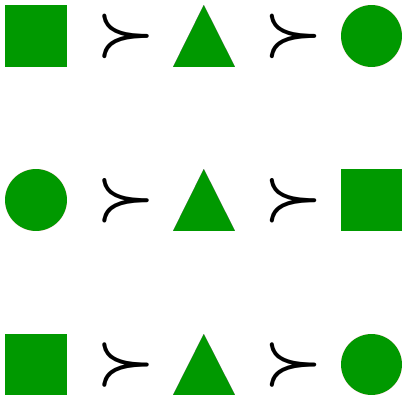
- suppose for profile $R^\star$ we want to justify the choice of outcome $X^\star$
- suppose $F(R^\star) = X^\star$ for some voting rule F
- suppose F is characterised by the set of axioms $\mathcal{A}$
- suppose we consider the axioms in $\mathcal{A}$ to be normatively appealing
- then we might say that we have an argument for choosing $X^\star$ in $R^\star$

But there are a number of problems here:

- *few characterisation results*, some with *unattractive axioms*
- some appealing axioms also feature in *impossibility results*
- we hardly can expect our audience to *understand* the results used
- overkill: we just care about $R^\star$, *not all profiles*

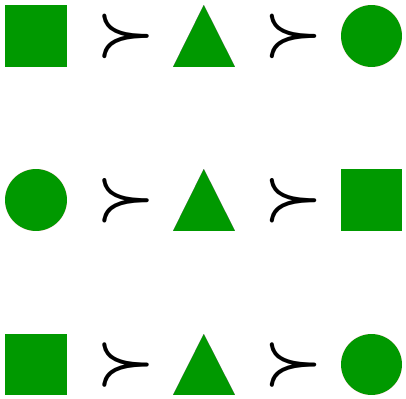
Can we instead justify outcomes by appealing to axioms directly?

Example



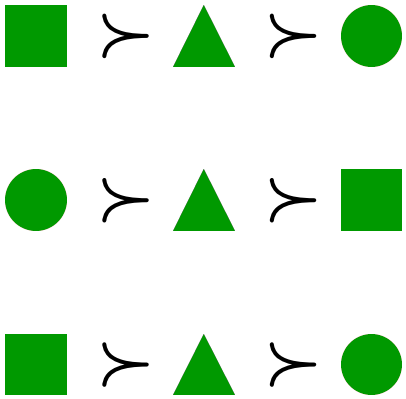
Exercise: *Can you think of a voting rule that makes* ■ *win?*

Example



Exercise: *Can you think of a voting rule that makes ▲ win?*

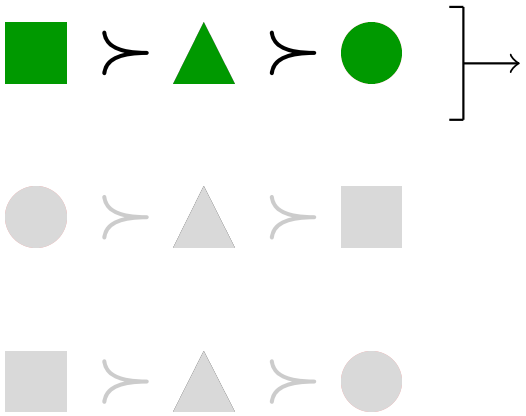
Example



What's a good outcome?

Why?

Example

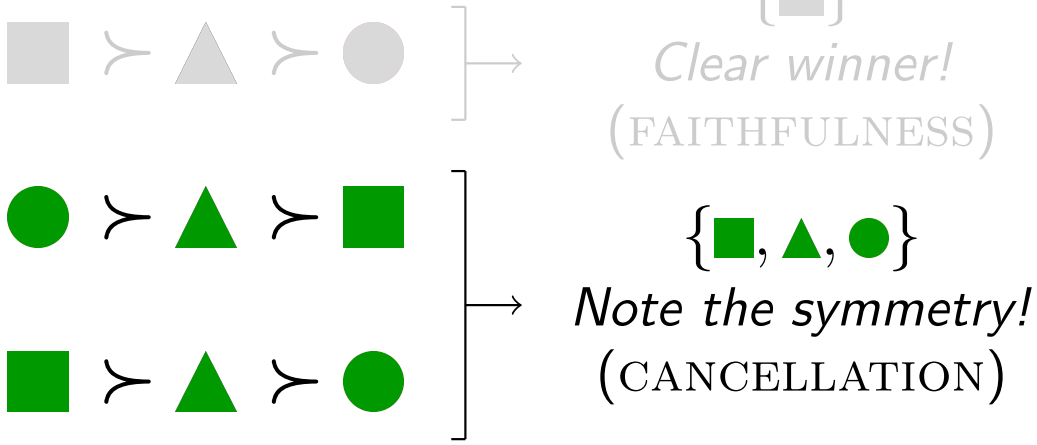


$\{\blacksquare\}$

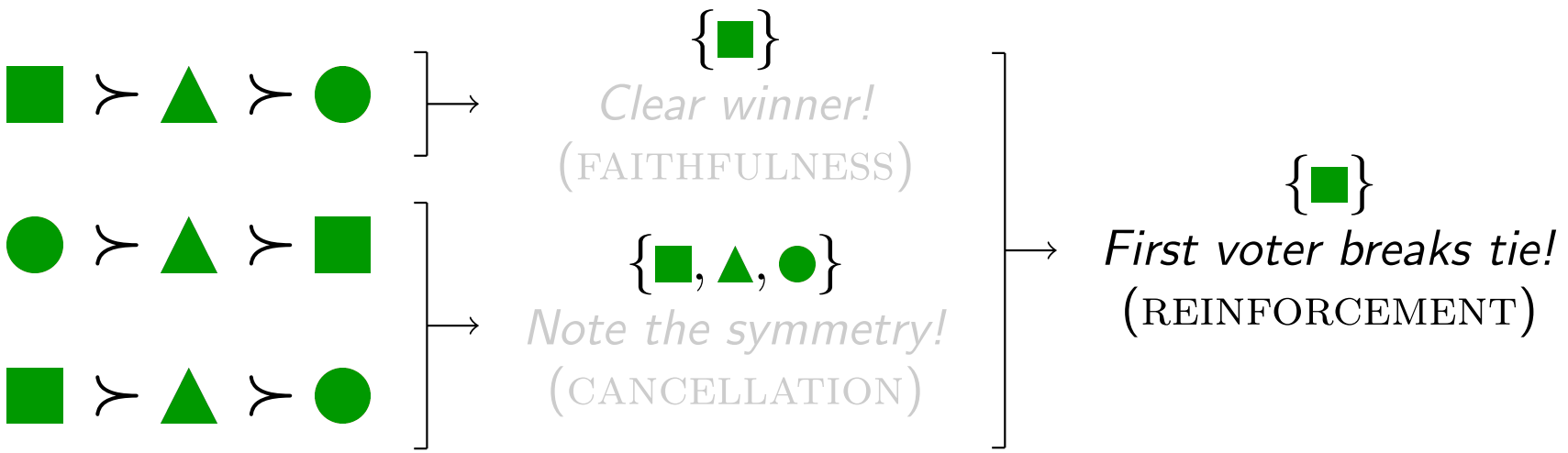
Clear winner!

(FAITHFULNESS)

Example


{■}
Clear winner!
(FAITHFULNESS)
{■, ▲, ●}
Note the symmetry!
(CANCELLATION)

Example


{■}
Clear winner!
(FAITHFULNESS)
{■, ▲, ●}
Note the symmetry!
(CANCELLATION)
{■}
First voter breaks tie!
(REINFORCEMENT)

Axioms: Interpretation and Instances

The *interpretation* of an axiom A is just a set of voting rules:

$$\mathbb{I}(A) \quad \subseteq \quad \mathcal{L}(X)^{N \subseteq N^\star} \to 2^X \setminus \{\emptyset\}$$

Example: $\mathbb{I}(\text{NEU}) = \{\, \text{BORDA}, \text{COPELAND}, \ldots, F_{4711}, \ldots \}$

An *instance* A' of axiom A (for a specific profile, etc.) is what you think it is, and itself an axiom, with $\mathbb{I}(A) = \bigcap_{A' \in \text{Inst}(A)} \mathbb{I}(A')$.

Example: $\text{Inst}(\text{PAR}) = \{$"*don't elect* c *in* $(abc^{[2]}, bca^{[5]})$*!*", $\ldots\}$

Proposal for a Definition

How can you justify outcome $X^\star$ given profile $R^\star$ (with electorate $N^\star$) using as arguments only axioms from a (large!) corpus $\mathbb{A}$? Slogan:

$$\textit{Justification} = \textit{Normative Basis} + \textit{Explanation}$$

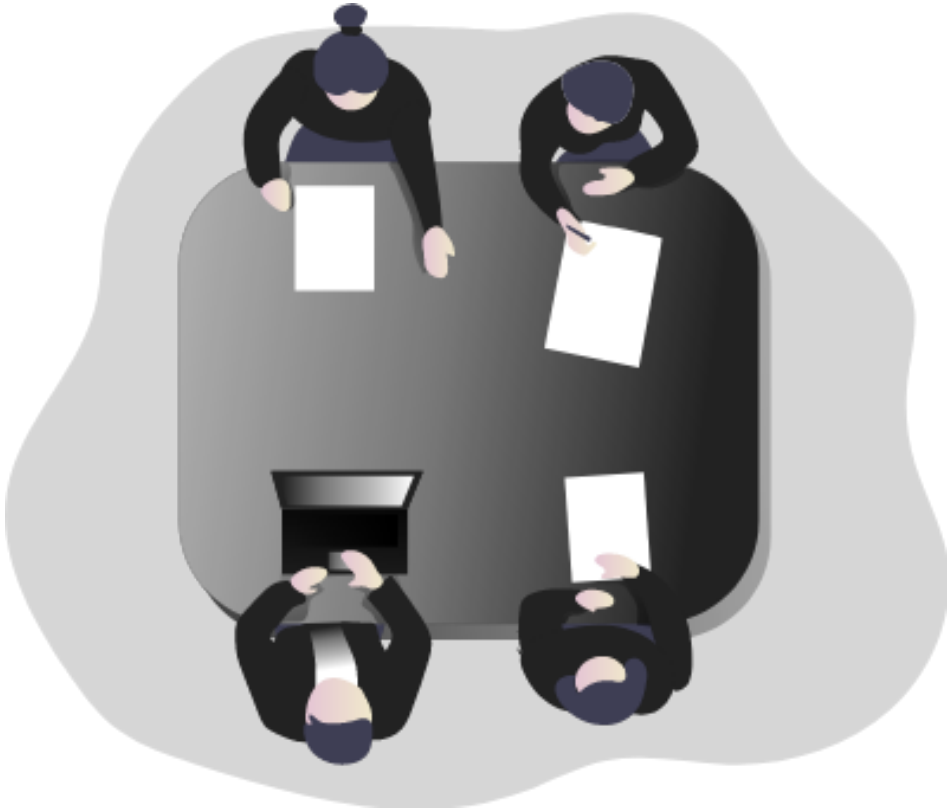
A pair $\langle \mathcal{A}^{\text{NB}}, \mathcal{A}^{\text{EX}} \rangle$ of sets of axioms is a *justification* if it satisfies:

- *Adequacy:* $\mathcal{A}^{\text{NB}} \subseteq \mathbb{A}$
- *Relevance:* $\mathcal{A}^{\text{EX}}$ is a set of instances of the axioms in $\mathcal{A}^{\text{NB}}$
- *Explanatoriness:* $F(R^\star) = X^\star$ for all rules $F \in \bigcap_{A' \in \mathcal{A}^{\text{EX}}} \mathbb{I}(A')$ and this is not the case for any proper subset of $\mathcal{A}^{\text{EX}}$
- *Nontriviality:* $\bigcap_{A \in \mathcal{A}^{\text{NB}}} \mathbb{I}(A) \neq \emptyset$ (*some* rule satisfies all axioms)

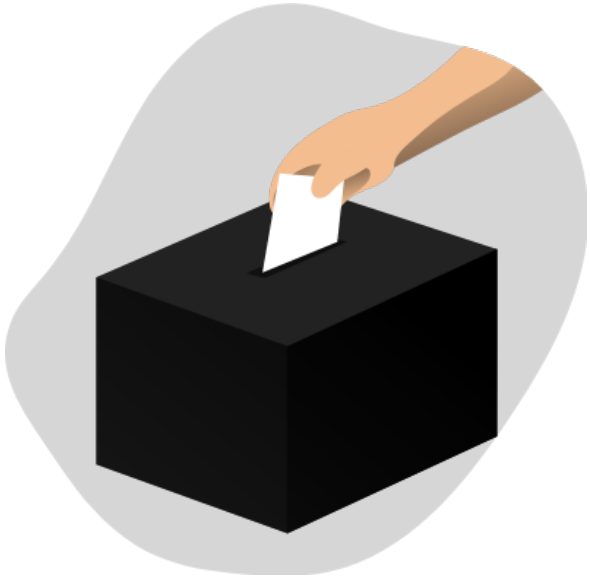
A. Boixel and U. Endriss. Automated Justification of Collective Decisions via Constraint Solving. AAMAS-2020.

Scenario 1: Confidence in Election Results

Scenario 2: Deliberation Support



Scenario 3: Justification Generation as Voting



Exercise: *What is the name of this well-known voting rule?*

$$F_{\{\mathrm{CON}\} \gg \{\mathrm{NEU}, \mathrm{REI}, \mathrm{FAI}, \mathrm{CAN}\}}$$

Remark: Schmidtlein (2022) has developed this application in depth.

M.C. Schmidtlein. Voting by Axioms. MSc Logic thesis, ILLC, University of Amsterdam, 2022.

Computing Justifications

We can encode axiom instances in propositional logic with variables of the form $p_{x,R}$ to say alternative x is amongst the winners in profile R.

Example: $\bigwedge_{y \in X} \bigwedge_{x \in X \setminus \{y\}} \bigwedge_{R \text{ s.t. } \forall i.(x,y) \in R(i)} \neg p_{y,R}$

Encode *instances* of axioms in $\mathbb{A}$ and *goal constraint* $F(R^\star) \neq X^\star$.
Then use a *SAT solver* to check whether this set is *satisfiable:*

- If *yes*, no justification exists.
- If *no*, a justification $\langle \mathcal{A}^{\text{NB}}, \mathcal{A}^{\text{EX}} \rangle$ exists if these steps succeed:
  - Find MUS (*minimal unsatisfiable subset*) including goal constraint.
    Let $\mathcal{A}^{\text{EX}}$ be MUS $\setminus$ {goal constraint}.
  - Let $\mathcal{A}^{\text{NB}}$ be the set of axioms in $\mathbb{A}$ with instances in $\mathcal{A}^{\text{EX}}$.
    Check that $\mathcal{A}^{\text{NB}}$ is *satisfiable* (for nontriviality).

Highly complex! But intractable tasks map to *well-studied problems* in automated reasoning. Challenge: generate only *relevant* instances.

O. Nardi, A. Boixel, and U. Endriss. A Graph-Based Algorithm for the Automated Justification of Collective Decisions. AAMAS-2022.

Aside: SAT Solving for Social Choice

This methodology of using SAT solvers to reason about social choice has been very successful also elsewhere in the field, particularly for finding new impossibility theorems. *Highly recommended!*

Consult the references below for tutorial-style material on this approach.

C. Geist and D. Peters. Computer-Aided Methods for Social Choice Theory. In U. Endriss (ed.), *Trends in Computational Social Choice*. AI Access, 2017.

U. Endriss. Slide set for "Advanced Topics in Computational Social Choice". ILLC, University of Amsterdam, 2021. Available at `http://bit.ly/adv-comsoc-21`.

Structured Explanations

For now, an *explanation* is a minimal set of axiom instances that forces the outcome we want. But *how* it does so is not (yet) captured.

Ultimately, we want to get a *structured explanation* that encodes an easily understandable proof for this claim of $\mathcal{A}^{\mathrm{EX}}$ forcing $X^\star$.

We have developed a *tableaux-style calculus* to reason about voting rules that can be used to construct such structured explanations.

The calculus manipulates statements of the form $\langle R, \mathcal{O} \rangle$, where R is a profile and $\mathcal{O}$ is the range of outcomes still considered possible for R. We use axioms to narrow down these ranges until we find $\langle R^\star, \{X^\star\} \rangle$.

This representation is reasonably close a *natural-language explanation*.

A. Boixel, U. Endriss, and R. de Haan. A Calculus for Computing Structured Justifications for Election Outcomes. AAAI-2022.

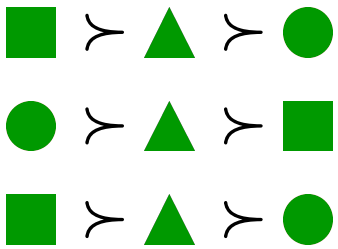
Demo

For small preference profiles, you can try it out for yourself:

http://bit.ly/xsoc-demo

Remark: For this demo the axiom of *anonymity* is always included, so we can express profiles more compactly (number of voters per ballot).

Exercise: *Generate a justification for our original example!*



A. Boixel, U. Endriss, and O. Nardi. Displaying Justifications for Collective Decisions. IJCAI-2022 (Demo Track).

Broader Considerations

First steps towards extending our approach from voting to the field of *matching* have recently been taken by Loustalot Knapp (2022).

Procaccia (2019) points out that in *fair division* axioms tend to more naturally lend themselves to explaining outcomes (e.g.: envy-freeness).

In earlier work with Olivier Cailloux we speculate about explanation and justification just being one aspect of providing (computer-enabled) support for people *arguing about voting rules*.

D. Loustalot Knapp. Justification of Matching Outcomes. MSc Logic thesis, ILLC, University of Amsterdam, 2022.

A.D. Procaccia. Axioms Should Explain Solutions. In J.-F. Laslier et al. (eds.), *The Future of Economic Design*. Springer, 2019.

O. Cailloux and U. Endriss. Arguing about Voting Rules. AAMAS-2016.

Last Slide

To approach the topic of *explainability in social choice*, I proposed a notion of *axiomatic justification* for election outcomes:

- Scenarios: Confidence Building | Deliberation Support | Voting
- Definition: Justification = Normative Basis + Explanation
- Algorithm: Graph Search + MUS Generation + SAT Solving
- Structured Explanations via Tableaux-style Calculus for Voting
- Opportunities: *lots of potential for follow-up research . . .*

demo: `http://bit.ly/xsoc-demo`
includes links to papers and outreach video