Explainability in Social Choice: Day 2

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

[Tutorial at *Formal Models of Democracy*, Rotterdam, April 2022
http://www.illc.uva.nl/~ulle/teaching/fmd-2022/]

Recap

The first part of this tutorial was a pretty classical introduction to the *axiomatic method* in *social choice theory*. We saw:

- Basic model of social choice functions (voting rules)
- Voting rules: plurality, Borda, STV, Copeland, cup rules, . . .
- Axioms: anonymity, Pareto, Condorcet, reinforcement, . . .
- Characterisations of voting rules in terms of axioms

Very nice stuff. Very useful stuff. But . . .

Explainability in Social Choice

Can the axiomatic method help us *explain* why a given outcome for a given profile of preferences might be *the right outcome*?

Yes, to a certain extent:

- let's say, given profile $\boldsymbol{R}^\star$, we want to explain the election of $X^\star$
- suppose $F(\boldsymbol{R}^\star) = X^\star$ 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 electing $X^\star$ in $\boldsymbol{R}^\star$

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 $\boldsymbol{R}^\star$, *not all profiles*

Exercise: *Any ideas for how to think about explainability instead?*

Example

■ ≻ ▲ ≻ ●

● ≻ ▲ ≻ ■

■ ≻ ▲ ≻ ●

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

Example

■ $\succ$ ▲ $\succ$ ●

● $\succ$ ▲ $\succ$ ■

■ $\succ$ ▲ $\succ$ ●

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

Example

■ $\succ$ ▲ $\succ$ ●

● $\succ$ ▲ $\succ$ ■

■ $\succ$ ▲ $\succ$ ●

What's a good outcome?
Why?

Example

■ ≻ ▲ ≻ ● ↦ {■}

Clear winner!

(FAITHFULNESS)

● ≻ ▲ ≻ ■

■ ≻ ▲ ≻ ●

Example

■ $\succ$ ▲ $\succ$ ● ↦ {■}

Clear winner!

(FAITHFULNESS)

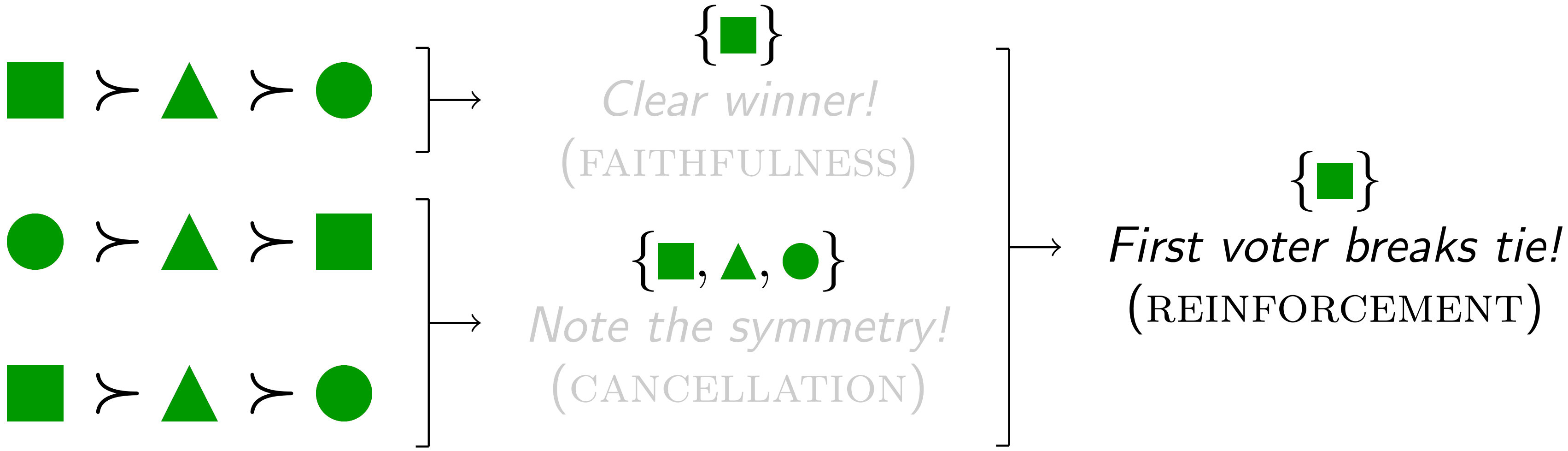
● $\succ$ ▲ $\succ$ ■

■ $\succ$ ▲ $\succ$ ● ↦ {■, ▲, ●}

Note the symmetry!

(CANCELLATION)

Example



■ ≻ ▲ ≻ ●

● ≻ ▲ ≻ ■

■ ≻ ▲ ≻ ●

{■}

Clear winner!

(FAITHFULNESS)

{■, ▲, ●}

Note the symmetry!

(CANCELLATION)

{■}

First voter breaks tie!

(REINFORCEMENT)

The Model

Remark: We initially used the letter A for the set of *alternatives*, but today using X instead will be more convenient (sorry!).

Remark: We need to generalise to a *variable-electorate model* to be able to formally deal with axioms such as reinforcement.

Suppose *voters* in $N^\star$ express *preferences* over *alternatives* in X. Consider *voting rules* defined on all *profiles* for subelectorates:

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

Axioms: Interpretation and Instances

Attractive rules might satisfy *axioms* such as *neutrality*, *Pareto*, ...

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: $\mathbb{I}(\textsc{neu}) = \{\,\textsc{borda}, \textsc{copeland}, \ldots, F_{4711}, \ldots\}$

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

Example: $\mathrm{Inst}(\textsc{par}) = \{$“*don't elect c in $(abc^{[2]}, bca^{[5]})$!*”, $\ldots\}$

Proposal for a Definition

How can you justify an election outcome $X^\star \subseteq X$ for a profile $\boldsymbol{R}^\star$ (with electorate $N^\star$) using axioms from a (large!) corpus $\mathbb{A}$?

Justification $=$ *Normative Basis* $+$ *Explanation*

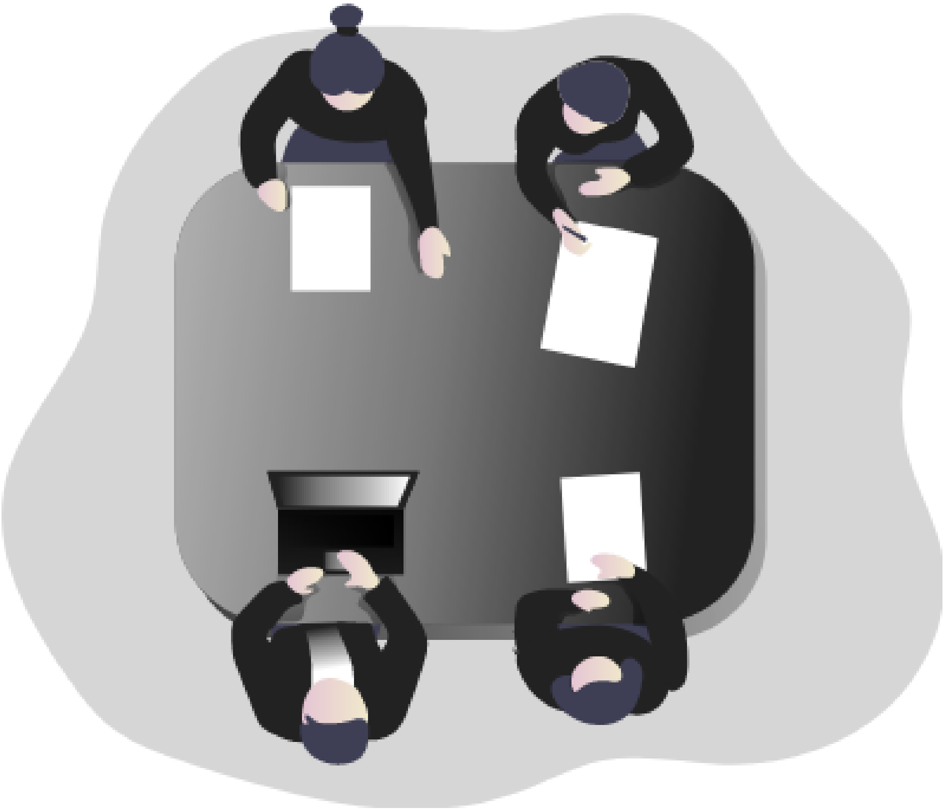
A pair $\langle \mathcal{A}^{\mathrm{NB}}, \mathcal{A}^{\mathrm{EX}} \rangle$ of sets of axioms is a *justification* if it satisfies:

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

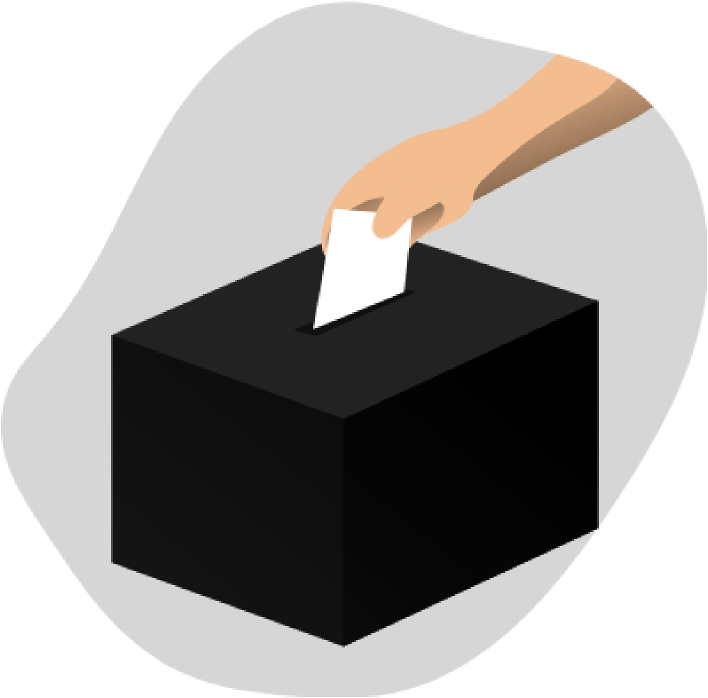
A. Boixel and U. Endriss. Automated Justification of Collective Decisions via Constraint Solving. AAMAS-2020.

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?*

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

Computing Justifications

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

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

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

Highly complex! But intractable tasks map to *well-studied problems* in automated reasoning. Must only generate *relevant* axiom 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 `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 $\langle \boldsymbol{R}, \mathcal{O} \rangle$, where $\boldsymbol{R}$ is a profile and $\mathcal{O}$ is the range of outcomes still considered possible for $\boldsymbol{R}$. We use axioms to narrow down these ranges until we find $\langle \boldsymbol{R}^\star, \{X^\star\} \rangle$.

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:

https://demo.illc.uva.nl/justify/

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!*

■ $\succ$ ▲ $\succ$ ●

● $\succ$ ▲ $\succ$ ■

■ $\succ$ ▲ $\succ$ ●

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

Need for Empirical Research: Good Explanations

What makes for a good/convincing/understandable explanation?

We don't really know (yet). This requires careful *empirical studies*, enabled by the tools we have developed so far.

Request: *When using the demo, please complete the feedback form!*

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.

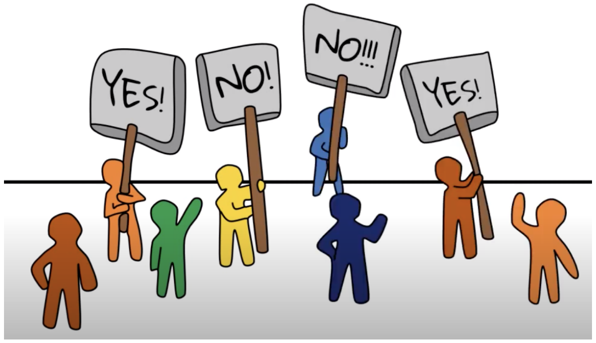
A.D. Procaccia. Axioms Should Explain Solutions. In J.-F. Laslier et al. (eds.), *The Future of Economic Design*. Springer, 2019.

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 . . .*



[`http://bit.ly/watch-our-movie`]