Algorithmic Explainability and Justifiability of Collective Decisions

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

[based on joint work with Arthur Boixel]

Preview

Given the *preferences* of several people and a *decision* they might take, we would like to *automatically* generate a *justification* for that decision.

Example: *How might we justify choosing A in the scenario below?*

$$\begin{aligned} \text{Agent 1:} \quad & A \succ_1 B \succ_1 C \\ \text{Agent 2:} \quad & B \succ_2 A \succ_2 C \\ \text{Agent 3:} \quad & A \succ_3 C \succ_3 B \end{aligned}$$

Two possible justifications (among many!):

- More than half of the agents say A is best ($\hookrightarrow$ *Majority Principle*).
- If only agents 1 and 2 vote, C is dominated ($\hookrightarrow$ *Pareto Principle*), so must declare tie between A and B ($\hookrightarrow$ *Anonymity* & *Neutrality*). Should let agent 3 break this tie, so select A ($\hookrightarrow$ *Reinforcement*).

Outline

I will argue for a concrete conception of the notion of "justification" and show how it can be realised in algorithmic terms:

- Definition
- Scenarios
- Automation

Most of this is based on unpublished joint work with Arthur Boixel. Some ideas go back to earlier work with Olivier Cailloux.

O. Cailloux and U. Endriss. Arguing about Voting Rules. Proc. 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2016).

The Model

Agents in N^* express *preferences* in $\mathcal{L}(X)$ over set of *alternatives* X.
Consider *voting rules* F defined on *profiles* for subelectorates $N \subseteq N^*$:

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

Attractive rules might satisfy *axioms* such as these:

- *Pareto Principle:* $y \notin F(\succ_N)$ if $\{i : x \succ_i y\} = N$ for some $x \in X$
- *Reinforcement**: $F(\succ_{N \uplus N'}) = F(\succ_N) \cap F(\succ_{N'})$ unless empty
- *Cancellation:* $F(\succ_N) = X$ if $|\{i : x \succ_i y\}| = \frac{|N|}{2}$ for all $x, y \in X$

Formally, the *interpretation* of an axiom A is just a set of voting rules:

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

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

Justification = Normative Basis + Explanation

Given a corpus of axioms $\mathbb{A}$, a profile $\succ_{N^*}$ in which the agents in N^* rank alternatives in X, and a subset $X^* \subseteq X$, we call a pair of sets of axioms $\langle NB, EX \rangle$, consisting of a *normative basis* and an *explanation*, a *justification* of outcome X^* under profile $\succ_{N^*}$ from corpus $\mathbb{A}$, if:

- *NB* is *adequate*: $NB \subseteq \mathbb{A}$.
- *EX* is *relevant*: *EX* is a set of instances of the axioms in *NB*
- *EX* is *explanatory*: $F(\succ_{N^*}) = X^*$ for all rules $F \in \bigcap_{A' \in EX} \mathbb{I}(A')$ and this is not the case for any proper subset of *EX*
- *NB* is *nontrivial*: $\bigcap_{A \in NB} \mathbb{I}(A) \neq \emptyset$ (*some* rule satisfies all axioms)

Scenario 1: Building Confidence in Election Results

You run the election of the new president for your professional society. The statutes of the society prescribe the use of voting rule F.

Most members were not directly involved with the decision to use F. Some even view this development with some suspicion.

You want to convince members the election outcome is "the right one". You could publish the ballots in anonymised form (for re-calculation).

But you could do more: publish a justification of the outcome in terms of axioms people might find convincing (not necessarily F's axioms!).

Scenario 2: Deliberation Support

You and your colleagues deliberate over what would be the best policy to adopt for your department. You hope to reach unanimity eventually.

Every now and then you conduct a straw poll to see what the most promising proposals are at that point.

After each poll you check which proposals can be justified from some *large* corpus of axioms, given the current preferences of people.

You then use these justifications (and their absence) as a basis of discussion for narrowing down the range of proposals. Repeat.

Scenario 3: Justification Generation as Voting

You do not manage to find a voting rule that satisfies all the axioms you care about. There are just too many impossibility results.

You could do this: *Rank* all *normative bases* you can possibly think of. Pick the outcome justifiable by the most preferred normative basis.

Result below ensures that the *voting rule induced* by any such ranking of bases is well defined (except that it might return $\emptyset$):

> **Theorem:** *It is impossible to justify two different outcomes for a given profile from the same normative basis.*

If at least one basis *NB* fully characterises a rule F, in the sense of $\bigcap_{A \in NB} \mathbb{I}(A) = \{F\}$, then induced rule always returns non-empty set.

Justification Problems as Constraint Networks

One *variable* for every possible profile involving (some) agents in N^*, taking *values* from $2^X \setminus \{\emptyset\}$. Axioms / axiom instances as *constraints*.

Example: For $X = \{A, B, C\}$ and $N^* = \{1, 2\}$ there are 48 profiles.

$\frac{1}{A}$	$\frac{1}{A}$	$\cdots$	$\frac{1}{C}$	$\frac{2}{A}$	$\frac{2}{A}$	$\cdots$	$\frac{2}{C}$	$\frac{1\ 2}{A\ A}$	$\frac{1\ 2}{A\ A}$	$\cdots$	$\frac{1\ 2}{C\ C}$
B	C	$\cdots$	B	B	C	$\cdots$	B	$B\ B$	$B\ C$	$\cdots$	$B\ B$
C	B		A	C	B		A	$C\ C$	$C\ B$		$A\ A$
V_1	V_2	$\cdots$	V_6	V_7	V_8	$\cdots$	V_{12}	V_{13}	V_{14}	$\cdots$	V_{48}

One of the instances of the *reinforcement* axiom:

$$(V_1 \cap V_8 \neq \emptyset) \rightarrow (V_{14} = V_1 \cap V_8)$$

Automated Search for Justifications

To generate justifications for X^* in profile $\succ_{N^*}$ from axiom corpus $\mathbb{A}$, prepare constraints for instances of axioms in $\mathbb{A}$ plus $F(\succ_{N^*}) \neq X^*$.

Then check whether the resulting constraint network is *satisfiable*.

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

Highly complex! But all computationally intractable tasks directly map to well-studied standard problems in constraint programming.

Example

Suppose you want to justify *only* A winning using only axioms in $\{\text{FAITHFULNESS}, \text{CANCELLATION}, \text{REINFORCEMENT}^\star\}$ for this profile:

$$\begin{aligned} &\text{Agent 1:} \quad A \succ_1 B \succ_1 C \\ &\text{Agent 2:} \quad A \succ_2 B \succ_2 C \\ &\text{Agent 3:} \quad C \succ_3 B \succ_3 A \end{aligned}$$

Suppose $p : \mathcal{L}(X)^+ \to \mathbb{N}$ maps profiles for $N \subseteq \{1, 2, 3\}$ to unique IDs.

Generating a justification amounts to finding an MUS such as this one:

$$\begin{aligned} &(\text{FAI}) \quad V_{p(\succ_1)} = \{A\} \\ &(\text{CAN}) \quad V_{p(\succ_2,\succ_3)} = \{A, B, C\} \\ &(\text{REI}) \quad \left[V_{p(\succ_1)} \cap V_{p(\succ_2,\succ_3)} \neq \emptyset\right] \to \left[V_{p(\succ_1,\succ_2,\succ_3)} = V_{p(\succ_1)} \cap V_{p(\succ_2,\succ_3)}\right] \\ &(\text{GC}) \quad V_{p(\succ_1,\succ_2,\succ_3)} \neq \{A\} \end{aligned}$$

$$EX = \{\text{three axiom instances}\} \qquad NB = \{\text{three axioms}\}$$

Nontriviality holds because some rules satisfy all three axioms.

Last Slide

I've presented an approach to automating the justification of collective decisions grounded in social choice theory and constraint programming:

- Justification = Normative Basis (*axioms*) + Explanation (*instances*)
- Justification Generation = MUS Generation + SAT
- Scenarios: Confidence Building | Deliberation Support | Voting