Computational Social Choice: Spring 2017

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

Plan for Today

References to "logic" in classical social choice theory are mostly about the axiomatic method, which is logic-like in spirit but doesn't make use of a formal language with an associated semantics and proof theory.

Today's lecture is about *logic for social choice:* embedding parts of the theory of social choice into a logical system.

We first review various arguments for *why this is useful* and then see three concrete approaches that use different kinds of logic to model the Arrovian framework of preference aggregation:

- an approach based on a specifically designed *modal logic*
- an approach using *classical first-order logic*
- an approach using *classical propositional logic*

U. Endriss. Logic and Social Choice Theory. In A. Gupta and J. van Benthem (eds.), *Logic and Philosophy Today*, College Publications, 2011.

Why?

Roughly speaking, for a given logic, *models* of that logic will encode *aggregation rules*, while *formulas* will encode their *properties*.

Why is this useful?

- Insight: formalisation to gain a *deeper understanding* of SCT.
- "Formal Minimalism": when considering an axiom in SCT, besides its *normative appeal* and its *mathematical strength*, we should also consider the *expressivity* of the language used to define it.
- Verification: formalisation can serve as a first step towards *automated verification*, both of *theoretical results* and of the correctness of *implementations* (i.e., of software).

Modelling the Arrovian Framework

Recall the Arrovian framework of *social welfare functions*, for a finite set N of individuals and an arbitrary set X of alternatives:

A SWF is a function $F : \mathcal{L}(X)^n \to \mathcal{L}(X)$ mapping any given profile of preference orders (i.e., linear orders) to a collective preference order.

F is defined on all profiles in $\mathcal{L}(X)^n$ (*universal domain* assumption).

Arrow suggested the following axioms (desirable properties of F):

- *Pareto:* if all individuals rank $x \succ y$, then so does society
- *IIA:* whether society ranks $x \succ y$ depends only on who ranks $x \succ y$
- *Nondictatorship:* F does not just copy the $\succ$ of a fixed individual

Arrow's Theorem establishes that no SWF F satisfies all three axioms, if there are $\geqslant 3$ alternatives. This holds for any finite set of individuals.

Can we express these things in a suitable logic?

Approach 1: Modal Logic

One approach to take is to develop a *new logic* specifically aimed at modelling the aspect of social choice theory we are interested in.

Modal logic looks like a useful technical framework for doing this.

It is intuitively clear that we can (somehow) devise a modal logic that can capture the Arrovian framework of SWF's, but how to do it exactly is less clear and finding a good way of doing this is a real challenge.

Adopting a semantics-guided approach, we first have to decide:

- what do we take to be our possible worlds?, and
- what accessibility relation(s) should we define?

Next, we shall review a specific proposal due to Ågotnes et al. (2011).

T. Ågotnes, W. van der Hoek, and M. Wooldridge. On the Logic of Preference and Judgment Aggregation. *J. Auton. Agents Multiagent Sys.*, 22(1):4–30, 2011.

Frames

Given: fixed (and finite) N (n individuals) and X (m alternatives)

Each *possible world* consists of

- a profile $\boldsymbol{R}$ and
- an ordered pair (x, y) of alternatives.

There are two *accessibility relations* defined on the possible worlds:

- Two worlds are related via relation $\textsc{prof}$ if their associated pairs are identical (i.e., only their profiles differ, if anything).
- Two worlds are related via relation $\textsc{pair}$ if their associated profiles are identical (i.e., only their pairs differ, if anything).

A *frame* $\langle \mathcal{L}(X)^n \times X^2, \textsc{prof}, \textsc{pair} \rangle$ consists of the set of worlds and the two accessibility relations (all induced by N and X).

Language

The language of the logic has the following *atomic* propositions:

- p_i for every individual $i \in N$
  Intuition: p_i is true at world $\langle \boldsymbol{R}, (x,y) \rangle$ if $x \succ y$ according to R_i
- $q_{(x,y)}$ for every pair of alternatives $(x,y) \in X^2$
  Intuition: $q_{(x',y')}$ is true at world $\langle \boldsymbol{R}, (x,y) \rangle$ if $(x,y) = (x',y')$
- a special proposition σ
  Intuition: σ is true at world $\langle \boldsymbol{R}, (x,y) \rangle$ if society ranks $x \succ y$

The set of *formulas* φ is defined as follows:

$$\varphi \;\; ::= \;\; p_i \mid q_{(x,y)} \mid \sigma \mid \neg\varphi \mid \varphi \wedge \varphi \mid [\text{PROF}]\varphi \mid [\text{PAIR}]\varphi$$

Disjunction, implication, and diamond-modalities are defined in the usual manner (e.g., $\langle \text{PROF} \rangle \varphi \equiv \neg[\text{PROF}]\neg\varphi$).

Semantics

In modal logic, a *valuation* determines which atomic propositions are true in which world, and a frame and a valuation together define a *model*.

For this logic, the valuation of p_i and $q_{(x,y)}$ is fixed and the valuation of σ will be defined in terms of a SWF F.

So, for given and fixed N and X (and thus for a fixed frame), we now define *truth* of a formula φ at a world $\langle \boldsymbol{R}, (x,y) \rangle$ w.r.t. a SWF F:

- $F, \langle \boldsymbol{R}, (x,y) \rangle \models p_i$ iff $(x,y) \in R_i$
- $F, \langle \boldsymbol{R}, (x,y) \rangle \models q_{(x',y')}$ iff $(x,y) = (x',y')$
- $F, \langle \boldsymbol{R}, (x,y) \rangle \models \sigma$ iff $(x,y) \in F(\boldsymbol{R})$
- $F, \langle \boldsymbol{R}, (x,y) \rangle \models \neg\varphi$ iff $F, \langle \boldsymbol{R}, (x,y) \rangle \not\models \varphi$
- $F, \langle \boldsymbol{R}, (x,y) \rangle \models \varphi \wedge \psi$ iff $F, \langle \boldsymbol{R}, (x,y) \rangle \models \varphi$ and $F, \langle \boldsymbol{R}, (x,y) \rangle \models \psi$
- $F, \langle \boldsymbol{R}, (x,y) \rangle \models [\textsc{prof}]\varphi$ iff $F, \langle \boldsymbol{R'}, (x,y) \rangle \models \varphi$ for all profiles $\boldsymbol{R'}$
- $F, \langle \boldsymbol{R}, (x,y) \rangle \models [\textsc{pair}]\varphi$ iff $F, \langle \boldsymbol{R}, (x',y') \rangle \models \varphi$ for all pairs (x',y')

That is, the operator $[\textsc{prof}]$ is a standard box-modality w.r.t. the relation $\textsc{prof}$ and $[\textsc{pair}]$ is a standard box-modality w.r.t. the relation $\textsc{pair}$.

Decidability

Formula φ is *satisfiable* if there are an F and a world w s.t. $F, w \models \varphi$.

The logic discussed here is *decidable*, i.e., there exists an effective algorithm that will decide whether a given formula is satisfiable:

- First, recall that *the frame is fixed:* to even write down a formula, we need to fix the language, which means fixing N and X.
- Second, observe that the number of possible SWF's is (huge but) *bounded:* there are exactly $m!^{(m!^n)}$ possibilities.
- Third, observe that *model checking is decidable:* there is an effective algorithm for deciding $F, w \models \varphi$ for given F, w, φ.
- Thus, for a given φ we can "just" try model checking for every possible SWF F and every possible world w.

Of course, this is not a practical algorithm. Ågotnes et al. consider complexity questions in more depth and also provide an axiomatisation.

Modelling: The Pareto Condition

We can model the *Pareto condition* as follows:

$$\textsc{Pareto} \quad := \quad [\textsc{Prof}][\textsc{Pair}](p_1 \wedge \cdots \wedge p_n \rightarrow \sigma)$$

That is, in every world $\langle \boldsymbol{R}, (x, y)\rangle$ it must be the case that, whenever all individuals rank $x \succ y$ (i.e., all p_i are true), then also society will rank $x \succ y$ (i.e., σ is true).

Write $F \models \varphi$ if $F, w \models \varphi$ for all worlds w.

We have: $F \models \textsc{Pareto}$ iff F satisfies the Pareto condition.

Remark: The nesting $[\textsc{Prof}][\textsc{Pair}]$ amounts to a *universal modality* (you can reach every possible world).

Modelling: Independence of Irrelevant Alternatives

Notation: For any coalition $C \subseteq N$, define p_C as

$$p_C \quad := \quad \bigwedge_{i \in C} p_i \;\wedge \bigwedge_{i \in N \setminus C} \neg p_i.$$

We can now express *IIA:*

$$\textsc{iia} \quad := \quad [\text{PROF}][\text{PAIR}] \bigwedge_{C \subseteq N} (p_C \wedge \sigma \rightarrow [\text{PROF}](p_C \rightarrow \sigma))$$

That is, in every world $\langle \boldsymbol{R}, (x,y) \rangle$ it must be the case that, if exactly the individuals in the group C rank $x \succ y$ (i.e., p_C is true) and society also ranks $x \succ y$ (i.e., σ is true), then for any other profile $\boldsymbol{R'}$ under which still exactly those in C rank $x \succ y$ society also must rank $x \succ y$.

We have $F \models \textsc{iia}$ iff F satisfies IIA.

Modelling: Dictatorships

Finally, we can model what it means for F to be *dictatorial:*

$$\textsc{Dictatorial} \quad := \quad \bigvee_{i \in N} [\textsc{Prof}][\textsc{Pair}](p_i \leftrightarrow \sigma)$$

That is, there exists an individual i (the dictator) such that it is the case that, to whichever world $\langle \boldsymbol{R}, (x, y) \rangle$ we move, society will rank $x \succ y$ (i.e., σ will be true) if and only if i ranks $x \succ y$ (i.e., p_i is true).

We have $F \models \neg\textsc{Dictatorial}$ iff F is nondictatorial.

Modelling Arrow's Theorem

Write $\models \varphi$ if $F \models \varphi$ for all SWF's F (for the fixed sets N and X).

We are now ready to state *Arrow's Theorem:*

$$\text{If } |X| \geqslant 3, \text{ then } \models \neg(\textsc{pareto} \wedge \textsc{iia} \wedge \neg\textsc{dictatorial}).$$

Note that this does *not* mean that we have a proof within this logic, although the completeness result of Ågotnes et al. regarding their axiomatisation means that such a proof is feasible in principle.

In recent work, we have been able to sketch such a proof for Arrow's Theorem for SCF's using a similar logic (Ciná and Endriss, 2016).

Remark: To be precise, the above is only a statement of Arrow's Theorem for a fixed (but arbitrary) choice of N and X.

G. Ciná and U. Endriss. Proving Classical Theorems of Social Choice Theory in Modal Logic. *J. Auton. Agents and Multiagent Systems*, 30(5):963–989, 2016.

Approach 2: First-Order Logic

Instead of designing a new logic specifically for our needs, we may ask whether what we want can be expressed in a given standard logic.

Next, we explore to what extent classical *first-order logic* can be used to model the Arrovian framework of social welfare functions.

Initial considerations:

- FOL is a natural logic to speak about *binary relations*, such as those used to model preference orders.
- Some aspects of the Arrovian framework (e.g., IIA speaking about *all* profiles with particular properties) seem to have a certain *higher-order feel* to them, which *could* be a problem.
- FOL cannot express *finiteness*, which *will* be a problem.

For details on the approach presented next, see the paper cited below.

U. Grandi and U. Endriss. First-Order Logic Formalisation of Impossibility Theorems in Preference Aggregation. *J. Phil. Log.*, 42(4):595-618, 2013.

Language

A key idea is to not talk about profiles (with their internal structure) directly, but to instead introduce the notion of *situation*.

Introduce these *predicate symbols* (with their intuitive meaning):

- $N(z)$: z is an individual
- $X(x)$: x is an alternative
- $S(u)$: u is a situation (referring to a profile)
- $p(z, x, y, u)$: individual z ranks x above y in situation/profile u
- $w(x, y, u)$: society ranks x above y in situation/profile u

Modelling: Social Welfare Functions

We can now write axioms forcing the intended interpretations, e.g.:

- Individual and collective preferences need to be *linear orders*. For instance, p must be interpreted as a *transitive* relation:

$$\forall z.\forall x_1.\forall x_2.\forall x_3.\forall u.\,[\,N(z) \wedge X(x_1) \wedge X(x_2) \wedge X(x_3) \wedge S(u) \;\rightarrow (p(z,x_1,x_2,u) \wedge p(z,x_2,x_3,u) \rightarrow p(z,x_1,x_3,u))\,]$$

- The predicates N, X and S must *partition* the domain. That is, any object must belong to exactly one of them:

$$\forall x.[N(x) \vee X(x) \vee S(x)] \;\wedge\; \forall x.[N(x) \rightarrow \neg X(x) \wedge \neg S(x)] \;\wedge\; \cdots$$

Together with a few other simple axioms like this, we can ensure that any model satisfying them must correspond to a SWF (see paper).

The only critical issue is to ensure that models are not too small: we need to ensure that the *universal domain* assumption gets respected.

Modelling: Universal Domain Assumption

The universal domain assumption can be modelled, but it's not pretty:

$$\begin{aligned}
&\forall z.\forall x.\forall y.\forall u.\,[p(z,x,y,u) \rightarrow \exists v.\,[S(v) \wedge p(z,y,x,v) \;\wedge \\
&\quad \forall x_1.[p(z,x,x_1,u) \wedge p(z,x_1,y,u) \rightarrow p(z,x_1,x,v) \wedge p(z,y,x_1,v)] \;\wedge \\
&\quad \forall x_1.[(p(z,x_1,x,u) \rightarrow p(z,x_1,y,v)) \wedge (p(z,y,x_1,u) \rightarrow p(z,x,x_1,v))] \;\wedge \\
&\quad \forall x_1.\forall y_1.[x_1 \neq x \wedge x_1 \neq y \wedge y_1 \neq y \wedge y_1 \neq x \rightarrow \\
&\qquad\qquad\qquad\qquad (p(z,x_1,y_1,u) \leftrightarrow p(z,x_1,y_1,v))] \;\wedge \\
&\quad \forall z_1.\forall x_1.\forall y_1.\,[z_1 \neq z \rightarrow (p(z_1,x_1,y_1,u) \leftrightarrow p(z_1,x_1,y_1,v))] \quad]]
\end{aligned}$$

That is, if there exists a situation u in which individual z ranks x above y, then there must exist a situation v where z ranks y above x and everything else remains the same. Once we ensure the existence of at least one situation, this generates a universal domain.

Modelling: Arrow's Axioms

Modelling Arrow's axioms is fairly easy.

The Pareto condition:

$$S(u) \wedge X(x) \wedge X(y) \rightarrow [\forall z.(N(z) \rightarrow p(z,x,y,u)) \rightarrow w(x,y,u)]$$

Independence of irrelevant alternatives (IIA):

$$\begin{aligned} &S(u_1) \wedge S(u_2) \wedge X(x) \wedge X(y) \;\rightarrow \\ &\quad [\forall z.(N(z) \rightarrow (p(z,x,y,u_1) \leftrightarrow p(z,x,y,u_2))) \;\rightarrow \\ &\qquad (w(x,y,u_1) \leftrightarrow w(x,y,u_2))] \end{aligned}$$

Being nondictatorial:

$$\neg\exists z.\, N(z) \wedge \forall u.\forall x.\forall y.\, [S(u) \wedge X(x) \wedge X(y) \wedge p(z,x,y,u) \rightarrow w(x,y,u)]$$

<u>Note:</u> All free variables are understood to be universally quantified.

Modelling: Arrow's Theorem

Let T_{SWF} be the set of axioms defining the theory of SWF's (those shown here and those only given in the paper, including one that ensure that there are $\geqslant 3$ alternatives). Let T_{ARROW} be the union of T_{SWF} and the three axioms on the previous slide.

We are now ready to state Arrow's Theorem:

T_{ARROW} does not have a finite model.

A shortcoming of this approach is that we cannot reduce this to a statement about some formula being a theorem of FOL. Only if we are willing to fix the number n of individuals, then we can do this (easily).

Thus, for fixed n this approach, in principle, permits a proof of Arrow's Theorem in FOL; and given the availability of complete theorem provers for FOL such a proof can, in principle, be found automatically. However, to date no such proof has been realised in practice.

Approach 3: Propositional Logic

For the special case of $n = 2$ and $m = 3$ (or indeed any fixed sizes) we can rewrite the FOL representation in propositional logic:

- predicates $p(z, x, y, u)$ become atomic propositions $p_{z,x,y,u}$
- predicates $w(x, y, u)$ become atomic propositions $w_{x,y,u}$
- universal quantifications become conjunctions and existential quantifications become disjunctions

That is, we need $2 \cdot 3^2 \cdot (3!)^2 + 3^2 \cdot (3!)^2 = 972$ propositional variables.

Direct rewriting of all axioms into CNF leads to an exponential blowup, but clever rewriting using auxiliary variables leads to a formula with around 35,000 variables and 100,000 clauses (Tang and Lin, 2009).

P. Tang and F. Lin. Computer-aided Proofs of Arrow's and other Impossibility Theorems. *Artificial Intelligence*, 173(11):1041–1053, 2009.

Computer-aided Proof of Arrow's Theorem

Tang and Lin (2009) prove two inductive lemmas:

- If there exists an Arrovian SWF for n individuals and $m+1$ alternatives, then there exists one for n and m (if $n \geqslant 2$, $m \geqslant 3$).
- If there exists an Arrovian SWF for $n+1$ individuals and m alternatives, then there exists one for n and m (if $n \geqslant 2$, $m \geqslant 3$).

That is, Arrow's Theorem holds iff its "*base case*" for 2 individuals and 3 alternatives is true—which we've modelled in propositional logic.

Despite being huge, a modern *SAT solver* can verify the inconsistency of the set of clauses corresponding to $\textsc{Arrow}(2,3)$ in < 1 second!

Further development of this technique has led to *discovery* of new results, beyond verification (see, e.g., Brandt and Geist, 2016).

P. Tang and F. Lin. Computer-aided Proofs of Arrow's and other Impossibility Theorems. *Artificial Intelligence*, 173(11):1041–1053, 2009.

F. Brandt and C. Geist. Finding Strategyproof Social Choice Functions via SAT Solving. *Journal of Artificial Intelligence Research (JAIR)*, 55:565–602, 2016.

Summary

We have seen three approaches to *modelling* certain aspects of social choice (here, the classical Arrovian framework) *in logic*, providing different degrees of support for *automated reasoning:*

- modal logic (specifically designed for this job)
- first-order logic (for arbitrary numbers of individuals/alternatives)
- propositional logic (for small sets of individuals/alternatives)

We are left with (at least) these questions and challenges:

- don't fix the *set of individuals* (and alternatives) in the language
- model the *universal domain* assumption in an elegant manner
- better support *automated reasoning*

What next? Final lecture on voting: multiwinner rules + reflection.