Towards a Logic of Social Welfare

Thomas Ågotnes[1]
joint work with Wiebe van der Hoek[2] and Michael Wooldridge[2]

[1]Department of Computer Engineering
Bergen University College, Norway

[2]Department of Computer Science
University of Liverpool, UK

1^{st} International Workshop on Computational Social Choice (COMSOC 06)

Motivation: Formal Reasoning about Social Choice

Social Choice Theory: Concept	Social Choice Theory: Example	Formal SCT: Concept
Social welfare function (SWF)		Model *M*
Possible property of SWFs	Pareto optimality	Formula ϕ
Fundamental property	Transitivity	Axiom ϕ
Theorem	Arrow's theorem	Derivable formula $\vdash \phi$
Proof		Formal derivation from axioms

Preferences and Social Welfare

- A: set of alternatives
- Preference relations $L(A)$: total orders $R \subseteq A \times A$ (antisymm., trans., refl.). R^s denotes the irreflexive version.
- Preference profiles for n agents: $L(A)^n$
- Social Welfare Function (SWF):

$$F : L(A)^n \to L(A)$$

Expressing IIA

Independence of Irrelevant Alternatives (IIA)

$$\forall_{(R_1,\ldots,R_n)\in L(A)^n}\forall_{(S_1,\ldots,S_n)\in L(A)^n}\forall_{a\in A}\forall_{b\in A}$$
$$(\forall_{i\in\Sigma}(aR_ib \Leftrightarrow aS_ib)) \Rightarrow (aF(R_1,\ldots,R_n)b \Leftrightarrow aF(S_1,\ldots,S_n)b)$$

Which constructs would we need in a logical language, in order to be able to express, e.g., IIA? It seems that we need to be able to express (in a single formula):

- Quantification over alternatives
- Quantification over preference relations, i.e., over sets of alternatives
- Properties of preference relations for different agents
- Properties of different preference relations for the same agent
- Comparisons of different preference relations
- The preference relation resulting from applying a SWF to other preference relations

Expressing IIA

Independence of Irrelevant Alternatives (IIA)

$$\forall_{(R_1,\ldots,R_n)\in L(A)^n}\forall_{(S_1,\ldots,S_n)\in L(A)^n}\forall_{a\in A}\forall_{b\in A}$$
$$(\forall_{i\in\Sigma}(aR_ib \Leftrightarrow aS_ib)) \Rightarrow (aF(R_1,\ldots,R_n)b \Leftrightarrow aF(S_1,\ldots,S_n)b)$$

Which constructs would we need in a logical language, in order to be able to express, e.g., IIA? It seems that we need to be able to express (in a single formula):

- Quantification over alternatives
- Quantification over preference relations, i.e., over sets of alternatives
- Properties of preference relations for different agents
- Properties of different preference relations for the same agent
- Comparisons of different preference relations
- The preference relation resulting from applying a SWF to other preference relations

A Logic of SWFs

$$\phi ::= r \mid r_i \mid \neg\phi \mid \phi \wedge \phi \mid \Box\phi \mid \boxdot\phi$$

where $r \in \Pi$ (propositions) and $i \in \Sigma$ (agents). Define $\Diamond\phi \equiv \neg\Box\neg\phi$, $\diamonddot\phi \equiv \neg \boxdot \neg\phi$.

Satisfaction: let F be a SWF, $\delta : \Pi \to L(A)^n$ and $a, b \in A$:

$$\begin{array}{lcl}
(A, F, \delta, (a, b)) \models r_i & \Leftrightarrow & (a, b) \in \delta_i(r) \\
(A, F, \delta, (a, b)) \models r & \Leftrightarrow & (a, b) \in F(\delta(r)) \\
(A, F, \delta, (a, b)) \models \Box\phi & \Leftrightarrow & \forall_{\delta'}(A, F, \delta', (a, b)) \models \phi \\
(A, F, \delta, (a, b)) \models \boxdot\phi & \Leftrightarrow & (\forall_{(a' \neq b') \in A \times A}(A, F, \delta, (a', b')) \models \phi)
\end{array}$$

$(A, F) \models \phi$ iff $(A, F, \delta, (a, b))$ for all $\delta, (a, b)$, etc.

A Logic of SWFs

$$\phi ::= r \mid r_i \mid \neg\phi \mid \phi \wedge \phi \mid \Box\phi \mid \boxdot\phi$$

where $r \in \Pi$ (propositions) and $i \in \Sigma$ (agents). Define $\Diamond\phi \equiv \neg\Box\neg\phi$, $\diamonddot\phi \equiv \neg \boxdot \neg\phi$.

Satisfaction: let F be a SWF, $\delta : \Pi \to L(A)^n$ and $a, b \in A$:

$$\begin{array}{lcl}
(A, F, \delta, (a, b)) \models r_i & \Leftrightarrow & (a, b) \in \delta_i(r) \\
(A, F, \delta, (a, b)) \models r & \Leftrightarrow & (a, b) \in F(\delta(r)) \\
(A, F, \delta, (a, b)) \models \Box\phi & \Leftrightarrow & \forall_{\delta'} (A, F, \delta', (a, b)) \models \phi \\
(A, F, \delta, (a, b)) \models \boxdot\phi & \Leftrightarrow & (\forall_{(a' \neq b') \in A \times A} (A, F, \delta, (a', b')) \models \phi)
\end{array}$$

$(A, F) \models \phi$ iff $(A, F, \delta, (a, b))$ for all $\delta, (a, b)$, etc.

Pareto Optimality

Pareto Optimality (PO)

$$\forall_{(R_1,\ldots,R_n)\in L(A)^n}\forall_{a\in A}\forall_{b\in A}((\forall_{i\in\Sigma} aR_i^s b) \Rightarrow aF(R_1,\ldots,R_n)^s b)$$

$$PO = \Box \boxdot ((r_1 \wedge \cdots \wedge r_n) \rightarrow r)$$

Proposition

$(A, F) \models PO$ iff F is pareto optimal

Pareto Optimality

Pareto Optimality (PO)

$\forall_{(R_1,\ldots,R_n)\in L(A)^n}\forall_{a\in A}\forall_{b\in A}((\forall_{i\in\Sigma} aR_i^s b) \Rightarrow aF(R_1,\ldots,R_n)^s b)$

$$PO = \Box \boxdot ((r_1 \wedge \cdots \wedge r_n) \to r)$$

Proposition

$(A, F) \models PO$ iff F is pareto optimal

Non-Dictatorship

Non-Dictatorship (ND)

$$\neg \exists_{i \in \Sigma} \forall_{(R_1,\dots,R_n) \in L(A)^n} F(R_1, \dots, R_n) = R_i$$

$$ND = \bigwedge_{i \in \Sigma} \Diamond \langle \cdot \rangle \neg (r \leftrightarrow r_i)$$

Proposition

$(A, F) \models ND$ iff F does not have a dictator

Non-Dictatorship

Non-Dictatorship (ND)

$\neg\exists_{i\in\Sigma}\forall_{(R_1,\ldots,R_n)\in L(A)^n} F(R_1,\ldots,R_n) = R_i$

$$ND = \bigwedge_{i\in\Sigma} \Diamond \diamond\!\!\!\!\cdot\;\neg(r \leftrightarrow r_i)$$

Proposition

$(A, F) \models ND$ iff F does not have a dictator

Independence of Irrelevant Alternatives

Independence of Irrelevant Alternatives (IIA)

$\forall_{(R_1,\dots,R_n)\in L(A)^n} \forall_{(S_1,\dots,S_n)\in L(A)^n} \forall_{a\in A} \forall_{b\in A}$
$(\forall_{i\in\Sigma}(aR_ib \Leftrightarrow aS_ib)) \Rightarrow (aF(R_1,\dots,R_n)b \Leftrightarrow aF(S_1,\dots,S_n)b)$

$$IIA = \Box \boxdot ((\bigwedge_{i\in\Sigma}(r_i \leftrightarrow s_i)) \rightarrow (r \leftrightarrow s))$$

Proposition

$(A, F) \models IIA$ iff F has the IIA property

Independence of Irrelevant Alternatives

Independence of Irrelevant Alternatives (IIA)

$\forall_{(R_1,\ldots,R_n)\in L(A)^n}\forall_{(S_1,\ldots,S_n)\in L(A)^n}\forall_{a\in A}\forall_{b\in A}$
$(\forall_{i\in\Sigma}(aR_ib \Leftrightarrow aS_ib)) \Rightarrow (aF(R_1,\ldots,R_n)b \Leftrightarrow aF(S_1,\ldots,S_n)b)$

$$IIA = \Box \boxdot ((\bigwedge_{i\in\Sigma}(r_i \leftrightarrow s_i)) \rightarrow (r \leftrightarrow s))$$

Proposition

$(A, F) \models IIA$ iff F has the IIA property

Arrow's Theorem

$$MT2 = \Diamond\,(\diamonddot(r_1 \wedge s_1) \wedge \diamonddot(r_1 \wedge \neg s_1))$$

Proposition

$(A, F) \models MT2$ iff $|A| > 2$

Theorem (Arrow)

$$\models MT2 \rightarrow \neg(PO \wedge ND \wedge IIA)$$

Arrow's Theorem

$$MT2 = \Diamond(\diamonddot(r_1 \wedge s_1) \wedge \diamonddot(r_1 \wedge \neg s_1))$$

Proposition

$(A, F) \models MT2$ iff $|A| > 2$

Theorem (Arrow)

$$\models MT2 \rightarrow \neg(PO \wedge ND \wedge IIA)$$

Judgment Aggregation

- Underlying logic **L** with language $\mathcal{L}$
- Agenda $\mathcal{A} \subseteq \mathcal{L}$ (closed under single negation)
- Judgment sets $J(\mathcal{A}, \mathbf{L})$: consistent and complete $A_i \subseteq \mathcal{A}$
- Judgment Aggregation Rule (JAR) f:
  $f(A_1, \ldots, A_n) \in J(\mathcal{A}, \mathbf{L})$

Interpretation of our language in JARs: let $\mathcal{A}$ be an agenda, f be a JAR, $\delta : \Pi \to J(\mathcal{A}, \mathbf{L})^n$ and $p \in \mathcal{A}$:

$$
\begin{array}{lcl}
(\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} r_i & \Leftrightarrow & p \in \delta_i(r) \\
(\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} r & \Leftrightarrow & p \in f(\delta(r)) \\
(\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} \Box\phi & \Leftrightarrow & \forall_{\delta'} (\mathcal{A}, f, \delta', p) \models_{\mathbf{L}} \phi \\
(\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} \boxdot\phi & \Leftrightarrow & (\forall_{p \in \mathcal{A}} (\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} \phi)
\end{array}
$$

Judgment Aggregation

- Underlying logic $\mathbf{L}$ with language $\mathcal{L}$
- Agenda $\mathcal{A} \subseteq \mathcal{L}$ (closed under single negation)
- Judgment sets $J(\mathcal{A}, \mathbf{L})$: consistent and complete $A_i \subseteq \mathcal{A}$
- Judgment Aggregation Rule (JAR) f: $f(A_1, \ldots, A_n) \in J(\mathcal{A}, \mathbf{L})$

Interpretation of our language in JARs: let $\mathcal{A}$ be an agenda, f be a JAR, $\delta : \Pi \to J(\mathcal{A}, \mathbf{L})^n$ and $p \in \mathcal{A}$:

$$\begin{array}{lcl}
(\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} r_i & \Leftrightarrow & p \in \delta_i(r) \\
(\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} r & \Leftrightarrow & p \in f(\delta(r)) \\
(\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} \Box\phi & \Leftrightarrow & \forall_{\delta'} (\mathcal{A}, f, \delta', p) \models_{\mathbf{L}} \phi \\
(\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} \boxdot\phi & \Leftrightarrow & (\forall_{p \in \mathcal{A}} (\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} \phi)
\end{array}$$

Example

Majority voting on a proposition:

$$MV = r \leftrightarrow \bigvee_{G \subseteq \Sigma, |G| > \frac{n}{2}} \bigwedge_{i \in G} r_i$$

The Discursive Dilemma

$\models_{\mathbf{L}} \neg \Box \boxdot MV$

Example

Majority voting on a proposition:

$$MV = r \leftrightarrow \bigvee_{G \subseteq \Sigma, |G| > \frac{n}{2}} \bigwedge_{i \in G} r_i$$

The Discursive Dilemma

$\models_{\mathbf{L}} \neg \Box \boxdot MV$

In order to achieve completeness, we extend the language

Extend the language with an atom

$$\mathbf{h}_p$$

for each $p \in \mathcal{A}$

$$(\mathcal{A}, f, \delta, p) \models_{\mathbf{L}} \mathbf{h}_q \quad \Leftrightarrow \quad p = q$$

Axiomatisation

Given underlying logic **L**, the logic *JAL*(**L**) is:

$\neg(\mathbf{h}_p \wedge \mathbf{h}_q)$ if $p \neq q$	*Atmost*		
$\bigvee_{p \in \mathcal{A}} \mathbf{h}_p$	*Atleast*	all inst. of prop. taut.	Taut
$\diamond \mathbf{h}_p$	*Agenda*	$\blacksquare(\psi_1 \rightarrow \psi_2) \rightarrow (\blacksquare\psi_1 \rightarrow \blacksquare\psi_2)$	*K*
$\diamond(\mathbf{h}_p \wedge \varphi) \rightarrow \boxdot(\mathbf{h}_p \rightarrow \varphi)$	*Once*	$\blacksquare\psi \rightarrow \psi$	*T*
$\diamond(\mathbf{h}_p \wedge x) \vee \diamond(\mathbf{h}'_p \wedge x)$	*CpJS*	$\blacksquare\psi \rightarrow \blacksquare\blacksquare\psi$	4
$(\Diamond i \wedge \Diamond\neg j) \rightarrow \bigwedge_{o \in O} \Diamond o$	*C*	$\neg\blacksquare\psi \rightarrow \blacksquare\neg\blacksquare\psi$	5
$\Box \boxdot \psi \leftrightarrow \boxdot\Box\psi$	(*COMM*)		

From $p_1, \ldots p_n \vdash_{\mathbf{L}} q$ infer $\diamond(\mathbf{h}_{p_1} \wedge x) \wedge \cdots \wedge \diamond(\mathbf{h}_{p_n} \wedge x) \rightarrow \boxdot(\mathbf{h}_q \rightarrow x) \wedge \boxdot(\mathbf{h}'_q \rightarrow \neg x)$	*Closure*
From $\varphi \rightarrow \psi$ and φ infer ψ	*MP*
From ψ infer $\blacksquare\psi$	*Nec*

where $\blacksquare \in \{\Box, \boxdot\}$, $x \in \{r, r_i\}$, $O = \{x_1, \ldots, x_k : x_j = (\neg)r_j\}$

Theorem

JAL(**L**) is sound and complete wrt. JARs over finite agendas.

Axiomatisation

Given underlying logic **L**, the logic $JAL(\mathbf{L})$ is:

$\neg(\mathbf{h}_p \wedge \mathbf{h}_q)$ if $p \neq q$	*Atmost*		
$\bigvee_{p \in \mathcal{A}} \mathbf{h}_p$	*Atleast*	all inst. of prop. taut.	Taut
$\Diamonddot \mathbf{h}_p$	*Agenda*	$\blacksquare(\psi_1 \to \psi_2) \to (\blacksquare\psi_1 \to \blacksquare\psi_2)$	*K*
$\Diamonddot(\mathbf{h}_p \wedge \varphi) \to \boxdot(\mathbf{h}_p \to \varphi)$	*Once*	$\blacksquare\psi \to \psi$	*T*
$\Diamonddot(\mathbf{h}_p \wedge x) \vee \Diamonddot(\mathbf{h}'_p \wedge x)$	*CpJS*	$\blacksquare\psi \to \blacksquare\blacksquare\psi$	4
$(\Diamond i \wedge \Diamond\neg j) \to \bigwedge_{o \in O} \Diamond o$	*C*	$\neg\blacksquare\psi \to \blacksquare\neg\blacksquare\psi$	5
$\Box \boxdot \psi \leftrightarrow \boxdot\Box\psi$	(*COMM*)		

From $p_1, \ldots p_n \vdash_{\mathbf{L}} q$ infer $\Diamonddot(\mathbf{h}_{p_1} \wedge x) \wedge \cdots \wedge \Diamonddot(\mathbf{h}_{p_n} \wedge x) \to \boxdot(\mathbf{h}_q \to x) \wedge \boxdot(\mathbf{h}'_q \to \neg x)$	*Closure*
From $\varphi \to \psi$ and φ infer ψ	*MP*
From ψ infer $\blacksquare\psi$	*Nec*

where $\blacksquare \in \{\Box, \boxdot\}$, $x \in \{r, r_i\}$, $O = \{x_1, \ldots, x_k : x_j = (\neg) r_j\}$

Theorem

$JAL(\mathbf{L})$ is sound and complete wrt. JARs over finite agendas.

Preference vs. Judgment aggregation

Dietrich and List (2006):

- PA can be embedded in JA
- Given a set of alternatives A, we can define the underlying logic $\mathbf{L}^A$ such that preference relations correspond to judgment sets

Corollary

$JAL(\mathbf{L}^A)$ is a sound and complete axiomatisation of SWFs over finite finite sets of alternatives A.

Preference vs. Judgment aggregation

Dietrich and List (2006):

- PA can be embedded in JA
- Given a set of alternatives A, we can define the underlying logic $\mathbf{L}^A$ such that preference relations correspond to judgment sets

Corollary

$JAL(\mathbf{L}^A)$ is a sound and complete axiomatisation of SWFs over finite finite sets of alternatives A.

Summary

- Language interpreted in SWFs or JARs
- Syntactically simple, yet expressive - can, e.g., express
  - Rules such as majority voting
  - Properties such as Pareto Optimality
  - Results such as Arrow's theorem, the discursive paradox, Condorcet's paradox
- Sound and complete axiomatisation (finite alternatives/agenda)
- Sheds light on the logical principles of judgment- and preference aggregation
- Sheds light on the differences between the logical principles behind judgement- and preference aggregation