J Philos Logic (2013) 42:595–618
DOI 10.1007/s10992-012-9240-8

First-Order Logic Formalisation of Impossibility Theorems in Preference Aggregation

Umberto Grandi · Ulle Endriss

Received: 27 July 2011 / Accepted: 15 June 2012 / Published online: 25 July 2012

Abstract In preference aggregation a set of individuals express preferences over a set of alternatives, and these preferences have to be aggregated into a collective preference. When preferences are represented as orders, aggregation procedures are called social welfare functions. Classical results in social choice theory state that it is impossible to aggregate the preferences of a set of individuals under different natural sets of axiomatic conditions. We define a first-order language for social welfare functions and we give a complete axiomatisation for this class, without having the number of individuals or alternatives specified in the language. We are able to express classical axiomatic requirements in our first-order language, giving formal axioms for three classical theorems of preference aggregation by Arrow, by Sen, and by Kirman and Sondermann. We explore to what extent such theorems can be formally derived from our axiomatisations, obtaining positive results for Sen's Theorem and the Kirman-Sondermann Theorem. For the case of Arrow's Theorem, which does not apply in the case of infinite societies, we have to resort to fixing the number of individuals with an additional axiom. In the long run, we hope that our approach to formalisation can serve as the basis for a fully automated proof of classical and new theorems in social choice theory.

Keywords Social choice theory · First-order logic · Axiomatisability · Preference aggregation · Automated reasoning

U. Grandi (✉) · U. Endriss
Institute for Logic, Language and Computation, University of Amsterdam,
Postbus 94242, 1090 GE Amsterdam, The Netherlands
e-mail: umberto.uni@gmail.com

U. Endriss
e-mail: ulle.endriss@uva.nl

1 Introduction

Social choice theory is a branch of mathematical economics that is concerned with the design and analysis of methods for collective decision making [3, 14]. Classical results in the field explore the possibility of aggregation in different settings. Perhaps the most famous one is Arrow's Theorem [2]; it states that it is impossible to aggregate the preferences of a finite set of individuals in a manner that would satisfy a small number of natural properties. In recent years there has been a growing interest in applications of logic to social choice theory [11]. In this paper we present a formalisation of several results from social choice theory in classical first-order logic (FOL). We define a language that enables us to formalise classical properties of aggregation procedures, we give axioms for various settings of aggregation, and we explore to what extent certain classical impossibility theorems can be derived in this formal framework.

There have been a number of recent contributions that address the formalisation of theorems in social choice theory using a variety of logical frameworks [1, 21, 22, 27, 29, 31]. There are several reasons for this broad interest in applying tools from mathematical logic to social choice theory. The first of them is of course that the full formalisation of a problem domain can help us gain a deeper understanding of that domain. More specifically, in social choice theory, it can clarify the exact nature of the assumptions that are being made to derive, for instance, a characterisation result [22]. Second, a complete formalisation together with an automatically derived (or automatically verifiable) proof can give additional assurances for the correctness of a result. As pointed out by Blau [5], Arrow's original proof contained an error; this has been acknowledged and corrected in the second edition of Arrow's book [2]. While there has been some discussion in the literature whether the standard proofs have been worked out in sufficient detail [21], we certainly do not want to suggest that the major results in social choice theory are not based on sound foundations. However, for verifying newer and less well studied results, full formalisation and automated reasoning could prove useful tools.

Previous work has discussed formalisations of Arrow's Theorem in modal logic [1], dependence logic,[1] and in the language of set theory [21, 31]. Here we explore to what extent it is possible to model the framework of preference aggregation in classical FOL. There are several reasons for focusing on FOL: it is a natural language for speaking about linear orders, which are central to the modelling of preferences, and the body of literature and results that a first-order formalisation enable us to apply is bigger than for most other logical systems. An informal first-order language was also used already in the work of social choice pioneers like Arrow and Sen,[2] and it constitutes a well

[1] J. Väänänen (personal communication, 2009); see also [30].

[2] Arrow once took a course with Tarski [7].

estabilished language that can be easily understood and used in the field of mathematical economics.

In this work we concentrate on three theorems, proved by Arrow [2], Sen [25] and Kirman and Sondermann [18]. We are able to show that for two of these theorems it is possible to completely describe the problem within a language of FOL based on the language of orders. The exception is Arrow's Theorem: for stating that it only applies to the case of a *finite* number of individuals we have to instantiate the statement for a fixed number of individuals (we will see that Arrow's Theorem is equivalent to a certain theory of FOL axioms not having a finite model). In particular, we will not require any form of second-order quantification, which may seem surprising given that several of the axioms used, for instance, in Arrow's Theorem certainly have a "second-order flavour". Our axiomatisation draws on several ideas from an important recent paper by Tang and Lin [27]. In this work, the two authors provide an alternative proof of Arrow's Theorem composed of two inductive mathematical lemmas and an automated proof of the impossibility for the base case of 3 alternatives and 2 individuals. This last step is performed utilising a propositional language based on situation calculus, constructing computer-generated formulas that instantiate Arrow's conditions. In our paper, we generalise this language to a complete axiomatisation of the Arrovian framework of social welfare functions in classical FOL. We obtain a logical language that is both human-readable and easy to implement, and we study its expressive power formalising several classical impossibility theorems in the field of preference aggregation using this language.

The remainder of the paper is organised as follows. In Section 2 we recall the framework of preference aggregation, stating the three theorems that constitute the object of our analysis. In Section 3 we define a first-order language for preference aggregation and we present first-order axioms for social welfare functions and for several conditions introduced in the social choice literature. The models of these first-order theories are studied in detail in Section 4, where we prove several axiomatisability results for the theorems introduced in Section 2. Section 5 concludes and discusses some preliminary results concerning the use of automated reasoning tools in this setting.

2 Social Welfare Functions and Impossibility Theorems

In this section we review the basic definitions of preference aggregation and we state three famous theorems [2, 18, 25] that we aim at formalising using a first-order language. We also present a recent proof method based on induction, introduced by Tang and Lin [27] to prove Arrow's Theorem, and we prove a generalisation of one of their lemmas.

Let $\mathcal{N}$ be a set of *individuals* expressing preferences over a set $\mathcal{X}$ of *alternatives*. We represent such preferences with a binary relation. In this paper we concentrate on two representations of preferences: linear and weak orders. A binary relation is a *linear* order if it is irreflexive, transitive and

complete. The term aP_ib stands for "individual i strictly prefers alternative a to alternative b". The choice of a preference relation P_i for each individual constitutes a *preference profile* $\boldsymbol{P} = (P_1, \ldots, P_n)$. A *weak* order is a binary relation that is reflexive, transitive and complete. We will denote weak orders with the letter R, thus aR_ib will stand for "individual i weakly prefers a to b" and call $\boldsymbol{R} = (R_1, \ldots, R_n)$ a profile of weak orders. Note that every weak order R induces an irreflexive and transitive binary relation, usually referred to as the strict part of R, namely the relation that holds between a and b whenever aRb holds but bRa does not.

If we denote with $\mathcal{L}(\mathcal{X})$ the set of all linear orders on $\mathcal{X}$, then the set of all *profiles* of (linear) preference orders is the set $\mathcal{L}(\mathcal{X})^{\mathcal{N}}$. A *social welfare function* (SWF) for $\mathcal{X}$ and $\mathcal{N}$ is a function $w : \mathcal{L}(\mathcal{X})^{\mathcal{N}} \to \mathcal{L}(\mathcal{X})$. A SWF associates with every *preference profile* $\boldsymbol{P} = (P_1, \ldots, P_n) \in \mathcal{L}(\mathcal{X})^{\mathcal{N}}$ a linear order $w(\boldsymbol{P})$, which in most interpretations is taken to represent the aggregation of the preferences of the individuals into a "social preference order" over $\mathcal{X}$. The same definition can be given using the set $\mathcal{R}(\mathcal{X})$ of all weak orders over $\mathcal{X}$ as the domain of aggregation, defining a SWF for $\mathcal{N}$ and $\mathcal{X}$ as a function $w : \mathcal{R}(\mathcal{X})^{\mathcal{N}} \to \mathcal{R}(\mathcal{X})$.

In most of the paper we shall assume that preferences are represented as linear orders, and we give details in Section 3.3 for the generalisation of our first-order formalisation to the case of SWFs defined on weak orders.

2.1 Arrow's Theorem

Since the seminal work of Arrow [2], social choice theory has made extensive use of the axiomatic method to classify and study aggregation procedures. There are several properties that an aggregation mechanism may satisfy, and some of them have been argued to be natural requirements for a SWF. In this section we will concentrate on three properties that have led Arrow to prove his famous theorem, stated here for the case of linear orders:

- **UN**: A SWF w satisfies *unanimity* if, whenever every individual strictly prefers alternative a to alternative b, so does society. Formally, if aP_ib for every individual $i \in \mathcal{N}$, then $a\, w(\boldsymbol{P})\, b$.
- **IIA**: A SWF w satisfies *independence of irrelevant alternatives* if the social ranking of two alternatives a and b depends only on their relative ranking by the individuals. The formal condition is that, given two preference profiles $\boldsymbol{P}$ and $\boldsymbol{P}'$, if for every individual $i \in \mathcal{N}$ we have that aP_ib if and only if aP'_ib, then $a\, w(\boldsymbol{P})\, b$ if and only if $a\, w(\boldsymbol{P}')\, b$.
- **ND**: A SWF w is *non-dictatorial* if there is no individual $i \in \mathcal{N}$ such that for every profile $\boldsymbol{P}$ the social preference order $w(\boldsymbol{P})$ is equal to P_i.

It is important to note that in our definition of SWFs there are two hidden conditions that could be stated as axioms, but that we have instead included as an integral part of the formal framework of preference aggregation. The first is usually called *unrestricted* or *universal domain*: it requires a SWF to be defined over *all* preference profiles in $\mathcal{L}(\mathcal{X})^{\mathcal{N}}$. Domain restrictions, such

as *single-peaked* preferences [4], are the most common escape from Arrow's impossibility (see, e.g., [13]). The second hidden condition is called *collective rationality*, as it is first stated in Arrow [2, Chapter VIII, Section V]. It requires the outcome of the aggregation to be a linear order, i.e., it requires the outcome to conform to the same rationality constraints as the input received from the individuals. Non-comparability or ties between alternatives are not allowed for a SWF at the collective level.

We are now ready to state Arrow's celebrated theorem:

Theorem 1 [2] *If $\mathcal{X}$ and $\mathcal{N}$ are finite and non-empty, and if $|\mathcal{X}| \geq 3$, then there exists no SWF for $\mathcal{X}$ and $\mathcal{N}$ that satisfies **UN**, **IIA** and **ND**.*

Various proofs of this theorem are known [14, 15], and several formulations can be found in the literature that differ in view of the assumptions on individual and collective preferences that are being made [28], starting from Arrow's original version for weak orders [2]. In Section 2.4 we present one recent proof of this result that is based on induction.

2.2 Infinite Societies

Given our interest in a logical formalisation of impossibility results such as Arrow's Theorem, we have to question the assumption of finiteness in Theorem 1. There are two parameters in Arrow's Theorem: the number of alternatives and that of individuals. If we relax the assumption of finiteness for the set of alternatives $\mathcal{X}$, then the statement continues to hold.[3] If instead the set of individuals is allowed to be infinite, then the impossibility does not hold anymore: there exists a unanimous and independent SWF that is also non-dictatorial on infinite societies, as has first been pointed out by Fishburn [12]. Independent and unanimous SWFs on infinite domains can nevertheless be characterised, and a general form of Arrow's Theorem be proved without restrictions on the cardinality of $\mathcal{N}$. This result is due to Kirman and Sondermann [18], and we now briefly review their theorem.

Call a subset $J \subseteq \mathcal{N}$ of individuals a winning coalition, if for every profile $\boldsymbol{P}$ and for every pair of alternatives x, y, if xP_jy for every $j \in J$ then $x\, w(\boldsymbol{P})\, y$. A winning coalition can force the outcome of the SWF over x and y by voting unanimously over these two alternatives.

Theorem 2 [18] *If a SWF satisfies **UN** and **IIA**, then the corresponding collection $\mathcal{J}$ of winning coalitions is an ultrafilter over the set $\mathcal{N}$.*[4]

[3] This result seems to be a folk theorem. We will nevertheless give a new proof of this generalisation in Section 2.4.

[4] A collection of subsets $\mathcal{J}$ is an ultrafilter if it contains the full set, is closed under finite intersections, and is maximal in the follwing sense: for every subset J of $\mathcal{N}$, exactly one of J and its complement $J^c = \mathcal{N} \setminus J$ is in $\mathcal{J}$ [8].

Arrow's Theorem comes as a straightforward corollary, since every ultrafilter over a finite set is *principal*, i.e., it contains a singleton $\{i\}$ [8]. This means that a coalition J is a winning coalition if and only if it contains i, which is therefore a dictator.

2.3 Impossibility of a Paretian Liberal

The Arrovian framework of SWFs can be extended by adding individual "spheres of influence" as a model for individual rights. This model was first proposed by Sen [25], who proved an impossibility result known as the Impossibility of a Paretian Liberal. We follow here the presentation of this result by Gaertner [14], adapting Sen's Theorem to the case of linear orders.

A *rights system* is a collection of pairs of alternatives for which an individual has the power to influence the result of the SWF. Formally, let D be a function $D : \mathcal{N} \to \mathcal{P}(\mathcal{X} \times \mathcal{X})$, such that if $(x, y) \in D(i)$ and $x P_i y$ then also $x\, w(\boldsymbol{P})\, y$. Sen's framework for individual rights is composed of a SWF w and a rights system D. The conditions that he argues to be minimal natural requirements for such a framework are the following:

- **UN**: A SWF w satisfies *unanimity* if, whenever every individual strictly prefers alternative a to alternative b, so does society. Formally, if $a P_i b$ for every individual $i \in \mathcal{N}$, then $a\, w(\boldsymbol{P})\, b$ (same condition as the one in Section 2.1).
- **MINLIB**: w is *minimally liberal* with respect to D if there exist two individuals i_1 and i_2 such that they are decisive in both ways with respect to two alternatives each, i.e., there exist two individuals i_1 and i_2 and four (not necessarily distinct) alternatives $\{x_1, y_1, x_2, y_2\}$ such that both (x_1, y_1) and (y_1, x_1) are in $D(i_1)$ and both (x_2, y_2) and (y_2, x_2) are in $D(i_2)$.

The Impossibility of a Paretian Liberal is the following theorem:

Theorem 3 [25] *There is no SWF that satisfies* ***UN*** *and* ***MINLIB****.*

The proof of this result is a straightforward reduction to a minimal case, and we present here the main idea for the sake of comparison with the inductive proofs we will introduce in the next section.

Proof Suppose individual i_1 is decisive over the pair (x_1, y_1) and individual i_2 over (x_2, y_2), and we assume that these 4 alternatives are pairwise distinct (the other cases being easier to prove). Consider a profile $\boldsymbol{P}$, where $x_1 P_1 y_1$ and $x_2 P_2 y_2$, and for both $i = i_1$ and $i = i_2$ it is the case that $y_2 P_i x_1$ and $y_1 P_i x_2$. Then by **MINLIB** we have that $x_1 w(\boldsymbol{P}) y_1$ and $x_2 w(\boldsymbol{P}) y_2$, and by **UN** that $y_2 w(\boldsymbol{P}) x_1$ and $y_1 w(\boldsymbol{P}) x_2$. This constitutes a cycle of $w(\boldsymbol{P})$, contradicting our assumption that $w(\boldsymbol{P})$ is a linear order. □

2.4 Inductive Proofs

Recall that no assumption of finiteness was made to prove Sen's Theorem, and the method employed in this proof is a reduction from a general impossibility to the base case of two individuals and four (or less) alternatives. We will now present a similar method devised by Tang and Lin [27] to prove Arrow's Theorem. Tang and Lin [27] prove Theorem 1 by means of two inductive lemmas, reducing the general statement to the base case of 3 alternatives and 2 individuals, and then verify this last step automatically with a computer. The first lemma is the inductive step on the number of alternatives: "if there exists a SWF for $m+1$ alternatives and n individuals that satisfies Arrow's conditions, then there exists a SWF for m alternatives and the same number of individuals that still satisfies Arrow's conditions." The contrapositive of this lemma spreads the impossibility from the base case to every finite set of alternatives: "if Arrow's Theorem holds for the case of 3 alternatives and n individuals, then it holds for every finite set of m alternatives and n individuals". We first generalise this result to also cover the case of an *infinite* number of alternatives:

Lemma 1 *If there exists a SWF w for $\mathcal{X}$ and $\mathcal{N}$, with $|\mathcal{X}| \geq 3$, that satisfies **UN**, **IIA** and **ND**, then there exists a set $\mathcal{X}' \subseteq \mathcal{X}$ with $|\mathcal{X}'| = 3$ and a SWF for $\mathcal{X}'$ and $\mathcal{N}$ that satisfies the same properties.*

Note that the contrapositive of Lemma 1 reads: "if Arrow's Theorem holds for the case of 3 alternatives and n individuals, then it also holds for any larger set $\mathcal{X}$ (including the infinite case) and n individuals".

Proof Let $\mathcal{X}' = \{a_1, a_2, a_3\}$ be any set containing three different alternatives in $\mathcal{X}$. Every linear order P over $\mathcal{X}'$ can be extended to a linear order P^e over the whole set $\mathcal{X}$ (though not in a unique way). Define a SWF w' for $\mathcal{X}'$ and $\mathcal{N}$ in the following way:

$$x\, w'(\boldsymbol{P})\, y \;:\Leftrightarrow\; x\, w(\boldsymbol{P}^e)\, y$$

where $\boldsymbol{P}$ is a preference profile over $\mathcal{X}'$ and $\boldsymbol{P}^e$ any extension of P to a preference profile over $\mathcal{X}$. By **IIA** this definition does not depend on the extension chosen. Furthermore, w' remains unanimous and independent of irrelevant alternatives by definition. It remains to be shown that w' is non-dictatorial. Suppose the contrary: we prove that w would then be dictatorial too, in contradiction with the assumptions. Let i be the dictator for w', and x and y two different alternatives in $\mathcal{X}$, and suppose that $x P_i y$ in a certain profile $\boldsymbol{P}$. We now show that also $x\, w(\boldsymbol{P})\, y$ must hold, thus i is a dictator on every pair of alternatives in $\mathcal{X}$. The case where both x and y are in $\mathcal{X}'$ is trivial. We can therefore restrict ourselves to the case where there are at least two distinct elements in $\mathcal{X}'$ different from x and y, a_1 and a_2. Let individual i change

her preference relation such that $a_1 P_i a_2$, obtaining profile $\boldsymbol{P}'$. Let now every individual j (including i) rearrange her preference such that $x P_j a_1$ and $a_2 P_j y$, and call this profile $\boldsymbol{P}''$. Both steps can be done without affecting the initial ranking of x and y, thus by **IIA** $x\, w(\boldsymbol{P})\, y$ if and only if $x\, w(\boldsymbol{P}'')\, y$. By unanimity of w we have $x\, w(\boldsymbol{P}'')\, a_1$ and $a_2\, w(\boldsymbol{P}'')\, y$. Since i is a dictator relative to $\mathcal{X}'$, it must be the case that $a_1\, w(\boldsymbol{P}'')\, a_2$ holds, and thus by transitivity also $x\, w(\boldsymbol{P}'')\, y$, which as previously observed implies $x\, w(\boldsymbol{P})\, y$. □

The second lemma of Tang and Lin [27] extends Arrow's impossibility from the case of two individuals to every finite set $\mathcal{N}$. A generalisation to an infinite set $\mathcal{N}$, analogous to our Lemma 1, cannot be proved, for Arrow's Theorem does not hold for infinite societies, as we have seen in Section 2.2.

The proof methods presented in this section inspire a new terminology for properties of SWFs. Let AX be a set of axioms or properties of SWFs. We say that AX has the *inductive property* with respect to alternatives if an inductive proof like that of Tang and Lin [27] can produce a SWF satisfying AX for $m-1$ alternatives starting from a SWF satisfying AX for m alternatives. We can define the same property with respect to individuals. If a certain set of axioms has the inductive property, then an impossibility result over a minimal case will spread over SWFs of any finite size. We say that a set of axioms AX satisfies the *finite model property* (FMP) with respect to alternatives (individuals) if from every SWF satisfying AX we can build a SWF over a finite set of alternatives (individuals) that satisfies the same axioms. If this property hold we can extend an impossibility result from the finite to the infinite case. Finally, putting these two properties together, we have the *reduction property*: from any SWF satisfying AX we can build another one that still satisfies the same axioms but with a set of alternatives (individuals) of minimal size. The proof of Sen's Theorem is a proof that the axioms of unanimity and minimal liberalism have the reduction property for both alternatives and individuals. Our Lemma 1 proves that the Arrovian axioms have the reduction property for alternatives. The inductive lemma by Tang and Lin [27] guarantees that the same axioms have the inductive property with respect to individuals, but Fishburn [12] shows that they do not have the finite model property.

3 Language for Axioms

In this section we present a formal system of axioms expressed in FOL to model the social choice framework of preference aggregation. Our approach borrows several ideas from Tang and Lin [27], whose main concern, however, is a different one and who do not provide a complete axiomatisation. We start by introducing a first-order language and we provide axioms to reason about SWFs. We then formalise Arrow's conditions in this language and, in a slightly extended language, Sen's conditions. We conclude by generalising our axiomatisation to the case of weak orders.

3.1 A Theory for Social Welfare Functions

The first step is to define a theory capable of reasoning about SWFs. In Section 2 we have introduced the main objects: individuals, alternatives and preference profiles. A closer look at Arrow's axioms suggests that if we aim at formalising such conditions with a first-order language we must be able to quantify over all three objects separately. While a unary predicate can serve the purpose by marking alternatives and individuals, problems arise when dealing with quantification over all possible linear orders (the set of preference profiles). At first sight this corresponds to a second-order quantification, but exploiting the finiteness of the domain and the fact that two linear orders can be generated from each other using a sequence of swaps, we are able to devise a version of the condition of universal domain that holds on finite models. Following Tang and Lin [27], we introduce a set of "situations" and consider them as names for different preference profiles. In our case the set of situations will be a subset of the domain marked by a unary predicate, allowing us to quantify over this set. We will indicate with $\boldsymbol{P}^u$ the preference profile associated with situation u. Call $\mathcal{L}_{\text{SWF}} = \{A^{(1)}, I^{(1)}, S^{(1)}, p^{(4)}, w^{(3)}\}$ the relational first-order signature consisting of the following components:

1. three unary predicates to mark alternatives (A), individuals (I), and situations (S).
2. a predicate p of arity 4 to represent, given an individual z and a situation u, the linear order P^u_z associated with situation u. Orders are represented as binary relations: $p(z, x, y, u)$ indicates that individual z prefers x over y in situation u.
3. a ternary relation w that stands for the SWF, producing the social preference relation $w(\boldsymbol{P}^u)$ for every situation u. $w(x, y, u)$ translates as x is preferred over y in the social order associated with situation u.

Formulas in this language express conditions for SWFs, and we now present an axiomatisation to characterise this class. We start from the axioms of linear order for $p(z, \cdot, \cdot, u)$:

LINp:
- $I(z) \wedge S(u) \wedge A(x) \wedge A(y) \rightarrow (p(z, x, y, u) \vee p(z, y, x, u) \vee x = y)$
- $I(z) \wedge S(u) \wedge A(x) \rightarrow \neg p(z, x, x, u)$
- $I(z) \wedge S(u) \wedge A(x_1) \wedge A(x_2) \wedge A(x_3) \wedge p(z, x_1, x_2, u) \wedge p(z, x_2, x_3, u) \rightarrow p(z, x_1, x_3, u)$

All axioms presented in this section are to be considered universally closed; therefore the first axiom should be read as: "for all z, u, x and y, if z is an individual, if u is a situation, and if x and y are alternatives, then either individual z in situation u prefers x to y, or she prefers y to x, or x is equal to y." This is the completeness (or connectedness) axiom, and the second and the third are the irreflexivity and transitivity axioms. Recall that a situation u encodes a preference profile, so the quantification over S-variables is a quantification over all preference profiles encoded in S. Further axioms will

ensure that these are all the logically possible profiles of linear orders over $\mathcal{X}$. The analogous axioms for $w(\cdot, \cdot, u)$ follow:

LINw:
- $S(u) \wedge A(x) \wedge A(y) \rightarrow (w(x, y, u) \vee w(y, x, u) \vee x = y)$
- $S(u) \wedge A(x) \rightarrow \neg w(x, x, u)$
- $S(u) \wedge A(x_1) \wedge A(x_2) \wedge A(x_3) \wedge w(x_1, x_2, u) \wedge w(x_2, x_3, u) \rightarrow w(x_1, x_3, u)$

These are axioms for *collective rationality*: they require the outcome of aggregation to be a linear order. The next two sets of axioms guarantee that there are at least 3 different alternatives, that A, I and S are non-empty and that they form a partition of the universe:

MIN:
- $\exists x_1.\exists x_2.\exists x_3.A(x_1) \wedge A(x_2) \wedge A(x_3) \wedge ((x_1 \neq x_2) \wedge (x_1 \neq x_3) \wedge (x_2 \neq x_3))$
- $\exists z.I(z)$
- $\exists u.S(u)$

PART:
- $A(x) \rightarrow (\neg I(x) \wedge \neg S(x))$
- $I(x) \rightarrow (\neg A(x) \wedge \neg S(x))$
- $S(x) \rightarrow (\neg I(x) \wedge \neg A(x))$
- $A(x) \vee I(x) \vee S(x)$

The next two axioms restrict the arguments of p and w to be of the correct type:

DEF:
- $p(z, x, y, u) \rightarrow (I(z) \wedge A(x) \wedge A(y) \wedge S(u))$
- $w(x, y, u) \rightarrow (A(x) \wedge A(y) \wedge S(u))$

We now turn our attention to the encoding of the set of all preference profiles into the set of elements marked by S. The first axiom guarantees that two distinct situations cannot encode the same preference profile, thus the encoding of situations into preference profiles must be injective:

INJ: $S(u) \wedge S(v) \wedge u \neq v \rightarrow$
$\exists z.\exists x.\exists y.[I(z) \wedge A(x) \wedge A(y) \wedge p(z, x, y, u) \wedge p(z, y, x, v)]$

To express the condition of universal domain in our language, and to be able to quantify over the entire set of situations, we use another idea due to Tang and Lin [27]: identify the set $\mathcal{L}(\mathcal{X})$ with the symmetric group $S(\mathcal{X})$ of all permutations over $\mathcal{X}$ and generate it via transpositions. This is the job of the next axiom:[5]

PERM: $p(z, x, y, u) \rightarrow \exists v.\{S(v) \wedge p(z, y, x, v) \wedge$
$\forall x_1.[p(z, x, x_1, u) \wedge p(z, x_1, y, u) \rightarrow p(z, x_1, x, v) \wedge p(z, y, x_1, v)] \wedge$
$\forall x_1.[(p(z, x_1, x, u) \rightarrow p(z, x_1, y, v)) \wedge (p(z, y, x_1, u) \rightarrow p(z, x, x_1, v))] \wedge$
$\forall x_1.\forall y_1.[x_1 \neq x \wedge x_1 \neq y \wedge y_1 \neq y \wedge y_1 \neq x \rightarrow (p(z, x_1, y_1, u) \leftrightarrow p(z, x_1, y_1, v))] \wedge$
$\forall z_1.\forall x_1.\forall y_1.\ [z_1 \neq z \rightarrow (p(z_1, x_1, y_1, u) \leftrightarrow p(z_1, x_1, y_1, v))]\}$

[5] Observe that in this axiom the variables x_1, y_1, and z_1 must be explicitly quantified, because they are within the scope of an existential quantifier; the other variables x, y, z, and u are (as before) implicitly bound by the universal closure of the axiom.

The complexity of this axiom is largely due to the fact that linear orders are being represented as binary relations. Given our representation of P_i not as a complete sequence of elements in $\mathcal{X}$ but as a subset of $\mathcal{X}^2$, we have to require that, given a situation u, an individual z, and two alternatives x and y, there exists another situation v such that (the following five items correspond to the five lines of the axiom):

1. the relative positions of x and y have been switched in P_z^v;
2. if an alternative x_1 was between x and y in P_z^u, then its relation with respect to x and y is switched in P_z^v;
3. if x_1 was more preferred than x in P_z^u, then in v it is more preferred than y (and thereby also x); if it was less preferred than y in P_z^u, then in v it is less preferred than x (and thereby also y).
4. for every pair of alternatives different from x and y the relative ranking is copied;
5. $P_{z'}^v = P_{z'}^u$ for every individual $z' \neq z$.

Call T_{SWF} the theory composed of all the axioms above. In Section 4 we will prove a completeness result with respect to the class of models that can be constructed from SWFs, providing a formal argument to the claim that T_{SWF} characterises the class of SWFs. It is worth noting that some of our axioms, such as **PART** or **INJ**, are not strictly required. Including these axioms gives us more "control" in the resulting models and improves the readability of the axiomatisation.

3.2 Arrow's Axioms

We are now able to formalise the conditions that lead to Arrow's impossibility result. Adding to T_{SWF} the next three axioms we obtain a theory that we shall call T_{ARROW}:

UN: $S(u) \wedge A(x) \wedge A(y) \rightarrow [(\forall z.(I(z) \rightarrow p(z, x, y, u))) \rightarrow w(x, y, u)]$

IIA: $S(u_1) \wedge S(u_2) \wedge A(x) \wedge A(y) \rightarrow$
$[\forall z.(I(z) \rightarrow (p(z, x, y, u_1) \leftrightarrow p(z, x, y, u_2))) \rightarrow (w(x, y, u_1)$
$\leftrightarrow w(x, y, u_2))]$

ND: $I(z) \rightarrow \exists x.\exists y.\exists u.[S(u) \wedge A(x) \wedge A(y) \wedge p(z, x, y, u) \wedge w(y, x, u)]$

Let us analyse in detail the axiom of independence of irrelevant alternatives. The first universal quantification provides us with two generic situations u_1 and u_2 and two alternatives x and y. The main implication then states that if all individuals do not change their preference about x and y when moving from situation u_1 to u_2, then the social outcome $w(x, y, u_1)$ in the first situation must be the same as $w(x, y, u_2)$. On a finite set $\mathcal{X}$ of alternatives the permutation axiom guarantees that this applies to all logically admissible profiles.

Several weaker versions of the axiom of independence have been proposed in the literature, in an attempt to escape Arrow's impossibility result. An axiomatisation of these frameworks can be obtained by simply replacing the axiom of independence presented in this section with a formalisation of the

weaker version. For instance, the notion of *ternary* and *m-ary independence* proposed by Blau [6] can be expressed in our language by modifying appropriately the antecedent of the current formalisation, to account for three (or more) alternatives.

3.3 Weak Orders and General Aggregation Procedures

Arrow's Theorem was initially formulated for weak orders [2], and in this section we provide a suitable modification of T_{SWF} to cover this and more general cases.

To allow for ties in the preferences of the individuals the first axioms to be modified are that of linear order **LINp**, changing irreflexivity into reflexivity:[6]

WEAKp:
- $I(z) \wedge S(u) \wedge A(x) \wedge A(y) \rightarrow (p(z, x, y, u) \vee p(z, y, x, u) \vee x = y)$
- $I(z) \wedge S(u) \wedge A(x) \rightarrow p(z, x, x, u)$
- $I(z) \wedge S(u) \wedge A(x_1) \wedge A(x_2) \wedge A(x_3) \wedge p(z, x_1, x_2, u) \wedge p(z, x_2, x_3, u) \rightarrow p(z, x_1, x_3, u)$

The same can be done for collective rationality formalised in **LINw**. Things get more complicated for what concerns the coding of situations. While the axiom **INJ** can be kept without modifications, the axiom of permutation has to be significantly changed to be able to construct the whole set of weak orders from a single situation. This can be done in the following way. First introduce an axiom that states the existence of a preference profile where all individuals are indifferent over all alternatives:

PERM1: $\exists u.S(u) \wedge (\forall z.\forall x.\forall y.\ I(z) \wedge A(x) \wedge A(y) \rightarrow p(z, x, y, u))$

The second step is to modify the permutation axiom to enable us to separate indifferent alternatives, putting one of the two at the bottom of the order:

PERM2: $p(z, x, y, u) \wedge p(z, y, x, u) \rightarrow$

$(\exists v_1.\{S(v_1) \wedge \forall x_1.[(x_1 \neq y) \rightarrow p(z, x_1, y, v_1) \wedge \neg p(z, y, x_1, v_1)] \wedge$
$\forall x_1.[x_1 \neq y \rightarrow p(z, x, x_1, v_1) \leftrightarrow p(z, x, x_1, u)] \wedge$
$\forall x_1.\forall y_1.[x_1 \neq x \wedge x_1 \neq y \wedge y_1 \neq y \wedge y_1 \neq x \rightarrow (p(z, x_1, y_1, v_1) \leftrightarrow p(z, x_1, y_1, u))] \wedge$
$\forall z_1.\forall x_1.\forall y_1.\ [z_1 \neq z \rightarrow (p(z_1, x_1, y_1, v_1) \leftrightarrow p(z_1, x_1, y_1, u))]\} \wedge$

$\exists v_2.\{S(v_2) \wedge \forall x_1.[p(z, x_1, y, v_2)] \wedge \forall x_1[(\forall x_2\ p(z, x_2, x_1, u)) \leftrightarrow p(z, y, x_1, v_2)] \wedge$
$\forall x_1.[x_1 \neq y \rightarrow p(z, x, x_1, v_2) \leftrightarrow p(z, x, x_1, u)] \wedge$
$\forall x_1.\forall y_1.[x_1 \neq x \wedge x_1 \neq y \wedge y_1 \neq y \wedge y_1 \neq x \rightarrow (p(z, x_1, y_1, v_2) \leftrightarrow p(z, x_1, y_1, u))] \wedge$
$\forall z_1.\forall x_1.\forall y_1.\ [z_1 \neq z \rightarrow (p(z_1, x_1, y_1, v_2) \leftrightarrow p(z_1, x_1, y_1, u))]\})$

[6] We will not change the name of the predicate p in the language, interpreting it as representing preference (weak or strict).

The first part of the axiom separates alternatives x and y that were clustered together, and sends y to the bottom of the order making it strictly dominated by every other alternative. The second part does the same job, but clusters y together with the alternatives that constituted the bottom of the initial situation u. It is easy to see that in this way we can generate all weak orders over a finite set of alternatives. At last, Arrow's conditions have to be adapted to conclude the axiomatisation of the framework. This can easily be done, paying particular attention to the unanimity axiom that is usually stated for the "strict part" of the order R_i (recall that a is strictly preferred to b iff aRb and $\neg b\, Ra$).

Analogously to what we have done for the case of weak orders, the definition of a SWF can be modified to cover the case of preferences represented as partial orders in the social output. More generally, we call *aggregation procedure* a function that associates a collective binary relation over $\mathcal{X}$ with a profile of binary relations over the same set $\mathcal{X}$ supplied by the individuals. This is the case of SWFs, where both input and output of the function are linear (or weak) orders. As we have seen in this case, with a suitable modification of the axioms **LINp** and **LINw** we can control, respectively, the properties of individual and collective preference relations. For instance, by removing the axiom of completeness from **LINw** we obtain SWFs which output an incomplete ranking of the alternatives. While classical social choice theory concentrates on total preference relations, either weak or linear orders, partial orders are attractive for both theoretical and computational reasons, for instance when the set of alternatives is too large to enable individuals to compare each pair of alternatives [23].

The main drawback of this approach is that a new axiom generating a universal domain, corresponding to **PERM**, has to be devised for any such system. In Section 4.4 we take one step further, proving that the condition of universal domain for linear orders (on both finite and infinite domains) is not first-order axiomatisable.

3.4 Sen's Framework of Individual Rights

In this section we provide additional axioms to formalise Sen's framework of individual rights. The language $\mathcal{L}_{\text{SWF}}$ has to be enriched with a new predicate d to represent decisive sets, obtaining the following signature $\mathcal{L}_{\text{SEN}} = \{A^{(1)}, I^{(1)}, S^{(1)}, p^{(4)}, w^{(3)}, d^{(2)}\}$. The interpretation of these symbols is the same as for $\mathcal{L}_{\text{SWF}}$, with d representing the decisive sets of player i in the following way: $d(i, x, y)$ holds iff $(x, y) \in D_i$.

We now define the theory T_{SEN} by adding to the theory T_{SWF} three axioms. First, since the relation d encodes the decisive sets of the individuals, in the next set of axioms we state that decisive sets are symmetric (if $(x, y) \in D_i$ then $(y, x) \in D_i$), irreflexive (if $(x, y) \in D_i$ then $x \neq y$), and in analogy to p and

w we require its arguments to be of the correct type. The final axiom in the following list encodes the meaning of decisiveness by relating d to the SWF w.

DEC:
- $I(z) \wedge A(x) \wedge A(y) \wedge d(z, x, y) \rightarrow d(z, y, x)$
- $I(z) \wedge A(x) \rightarrow \neg d(z, x, x)$
- $d(z, x, y) \rightarrow (I(z) \wedge A(x) \wedge A(y))$
- $d(z, x, y) \rightarrow \forall u.p(z, x, y, u) \rightarrow w(x, y, u)$

Second, we formalise the conditions of Sen's Theorem:

UN: • $S(u) \wedge A(x) \wedge A(y) \rightarrow [\forall z.(I(z) \rightarrow p(z, x, y, u)) \rightarrow w(x, y, u)]$

MINLIB: • $\exists z_1.\exists z_2.\exists x_1.\exists y_1.\exists x_2.\exists y_2.[(d(z_1, x_1, y_1) \wedge (d(z_2, x_2, y_2) \wedge (z_1 \neq z_2)]$

The first axiom, **UN**, is the same axiom of unanimity as for the Arrovian framework, and the second axiom, **MINLIB**, formalises minimal liberalism, stating that for at least two distinct individuals there are two alternatives on which they are decisive.

3.5 Formalisations in Other Logical Languages

While we are not aware of any other work exploring the limits of classical FOL in expressing the Arrovian framework of SWFs, there have been several contributions to the literature making proposals for a full formalisation of Arrow's Theorem, using a variety of logical frameworks. In this section, we briefly review some of them.

A number of results in social choice theory have been proved by Tang [26] using the inductive method we sketched in Section 2 (see also Lin [19] for a more general view). Tang and Lin [27] use a formalisation in the style of the Situation Calculus to model and ultimately automatically prove or even discover theorems in social choice theory. This language proves very useful for the purpose of automatically checking base cases of theorems such as Arrow's. In these small domains it is possible to list all instances of their formalisation of the axioms in propositional logic, and later check the (un)satisfiability of these formulas using a SAT solver. While our first-order language borrows several ideas from their approach, it constitutes a language in the logical sense of this term, and enables us to study results concerning axiomatisability and expressivity. Moreover, our language requires less mathematical fatigue to support the automation, since no inductive lemmas have to be proven before an implementation can take place. In addition to that, it does not require us to specify the number of alternatives and of individuals explicitly in the language. On the other hand, as we shall briefly discuss in Section 5, automatically proving an impossibility theorem from axioms expressed in our language, while possible in principle, is highly demanding in practice.

Second-order logic of orders is the natural candidate to write axioms like **PERM**, and it has indeed been employed by Nipkow [21] and Wiedijk [31]

to formalise the proof of Arrow's Theorem using automatic theorem checkers like Mizar and Isabelle. In Section 5 we review this approach more in detail.

The labelling of variables with unary predicates like I, S and A immediately suggest an alternative formalisation in many-sorted first-order logic [10]. This approach has been followed by Geist and Endriss [16] in the related field of *ranking sets of objects*, i.e., the study of how to extend preferences from alternatives to set of alternatives, and provides a more readable axiomatisation.

Agotnes et al. [1] and Troquard et al. [29] develop *modal logics* for expressing concepts from social choice theory, including Arrow's Theorem. In the first paper the authors provide a modal framework capable of reasoning about preference and judgment aggregation, providing a formal proof in their language of a key lemma in the proof of Arrow's Theorem. The second work concentrates on social choice *functions* (i.e., procedures that associate a subset of alternatives to every profile), and provides a sound and complete axiomatisation of this class. The authors present logical formalisations of several axioms, with a focus on strategy-proofness. Both these approaches obtain interesting and useful results for the specification and verification of properties of aggregation procedures. However, the logical systems they introduce are specifically built for this purpose, and their potential for a full formalisation of impossibility theorems is limited by the fact that the number of individuals is fixed in their language.

A formalisation in dependence logic has been sketched by Väänänen (personal communication, 2009). It represents an interesting approach in which it relates the Arrovian axiom of independence of irrelevant alternatives with concepts of dependence embedded in this logic. The drawback of this axiomatisation is, again, that the number of alternatives and of individuals appears explicitly in the axioms.

4 Formalisation of Impossibility Theorems

In Section 3 we have referred to T_{SWF} as the theory of SWFs, and in this section we justify this choice by proving that T_{SWF} axiomatises the class of SWFs. Using the terminology introduced by Pauly [22], we prove that T_{SWF} *absolutely axiomatises* the set of SWFs, i.e., a model for $\mathcal{L}_{\text{SWF}}$ represents a SWF if and only if it satisfies the theory T_{SWF}. To be precise, this is true for the finite case. In the general case, T_{SWF} axiomatises a set of *partial* SWFs defined on subdomains satisfying a certain condition of closure. This translates in the finite case into an absolute axiomatisation of all SWFs. To do so we will associate with every SWF w a model $\mathcal{M}_w$ of T_{SWF}, and then prove a completess result. This enables us to determine precisely to what extent the three theorems we have introduced in Section 2 can be formally derived from our axioms. We shall assume for the rest of the section that the set of alternatives is non-empty and contains at least 3 elements, and that the set of individuals is non-empty.

A model of T_{SWF} is a structure $\mathcal{M} = (M, A, I, S, p, w)$, specifying the interpretation of every symbol in the language presented in Section 3.

Definition 1 If w is a SWF for $\mathcal{X}$ and $\mathcal{N}$, then $\mathcal{M}_w$ is the following $\mathcal{L}_{\text{SWF}}$-model:

1. the universe $M = \mathcal{X} \sqcup \mathcal{N} \sqcup \mathcal{L}(\mathcal{X})^{\mathcal{N}}$, the disjoint union of the sets corresponding to the three unary predicates A, I and S (in particular the set S is equal to the set of all preference profiles $\mathcal{L}(\mathcal{X})^{\mathcal{N}}$);
2. $(z, x, y, u) \in p \Leftrightarrow x \, P_z^u \, y$, where P_z^u is the preference relation of z in profile u; and
3. $(x, y, u) \in w \Leftrightarrow x \, w(\boldsymbol{P}^u) \, y$.

If $\mathcal{X}$ is finite, then the resulting model $\mathcal{M}_w$ is unique. In the case where $\mathcal{X}$ is infinite, on the other hand, this is not the only model that can be built from w. To obtain a full characterisation we need the following definition:

Definition 2 Given a set $\mathcal{X}$, let $S(\mathcal{X})$ denote the set of permutations over $\mathcal{X}$. A transposition is a permutation that switches just two elements of the set. $G \subseteq S(\mathcal{X})$ is closed under transpositions if whenever $g \in G, g \circ \tau \in G$ for every transposition τ.

Observe that if $\mathcal{X}$ is finite, then the only subset of $S(\mathcal{X})$ closed under transpositions is $S(\mathcal{X})$ itself.

Let now w be a SWF on an infinite set of alternatives $\mathcal{X}$. We have already remarked that we can identify the set $\mathcal{L}(\mathcal{X})$ with the set $S(\mathcal{X})$ of all permutations over $\mathcal{X}$. With every choice of $G_i \subseteq S(\mathcal{X})$ closed under transpositions for every individual $i \in \mathcal{N}$ we can associate a model of T_{SWF}, using the same construction as in Definition 1, except that the set of situations is now the Cartesian product $S = \prod_{i \in \mathcal{N}} G_i$. In the finite case this definition boils down to Definition 1, because $\mathcal{L}(\mathcal{X})$ is the only possible choice for G_i for every individual. The following completeness result shows that these are all possible models of T_{SWF}:

Proposition 1 *$\mathcal{M} \models T_{\text{SWF}}$ if and only if there exist two non-empty sets $\mathcal{X}$ and $\mathcal{N}$, with $|\mathcal{X}| \geq 3$, and a SWF w for $\mathcal{X}$ and $\mathcal{N}$ such that $\mathcal{M} = \mathcal{M}_w$.*

Proof It is easy to prove that $\mathcal{M}_w$ is a model of T_{SWF}. By definition, for every z and u the relations $p(z, \cdot, \cdot, u)$ and $w(\cdot, \cdot, u)$ are linear orders over $\mathcal{X}$, so the **LINp** axioms are satisfied as well as **LINw**. The axioms **MIN**, **PART** and **INJ** are valid by virtue of items (*i*) and (*ii*) in Definition 1. The set of situations S is either the set of all preference profiles or a Cartesian product $\prod_{i \in \mathcal{N}} G_i$ of subsets of $\mathcal{L}(\mathcal{X})$ closed under transpositions. This is sufficient to validate axiom **PERM**. To see this, let u be a situation in S and i an individual, and

consider for every pair of alternatives the linear order obtained by switching these two alternatives in the order of individual i in situation u. This procedure is equivalent to composing an element in G_i (the order of individual i in u) with a transposition. Since G_i is closed under transpositions, the new profile we obtain is still an element of S, i.e., there exists a situation v that represents it. Thus, the axiom of permutation is satisfied.

Suppose now that $\mathcal{M} \models T_{\text{SWF}}$. We can define the two sets $\mathcal{N}$ and $\mathcal{X}$ as the subsets of the universe indicated by the unary predicates. With every element in S we can associate a preference profile, the one encoded by the relation $p^{\mathcal{M}}$. From the relation $w^{\mathcal{M}}$ we can define a *partial* SWF, whose domain is the set of all preference profiles encoded in S, a subset $G \subseteq \mathcal{L}(\mathcal{X})^{\mathcal{N}}$. By **PERM**, if we take the projection of G on every component i, denoted with G_i, we obtain a set of linear orders that is closed under transpositions: for every individual i, if $g \in G_i$ then g composed with every transposition (a swap of a pair of alternatives) is still in G_i. Thus G is of the form $\prod_{i \in \mathcal{N}} G_i$, and $\mathcal{M} = \mathcal{M}_w$ as defined in Definition 1. □

4.1 Arrow's Theorem

As we have seen, if the set of alternatives is finite we can associate a unique SWF with every model of T_{SWF}. Therefore, by virtue of Proposition 1, we can restate Arrow's Theorem as follows:

Theorem 4 *T_{ARROW} has no finite models.*

Despite its theoretical interest, a result like Theorem 4 is of little practical use for a potential application to automated reasoning. What should be sought is a formalisation of Arrow's theorem in a sentence that can be derived formally from our theory. The first attempt of proving the inconsistency of T_{ARROW} fails, because Arrow's Theorem does *not* hold in the case of an infinite number of individuals, as we have seen in Section 2.2. (The issue of an infinite number of *alternatives*, on the contrary, is fully resolved by Lemma 1.) Fishburn's result [12] translates in our framework into the existence of an infinite model $\mathcal{M}$ of T_{SWF} such that $\mathcal{M} \models (\mathbf{UN} \wedge \mathbf{IIA} \wedge \mathbf{ND})$. Since there is no first-order formula that characterises finite models (see e.g. Enderton [10]), we have to somehow circumvent this problem.

One possibility is to give up some generality and to fix the number of individuals with a set of additional axioms. Call T^n_{SWF} the theory composed of all axioms of T_{SWF} plus the following axiom:

$$\exists z_1, \ldots, z_n . I(z_1) \wedge \ldots \wedge I(z_n) \wedge (\textstyle\bigwedge_{k \neq j} z_k \neq z_j) \wedge [I(z) \rightarrow (z = z_1) \vee \cdots \vee (z = z_n)]$$

With a proof analogous to that of Proposition 1 we obtain a completeness result for T^n_{SWF} with respect to SWFs defined for a set $\mathcal{N}$ of n individuals. Now the following proposition holds:

Proposition 2 *If w is a SWF for $\mathcal{X}$ and $\mathcal{N}$ with $|\mathcal{X}| \geq 3$ and $|\mathcal{N}| = n$, and if $\mathcal{M}_w$ is the corresponding model, then $\mathcal{M}_w \models \neg(\boldsymbol{UN} \wedge \boldsymbol{IIA} \wedge \boldsymbol{ND})$. Therefore, for every n, $T^n_{\text{SWF}} \vdash \neg(\boldsymbol{UN} \wedge \boldsymbol{IIA} \wedge \boldsymbol{ND})$.*

Proof If the number of alternatives is finite, then the first part of this result is a direct consequence of Arrow's Theorem. In case there are an infinite number of alternatives, we can resort to a proof similar to that of Lemma 1 to reduce a model $\mathcal{M}_w$, constructed from a SWF w, to a base model for only 3 alternatives that agrees with the initial model on the three Arrow's conditions. The key observation is that in the proof of Lemma 1 we never used the condition of universal domain in its full generality: every time we defined a new profile, it was always constructible with a finite sequence of switches between pairs of alternatives. The condition of closure under transpositions therefore guarantees that the result extends to every $\mathcal{M}_w$ defined on a finite set $\mathcal{N}$. Since the conjunction of Arrow's conditions is falsified on the base model, then it is falsified also on the initial model constructed on an infinite number of alternatives. The second part of the statement follows by completeness of FOL. □

4.2 The Kirman-Sondermann Theorem

To formalise the Kirman-Sondermann Theorem we have to first encode in our language the statement that the set $\mathcal{J}$ of "winning coalitions" is an ultrafilter:

- $\mathcal{N} \in \mathcal{J}$: $[\forall z.I(z) \rightarrow p(z, x, y, u)] \rightarrow w(x, y, u)$
- Closure under intersections: $w(x, y, u_1) \wedge w(x, y, u_2) \rightarrow [(\forall z.I(z) \rightarrow (p(z, x, y, u_1) \wedge p(z, x, y, u_2) \leftrightarrow p(z, x, y, v))) \rightarrow w(x, y, v)]$
- Maximality: $[\forall z.I(z) \rightarrow (p(z, x, y, u) \leftrightarrow \neg p(z, x, y, v))] \rightarrow (w(x, y, u) \leftrightarrow \neg w(x, y, v))$

Call **UF** the conjunction of these axioms. It is important to note that these axioms characterise the notion of ultrafilter for this particular framework only. We now prove the following restatement of the Kirman-Sondermann Theorem:

Theorem 5 $T_{\text{SWF}} \cup \{\boldsymbol{UN}, \boldsymbol{IIA}\} \vdash \boldsymbol{UF}$*:*

Proof We will prove that all models $\mathcal{M}_w$ of $T_{\text{SWF}} \cup \{\textbf{UN, IIA}\}$ verify all axioms in **UF**, and conclude using completeness of FOL to obtain provability. By the Kirman-Sondermann Theorem, if w satisfies **UN** and **IIA** then the collection of winning coalitions is an ultrafilter. Let then $\mathcal{M}_w$ be a model built from w. The

first axiom of **UF** is clearly satisfied, since the SWF is unanimous. The second axiom states that whenever in two situations x is ranked higher than y, then in every other situation, if the intersection of the individuals who ranked x higher than y in the two previous situations continue to do so, then x should still be ranked higher than y. This axiom is valid by closure under intersections of the set of winning coalitions of w. The only detail requiring attention is that M_w could be defined over a (transposition-closed) subset of the universal domain. This constitutes no problem, as no axioms require the existence of particular profiles. With similar reasoning we can prove that the last axiom of maximality is also valid in $\mathcal{M}_w$. □

Note that the condition of non-dictatorship included in T_{ARROW} corresponds to requiring the ultrafilter to be *free* (i.e., non-principal): the existence of a dictator is equivalent to characterising the set of winning coalitions as those subsets containing an element i of $\mathcal{N}$. This gives a formal proof that the set of winning coalitions under Arrow's conditions must be a free ultrafilter. Since it is not possible to build a free ultrafilter over a finite set [8], we get an indirect formalisation of the argument presented by Fishburn [12]: if a SWF satisfies **UN**, **IIA** and **ND**, then the number of individuals must be infinite.

4.3 Sen's Theorem

The case of Sen's Theorem is easier. The theorem does not presuppose the finiteness of the domain of aggregation, and its proof works by reduction: given a SWF satisfying Sen's axioms, by restricting this function to the two decisive individuals on to a restricted set of three alternatives we derive a contradiction. The proof of Proposition 1 can be easily adapted to Sen's framework, and Theorem 3 is therefore equivalent to the following:

Theorem 6 *T_{SEN} is inconsistent (it has no models).*

We conclude this section with some general statements about the formalisation of axioms for SWFs. Using the terminology introduced in Section 2.4, we can state that if a set of axioms has the reduction property with respect to both alternatives and individuals, then an impossibility result corresponds to the theory formalising these axioms being inconsistent (cf. Sen's Theorem). If they only satisfy the inductive property instead, an impossibility result corresponds to the inconsistency of the theory in the finite case (cf. Arrow's Theorem). In the other direction, exploiting results in logic to obtain properties of the axiomatic requirements, the finite model property of a set of axioms could be obtained by analysing the shape of the first-order formulas used to translate them (see, e.g., [9]).

4.4 Universal Domain

In Section 3 we put forward two axioms to formalise the condition of universal domain for linear and weak orders. These axioms are rather complex, and rely

heavily on the assumption of finiteness of a model to generate all possible profiles of preferences. In this section we prove a non-axiomatisability result for the class of SWFs satisfying the axiom of universal domain over arbitrary sets of alternatives, thus justifying our choice of the axiom **PERM** as the best approximation to formalise the condition of universal domain in our first-order language.

If $\mathcal{M}$ is a model of T_{SWF}, we say that $\mathcal{M}$ satisfies the condition of universal domain if for every possible profile of linear orders there is a situation u that encodes it. Call $\mathcal{U}$ this class of models. What we seek is a $\mathcal{L}_{\text{SWF}}$-formula φ that axiomatises this class. This turns out to be impossible, as we show next:

Proposition 3 *There is no $\mathcal{L}_{\text{SWF}}$-formula φ such that for all models $\mathcal{M}$ of T_{SWF}, $\mathcal{M} \models \varphi$ if and only if $\mathcal{M}$ satisfies the condition of universal domain. That is, the classs $\mathcal{U}$ is not $\mathcal{L}_{\text{SWF}}$-axiomatisable.*

Proof For the sake of contradiction, suppose such a formula φ does exist. Since $\mathcal{L}_{\text{SWF}}$ is finite, using the downward Löwenheim-Skolem Theorem, we can construct a countable model $\mathcal{M}_1$ of $T_{\text{SWF}} \cup \varphi$. Recall that the universe of $\mathcal{M}_1$ is partitioned into three sets $\mathcal{X}_1$, $\mathcal{N}_1$ and S_1. Let $\mathcal{X}_1$ be the set of elements of $\mathcal{M}_1$ marked by predicate A. If $\mathcal{X}_1$ is finite, then the set $\mathcal{L}(\mathcal{X}_1)$ of all linear orders over $\mathcal{X}_1$ is also finite. The universal domain $\mathcal{L}(\mathcal{X}_1)^{\mathcal{N}_1}$ can therefore be either finite, if $\mathcal{N}_1$ is also finite, or uncountable in the case of an infinite set of individuals. Since S_1 encodes the universal domain and $\mathcal{M}_1$ is countable we conclude that $\mathcal{X}_1$ cannot be finite. Suppose then that $\mathcal{X}_1$ is countable. Then the set $\mathcal{L}(\mathcal{X}_1)$ is not countable. (This can be seen in the following way: every countable ordinal induces a non-isomorphic linear order over $\mathcal{X}_1$, therefore the cardinality of $\mathcal{L}(\mathcal{X}_1)$ is at least the cardinality of β_1, which is uncountable.)[7] This is a contradiction, since $\mathcal{M}_1$ is countable and S_1 is a subset of the domain of $\mathcal{M}_1$. □

5 Conclusion

In this work we have presented a first-order axiomatisation of social welfare functions, formalising successfully three important results in social choice theory. First, we have presented a first-order language and a theory for SWFs, we have formalised Arrow's conditions, and we have extended the language to cover the model of individual rights proposed by Sen [25]. We have been able to reduce non-trivial conditions to first-order statements, such as independence of irrelevant alternatives and the universal domain condition. A thorough study of the formalisation of the universal domain condition has been carried out throughout the paper, especially in Section 3.3 and Section 4.4, to cover the case of weak orders and of general sets of alternatives.

[7] More precisely, $\aleph_1 = |\omega_1| \leq |\mathcal{L}(\mathcal{X}_1)| \leq |\mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_1)| = 2^{\aleph_0}$.

In Section 4 we have focused on three famous theorems in social choice theory, namely those of Arrow [2], Sen [25], and Kirman and Sondermann [18]. We have explored to which extent they can be formalised and formally derived from the first-order axioms presented in Section 3. Sen's Theorem stands out as the easiest case, whose proof (a reduction proof, using the terminology introduced in Section 2.4) enables us to state an easy correspondence between the original statement and the inconsistency of the FOL axioms formalising Sen's conditions. For the case of Arrow's Theorem we have solved the issue of an infinite number of alternatives by proving Lemma 1, which reduces the impossibility to the case of 3 alternatives. Arrow's statement is therefore equivalent to the unsatisfiability of our axioms in finite models. We have also proved that, if the number of individuals is fixed in our language, then there is a formal derivation of Arrow's Theorem from our axioms. For the most general case of a possibly infinite number of individuals we have proved that a statement inconsistent with the assumption of an infinite society can be formally derived from Arrow's conditions, formalising in this way the Kirman-Sondermann Theorem. In Section 3.5 we have discussed related work that deals with formalising results in social choice theory in languages other than FOL.

Here lies the first of several ideas for future work. A comparison study between our formalisation and that in stronger logics or languages might lead to the use of more powerful theoretical results, for instance from model theory. A closer study of the relation between finite model properties and the "shape" of some of our axioms might lead to simpler proofs of our axiomatisability results, and might lead to interesting results by using methods from descriptive complexity theory [9]. In this direction, interesting connections with the work of Herzberg and Eckert [17] might be expected.

The results proved in Section 4 support the belief that automated reasoning can play a role in proving theorems of social choice theory, and we carried out some preliminary experiments using an automated theorem prover. The system we have chosen is **Prover9**, the successor of the well-known and widely used **Otter** theorem prover [20]. The task of writing an input file containing our axiomatisation does not pose a serious challenge, thanks to the simplicity of the syntax and the high readability of our axioms. However, to date we have not been able to automatically prove the theorems formalised in this paper. It is very likely that a suitable reformulation of the axioms, in a way that can help and guide the work of the theorem prover, would prove successful in increasing its speed and efficiency. Readers interested in this problem can find the list of all the axioms for Sen's Theorem in Appendix A. We have tested **Prover9**, as well as the equational theorem prover **E** [24], on this list of axioms, without obtaining a result after a reasonable amount of time, except for a minimal case with just two individuals and three alternatives where we instantiated the axiom of permutation for the 36 situations.

There is a growing literature concerning the use of automated reasoning in social choice theory, and we conclude this paper by reviewing some of these results. As mentioned before, Tang and Lin [27] have shown that Arrow's

Theorem in its general form (for finite $\mathcal{X}$ and $\mathcal{N}$) follows from Arrow's Theorem for 3 alternatives and 2 individuals. For this base case, these authors give a formalisation in *propositional logic*. While the number of SWF's is already prohibitively large in this case (namely $6^{36} \approx 10^{28}$), a complete instantiation of Arrow's conditions for 36 profiles in the base case is still feasible, and Tang and Lin [27] report that unsatisfiability can be verified using a state-of-the-art SAT solver in less than 1 second. This approach, altough successfull in providing new proofs of several classical theorems of social choice theory, has the drawback of not being easily generalised and adapted to other frameworks, since for every new application new inductive lemmas have to be proved, and new instanciations have to be generated.

The same method was employed and enhanced by Geist and Endriss [16] in the related field of *ranking sets of objects*. In this work the authors are able to prove a general inductive lemma for a set of axioms sharing a common structure, and they devise a complete procedure to automatically discover (im)possibility theorems by listing the formalisation of several of these axioms, and automatically going through all combinations.

A different approach is the one adopted by Nipkow [21] and Wiedijk [31]. These authors verify formally two proofs of Arrow's Theorem given by Geanakoplos [15] using *proof checkers for higher-order logic* (the Isabelle and Mizar system, respectively). The condition of finiteness of the set of individuals is expressible in these higher-order languages (and for these particular proofs, this condition must be stated also for the set of alternatives), making it possible to prove the full statement of Arrow's Theorem. This is the only approach so far where neither the number of individuals nor the number of alternatives is specified in the language.

Acknowledgements Earlier versions of this paper have been presented at the Third Workshop on Decision, Games and Logic in 2009 in Lausanne and at the Second Workshop of Logic, Rationality and Interaction in 2009 in Chongqing. We want to thank the audiences and reviewers of these two workshops for their kind and useful comments, as well as the JPL reviewers. We would also like to thank Stéphane Airiau, Daniele Porello and Joel Uckelman for their useful comments. A particular thanks to Joel for helping us getting started with Prover9.

Appendix A: T_{SEN} in Prover9 Syntax

```
 formulas(sos).

% LIN_p
(I(z) & S(u) & A(x) & A(y)) -> (p(z,x,y,u)|p(z,y,x,u)|x=y).
(I(z) & S(u) & A(x)) -> -p(z,x,x,u).
(I(z) & S(u) & A(x) & A(y) & A(v) & p(z,x,y,u) & p(z,y,v,u) )
 -> p(z,x,v,u).
```

```
% LIN_w
(S(u) & A(x) & A(y)) -> (w(x,y,u)|w(y,x,u)|x=y).
(S(u) & A(x) & A(y)) -> -w(x,x,u).
(S(u) & A(x) & A(y) & A(v) & w(x,y,u) & w(y,v,u)) -> w(x,v,u).

% DEC
d(z,x,y)->(I(z) & A(x) & A(y)).
d(z,x,y)->d(z,y,x).
d(z,x,y)->(x!=y).
d(z,x,y) -> (all u p(z,x,y,u) -> w(x,y,u)).

% PART
A(x) -> (-I(x) & -S(x)).
I(x) -> (-A(x) & -S(x)).
S(x) -> (-I(x) & -A(x)).
A(x) | I(x) | S(x).

% DEF
p(z,x,y,u)->(I(z) & A(x) & A(y) & S(u)).
w(x,y,u)->(A(x) & A(y) & S(u)).

% INJ
S(u) & S(v) & (u!=v) -> exists z exists x exists y
(I(z) & A(x) & A(y) & p(z,x,y,u) & p(z,y,x,v)).

% PERM
p(z,x,y,u) -> exists v (S(v) & p(z,y,x,v) &
(all x1 (p(z,x,x1,u) & p(z,x1,y,u) -> p(z,x1,x,v)
& p(z,y,x1,v))) &
(all x2 (p(z,x2,x,u) -> p(z,x2,y,v))) &
 (all x3 (p(z,y,x3,u) -> p(z,x,x3,v))) &
 (all x4 all y1 (x4 != x & x4 != y & y1 != y &
 y1 != x -> (p(z,x4,y1,u) <-> p(z,x4,y1,v)))) &
 (all z1 all x5 all y2 (z1 != z ->
(p(z1,x5,y2,u) <-> p(z1,x5,y2,v))))).

% UN
(S(u) & A(x) & A(y)) ->
(( all z (I(z) -> p(z,x,y,u))) -> w(x,y,u)).

% MINLIB
exists z1 exists z2 exists x1 exists y1 exists x2
exists y2 (d(z1,x1,y1) & d(z2,x2,y2) & z1!=z2).

end_of_list.
```

References

1. Ågotnes, T., van der Hoek, W., & Wooldridge, M. (2009). On the logic of preference and judgment aggregation. *Autonomous Agents and Multi-Agents Systems, 22*(1), 4–30.
2. Arrow, K. J. (1963). *Social choice and individual values*, 2nd edition. John Wiley & Sons.
3. Arrow, K. J., Sen, A. K., & Suzumura, K. (Eds.) (2002). *Handbook of social choice and welfare*. North-Holland.
4. Black, D. (1948). On the rationale of group decision-making. *The Journal of Political Economy, 56*(1), 23–34.
5. Blau, J. H. (1957). The existence of social welfare functions. *Econometrica, 25*(2), 302–313.
6. Blau, J. H. (1971). Arrow's Theorem with weak independence. *Economica, 38*(152), 413–420.
7. Burdman Feferman, A., & Feferman, S. (2004). *Alfred Tarski: Life and logic*. Cambridge University Press.
8. Davey, B. A., & Priestley, H. A. (1990). *Introduction to lattices and orders*. Cambridge University Press.
9. Ebbinghaus, H. D., & Flum, J. (1999). *Finite model theory*. Springer.
10. Enderton, H. B. (1972). *A mathematical introduction to logic*. Academic Press.
11. Endriss, U. (2011). Logic and social choice theory. In A. Gupta & J. van Benthem (Eds.), *Logic and philosophy today* (Vol. 2, pp. 333–377). College Publications.
12. Fishburn, P. C. (1970). Arrow's impossibility theorem: Concise proof and infinite voters. *Journal of Economic Theory, 2*(1), 103–106.
13. Gaertner, W. (2001). *Domain conditions in social choice theory*. Cambridge University Press.
14. Gaertner, W. (2006). *A primer in social choice theory*. Oxford University Press.
15. Geanakoplos, J. (2005). Three brief proofs of Arrow's impossibility theorem. *Economic Theory, 26*(1), 211–215.
16. Geist, C., & Endriss, U. (2011). Automated search for impossibility theorems in social choice theory: Ranking sets of objects. *Journal of Artificial Intelligence Research, 40*, 143–174.
17. Herzberg, F., & Eckert, D. (2012). Impossibility results for infinite-electorate abstract aggregation rules. *Journal of Philosophical Logic, 41*, 273–286.
18. Kirman, A., & Sondermann, D. (1972). Arrow's Theorem, many agents, and invisible dictators. *Journal of Economic Theory, 5*(2), 267–277.
19. Lin, F. (2007). Finitely-verifiable classes of sentences. Presented at the 8th international symposium on logical formalizations of commonsense reasoning, Stanford.
20. McCune, W. (2003). ANL/MCS-TM-263 OTTER 3.3 reference manual. Technical memo, Argonne National Laboratory, Argonne.
21. Nipkow, T. (2009). Social choice theory in HOL: Arrow and Gibbard-Satterthwaite. *Journal of Automated Reasoning, 43*(3), 289–304.
22. Pauly, M. (2008). On the role of language in social choice theory. *Synthese, 163*(2), 227–243.
23. Pini, M. S., Rossi, F., Venable, K. B., & Walsh, T. (2009). Aggregating partially ordered preferences. *Journal of Logic and Computation, 19*, 475–502.
24. Schulz, S. (2004). System description: E 0.81. In *Proceedings of the 2nd International Joint Conference on Automated Reasoning (IJCAR-2004)*.
25. Sen, A. K. (1970). The impossibility of a Paretian liberal. *The Journal of Political Economics, 78*(1), 152–157.
26. Tang, P. (2010). Computer-aided theorem discovery: A new adventure and its application to economic theory. Ph.D. Thesis, Hong Kong University of Science and Technology.
27. Tang, P., & Lin, F. (2009). Computer-aided proofs of Arrow's and other impossibility theorems. *Artificial Intelligence, 173*(11), 1041–1053.
28. Taylor, A. D. (2005). *Social choice and the mathematics of manipulation*. Cambridge University Press.
29. Troquard, N., van der Hoek, W., & Wooldridge, M. (2011). Reasoning about social choice functions. *Journal of Philosophical Logic, 40*, 473–498.
30. Väänänen, J. (2007). *Dependence logic*. Cambridge University Press.
31. Wiedijk, F. (2007) Arrow's impossibility theorem. *Formalized Mathematics, 15*(4), 171–174.