First-Order Logic Formalisation of Arrow's Theorem

Umberto Grandi and Ulle Endriss

Institute for Logic, Language and Computation
University of Amsterdam

Abstract. Arrow's Theorem is a central result in social choice theory. It states that, under certain natural conditions, it is impossible to aggregate the preferences of a finite set of individuals into a social preference ordering. We formalise this result in the language of first-order logic, thereby reducing Arrow's Theorem to a statement saying that a given set of first-order formulas does not possess a finite model. In the long run, we hope that this formalisation can serve as the basis for a fully automated proof of Arrow's Theorem and similar results in social choice theory. We prove that this is possible in principle, at least for a fixed number of individuals, and we report on initial experiments with automated reasoning tools.

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 [1]. One of the classical results in the field is Arrow's Theorem [2]; it states that 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 this paper we propose a formalisation of Arrow's Theorem in classical first-order logic (FOL), which may eventually pave the way for an automated proof of this important result.

There have been a number of recent contributions that address the formalisation of theorems in social choice theory (e.g., Pauly [3], Ågotenes et al. [4], Tang and Lin [5], Wiedijk [6], and Nipkow [7]). There are several reasons for this broad interest in applying tools from mathematical logic and automated reasoning to social choice theory. One 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 [3]. 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 [8], 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 [7], 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, automated reasoning could prove a very useful tool. Finally, the use of automated reasoning in social choice theory has the potential to unveil entirely new results. For instance, we can imagine that it may soon become possible to use automated theorem provers to check whether a known impossibility result persists when we weaken or otherwise alter some of the axioms, or to use model generators to automatically derive counterexamples. To a limited extent, such results have already been achieved in recent work by Tang and Lin [5].

Previous work has discussed formalisations of Arrow's Theorem in modal logic [4], dependence logic,[1] and in the language of set theory [6, 7]. Here we explore to what extent it is possible to model the framework underlying Arrow's Theorem in classical FOL. There are two reasons for focusing on FOL: it is a natural language for speaking about linear orders, which are central to the modelling of preferences, and automated theorem proving is more developed for FOL than it is for other systems. We are able to show that it is possible to completely describe the problem within a language of FOL based on the language of linear orders, with one exception: for stating that Arrow's Theorem only applies to the case of a *finite* number of individuals we have to resort to a statement outside the language (we will see that Arrow's Theorem is equivalent to a certain finite 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 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 [5], but goes beyond that work in providing a complete axiomatisation of the Arrovian framework of social welfare functions in classical FOL.

The remainder of the paper is organised as follows. In Section 2 we recall Arrow's Theorem and prove a useful lemma. Section 3 presents our axioms and ends with the restatement of Arrow's Theorem in our framework. The models of our axiomatisation are studied in detail in Section 4, with particular attention being paid to the issue of an infinite number of individuals. Related work is discussed in Section 5; and in Section 6 we discuss our preliminary results with an automated theorem prover and conclude. For the rest of the paper we shall assume familiarity with the basic concepts of first-order logic (see e.g. [10]).

2 Social Welfare Functions and Arrow's Theorem

In this section we review Arrow's Theorem and the framework of social welfare functions in which it is stated. We also discuss a recent contribution by Tang and Lin [5], who give a new proof of Arrow's Theorem based on an inductive argument, in which the base case can be checked automatically using automated reasoning tools, and we show how to generalise a lemma proved by these authors

[1] J. Väänänen (personal communication, 2009); see also [9].

so as to also cover the case where there are an infinite number of alternatives that need to be ranked.

Let I be a set of individuals expressing preferences over a set A of alternatives. We assume that these preferences are represented by linear orders[2] P_i, so that aP_ib holds if individual i strictly prefers a to b. We denote with $\mathcal{L}(A)$ the set of all linear orders on A, and call a *social welfare function* (SWF) for A and I a function $w : \mathcal{L}(A)^I \to \mathcal{L}(A)$. A SWF associates with every *preference profile* $\underline{P} = (P_1, \ldots, P_n) \in \mathcal{L}(A)^I$ a linear order $w(\underline{P})$, that in most interpretations is taken to represent the aggregation of the preferences of the individuals into a "social preference order" over A.

There are several properties that such an aggregation mechanism may satisfy, and some of them have been argued to be natural requirements for a SWF. The fact that in our definition w is defined on *all* preference profiles in $\mathcal{L}(A)^I$ represents what is often stated as a first such property, the *universal domain* condition. The three additional properties that lead to the statement of Arrow's Theorem are the following:

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

Arrow's Theorem [2] states that:

Theorem 1. *If A and I are finite and non-empty, and if $|A| \geq 3$, then there exists no SWF for A and I that satisfies **UN**, **IIA** and **ND**.*

Several proofs of the theorem are known (see e.g. [11]), and most of them give a general argument that works for any number of individuals and any number of alternatives. A new inductive proof has recently been given by Tang and Lin [5]: the authors prove two lemmas to reduce the general statement to a base case with 3 alternatives and 2 individuals, and verify this last step with a computer, using either constraint programming or a satisfiability solver. 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

[2] The original statement of Arrow's Theorem assumes weak orders, although many proofs in the literature are restricted to this simpler case. We will discuss how our framework can be extended to the more general case in Section 6.

for the case of 3 alternatives and n individuals, then it holds for every finite set of m alternatives and n individuals." We now prove a generalisation of this lemma that also covers the case of an *infinite* number of alternatives:

Lemma 1. *If there exists a SWF w for A and I, with $|A| \geq 3$, that satisfies **UN**, **IIA** and **ND**, then there exists a set A' with $|A'| = 3$ and a SWF for A' and I 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 A (including the infinite case) and n individuals".

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

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

where $\underline{P}$ is a preference profile over A' and $\underline{P}^e$ any extension to a preference profile over A. By **IIA** this definition does not depend on the extension chosen; w' remains unanimous and independent of irrelevant alternatives by definition.

It remains to show 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 A, and suppose that xP_iy in a certain profile $\underline{P}$. We now show that also $x\, w(\underline{P})\, y$ must hold, thus i is a dictator on every pair of alternatives in A. The case where both x and y are in A' is trivial. We can therefore restrict ourselves to the case where there are at least two distinct elements in A' different from x and y, a_1 and a_2. Let individual i change her preference relation such that $a_1P_ia_2$, obtaining profile $\underline{P}'$. Let now every individual (including i) rearrange her preference such that xP_ja_1 and a_2P_jy, and call this profile P''. Both steps can be done without affecting the initial ranking of x and y, thus by **IIA** $x\, w(\underline{P})\, y$ if and only if $x\, w(\underline{P}'')\, y$. By unanimity of w we have $x\, w(\underline{P}'')\, a_1$ and $a_2\, w(\underline{P}'')\, y$. Since i is a dictator relative to A', it must be the case that $a_1\, w(\underline{P}'')\, a_2$ holds, and thus by transitivity also $x\, w(\underline{P}'')\, y$, which as previously observed implies $x\, w(\underline{P})\, y$. □

3 Axiomatisation

In this section we present a formal system that can model the social choice framework of Arrow's Theorem. Our approach borrows several ideas from Tang and Lin [5], whose main concern, however, is a different one and who do not provide a complete axiomatisation. Arrow's conditions suggest a formalisation in second-order logic, due to the quantification over preference profiles. Following Tang and Lin [5], we instead 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, thus allowing us to quantify over this

set, which in turn enables us to give a first-order axiomatisation. We will indicate with $\underline{P}^u$ the preference profile associated to situation u. We first define the following first-order signature $\mathcal{L} = \{a_1, a_2, a_3, i_1, s_1, A^{(1)}, I^{(1)}, S^{(1)}, p^{(4)}, w^{(3)}\}$:

- a_1, a_2, a_3 are constants indicating three alternatives, i_1 indicates an individual, and s_1 indicates a situation;
- the three unary predicates mark alternatives (A), individuals (I), and situations (S);
- the predicate p represents, given an individual z and a situation u, the linear order P^u_z associated with situation u; and
- w stands for the social welfare function, representing with a ternary predicate the social preference relation $w(\underline{P}^u)$ for every situation u.

Using this language, we start by introducing the axioms of linear order for p:

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

All axioms presented in this paper 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.

The analogous axioms for $w(\cdot,\cdot,u)$ follow:

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

The next two sets of axioms guarantee that there are at least 3 different alternatives, that i_1 is an individual, s_1 is a situation, and that A, I and S form a partition of the universe of a model:

MIN:
- $A(a_1) \wedge A(a_2) \wedge A(a_3) \wedge I(i_1) \wedge S(s_1)$
- $\neg(a_1 = a_2) \wedge \neg(a_1 = a_3) \wedge \neg(a_2 = a_3)$

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

The next axiom guarantees that two distinct situations cannot encode

the same preference profile, thus the encoding of situations into preference profiles must be injective:

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

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 from the same paper by Tang and Lin [5]: identify the set $\mathcal{L}(A)$ with the symmetric group $S(A)$ of all permutations over A and generate it via transpositions. This is the job of the next axiom:[3]

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

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 A but as a subset of A^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):

- the relative positions of x and y have been switched in P_z^v;
- 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;
- 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).
- for every pair of alternatives different from x and y the relative ranking is copied;
- $P_{z'}^v = P_{z'}^u$ for every individual $z' \neq z$.

Call T_{SWF} the theory composed of all the axioms above, as it summarises the properties of social welfare functions. Adding the next three axioms we obtain a theory that we shall call T_{ARROW}:

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

Arrow's Theorem can now be restated as:[4]

[3] 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 instead implicitly bound by the universal closure of the axiom.

[4] This equivalence is a straightforward consequence of Proposition 1 that will be stated in the following section. Once we have proved that every model of T_{SWF} is associated

Theorem 2. *T_{ARROW} has no finite models.*

It is worth noting that some of our axioms, such as **PART** or **INJ**, are not strictly required. Including these axioms permits to have more "control" in the resulting models and improves the readability of the axiomatisation.

4 Dealing with the Infinite

In Section 3 we have referred to T_{SWF} as the theory of social welfare functions, and in this section we justify this choice of words by proving that T_{SWF} axiomatises this class.[5] We will do so by associating with every SWF w a model $\mathcal{M}_w$ of T_{SWF}, and proving a completeness result. This enables us to determine precisely to what extent Arrow's Theorem can be proved automatically. Special attention will be devoted to the issue of an infinite domain, where Arrow's Theorem does not hold. We will present two different approaches to overcome this difficulty, first by fixing the number of individuals directly in the language, and then a second one based on results by Kirman and Sondermann [12]. From now on we shall assume 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 $\mathcal{M} = (M, a_1, a_2, a_3, i_1, s_1, A, I, S, p, w)$, specifying the interpretation of every symbol in the language presented in Section 3.

Definition 1. *If w is a SWF for A and I, then $\mathcal{M}_w$ is the following $\mathcal{L}$-model:*

(i) the universe $M = A \sqcup I \sqcup \mathcal{L}(A)^I$; the disjoint union of the sets corresponds to the three unary predicates A, I and S (in particular the set S is equal to the set of all preference profiles $\mathcal{L}(A)^I$);

(ii) a_1, a_2, a_3 are three different alternatives, i_1 is an individual, and s_1 is a preference profile;

(iii) $(z, x, y, u) \in p \Leftrightarrow x\, P_z^u\, y$, where P_z^u is the preference relation of z in profile u; and

(iv) $(x, y, u) \in w \Leftrightarrow x\, w(\underline{P}^u)\, y$.

If A is finite, then the resulting model $\mathcal{M}_w$ is in some sense unique, depending only on the choice of the constants. In the case where A 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 A, let $S(A)$ denote the set of all permutations over A. A transposition is a permutation that switches just two elements of the set. $G \subseteq S(A)$ is closed under transpositions if whenever $g \in G$, $g \circ \tau \in G$ for every transposition τ.*

---

with a SWF, it will be sufficient to check that our last three axioms correspond to Arrow's conditions to prove that Theorem 2 is equivalent to Arrow's Theorem.

[5] Using the terminology introduced by Pauly [3], we will prove that T_{SWF} *absolutely axiomatises* the set of partial SWFs satisfying a condition of closure on the domain. This translates in the finite case into an absolute axiomatisation of all SWFs.

Observe that if A is finite, then the only subset of $S(A)$ closed under transpositions is $S(A)$ itself.

Let now w be a SWF on an infinite set of alternatives A. We have already remarked that we can identify the set $\mathcal{L}(A)$ with the set $S(A)$ of all permutations over A. With every choice of $G_i \subset S(A)$ closed under transpositions for every individual $i \in I$ 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 $S = \prod_{i \in I} G_i$. In the finite case this definition boils down to Definition 1, because $\mathcal{L}(A)$ is the only possible choice 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 A and I, with $|A| \geq 3$, and a SWF w for A and I 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 A, so the **LINp** axioms are satisfied as well as **LINw**. The axioms **MIN**, **PART** and **INJ** are valid thanks to (i) and (ii) in Definition 1. The set of situations S is either the set of all preference profiles or a Cartesian product $\prod_{i \in I} G_i$ of subsets of $\mathcal{L}(A)$ closed under transpositions. This is sufficient to validate axiom **PERM**: given a situation u in S, for every individual and for every pair of alternatives the linear order obtained by switching these two alternatives is the composition of an element in G_i with a transposition. Therefore the new profile is still an element of S, i.e., there exists a situation v that represents this profile.

Suppose now that $\mathcal{M} \models T_{\text{SWF}}$. We can define the two sets I and A as the subsets of the universe indicated by the unary predicates. To every element in S we can associate a preference profile, the one encoded in 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 $G \subseteq \mathcal{L}(A)^I$. 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 $\prod_{i \in I} G_i$, and $\mathcal{M} = \mathcal{M}_w$ as defined in Definition 1. □

In view of our ultimate goal of using automated reasoning in social choice theory, a result like Theorem 2 is of little practical use, despite its theoretical interest. 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 has first been pointed out by Fishburn [13]. (The issue of an infinite number of *alternatives*, on the contrary, is fully resolved by Lemma 1.) Fishburn's result 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. [10]), we have to somehow circumvent this problem.

One possibility is to give up some generality and to fix the number of individuals in the language. Let therefore the new language $\mathcal{L}_n$ be $\mathcal{L} \cup \{i_2, \ldots, i_n\}$

with $n-1$ new constants, and call T^n_{SWF} the theory composed of all axioms of T_{SWF} plus the following axioms:

- $i_k \neq i_j$ for every $k \neq j$
- $I(i_2) \wedge \cdots \wedge I(i_n)$
- $I(z) \rightarrow (z = i_1) \vee \cdots \vee (z = i_n)$

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 I of n individuals. Now the following automated-reasoning friendly proposition holds:

Proposition 2. *If w is a SWF for A and I with $|A| \geq 3$ and $|I| = n$, 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 there exists a proof of $\neg(\boldsymbol{UN} \wedge \boldsymbol{IIA} \wedge \boldsymbol{ND})$ in T^n_{SWF}.*

Proof. The proof follows closely that of Lemma 1. In that proof, 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 I. □

The second approach we present is an indirect one: derive a consequence of T_{ARROW} that forces the resulting models to be infinite. Following the presentation of Arrow's Theorem in the case of an infinite number of individuals given by Kirman and Sondermann [12], this statement is the following: if a SWF satisfies **UN** and **IIA**, then the collection of "winning coalitions", those subsets $J \subseteq I$ such that if xP_jy for every $j \in J$ then $x\,w(\underline{P})\,y$, is an ultrafilter over I. A full axiomatisation of this statement can be given in the same language of T_{SWF} and is sketched in Appendix A. The condition of non-dictatorship corresponds to requiring the ultrafilter to be free: an unsatisfiable requirement if the set of individuals is finite. This finally formalises the argument of Fishburn [13] we presented in this section: if a SWF satisfies **UN**, **IIA** and **ND**, then the number of individuals must be infinite.

In conclusion, we have proved that an automated proof of Arrow's Theorem is possible, despite not in its most general form: for every finite number of individuals there is a (possibly different) first-order proof of the theorem.[6] The general case can be proved indirectly by deducing a set of statements about the sets of winning coalitions that force the set of individuals to be infinite. We report on our preliminary results with automated theorem prover in the last section.

5 Related Work

While we are not aware of any other work exploring the limits of classical first-order logic in expressing the Arrovian framework of social welfare functions,

[6] And since the set of theorems of a first-order theory is recursively enumerable it will eventually be found by an automated theorem prover.

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.

As mentioned before, Tang and Lin [5] have shown that Arrow's Theorem in its general form (for finite A and I) follows from Arrow's Theorem for 3 alternatives and 2 individuals. For this base case, these authors give a formalisation in *propositional logic*. This is possible, because the number of possible situations (preference profiles) is finite (namely $3! \times 3! = 36$) for this scenario. While the number of SWFs is already prohibitively large in this case (namely $6^{36} \approx 10^{28}$), a complete instantiation of Arrow's conditions for 36 situations is still feasible, and Tang and Lin [5] report that unsatisfiability can be verified using a state-of-the-art SAT-solver in less than 1 second. While our implementation of the same base case in FOL cannot compete with this performance, it arguably has the advantage of being more easily extended. The propositional language presented in [5] has the advantage of being rapidly solved, but can only be used to verify a base case. Building on this language, we aim instead at providing a fully automated proof of Arrow's Theorem without relying on any inductive lemma. (Note that the role of Lemma 1 is that of a theoretical guarantee for the existence of such a proof, at least for a fixed number of individuals, and it would not be part of any eventual automated derivation.) Also, our axiomatisation in Prover9 syntax is human-readable and easily fits on a single page (see Section 6 and Appendix B), while Tang and Lin's input to the SAT-solver is very large and has to be computer-generated (it consists of 106354 clauses).

Kaneko and Suzuki [14] discuss bounds on the size of a potential proof of Arrow's Theorem in a Gentzen-style *sequent calculus*, for the special case of 2 individuals and 3 alternatives.

Ågotnes et al [4] develop a *modal logic* for expressing concepts from social choice theory, including Arrow's Theorem. This logic is specifically designed for this purpose, and to date no automated procedure has been developed. The potential of the approach is limited by the fact that the number of individuals as well as the number of alternatives is fixed in the language.

Yet another approach is the one adopted by Nipkow [7] and Wiedijk [6]. These authors verify formally two proofs of Arrow's Theorem given by Geanakoplos [11] using *proof checkers* (Isabelle and Mizar, respectively). Their language is the language of set theory and their objects are sets; the condition of finiteness of the set of individuals is expressible in this language and this makes it possible to formalise and check the full statement of Arrow's Theorem. However, this approach requires a substantial amount of work in the process of rewriting an existing proof and then allows us to check every single simple step automatically.

The FOL framework developed by Rubinstein [15], while working with FOL, is different from ours. It aims at proving the existence of *single-profile analogues* of various results in social choice theory using social welfare functions defined on models of a suitable first-order theory. The single-profile approach avoids quantification over preference profiles from the outset. The exact relationship between these two frameworks certainly deserves future investigation.

6 Conclusions and Future Work

In this work we have given a first-order axiomatisation of social welfare functions, formalising the framework in which Arrow's Theorem is stated. We have been able to reduce non-trivial conditions to first-order statements, such as the universal domain condition and **IIA**. The issue of an infinite number of alternatives has been solved by proving a lemma that reduces the impossibility to the case of 3 alternatives. We have 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, and we have suggested an indirect approach to formalise the general case with a possibly infinite number of individuals.

All these results 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 used is Prover9, the successor of the well-known and widely used Otter theorem prover [16]. The task of writing an input file containing our axiomatisation does not pose any challenge, thanks to the simplicity of the syntax and the high readability of our axioms (see Appendix B). However, to date we have not been able to generate an automated proof of Arrow's Theorem. We designed a step-by-step proof of the simplest case of Arrow's Theorem for 2 individuals and 3 alternatives, following the formalisation of a simple proof of Arrow's Theorem by Nipkow [7]. At each step we received a negative response, with the prover exceeding the search space limits or not providing an answer in a reasonable amount of time.

A critical point, that may go some way towards explaining the difficulty of automatically deriving a proof, is that all of the intermediate lemmas we formalised rely on some steps where the existence of a particular preference profile has to be shown, using the condition of universal domain. This seems to require a clever use of the axiom of permutation, guessing the correct sequence of swaps to get from a profile to another, and it is likely to be the cause of the failure of Prover9 on these tasks. 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.

Despite these difficulties we were able to obtain some simple results, mainly by restricting the domain to the case of 2 individuals and 3 alternatives. For instance, we were able to generate an automated proof for the fact that the unanimity condition entails a weaker condition known as the *non-imposition* property [1]. A SWF satisfies non-imposition if for every pair of distinct alternatives x and y there is a profile $\underline{P}$ such that $x\, w(\underline{P})\, y$. In the syntax of Prover9 this condition can be written as follows:

```
A(x) & A(y) & x!=y -> (exists u (S(u) & w(x,y,u))).
```

We added two axioms to those in Appendix B in order to fix the number of alternatives and individuals, and we instantiated the axiom of permutation to this restricted domain. This still produces a readable axiomatisation, and Prover9 succeeds in providing a proof after about 3 hours on a standard desktop machine (with a memory limit of 500 Mb). The proof consists of 193 steps, and

the number of clauses generated is 20623974, of which 257685 have been kept to arrive at the proof. We have also run the same problem using another automated theorem prover, the E prover [17], and we obtained a positive response in a few seconds (on the other hand, for more difficult problems, E tends to run out of memory faster than Prover9).

This work can be extended in a number of ways. First, it is likely that a reformulation of the axioms and a guided use of the theorem prover will significantly improve performance and lead to the creation of a usable tool for social choice theorem proving. Second, it would be interesting to extend the axiomatisation to allow for preferences that are weak orders, allowing both the individual and the social order to express ties between alternatives. This can be achieved by replacing the irreflexivity axiom for both p and w with a reflexivity axiom, and adjusting the axiom of permutation to entail the condition of universal domain: starting from a linear order over alternatives, added "by default" in a model, it is possible to generate all weak orders requiring that in every situation every two alternatives can not only be swapped, but also ranked the same in the preference relation of every individual. Third, a large number of other results in social choice theory are likely to also be expressible in first order-logic. Examples include Sen's theorem on the impossibility of a Paretian Liberal and the Gibbard-Satterthwaite Theorem on the impossibility of strategy-proof voting rules that are non-dictatorial. In this direction, as already remarked, lies the main potential of this method: the use of automated reasoning as a tool for an easier exploration of new results in social choice theory.

Acknowledgements. We would like to thank Daniele Porello, Joel Uckelman, Stéphane Airiau, and two anonymous reviewers for their useful comments. A particular thanks to Joel for helping us getting started with Prover9.

References

1. Arrow, K.J., Sen, A.K., Suzumura, K., eds.: Handbook of Social Choice and Welfare. North-Holland (2002)
2. Arrow, K.J.: Social Choice and Individual Values. 2nd edn. John Wiley & Sons (1963)
3. Pauly, M.: On the role of language in social choice theory. Synthese **163**(2) (2008) 227–243
4. Ågotnes, T., van der Hoek, W., Wooldridge, M.: Reasoning about judgment and preference aggregation. In: Proc. 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2007), IFAAMAS (2007)
5. Tang, P., Lin, F.: Computer-aided proofs of Arrow's and other impossibility theorems. Artificial Intelligence **173**(11) (2009) 1041–1053
6. Wiedijk, F.: Arrow's impossibility theorem. Formalized Mathematics **15**(4) (2007) 171–174
7. Nipkow, T.: Social choice theory in HOL: Arrow and Gibbard-Satterthwaite. Technical report, Technische Universität München (2008)
8. Blau, J.H.: The existence of social welfare functions. Econometrica **25**(2) (1957) 302–313

9. Väänänen, J.: Dependence Logic. Cambridge University Press (2007)
10. Shoenfield, J.R.: Mathematical Logic. Addison-Wesley (1967)
11. Geanakoplos, J.: Three brief proofs of Arrow's impossibility theorem. Economic Theory **26**(1) (2005) 211–215
12. Kirman, A., Sondermann, D.: Arrow's theorem, many agents, and invisible dictators. Journal of Economic Theory **5**(2) (1972) 267–277
13. Fishburn, P.C.: Arrow's impossibility theorem: Concise proof and infinite voters. Journal of Economic Theory **2**(1) (1970) 103–106
14. Kaneko, M., Suzuki, N.Y.: A proof-theoretic evaluation of Arrow's theorem (2008) Working Paper presented at the 9th International Meeting of the Society for Social Choice and Welfare, Montréal.
15. Rubinstein, A.: The single profile analogues to multi profile theorems: Mathematical logic's approach. International Economic Review **25**(3) (1984) 719–730
16. McCune, W.: OTTER 3.3 Reference Manual. Technical Memo ANL/MCS-TM-263, Argonne National Laboratory, Argonne, IL (2003)
17. Schulz, S.: System Description: E 0.81. In: Proc. 2nd nternational Joint Conference On Automated Reasoning (IJCAR-2004), Springer (2004) 223–228

Appendix A: Axioms for Kirman-Sondermann Theorem

The set $\mathcal{J}$ of "winning coalitions" is an ultrafilter:[7]

- $I \in \mathcal{J}$ (**UN**):
  $\exists u.\exists x.\exists y.(\forall z.(I(z) \rightarrow p(z,x,y,u)) \rightarrow w(x,y,u))$
- $J \in \mathcal{J}$ and $J \subseteq K$ then $K \in \mathcal{J}$:
  $w(x,y,u) \rightarrow [\forall z.((I(z) \wedge p(z,x,y,u)) \rightarrow p(z,x,y,v)) \rightarrow w(x,y,v)]$
- $J_1, J_2 \in \mathcal{J}$ then $J_1 \cup J_2 \in \mathcal{J}$:
  $w(x,y,u_1) \wedge w(x,y,u_2) \rightarrow$
  $[\forall z.(I(z) \wedge p(z,x,y,u_1) \wedge p(z,x,y,u_2) \leftrightarrow p(z,x,y,v)) \rightarrow w(x,y,v)]$
- $J \subset I$ then $J \in \mathcal{J}$ or $J^c \in \mathcal{J}$:
  $\forall z.(I(z) \rightarrow (p(z,x,y,u) \leftrightarrow \neg p(z,x,y,v))) \rightarrow (w(x,y,u) \vee w(x,y,v))$
- Free ultrafilter (**ND**):
  $\neg\exists z.(I(z) \wedge \forall x.\forall y.\forall u.(w(x,y,u) \leftrightarrow p(z,x,y,u)))$

Call **FUF** the conjunction of these axioms. With an analogous proof to that of Proposition 2 we obtain that $T_{\text{ARROW}} \vdash \mathbf{FUF}$. This gives a formal proof that the set of winning coalitions under Arrow's conditions must be a free ultrafilter (i.e., the Kirman-Sonderman Theorem). Since it is not possible to build a free ultrafilter over a finite set, this formal proof is an indirect formalisation of Fishburn's generalisation of Arrow's Theorem.

[7] The axioms that follow formalise the notion of ultrafilter in this particular case only. Their formulation use a definition of "winning coalitions" that strongly relies on Lemma A by Kirman and Sondermann [12].

Appendix B: T_{ARROW} in Prover9 Syntax

```
% LINp
(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).

% LINw
(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).

% MIN
A(a1) & A(a2) & A(a3) & I(b1) & S(c1) & a1!=a2 & a2!=a3 & a1!=a3.

% 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)).

% IIA
(S(u1) & S(u2) & A(x) & A(y)) ->
((all z (I(z) -> (p(z,x,y,u1)<->p(z,x,y,u2)))) -> (w(x,y,u1)<->w(x,y,u2))).

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