Journal of Automated Reasoning manuscript No.
(will be inserted by the editor)

Social Choice Theory in HOL

Arrow and Gibbard-Satterthwaite

Tobias Nipkow

Dedicated to Mike Gordon on the occasion of his 60th birthday

Received: date / Accepted: date

Abstract This article presents formalizations in higher-order logic of two proofs of Arrow's impossibility theorem due to Geanakoplos. The Gibbard-Satterthwaite theorem is derived as a corollary. Lacunae found in the literature are discussed.

Keywords Social choice theory · Arrow's theorem · Gibbard-Satterthwaite theorem · Higher-order logic · Theorem proving

1 Introduction

In 1950, Kenneth Arrow proved his famous impossibility theorem [1,2] about voting systems: under certain minimal plausible fairness conditions, the only possible voting system is a dictatorship. In 1973, Gibbard [6] and Satterthwaite [11] proved a similar result where non-manipulability forces a dictatorship. These two theorems are closely related staples of social choice theory [13], and numerous proofs for them have been given in the literature. Yet logicians, as usual, have found fault with (some of) these proofs. Routley [10] writes:

> The standard and textbook proofs of Arrow's general impossibility theorem are, like the original proofs, invalid.

This was written 30 years ago but has not attracted much attention. According to Google Scholar, Routley's paper is cited 3 times (once in German and twice in Serbian); in contrast, Arrow's book [2], which Routley refers to, has been cited more than 5000 times.

In this paper I formalize some recently published proofs of Arrow's theorem [4,5,8] in HOL and find that in places they still suffer from opaqueness and missing cases. Although this makes it sound like a fault finding expedition, it should really be seen as an experiment in formalization in an area that had been *terra incognita* for theorem provers when I started this work.

The bulk of this work is divided into three parts. Section 4 describes a formalized proof of Arrow's theorem based on the first proof by Geanakoplos [4]. I had formalized this proof in the summer of 2002 but never published it, partly because I had not been able to formalize

T. Nipkow
Institut für Informatik, Technische Universität München
E-mail: www.in.tum.de/~nipkow

the third proof by Geanakoplos [4], as I had originally intended. The reasons for this will be explained in section 5, where I present a formalization of Geanakoplos' third proof based on a revised version of his proof published only in 2005 [5]. Missing cases in various versions of this proof are analyzed. Finally, in section 6, I present a formalized proof of the Gibbard-Satterthwaite theorem from Arrow's theorem.

All proofs are formalized in the logic HOL of the theorem prover Isabelle [7]. The reader should be warned that we present those proofs in great detail, although at least in English. Thus this paper is not meant to replace the original informal proofs — those are still preferable for a compact description of the key ideas. But since I take exception to missing details and incomplete case distinctions in those proofs, I had better present the full details myself, for future reference for others, although no doubt the manual transcription from the Isabelle sources has introduced new errors. For the complete Isabelle proof see the Archive of Formal Proofs at `afp.sf.net`.

2 Arrow's Impossibility Theorem

We start by giving an informal exposition of the statement of Arrow's impossibility theorem. Social choice theory is about how preferences by individuals are combined into an overall *social preference*, for example in voting systems. A *preference* is a complete ranking of some fixed set of alternatives, with ties allowed. A *profile* is a mapping from individuals to preferences; if individuals are numbered from 1 to N, profiles are often viewed as lists or tuples. A *social welfare function* maps profiles to social preferences. Note that "voting" is not anonymous: the social welfare function "knows" who has which preference because profiles are not merely (multi)sets of preferences. A social welfare function satisfies

unanimity if, whenever everybody puts a below b in their ranking, the social ranking of a is also below b;

independence of irrelevant alternatives (IIA) if the social ranking of two alternatives a and b depends only on their relative ranking by every individual. That is, given two profiles such that each individual ranks a relative to b the same way in both profiles, the social preferences induced by both profiles must also rank a relative to b the same way.

IIA is best explained in pictures. Consider the following two profiles, where the columns are the preferences of the three individuals:

a	b	c
b	a	a
c	c	b

c	b	a
a	c	b
b	a	c

Comparing the profiles column-wise we find that the ranking of a relative to b is the same. Hence IIA would require that the relative social ranking of a relative to b must also be the same. That is, if one profile induces $a < b$ (or $b < a$), so must the other.

Theorem 1 (Arrow) *Let the set of individuals and alternatives be finite and let there be at least 3 alternatives. Then any social welfare function that satisfies unanimity and independence of irrelevant alternatives is a dictatorship.*

A *dictatorship* is a function that always selects the preference of one fixed individual, the dictator.

The reader should beware that the literature abounds with different versions of Arrow's theorem. As Taylor [13] puts it:

Everybody has his or her own take on Arrow's impossibility theorem.

I follow the formulation by Geanakoplos, except for some of the terminology.

3 Basic Notation

HOL conforms largely to everyday mathematical notation, with a few modifications and additions. The notation $t::\tau$ means that term t has type τ. The space of total functions is denoted by $\Rightarrow$. Functions come with an update operation: $f(x := y)$ is function f redefined at point x to return result y. Symbol $\equiv$ stands for definitional equality.

Occasionally $\longrightarrow$ is displayed as *If-then*.

4 Geanakoplos' First Proof

4.1 Alternatives, Individuals, Preferences and Profiles

Geanakoplos starts from a finite set of at least three alternatives. We formalize this as a type *alt* with the axioms of finiteness and

$\exists a\ b\ c::alt.\ distinct\ [a, b, c]$

where *distinct* $[a, b, c]$ is merely a shorthand for $a \neq b \wedge a \neq c \wedge b \neq c$. As is customary in social choice theory, Geanakoplos simply numbers the individuals from 1 through N. To avoid the implicit ordering we turn this into a type *indi* and assume its finiteness.

A preference is a "ranking of the alternatives from top to bottom, with ties allowed" [5]. Mathematically speaking this is a total preorder. Our formalization of a preference in this first proof is a *utility function*, that is, a function from alternatives to real numbers. Clearly, every utility function induces a total preorder. Conversely, it is easy to see that every finite total preorder can be modelled as some utility function. In fact, the converse works as long as the underlying set is not larger than the reals. We define two type abbreviations:

$$pref = alt \Rightarrow real \qquad prof = indi \Rightarrow pref.$$

Our choice of modelling preferences as utility functions was mainly based on the ease with which a number of operations on preferences can be expressed (see below), and because reasoning about preferences often reduces to linear arithmetic, which is handled automatically by Isabelle. In our second proof (see section 5) we employed the standard set-theoretic model of linear orders instead and did not find that this caused significant complications in definitions or proofs.

We write a, b, c, d, e for alternatives, i, j for individuals, p for preferences and P, Q for profiles.

In order to hide the representation of preferences we introduce auxiliary functions for the key operations on preferences required in the proofs. To express that b is at the top, bottom or either position of p we define

$$\begin{array}{lcl} p <\cdot\ b & \equiv & \forall a.\ a \neq b \longrightarrow p\ a < p\ b \\ b \cdot<\ p & \equiv & \forall a.\ a \neq b \longrightarrow p\ b < p\ a \\ extreme\ p\ b & \equiv & b \cdot<\ p \vee p <\cdot\ b \\ Extreme\ P\ b & \equiv & \forall i.\ extreme\ (P\ i)\ b \end{array}$$

To move b into top or bottom position or between two other positions we define

$$\begin{array}{lcl} \mathit{mktop}\ p\ b & \equiv & p(b := \mathit{Max}\ (\mathit{range}\ p) + 1) \\ \mathit{mkbot}\ p\ b & \equiv & p(b := \mathit{Min}\ (\mathit{range}\ p) - 1) \\ \mathit{between}\ p\ a\ b\ c & \equiv & p(b := (p\ a + p\ c)\ /\ 2) \end{array}$$

Note that this works only because *range p* is finite because the set of alternatives is finite. Otherwise the maximum or minimum need not exist.

4.2 The Proof

The two assumptions of Arrow's theorem, unanimity (U) and irrelevance of independent alternatives (IIA), are global assumptions during the proof. That is, we work in the context of a fixed social welfare function $F :: \mathit{prof} \Rightarrow \mathit{pref}$ subject to these assumptions:

(U) $\mathit{If}\ \forall i.\ P\ i\ a < P\ i\ b\ \mathit{then}\ F\ P\ a < F\ P\ b.$
(IIA) $\mathit{If}\ \forall i.\ P\ i\ a < P\ i\ b \longleftrightarrow P'\ i\ a < P'\ i\ b\ \mathit{then}\ F\ P\ a < F\ P\ b \longleftrightarrow F\ P'\ a < F\ P'\ b.$

This is the direct formalization of the informal versions of unanimity and IIA in section 2. The formal version of IIA is strikingly more concise than its informal counterpart.

The overall goal is to show that there is a dictator for F. The Isabelle proof follows Geanakoplos closely. First we need the *Extremal Lemma*: If b is extreme in every individual's preference, it must be extreme in the social preference.

Lemma 1 (Extremal Lemma) $\mathit{If}\ \mathit{Extreme}\ P\ b\ \mathit{then}\ \mathit{extreme}\ (F\ P)\ b.$

Proof by contradiction. If not $\mathit{extreme}\ (F\ P)\ b$ then there must be a and c such that $F\ P\ a \geq F\ P\ b$ and $F\ P\ b \geq F\ P\ c$ and $\mathit{distinct}\ [a, b, c]$ (which requires a sub-proof, see below). We create P' from P by moving c above a as follows

$$P'\ i = (\mathit{if}\ P\ i <\cdot b\ \mathit{then}\ \mathit{between}\ (P\ i)\ a\ c\ b\ \mathit{else}\ (P\ i)(c := P\ i\ a + 1))$$

taking care not to change the relative ranking of a and c w.r.t. b. By IIA we still have $F\ P'\ a \geq F\ P'\ b$ and $F\ P'\ b \geq F\ P'\ c$. But by unanimity we have $F\ P'\ a < F\ P'\ c$, a contradiction.

The sub-proof mentioned above assumes $\neg\ \mathit{extreme}\ p\ b$ (where $p = F\ P$), which implies the existence of a and c distinct from b such that $p\ a \geq p\ b$ and $p\ b \geq p\ c$. In case $a \neq c$, we are done. If $a = c$ (and thus $p\ a = p\ b = p\ c$) we pick some d distinct from a and b and make a case distinction: if $p\ b < p\ d$ then $p\ d \geq p\ b \wedge p\ b \geq p\ c$, and if $p\ b \geq p\ d$ then $p\ c \geq p\ b \wedge p\ b \geq p\ d$. □

The rest of the proof relies on the notion of an (extremely) *pivotal* individual i

$$\begin{array}{l} \mathit{pivotal}\ i\ b \equiv \\ \exists P.\ \mathit{Extreme}\ P\ b \wedge b <\cdot P\ i\ b \wedge b <\cdot F\ P \wedge F\ (P(i := \mathit{mktop}\ (P\ i)\ b)) <\cdot b \end{array}$$

who can move b from the bottom to the top of the social preference by moving b from the bottom to the top of his preference, for some profile where b is always at the top or bottom.

First we show that for any b such a pivotal individual exists. The construction is inductive: Start with a profile where all individuals place b at the bottom, and thus, by unanimity, b is also at the bottom of the social preference; let the individuals one by one flip b from the bottom to the top of their preference. By the extremal lemma, b must be at the top or bottom of the social preference throughout this process. In the final profile everybody puts b at the top, and thus, by unanimity, b is also at the top of the social preference. Hence there must be some individual who caused b to flip from the bottom to the top of the social preference. Formally we prove the following proposition by induction:

Lemma 2 If *Extreme P b* and $b \cdot\!< F\ P$ then $\exists i.$ *pivotal i b*.

Proof Since our individuals are not numbers, we induct on the size of $D = \{i \mid b \cdot\!< P\ i\}$, the set of individuals who (still) have b at the bottom of their preference. If D is empty, the premises are contradictory: from $D = \emptyset$ we have that b must be at the top of every preference in P (because *Extreme P b*) and hence, by unanimity, that $F\ P <\cdot b$. If $D = D' \cup \{i\}$ and $i \notin D'$, let $P' = P(i := mktop\ (P\ i)\ b)$. If $F\ P' <\cdot b$ then i is the pivotal individual. Otherwise, by induction hypothesis, there is a pivotal individual for $D' = \{i \mid b \cdot\!< P'\ i\}$ (because, by the extremal lemma, $\neg\ F\ P' <\cdot b$ means $b \cdot\!< F\ P'$). □

By instantiating P in Lemma 2 with a profile where everybody puts b at the bottom (for example by assigning it utility 0 and assigning utility 1 to all other alternatives), unanimity forces $b \cdot\!< F\ P$, and we obtain the existence of a pivotal individual:

$\exists i.$ *pivotal i b*

Now we show in stages that a pivotal individual is a dictator. Geanakoplos introduces (implicitly) the following degrees of dictatorship:

$$\begin{array}{lcl} i\ \textit{dictates}\ a < b & \equiv & \forall P.\ P\ i\ a < P\ i\ b \longrightarrow F\ P\ a < F\ P\ b \\ i\ \textit{dictates-except}\ c & \equiv & \forall a\ b.\ c \notin \{a, b\} \longrightarrow i\ \textit{dictates}\ a < b \\ \textit{dictator}\ i & \equiv & \forall a\ b.\ i\ \textit{dictates}\ a < b \end{array}$$

where i *dictates* $a < b$ is a ternary predicate in mixfix notation and i *dictates-except* c a binary predicate in infix notation.

First we show that a pivotal individual for b dictates every pair not involving b:

If pivotal i b then i dictates-except b.

By definition we must show $F\ Q\ a < F\ Q\ c$ assuming $Q\ i\ a < Q\ i\ c$ for an arbitrary Q and arbitrary a, c distinct from b. From *pivotal i b* we obtain a profile P such that *Extreme P b* and $b \cdot\!< F\ P$ and $F\ P' <\cdot b$ where $P' = P(i := mktop\ (P\ i)\ b)$. We turn Q into a Q' by moving b to the top or bottom like in P, for all $j \neq i$, and moving b between a and c for i:

$$\begin{array}{l} Q'\ j = \\ (\textit{if}\ j = i\ \textit{then between}\ (Q\ j)\ a\ b\ c \\ \quad \textit{else if}\ P\ j <\cdot b\ \textit{then mktop}\ (Q\ j)\ b\ \textit{else mkbot}\ (Q\ j)\ b) \end{array}$$

Because a and c are not moved, IIA implies $F\ Q\ a < F\ Q\ c \longleftrightarrow F\ Q'\ a < F\ Q'\ c$. Hence our goal has become $F\ Q'\ a < F\ Q'\ c$, which we show in two steps: $F\ Q'\ a < F\ Q'\ b$ and $F\ Q'\ b < F\ Q'\ c$. Step $F\ Q'\ a < F\ Q'\ b$ follows by IIA from $F\ P'\ a < F\ P'\ b$ because $\forall j.\ P'\ j\ a < P'\ j\ b \longleftrightarrow Q'\ j\ a < Q'\ j\ b$ because *Extreme P b*. Step $F\ Q'\ b < F\ Q'\ c$ also follows by IIA: we have $F\ P\ b < F\ P\ c$ because $b \cdot\!< F\ P$ and $\forall j.\ P\ j\ b < P\ j\ c \longleftrightarrow Q'\ j\ b < Q'\ j\ c$ because *Extreme P b* and $b \cdot\!< P\ i$.

Second, we show that a pivotal individual i for b (*pivotal i b*) is a dictator, thus concluding the proof of

$\exists i.$ *dictator i*

We already know i *dictates-except b*. Hence it remains to show i *dictates* $a \lesseqgtr b$ (which abbreviates i *dictates* $a<b \wedge i$ *dictates* $b<a$) for any $a \neq b$. Let c be distinct from a and b. Hence there is a pivotal individual j for c (*pivotal j c*) which therefore dictates every pair not involving c (j *dictates-except* c). Hence j *dictates* $a \lesseqgtr b$. Thus it remains to show $i = j$, which we prove by contradiction, assuming $i \neq j$. Now we recall that *pivotal i b* yields a profile P such that *Extreme P b* and $b \cdot\!< F\ P$ and $F\ P' <\cdot b$ where $P' = P(i := mktop\ (P\ i)\ b)$. We

cannot have $P\ j\ a < P\ j\ b$ because this would imply $F\ P\ a < F\ P\ b$ (using j *dictates* $a \leqslant b$), which contradicts $b \cdot< F\ P$. This in turn implies $P\ j\ b < P\ j\ a$ because of *Extreme* $P\ b$ and $a \neq b$. By definition of P' and because $i \neq j$ this implies $P'\ j\ b < P'\ j\ a$, which contradicts $F\ P' <\cdot b$. This finishes the proof.

4.3 Comparison

Although the proof by Geanakoplos and the above translation into Isabelle are morally the same proof there are a number of differences:

- The Isabelle proof is 350 lines, about 6 pages, whereas Geanakoplos' proof is about about 1.5 pages, in both cases including definitions and statements.
- In Isabelle, the main proof is preceded by a prelude of auxiliary lemmas of about 100 lines not considered here.
- Geanakoplos implicitly turns a procedure for finding a pivotal individual into a function; we have expressed this as an existential quantifier.

Although we did not find any errors in Geanakoplos' proof, two statements required a more detailed proof. The first one is also commented on by Wiedijk [15]. In the proof of the extremal lemma we write above:

> If not *extreme* $(F\ P)\ b$ then there must be a and c such that $F\ P\ a \geq F\ P\ b$ and $F\ P\ b \geq F\ P\ c$ and *distinct* $[a, b, c]$.

This paraphrases Geanakoplos but it requires an auxiliary proof of about 20 lines.

The second argument that needs a bit of elaboration is the proof that the pivotal individual i for b equals the dictator j for $a \leqslant b$. Here the argument is fairly straightforward and hinted at in Geanakoplos' text.

5 Geanakoplos' Third Proof

This proof has a similar flavour as the first one but proceeds via an intuitive lemma whose proof turns out to be more involved than Geanakoplos' text leads one to expect. For added interest we also vary the model (relations instead of utility functions) to study the impact on the formalization.

5.1 Alternatives, Individuals, Preferences and Profiles

We again formalize the set of alternatives as a type *alt* and assume

$$\exists a\ b\ c{::}alt.\ distinct\ [a, b, c]$$

However, we drop the finiteness assumption, which turns out to be unnecessary, something also noticed by Routley [10]. Geanakoplos does not state this and many other accounts of Arrow's theorem make the same finiteness assumption [9,8].

Individuals are again formalized by a finite type *indi* whose cardinality we abbreviate by N. Generalizations to infinite sets of individuals exist but require more work [13].

Preferences are now formalized as sets of pairs:

$$pref = (alt \times alt)\ set$$

The set of strict linear preferences, i.e. transitive, irreflexive and total relations, is called *Lin*. Variable L is by default a preference. In contrast to Geanakoplos we focus on strict linear orders in this proof, primarily because later we will derive the Gibbard-Satterthwaite theorem from Arrow's theorem following a proof that is based on strict linear orders. Remember that utility functions had the pleasant property that they encoded exactly the desired preorders, whereas now we are forced to reason explicitly with typing properties like $L \in Lin$ to express that L is a strict linear order. This overhead complicates lemmas and proofs somewhat. What typically happens during proof construction is that some step fails to go through until one realizes that a typing assumption is missing. Once that is included, predicate calculus automation will usually do the rest.

For readability we introduce the abbreviation $a <_L b \equiv (a,b) \in L$.

The following functions manipulate preferences by moving alternatives around: *mktop* and *mkbot* resemble their namesakes in the previous section, *rem* is auxiliary.

$$mktop\ L\ b \equiv \{(x,y) \mid x <_L y \wedge x \neq b \wedge y \neq b\} \cup \{(x,b) \mid x \neq b\}$$
$$mkbot\ L\ b \equiv \{(x,y) \mid x <_L y \wedge x \neq b \wedge y \neq b\} \cup \{(b,y) \mid y \neq b\}$$
$$rem\ L\ a \equiv \{(x,y) \mid x <_L y \wedge x \neq a \wedge y \neq a\}$$

Function *below* moves a immediately below b:

$$below\ L\ a\ b \equiv$$
$$\{(x,y) \mid x <_L y \wedge x \neq a \wedge y \neq a\} \cup \{(a,b)\} \cup \{(x,a) \mid x <_L b \wedge x \neq a\} \cup$$
$$\{(a,y) \mid b <_L y \wedge y \neq a\}$$

Function *above* moves b immediately above a:

$$above\ L\ a\ b \equiv$$
$$\{(x,y) \mid x <_L y \wedge x \neq b \wedge y \neq b\} \cup \{(a,b)\} \cup \{(x,b) \mid x <_L a \wedge x \neq b\} \cup$$
$$\{(b,y) \mid a <_L y \wedge y \neq b\}$$

Profiles and *social welfare functions* need to be defined as sets as well:

$$Prof \equiv I \to Lin$$
$$SWF \equiv Prof \to Lin$$

where I is defined as the set of all individuals and $\to$ is the predefined set-level function space constructor: $A \to B \equiv \{f \mid \forall a{\in}A.\ f\,a \in B\}$.

The notion of a dictator w.r.t. a social welfare function is defined in the obvious manner:

$$dictator\ F\ i \equiv \forall P{\in}Prof.\ F\,P = P\,i$$

5.2 The Proof

Given a social welfare function F subject to unanimity and IIA

(U) $\forall P{\in}Prof.\ \forall a\ b.\ (\forall i.\ a <_{P\,i} b) \longrightarrow a <_{F\,P} b$

(IIA) $\forall P,P'{\in}Prof.\forall a\ b.\ (\forall i.\ a <_{P\,i} b \longleftrightarrow a <_{P'\,i} b) \longrightarrow (a <_{F\,P} b \longleftrightarrow a <_{F\,P'} b)$

we need to show that there is a dictator w.r.t. F. The key lemma is *Pairwise Neutrality*:

Lemma 3 (Pairwise Neutrality) *If $a \neq b$ and $a' \neq b'$ and $P \in Prof$ and $P' \in Prof$ and $\forall i.\ a <_{P\,i} b \longleftrightarrow a' <_{P'\,i} b'$ then $a <_{F\,P} b \longleftrightarrow a' <_{F\,P'} b'$.*

Proof I could not prove this lemma outright but needed case distinctions on $a = b'$ and $b = a'$. This results in a sequence of similar but more specialized propositions. To emphasize the similarity, we abbreviate the premises of the pairwise neutrality lemma by H. The first case is $a \neq b'$ and $b \neq a'$, first in one direction only:

(1) *If* $a \neq b'$ *and* $b \neq a'$ *and* H *and* $a <_{F\,P} b$ *then* $a' <_{F\,P'} b'$.

For the proof we define a new profile Q from P by moving, for each individual, a' just below a (if $a \neq a'$) and b' just above b (if $b \neq b'$):

$Q\ i =$
(*let* $L =$ *if* $a = a'$ *then* $P\ i$ *else below* $(P\ i)\ a'\ a$ *in if* $b = b'$ *then* L *else above* $L\ b\ b'$)

This does not change the ranking of a relative to b and hence $a <_{F\,Q} b$ by IIA. Unanimity implies $a' <_{F\,Q} a$ if $a \neq a'$ and $b <_{F\,Q} b'$ if $b \neq b'$ (by definition of Q). Therefore $a' \leq_{F\,Q} a$ and $b \leq_{F\,Q} b'$. Transitivity implies $a' <_{F\,Q} b'$. By definition of Q we have $\forall i.\ a <_{P\,i} b \longleftrightarrow a' <_{Q\,i} b'$ because $a \neq b$, $a' \neq b'$, $a \neq b'$ and $a' \neq b$. From assumption $\forall i.\ a <_{P\,i} b \longleftrightarrow a' <_{P'\,i} b'$ we conclude $\forall i.\ a' <_{Q\,i} b' \longleftrightarrow a' <_{P'\,i} b'$. Because of $a' <_{F\,Q} b'$, IIA implies the desired $a' <_{F\,P'} b'$.

From (1) it trivially follows by symmetry that

(2) *If* $a \neq b'$ *and* $b \neq a'$ *and* H *then* $a <_{F\,P} b \longleftrightarrow a' <_{F\,P'} b'$.

Next we consider the case where $a = b'$ and $b = a'$:

(3) *If* $a \neq b$ *and* $P \in$ *Prof and* $P' \in$ *Prof and* $\forall i.\ a <_{P\,i} b \longleftrightarrow b <_{P'\,i} a$ *then* $a <_{F\,P} b \longleftrightarrow b <_{F\,P'} a$.

For the proof we pick some distinct third alternative c. Taking P, move c directly below b, obtaining Q. Thus $\forall i.\ a <_{P\,i} b \longleftrightarrow a <_{Q\,i} c$ and hence $a <_{F\,P} b \longleftrightarrow a <_{F\,Q} c$. Taking Q, move b directly below a, obtaining R. Taking R, move a directly below c, obtaining S. Just like above, it follows that $a <_{F\,Q} c \longleftrightarrow b <_{F\,R} c$ and $b <_{F\,R} c \longleftrightarrow b <_{F\,S} a$. Comparing S with the original P we find $\forall i.\ b <_{S\,i} a \longleftrightarrow a <_{P\,i} b$ and hence by assumption we have $\forall i.\ b <_{S\,i} a \longleftrightarrow b <_{P'\,i} a$ and thus by IIA $b <_{F\,S} a \longleftrightarrow b <_{F\,P'} a$. Altogether this yields the desired $a <_{F\,P} b \longleftrightarrow b <_{F\,P'} a$.

Now consider the case where $|\{a,b,a',b'\}| = 3$:

(4) *If distinct* $[a, b, c]$ *and* $P \in$ *Prof and* $P' \in$ *Prof and* $\forall i.\ a <_{P\,i} b \longleftrightarrow b <_{P'\,i} c$ *then* $a <_{F\,P} b \longleftrightarrow b <_{F\,P'} c$.

For the proof define Q as the pointwise converse of P: $Q\ i = (P\ i)^{-1}$. The assumptions imply $b <_{F\,Q} a \longleftrightarrow b <_{F\,P'} c$ by (2). Lemma (3) yields $a <_{F\,P} b \longleftrightarrow b <_{F\,Q} a$ and transitivity of $\longleftrightarrow$ yields the desired $a <_{F\,P} b \longleftrightarrow b <_{F\,P'} c$.

From (1–4) we can prove Pairwise Neutrality by an exhaustive case distinction: If $a \neq b'$ and $b \neq a'$, use (2), if $a = b'$ and $b \neq a'$ or if $a \neq b'$ and $b = a'$, use (4), and if $a = b'$ and $b = a'$, use (3). □

Now we begin the actual proof of Arrow's theorem. In the Isabelle proof we start by obtaining a bijection h between type *indi* and the set $\{0..N-1\}$ of natural numbers below N. In this presentation, however, we omit h and identify individuals with elements of the set $\{0..N-1\}$. We fix two distinct alternatives a and b and some $L \in Lin$ such that $a <_L b$. The existence of some such L appears nontrivial if *alt* is arbitrarily large, a complication Geanakoplos does not have to face because he assumes *alt* is finite. In the end I used the well-ordering theorem for HOL (which may be overkill): there is a well-order of type $(\alpha \times \alpha)$*set*

for any type α. Hence there is in particular a strict linear order, from which one obtains L by moving a directly below b. Similarly we obtain some $L' \in Lin$ such that $b <_{L'} a$. Now we define a sequence of profiles $\Phi\ n = (\lambda i.\ \mathit{if}\ i < n\ \mathit{then}\ L\ \mathit{else}\ L')$, $n \leq N$. That is, in $\Phi\ 0$ every individual has preference L', and then they change to L one by one, until all of them have preference L in $\Phi\ N$. By unanimity we have $\neg\ a <_{F\ (\Phi\ 0)} b$ but $a <_{F\ (\Phi\ N)} b$. Hence there must exist some $n < N$ such that $\forall m \leq n.\ b <_{F\ (\Phi\ m)} a$ but $a <_{F\ (\Phi\ (n+1))} b$. Now we will see that this n did not just force the flip but is actually the dictator.

It is easy to see that dictatorship of some n can be shown as follows: Let P be a profile; let c and d be two distinct alternatives such that $c <_{P\ n} d$; show $c <_{F\ P} d$. For this proof let e be a third alternative distinct from c and d. Modify $P\ i$ as follows: move e to the top if $i < n$, move e directly above c if $i = n$, and move e to the bottom if $i > n$. Call the result Q. Since we have not moved c or d we have $\forall i.\ c <_{P\ i} d \longleftrightarrow c <_{Q\ i} d$ and hence $c <_{F\ P} d \longleftrightarrow c <_{F\ Q} d$ by IIA. Thus it now suffices to show $c <_{F\ Q} d$. By definition we have $\forall i.\ c <_{Q\ i} e \longleftrightarrow a <_{\Phi\ (n+1)\ i} b$ and hence $c <_{F\ Q} e \longleftrightarrow a <_{F\ (\Phi\ (n+1))} b$ by Pairwise Neutrality. Because $a <_{F\ (\Phi\ (n+1))} b$ by construction, we have $c <_{F\ Q} e$. Similarly we have $\forall i.\ e <_{Q\ i} d \longleftrightarrow b <_{\Phi\ n\ i} a$ by definition, and hence $e <_{F\ Q} d \longleftrightarrow b <_{F\ (\Phi\ n)} a$ by Pairwise Neutrality. Because $b <_{F\ (\Phi\ n)} a$ by construction, we have $e <_{F\ Q} d$. By transitivity the desired $c <_{F\ Q} d$ follows. This concludes the proof that n is the dictator.

5.3 Comparison

It turns out that both Isabelle proofs have roughly the same size: 350 lines for the first, 300 lines for the third proof. This is what the size of the journal proofs would lead us to expect: both take roughly 1 page. More interesting is the comparison with the proofs in the literature.

5.3.1 Geanakoplos 2001

My first attempt, in 2002, to prove Arrow's theorem in Isabelle was based on Geanakoplos' technical report [4]. Initially I had been drawn to his third proof, which seemed the briefest, but was something of a riddle. Hence I eventually turned towards his first proof. The result was shown in the previous section. The third proof [4] suffered from two problems I emailed to Geanakoplos in July 2002:

- The pairwise neutrality lemma talks about two profiles, but the text leaves this implicit, which is confusing.
- The case distinction in the proof of the pairwise neutrality lemma is incomplete. It does not cover the cases $a = b'$ or $b = a'$ (which give rise to (3) and (4) above).

5.3.2 Geanakoplos 2005

In the published version [5], the pairwise neutrality lemma has been revised. It talks about the two profiles explicitly. And the proof distinguishes two cases. The second case is clear enough: $a = b' \wedge b = a'$, and proposition (3) above corresponds to it. The first case is:

> Take $c \notin \{a, b\}$ and suppose first that $(a', b') = (a, c)$ or $(a', b') = (c, b)$ or $(a', b') = (c, d)$ with $d \notin \{a, b, c\}$.

After a lot of head-scratching and with the help of Isabelle I decided that this is just a complicated way of saying $a \neq b' \wedge b \neq a'$, under the assumptions $a \neq b$, $a' \neq b'$ and $(a, b) \neq (a', b')$ (the trivial case). Thus the second case in the case distinction should have been $a = b' \vee b = a'$ rather than $a = b' \wedge b = a'$. Put differently, proposition (4) above is missing.

5.3.3 Nisan 2007

Nisan follows Geanakoplos' third proof [5]. His proof of the pairwise neutrality lemma is particularly brief. It starts off by asserting that

> By renaming we can assume without loss of generality that $a' \neq b$.

As it was unclear to me how this could cover the case $(a, b) = (b', a')$, I asked Nisan by email and received the reply that

> $(a, b) = (b', a')$ cannot happen since the assumption is that $a <_{P\,i} b \longleftrightarrow a' <_{P\,i} b'$.

Unfortunately the correct assumption is $a <_{P\,i} b \longleftrightarrow a' <_{P'\,i} b'$, which is not inconsistent with $(a, b) = (b', a')$. Moreover, the rest of his proof fails in case $b' = a$. Another case of missing cases.

6 The Gibbard-Satterthwaite Theorem

The Gibbard-Satterthwaite theorem [6,11] is another staple of social choice theory. It concerns manipulability of social choice functions, for example voting systems. In a nutshell, the only non-manipulable social choice function is a dictatorship. Traditionally, the Gibbard-Satterthwaite theorem is proved as a corollary to Arrow's theorem, and this is also the path that we will follow. At the same time it should be mentioned that both theorems are often regarded as "equivalent". Taylor [13] writes:

> We say that two theorems are equivalent if each is "easily derivable" from the other, where the ease of the derivation is measured (intuitively) relative to the difficulty of the stand-alone proofs of the theorems whose equivalence is being asserted.

If difficulty is measured in terms of the length of a computer-checked proof, it is not clear whether the two theorems are equivalent in the above sense: my derivation of Gibbard-Satterthwaite from Arrow is only 20% shorter than my direct proofs of Arrow. And although I have not proved Gibbard-Satterthwaite directly as well, it is quite likely that such a proof would be roughly as long as my proofs of Arrow: adapting ideas of Geanakoplos' first proof, Reny [9] presents two proofs of the two theorems side by side which are "essentially identical".

As with Arrow's theorem, formulations of the Gibbard-Satterthwaite theorem differ; we roughly follow Nisan [8]. He phrases it in terms of *social choice functions*, i.e. functions from profiles to single alternatives. Such a choice function is *manipulable* if there is a profile where some individual can enforce an alternative more to his liking by modifying his preferences, i.e. by misrepresenting his preferences. That is, assume there is a situation/profile where some individual i ranks a below b and the social choice is a. If i can, by modifying his original and true preferences somehow, cause the social choice to become b, the social choice function is *manipulable*.

Theorem 2 (Gibbard-Satterthwaite) *Any non-manipulable social choice function onto a set of at least three alternatives is a dictatorship.*

We prove this theorem in the context of the same setup of individuals, alternatives and profiles as in Geanakoplos' third proof in the previous section. The letter f will denote a social choice function (of type $(\mathit{indi} \Rightarrow \mathit{pref}) \Rightarrow \mathit{alt}$).

The precise definitions of ontoness and manipulability are as follows:

$$\begin{array}{lcl} \mathit{onto}\, f & \equiv & \forall a.\, \exists P{\in}\mathit{Prof}.\, a = f\, P \\ \mathit{manipulable}\, f & \equiv & \exists P{\in}\mathit{Prof}.\, \exists i.\, \exists L{\in}\mathit{Lin}.\, f\, P <_{P\, i} f\, (P(i := L)) \end{array}$$

We immediately prove the following characterization of *non-manipulability*:

Lemma 4 (Non-manipulability)

$\neg\, \mathit{manipulable}\, f \longleftrightarrow$
$(\forall P{\in}\mathit{Prof}.$
$\quad \forall i.\, \forall L{\in}\mathit{Lin}.\, f\, P \neq f\, (P(i := L)) \longrightarrow f\, (P(i := L)) <_{P\, i} f\, P \wedge f\, P <_L f\, (P(i := L)))$

Proof In the $\longrightarrow$ direction we assume $\neg$ *manipulable* f and show that if $P \in \mathit{Prof}$, $L \in \mathit{Lin}$ and $f\, P \neq f\, (P(i := L))$, then

$f\, (P(i := L)) <_{P\, i} f\, P$ follows from $\neg$ *manipulable* f by strict linearity and
$f\, P <_L f\, (P(i := L))$ also follows from $\neg$ *manipulable* f, where P is $P(i := L)$ and L is $P\, i$, by strict linearity because $P(i := L, i := P\, i) = P$.

In the $\longleftarrow$ direction we assume the right-hand side (RHS) of the equivalence and show $\neg$ *manipulable* f indirectly. Assume *manipulable* f, that is, $P \in \mathit{Prof}$, $L \in \mathit{Lin}$ and $f\, P <_{P\, i} f\, (P(i := L))$. Because $f\, P <_{P\, i} f\, (P(i := L))$ implies $f\, P \neq f\, (P(i := L))$, (RHS) implies $f\, (P(i := L)) <_{P\, i} f\, P$, which contradicts $f\, P <_{P\, i} f\, (P(i := L))$. □

6.1 The Proof

The following proof of the Gibbard-Satterthwaite theorem takes place in the context of a social choice function f that is non-manipulable and onto. We need to show that it is a dictatorship, which is defined for social choice functions as follows:

$$\mathit{dict}\, f\, i \equiv \forall P{\in}\mathit{Prof}.\, \forall a.\, a \neq f\, P \longrightarrow a <_{P\, i} f\, P$$

The winner must be at the top of the dictator's ranking.

As a first lemma we prove *monotonicity* from non-manipulability: if nobody downgrades the winner, he stays the winner.

Lemma 5 (Monotonicity) *If* $P \in \mathit{Prof}$ *and* $P' \in \mathit{Prof}$ *and* $\forall i\, a.\, a <_{P\, i} f\, P \longrightarrow a <_{P'\, i} f\, P$ *then* $f\, P' = f\, P$.

Proof As in the proof of Arrow's theorem we identify individuals with elements of the set $\{0..N{-}1\}$ for reasons of readability. We define a sequence of profiles $\Phi\ 0, \ldots, \Phi\ N$ such that $\Phi\ 0 = P$, $\Phi\ N = P'$, and $\Phi\ (n+1)$ is $\Phi\ n$ where individual n has changed his mind from $P\ n$ to $P'\ n$. A direct definition is $\Phi\ n = (\lambda i.\ \mathit{if}\ i < n\ \mathit{then}\ P'\ i\ \mathit{else}\ P\ i)$. We show by induction on n that $n \leq N \longrightarrow f\, (\Phi\ n) = f\, P$. Setting $n = N$ yields the desired $f\, P' = f\, P$ because $i < N$ for every individual. The base case $n = 0$ of the induction is trivial. In the induction step let $Q = (\Phi\ n)(n := P'\ n)$. It follows that $Q = \Phi\ (n+1)$. We prove $f\, (\Phi\ (n+1)) = f\, P$ by contradiction. From $f\, (\Phi\ (n+1)) \neq f\, P$ and $Q = \Phi\ (n+1)$ and the induction hypothesis $f\, (\Phi\ n) = f\, P$ we obtain $f\, (\Phi\ n) \neq f\, Q$. By the non-manipulability lemma we have $f\, Q <_{\Phi\ n\ n} f\, (\Phi\ n)$ and $f\, (\Phi\ n) <_{P'\ n} f\, Q$. Because $\Phi\ n\ n = P\ n$ and $f\, (\Phi\ n) = f\, P$, the assumption of the monotonicity lemma and $f\, Q <_{\Phi\ n\ n} f\, (\Phi\ n)$ yield $f\, Q <_{P'\ n} f\, (\Phi\ n)$, thus contradicting $f\, (\Phi\ n) <_{P'\ n} f\, Q$. □

Now we come to the main construction in the proof. In order to apply Arrow's theorem, we turn our social choice function into a social welfare function. The idea is simple: in order to determine if socially $a < b$, move a and b to the top of each individual's preference and check if that leads to b being chosen. First we define how to move a whole set of alternatives to the top of an ordering without reordering them w.r.t. each other:

$$Top\ S\ L \equiv$$
$$\{(a,b) \mid a <_L b \wedge (a \in S \wedge b \in S \vee a \notin S \wedge b \notin S)\} \cup \{(a,b) \mid a \notin S \wedge b \in S\}$$

For example, if $a <_L b <_L c <_L d$ and $L' = Top\ \{a, c\}\ L$ then $b <_{L'} d <_{L'} a <_{L'} c$.

Extending *Top S* from a single ordering to a profile P is easy: $Top\ S \circ P$. This allows a compact formalization of the above sketch of the translation of social choice to social welfare functions:

$$swf\ f \equiv \lambda P.\ \{(a,b) \mid a \neq b \wedge f\ (Top\ \{a,b\} \circ P) = b\}$$

Note that *swf* is a function that is applied to a social welfare function f.

The following set-valued unanimity lemma can be formalized similarly concisely: if everybody puts the elements of S at the top of their preference, f will choose a member of S:

Lemma 6 (Top-unanimity) *If* $P \in Prof$ *and* $S \neq \emptyset$ *then* $f\ (Top\ S \circ P) \in S$.

Proof Since $S \neq \emptyset$ there is some $a \in S$. Because f is onto, there is some profile P_a such that $f\ P_a = a$. We define a sequence of profiles $\Phi\ 0, \ldots, \Phi\ N$:

$$\Phi\ n = (\lambda i.\ \mathit{if}\ i < n\ \mathit{then}\ Top\ S\ (P\ i)\ \mathit{else}\ P_a\ i)$$

That is, $\Phi\ 0 = P_a$, $\Phi\ N = Top\ S \circ P$, and in between individuals change their mind one by one. We show by induction on n that $n \leq N \longrightarrow f\ (\Phi\ n) \in S$. Setting $n = N$ yields the desired $f\ (Top\ S \circ P) \in S$. The base case $n = 0$ is trivial because $a \in S$. In the induction step we make a case distinction. If $f\ (\Phi\ n) = f\ (\Phi\ (n+1))$, the claim holds by induction hypothesis. If $f\ (\Phi\ n) \neq f\ (\Phi\ (n+1))$, then the non-manipulability lemma yields $f\ (\Phi\ n) <_{Top\ S\ (P\ n)} f\ (\Phi\ (n+1))$ (because $\Phi(n+1) = (\Phi\ n)(n := Top\ S\ (P\ n))$). Together with the induction hypothesis $f\ (\Phi\ n) \in S$ this implies that also $f\ (\Phi\ (n+1)) \in S$. □

Finally we can prove that *swf f* satisfies the premises of Arrow's theorem. First we show that *swf f* is actually a social welfare function. That is, given $P \in Prof$ we must prove $swf\ f\ P \in Lin$. Totality, i.e. $a <_{swf\ f\ P} b \vee b <_{swf\ f\ P} a$ for arbitrary a,b, follows from Top-unanimity because $f\ (Top\ \{a,b\} \circ P) \in \{a,b\}$. Irreflexivity, i.e. $\neg\ a <_{swf\ f\ P} a$, is trivial by definition of *swf*. For transitivity assume $a <_{swf\ f\ P} b$ and $b <_{swf\ f\ P} c$, i.e. $a \neq b, f\ (Top\ \{a,b\} \circ P) = b$, $b \neq c$ and $f\ (Top\ \{b,c\} \circ P) = c$. We need to prove $a <_{swf\ f\ P} c$, i.e. $a \neq c$ and $f\ (Top\ \{a,c\} \circ P) = c$. Clearly $a = c$ contradicts the assumptions. Now observe that $f\ (Top\ \{a,b,c\} \circ P) = a$ is contradictory because monotonicity implies $f\ (Top\ \{a,b\} \circ P) = a$ (because $a \in \{a,b\} \subseteq \{a,b,c\}$, the premise of monotonicity holds), which contradicts the assumptions. Similarly one shows that $f\ (Top\ \{a,b,c\} \circ P) = b$ is contradictory. By Top-unanimity this leaves only $f\ (Top\ \{a,b,c\} \circ P) = c$. By monotonicity we obtain the desired $f\ (Top\ \{a,c\} \circ P) = c$. This concludes the proof that *swf f* is what it claims to be, a social welfare function.

Next we have to show that *swf f* satisfies unanimity. Assume $P \in Prof$ and $\forall i.\ a <_{P\ i} b$ (and thus $a \neq b$). Therefore $Top\ \{a,b\} \circ P = Top\ \{b\} \circ Top\ \{a,b\} \circ P$ and $f(Top\ \{b\} \circ Top\ \{a,b\} \circ P) = b$ by Top-unanimity. Hence $f(Top\ \{a,b\} \circ P) = b$ and thus $a <_{swf\ f\ P} b$.

We also need to show that *swf f* satisfies IIA. Assume $P,P' \in Prof$ and $\forall i.\ a <_{P\ i} b \longleftrightarrow a <_{P'\ i} b$. Let $Q = Top\ \{a,b\} \circ P$ and $Q' = Top\ \{a,b\} \circ P'$. Let us first prove

$$(*)\ \ \forall i\ c.\ c <_{Q\ i} f\ Q \longrightarrow c <_{Q'\ i} f\ Q.$$

By Top-unanimity we have $f\ Q \in \{a,b\}$. Thus, if $c \notin \{a,b\}$ then $c <_{Q'\ i} f\ Q$. If $c = a$, then $c <_{Q\ i} f\ Q$ implies $f\ Q = b$ and $a <_{P\ i} b$ and thus (by assumption) $a <_{P'\ i} b$ and hence $c <_{Q'\ i} f\ Q$. Symmetrically for $c = b$. Thus $(*)$ holds and we can appeal to monotonicity to deduce $f\ Q = f\ Q'$. By definition of *swf* we obtain the desired $a <_{swf\ f\ P} b \longleftrightarrow a <_{swf\ f\ P'} b$.

Having proved that *swf f* satisfies the premises of Arrow's theorem we obtain a dictator for *swf f*. As a last step we prove that this individual is also a dictator for f:

If dictator $(swf\,f)$ i *then dict* $f\,i$.

We assume *dictator* $(swf\,f)$ i, $P \in Prof$ and $a \neq f\,P$. By monotonicity this implies $f\ (Top\ \{a, f\,P\} \circ P) = f\,P$. Because i is a dictator for $swf\,f$ we get $P\,i = \{(a, b) \mid a \neq b \wedge f\ (Top\ \{a, b\} \circ P) = b\}$. The two facts together with $a \neq f\,P$ imply the desired $a <_{P\,i} f\,P$. Thus we have finally proved the conclusion of the Gibbard-Satterthwaite theorem:

$\exists i.\ dict\,f\,i$

6.2 Comparison

The proof above differs from Nisan's in a few points.

- What Nisan calls monotonicity and proves equivalent to non-manipulability I have called non-manipulability in the first place. My notion of monotonicity agrees with that of Reny [9] and others.
- Nisan (re)proves instances of the monotonicity lemma wherever I use the monotonicity lemma.
- There are a few smaller omissions in Nisan's proof: for example, the precondition $S \neq \emptyset$ in Top-unanimity, and the requirement $a \neq b$ in the definition of the social welfare function from a social choice function (making Nisan's social welfare function return reflexive orderings). The only larger omission is in the proof of the implication *dictator* $(swf\,f)$ $i \longrightarrow dict\,f\,i$ where he just writes "obvious". However, it is not clear to me how one can avoid the argument I made above, which requires monotonicity, and hence in his setting another small inductive proof. Since he spells these inductions out in all other cases, it is unclear why he did not do so in this instance, too.

7 Related work

Wiedijk [14, 15] independently formalized Geanakoplos' first proof with the help of the Mizar system. Of course he also discovered the omission we discussed in section 4.3, and fixed it the same way. His formalization takes about 1000 lines of Mizar text, roughly 3 times as much as the Isabelle version. This is unsurprising because of Isabelle's higher level of automation.

Peter Gammie independently formalized Arrow's theorem and related results in Isabelle in 2006. This work has recently been made available in the Archive of Formal Proofs [3]. He draws on Sen's landmark work [12].

8 Conclusion

Social choice theory turns out to be perfectly suitable for mechanical theorem proving. The concepts are directly representable in set theory and the traditional formulations leave something to be desired in thoroughness. There is a large body of knowledge that should be translated systematically instead of diving in to prove a few main results. However, it is unclear if this will lead to new insights into either social choice theory or theorem proving.

Acknowledgements Jasmin Christian Blanchette suggested $\lesssim$ and many other improvements of this article. Two anonymous referees contributed considerably to an improved presentation of the paper with their painstaking reviews. Peter Gammie commented on and discussed a draft version.

I would like to thank Mike Gordon for TR 68, which was a profound revelation and inspiration for me.

References

1. Kenneth Arrow. A difficulty in the concept of social welfare. *The Journal of Political Economy*, 58:328–346, 1950.
2. Kenneth Arrow. *Social Choice and Individual Values*. 1951.
3. Peter Gammie. Some classical results in social choice theory. In Gerwin Klein, Tobias Nipkow, and Lawrence Paulson, editors, *The Archive of Formal Proofs*. `http://afp.sf.net/entries/SenSocialChoice.shtml`, 2008. Formal proof development.
4. John Geanakoplos. Three brief proofs of Arrow's impossibility theorem. Technical report, Cowles Foundation Discussion Paper No. 1123RRR, 2001.
5. John Geanakoplos. Three brief proofs of Arrow's impossibility theorem. *Economic Theory*, 26:211–215, 2005.
6. Allan Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41:587–601, 1973.
7. Tobias Nipkow, Lawrence Paulson, and Markus Wenzel. *Isabelle/HOL — A Proof Assistant for Higher-Order Logic*, volume 2283 of *Lecture Notes in Computer Science*. Springer, 2002.
8. Noam Nisan. Introduction to mechanism design (for computer scientists). In N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, editors, *Algorithmic Game Theory*. Cambridge University Press, 2007.
9. Philip Reny. Arrow's theorem and the Gibbard-Satterthwaite theorem: A unified approach. *Economics Letters*, 70:99–105, 2001.
10. Richard Routley. Repairing proofs of Arrow's general impossibility theorem and enlarging the scope of the theorem. *Notre Dame Journal of Formal Logic*, 20:879–890, 1979.
11. Mark Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.
12. Amartya Sen. *Collective Choice and Social Welfare*. Holden-Day, 1970.
13. Alan D. Taylor. *Social Choice and the Mathematics of Manipulation*. Cambridge University Press, 2005.
14. Freek Wiedijk. Arrow's impossibility theorem. *Formalized Mathematics*, 15:171–174, 2007.
15. Freek Wiedijk. Formalizing Arrow's theorem. *Sadhana*, 34:193–220, February 2009.