universität innsbruck

Seminar Report

Formalization of Social Choice Theory

Julian Parsert
julian.parsert@uibk.ac.at

26 February 2018

Supervisor: Cezary Kaliszyk

Abstract

Kenneth Arrow showed that every voting scheme with desired properties must necessarily adhere to some undesired restrictions. This impossibility theorem was accompanied by a foundational approach of reasoning about human behavior and interaction in a mathematical way. In order to — among other reasons — attract attention of the social choice community Wiedijk formalized this theorem. We will discuss social choice theory and formalizations thereof as well as future work in the formalization of social choice theory and related areas.

Contents

1 Introduction **1**

2 Individual Preferences and Related Concepts **1**

3 Social Choice Theory **3**

4 Arrows' Impossibility Theorem **5**

5 Discussion **7**

Bibliography **9**

1 Introduction

Mechanisms of how a group of individuals come to a decision have always been of high interest. Already in ancient Rome the senator Pliny describes an issue with voting, which nowadays is known as *election control by deleting of candidates*. As the name suggests, this describes how strategically voting for a less than optimal option might improve the outcome for an individual. Later, the french revolutionary Marquis de Condorcet also described multiple problems with different election systems, which will be discussed in subsequent sections.

Up to the middle of the twentieth century, social choice theory consisted mostly of describing voting systems and finding problems with them. However, with the development of formal game theory by von Neumann and Morgenstern social choice theory also went through a paradigm shift. This change is mostly due to the work of Arrow and others. In 1951 he showed that the problems described by Pliny, Condorcet, etc. have a more fundamental nature. This result would later be known as *Arrow's Impossibility Theorem* and will be the main topic of Section 4. However, Arrow's work did not only provide a fundamental result, but also provided a mathematical framework for reasoning about social choice. Indeed, all of these groundbreaking works in game theory and social choice theory specified mathematical terms and definitions for social, political, and economic concepts. Using an axiomatic approach logical consequences were rigorously derived and interpreted [2].

In the 21st century social choice has mainly sparked interest for computer scientists in two ways. First of all, the formalizing of social choice theory results in proof assistants, and secondly the founding of computational social choice [2]. Here, we will discuss the former. In particular we consider formalizations of *Arrow's theorem*. This includes Freek Wiedijk's [13] formalization in Mizar. The same theorem was also formalized by Nipkow in Isabelle [6] and Gammie [4]. The latter also formalized *May's Theorem* and *Sen's Theorem* as well as other *Social Choice Theory* results, which we will not discuss here. Same holds for Eberl [3], who formalized multiple concepts of *Randomised Social Choice Theory*. A computer aided proof has also been conducted by Tang and Lin [10].

Content In the next section, we will discuss preference relations in general. Subsequently, we will introduce the notion of social choice and definitions of social choice theory in Section 3. In Section 4 we will introduce Arrow's impossibility theorem. We will discuss formalizations in all sections separately. Finally, in Section 5 we discuss where further work could go and why formalization of social choice theory deserves further exploration.

2 Individual Preferences and Related Concepts

Preference relations are a fundamental concept in many areas related to game theory and social choice theory. In essence preference relations create a ranking between multiple alternatives and are supposed to model the preference or taste of a certain individual. Hence, they play a fundamental role in areas such as economics, game theory, and

of course social choice theory, where a the modeling of human behavior is intended. Formally, a preference relation is a preorder. Often, we assume a rational preference relation, which adds the assumption of totality to the preorder. The notion of preference relations has been formalized in some instances. One can be found in [8]. A definition of such preferences in Isabelle/HOL could look similar to the following.

locale *rational-preference* =
  fixes *carrier* :: δ *set*
  fixes *relation* :: δ *relation*
  assumes *preorder-on carrier*
  assumes *total-on carrier*
  assumes $x \succeq[relation]\ y \implies x \in carrier$
  and $x \succeq[relation]\ y \implies y \in carrier$

In fact this is the how it is formalized in [8]. Hence, any additional information about Isabelle mechanics and design choices can be found there. We will continue using the notation $\succeq$ for a preference relation and $\succeq_i$ to denote the preference relation of individual i in case there is more than one individual. In economic practice however, preference relations are a cumbersome concepts and highly impractical. Instead, the notion of utility functions is used. Utility functions allow for the use of mathematical machinery that is not possible with binary relations, such as optimization etc. Utility functions are defined as follows:

Definition 2.1 (Utility function)**.** A function $u : C \mapsto \mathbb{R}$ is said to represent a preference relation P, if

$$\forall x\ y \in C.\ x \succeq [P]\ y \iff u(x) \geq u(y).$$

The function u is a utility function representing the preference relation P.

Interestingly, we do not need to specify any of the axioms of rational preferences for *utility-function*, since reflexivity, transitivity, and totality implicitly follow from the function u.

locale *utility-function* =
  fixes *carrier* :: δ *set*
  fixes *relation* :: δ *relation*
  fixes u :: $\delta \Rightarrow \mathbb{R}$
  assumes $x \in carrier \implies y \in carrier \implies$
    $x \succeq[relation]\ y \iff u\ x \geq u\ y$
  assumes $x \succeq[relation]\ y \implies x \in carrier$
  and $x \succeq[relation]\ y \implies y \in carrier$

The main idea behind these concepts is, that behavior can be modeled using different properties of utility functions and preference relations. For example, the human behavior that could adequately be described as “more is always better” corresponds to the mathematical notion of monotonicity. Hence, this can be modeled by a monotonic utility

function. By assuming that agents will always try to maximize their utility, we can derive what decision an agent is going to make, given a utility function.

If one would continue further in the direction of game theory, and algorithmic game theory, the detailed study of utility functions and different forms of utility functions would be a necessity. However, since we are interested in social choice theory, we are not going to explain further details of utility theory.

3 Social Choice Theory

In the previous section we discussed how individuals come to decisions and how we can compare potential decisions and related outcomes. Now, we will consider the concept of social choice. As the name suggests, in social choice theory we are concerned with how a group of individuals come to a decision. Clearly, the most prominent practical example is that of voting for a political party, where each individual has a personal ranking of the options and a voting mechanism is supposed to aggregate all these rankings and combine them to a single one. We will introduce important definitions of Social Choice Theory before we will focus on Arrow's Impossibility theorem in the next section.

In many books such as[7] the authors consider.

Important Definitions

Previously we discussed preference relations. However, now we want to aggregate a set of preference relations that belong to a set of individuals $\mathcal{I}$. We will call the "set of preferences" a preference profile. Such a function $\mathcal{F}$ maps a preference profile of size n[1] to a single preference relation. We call such a function a social welfare function. Note that the set $\mathcal{P}$ is the set of all rational preference relations (cf. total preorder) on the carrier set $\mathcal{C}$. Hence a social welfare function is defined as follows:

Definition 3.1 (Social Welfare Function (SWF))**.** A social welfare function, $\mathcal{F}$, is a mapping $\mathcal{F} : \mathcal{P}^n \to \mathcal{P}$. We will use the shorthand notation $\succeq_{\mathcal{F}}$ denote the binary (preference) relation returned by $\mathcal{F}$.

A related concept is that of social choice functions. In contrary to SWFs, a social choice function does not map to a preference relation, but rather a single element $e \in \mathcal{C}$.

Definition 3.2 (Social Choice Function (SCF))**.** A social choice function, $\boldsymbol{f}$, is a mapping $\boldsymbol{f} : \mathcal{P}^n \to \mathcal{C}$, where n is the number of individuals.

These functions are also sometimes called constitution, since they define a set of rules on how to calculate the output based on the input — much like the constitution of a country defines how elections take place. Like any function we can now define properties of these mechanisms. Since Arrow's theorem concerns itself mostly with social welfare functions, we are going to focus on them. Rest assured, similar results also hold for social choice functions.

[1] n is the number of individuals, hence $|\mathcal{I}| = n$.

Since dictatorial SWFs are the most natural, we will start with them. As the name suggests a SWF is dictatorial, if there is a single individual that always determines the result of the social welfare function.

Definition 3.3 (Dictatorial SWF)**.** A Social Welfare Function $\mathcal{F}$ is dictatorial if there exists an individual $i \in \mathcal{I}$ such that

$$\forall x\ y \in \mathcal{C}.\ x \succ_i y \rightarrow x \succ y.$$

Alternatively, $\mathcal{F}$ is dictatorial for i if $\mathcal{F}$ is a projection π_i^n.

Clearly, this corresponds to the natural intuition what a dictatorial constitution looks like. Example 1 illustrates such a dictatorial SWF.

Example 3.4. If there exists an individual $i \in \mathcal{I}$ such that i is a dictator, if $\mathcal{F}(1, \ldots, n)$ mirrors $\succeq_i$. Table 1 describes such a dictator i.

1	...	i	...	n	$\mathcal{F}$
...	...	b	⋮	...	b
⋮	⋮	X	⋮	⋮	X
...	...	a	...	...	a
⋮	⋮	Y	⋮	⋮	Y
...	...	c	...	...	c
⋮	...	Z	...	...	Z

Table 1: Table describing a $\mathcal{F}$ mapping inputs $1 \ldots n$ to $\mathcal{F}(1, \ldots, n)$.

A similar concept is that of decisiveness. Essentially, decisiveness describes an individual that is a "dictator" for only one pair of alternatives.

Definition 3.5 (Decisive/Pivotal Individual)**.** An agent i is pivotal for $a, b \in \mathcal{C}$ if by changing the relation between a and b the SWF changes this relation as well. More formally, $a \succeq_i b \iff a \succeq_{\mathcal{F}} b$.

One can observe that an individual that is a dictator also is *decisive* for any pair $x\ y \in \mathcal{C}$. Conversely, if an individual is decisive for all pairs $x\ y \in \mathcal{C}$, that individual is a dictator.

Next, we define unanimity, which is also an intuitive property of preference relations. It simply states, that in case everyone agrees on a certain ranking between two alternatives, then $\mathcal{F}$ must incorporate this ranking.

Definition 3.6 (Unanimity)**.** If all individuals have the same relation between two alternatives a and b, the SWF has to reflect that relation. More formally, $\forall a\ b \in \mathcal{C}.\ (\forall i \in I.\ a \succeq_i b) \rightarrow a \succeq_{\mathcal{F}} b$

The last property of social welfare functions that we consider, is less intuitive. The assumption of independence of irrelevant alternatives states that the ranking between $a\ b \in \mathcal{C}$ depends only on how a and b are ranked by each individual.

Definition 3.7 (Independence of Irrelevant Alternatives (IIA))**.** The social preference between two alternatives a and b only depends on the voters' preferences between a and b. Formally, given two preference profiles $(P_1, \ldots, P_n) \in \mathscr{P}^n$ and $(Q_1, \ldots, Q_n) \in \mathscr{P}^n$ then

$$\begin{gathered}\forall a\ b \in \mathcal{C}.\ (\forall i \in I.\ a \succ_{P_i} b \iff a \succ_{Q_i} b) \\ \Longrightarrow \\ (a \succ_{\mathscr{F}(P_1,\ldots,P_n)} b \iff a \succ_{\mathscr{F}(Q_1,\ldots,Q_n)} b)\end{gathered}$$

Indeed, IIA has been highly disputed and has been subject to scrutiny. Hence, many alternative weaker definitions have been created [9]. Since we are interested in Arrow's impossibility result, we assume the concept as shown in Definition 3.7.

4 Arrows' Impossibility Theorem

In Section 1 we showed that throughout history multiple attempts were made at describing an adequate voting scheme. However, non of them succeeded. As it turns out this problem is more fundamental in nature as shown in [1].

With his work on social choice theory in [1], Kenneth Arrow essentially founded the field of social choice theory. The most famous result is Arrow's Impossibility Theorem, which states that there does not exist a voting scheme that adheres to the previously defined properties. The formal statement is Theorem 4.1.

Theorem 4.1 (Arrow's Impossibility Theorem)**.** *Any Social Welfare Function with more than 2 alternatives that respects rationality*[2]*, independence of irrelevant alternatives (IIA) and unanimity is dictatorial.*

The significance of this result undoubtedly is the main contributing factor to the significant amount of formalization attempts. Indeed, this theorem has been formalized by multiple people. Most notably, Wiedijk [13] who even motivates his Mizar formalization with the intend to spark interest for theorem proving in the social choice community. There has also been a formalization in the proof assistant Isabelle [11] by Nipkow [6] as well as [4]. An additional computer aided proof was conducted by Tang and Lin [10]. It is worth noting, that while the former formalizations mainly formalized existing paper proofs, Tang and Lin developed a new proof. Their method reduced the problem to a base case of two individuals and three choices which was proven using a computer aided approach.

We will look at several common voting schemes and discuss certain properties and potential flaws.

[2]We did not specifically introduce this condition for SWFs, instead we use the definition for standard preference relations.

Example 4.2 (Naive implementation)**.** We consider the naive idea of iteratively letting two options "compete" against each other until a winner is found. Given three choices a, b, and c we can compare a and b, b and a, and finally a and c. In case we have a preference profile such as in Table 2 we would derive a SWF that contradicts transitivity.

c	a	b	$c \succ b$
b	c	a	$b \succ a$
a	b	c	$a \succ c$
			$\lightning$

Table 2: Preference profile that would lead to a contradiction, since $c \succ c$.

As seen in the previous example, the naive implementation is already flawed. However, one would think that simply adding the possibility of indifferent preferences would solve this problem.

Example 4.3 (Allow Indifference)**.** We start with the same configuration as in the previous example. Now we allow for an indifference, thus circumventing the inconsistency shown in Example 4.2. However, this still leads to a contradiction with the assumption of independence of irrelevant alternatives. By switching the relation between a and b the social welfare function also switches the relation with c which by IIA is supposed to be an irrelevant alternative.

c	a	b	$c \approx b$
b	c	a	$b \approx a$
a	b	c	

c	a	a	$a \succ b$
b	c	b	$c \succ b$
a	b	c	$a \succ c$
			$\lightning$(IIA)

Table 3: By switching the preference for a and b the SWF also switches the relation between c a and c b, which contradicts IAA.

The formal statement in the Isabelle/HOL formalization of Nipkow [6] can be found in the Archive of Formal Proofs[3]. An excerpt is shown here:

locale *arrow* =
  fixes $F :: prof \Rightarrow pref$
  assumes *unanimity* : $(\bigwedge i.\ P\ i\ a < P\ i\ b) \Longrightarrow F\ P\ a < F\ P\ b$
  assumes IAA : $(\bigwedge i.\ (P\ i\ a < P\ i\ b) = (F\ P'\ a < F\ P'\ b)) \Longrightarrow$
    $(F\ P\ a < F\ P\ b) = (F\ P'\ a < F\ P'\ b)$
begin
...
theorem $\exists i.\ dictator\ i$
...
end

[3] https://www.isa-afp.org/entries/ArrowImpossibilityGS.html

The function F takes a preference profile and returns a function of type $\mathcal{C} \Rightarrow \mathbb{R}$. Therefore, F essentially returns a utility function. Hence, as explained in Section 2 the assumption of rationality (reflexivity, transitivity, and totality) is not necessary, since it is already implied by F. As mentioned, the theorem has also been formalized by Wiedijk in Mizar. The statement of the theorem is as follows:

theorem
  for f **being** *Constitution of* A, N **st**
    f **is** *independent_of_irrelevant_alternatives* &
    f **is** *respecting_unanimity*
  holds f **is** *a_dictatorship*;

It is worth noting that this is only the "pretty" version of the statement. The actual statement is expressed in a so-called cluster, which is less readable [13].

Regarding the proof structure Wiedijk and Nipkow chose different proofs. Geanakoplos [5] wrote a short paper showing three different proofs of Arrow's theorem. The third of which, is the proof which Nipkow [6] follows in his formalization, while Wiedijk [13] follows the first one. Geanakoplos [5] introduces two lemmas, the *extremal lemma* and the *strict neutrality lemma*. Even though they are equivalent, the former is used in the first proof and the latter in the third proof.

Lemma 4.4 (Extremal Lemma)**.** *Given the same assumptions as in Theorem 4.1, if every individual puts the same alternative $x \in \mathcal{C}$ at an extreme position (either very top or very bottom), then the SWF $\mathcal{F}$ must also put x at an extreme position.*

Since Wiedijk follows the structure of the first proof in [5], he makes use of the extremal lemma. After defining the notion of extremal, the Mizar statement reads as follows:

for p, b **st for** i **holds** $extreme[p.i, b]$ **holds** $extreme[f.p, b]$

As one can see, it simply shows an implication using a preference profile p which maps every i to a preference relation $p.i$ and a SWF f. With this we will conclude the discussion of the formalizations and refer to the actual proof scripts [6, 12] for further information and details.

5 Discussion

The formalization of mathematics is an important task that is gaining more and more traction. This can be seen in Tao's recent blog post[4] where the importance of computer assistance is mentioned.

With the importance that social choice, game theory, and especially economics, we believe that rigorously defined and proven theory is inevitable. Indeed, economic policy and decision making largely depends on economic or game theoretic models that are

[4] https://terrytao.wordpress.com/2017/12/21/metrics-of-linear-growth-the-solution/

backed by mathematical models. Hence, it only seems natural to have them rigorously defined and its proofs rigorously checked by a proof assistant.

Even though we only focused on one result here, there are multiple formalizations of theoretical and abstract results in game and social choice theory, as well as economics. However, very little work has been done when it comes to actual computation. Indeed, we see quite some potential in the formalization of computational social choice as well as algorithmic game theory. Where code generation and extraction mechanisms can be utilized to obtain correct and verified code. To this end, one would need to formalized concepts such as Nash equilibria and the famous existence theorem for Nash equilibria [7]. Since this theorem is based on expected utility functions, a formalization thereof is inevitable. This is work that is currently being conducted.

References

[1] K. J. Arrow. *Social choice and individual values*. Wiley New York, 1951.

[2] F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia. *Handbook of Computational Social Choice*. Cambridge University Press, New York, NY, USA, 1st edition, 2016.

[3] M. Eberl. Randomised social choice theory. *Archive of Formal Proofs*, May 2016. `http://isa-afp.org/entries/Randomised_Social_Choice.shtml`, Formal proof development.

[4] P. Gammie. Some classical results in social choice theory. *Archive of Formal Proofs*, Nov. 2008. `http://isa-afp.org/entries/SenSocialChoice.html`, Formal proof development.

[5] J. Geanakoplos. Three brief proofs of arrow's impossibility theorem. *Economic Theory*, 26(1):211–215, Jul 2005.

[6] T. Nipkow. Arrow and Gibbard-Satterthwaite. *Archive of Formal Proofs*, 2008.

[7] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA, 2007.

[8] J. Parsert and C. Kaliszyk. Formal Microeconomic Foundations and the First Welfare Theorem. In *Proceedings of the 7th ACM SIGPLAN International Conference on Certified Programs and Proofs*, CPP 2018, pages 91–101. ACM, 2018.

[9] A. Sen. Internal consistency of choice. *Econometrica*, 61(3):495–521, 1993.

[10] P. Tang and F. Lin. Computer-aided proofs of arrow's and other impossibility theorems. *Artificial Intelligence*, 173(11):1041 – 1053, 2009.

[11] M. Wenzel, L. C. Paulson, and T. Nipkow. The Isabelle framework. In O. A. Mohamed, C. A. Muñoz, and S. Tahar, editors, *Theorem Proving in Higher Order Logics, 21st International Conference, TPHOLs 2008*, volume 5170 of *LNCS*, pages 33–38. Springer, 2008.

[12] F. Wiedijk. Arrow's impossibility theorem. *Formalized Mathematics*, 15(4):171–174, 2007.

[13] F. Wiedijk. Formalizing Arrow's theorem. *Sadhana*, 34(1):193–220, Feb 2009.