AN

INFORMAL INTRODUCTION TO

RATIONAL CHOICE THEORY

BEING AN EXCEEDINGLY EXCELLENT

AND EXPEDIENT EXPOSITION

OF THE THEORIES OF

SOCIAL CHOICE

AND

GAMES

INCLUDING MANY APPLICATIONS TO THE

POLITICAL SCIENCES

WITH COMMENTARY BOTH

COLORFUL AND ENLIGHTENING

BY THE AUTHOR

Contents

1 Introduction

Rational choice theory is a family of mathematical theories that investigate collective decision-making. There are two main areas that these notes cover: social choice theory and game theory. The first of these might be characterized as the study of preference aggregation. What methods can be used to distill a group preference from the preferences of the individuals that comprise it? The second of these incorporates something more, both preferences and beliefs. As a model, it includes strategy as well as simple preferences.

> ⚠ **WARNING:** These notes are neither exhaustive nor comprehensive nor serious. The content is my interpretation of a course I took that emphasized an overview of the field, not a deep theoretical dive into any particular area.

If you find any errors, let me know!

2 Social choice theory

The formalism for social choice theory starts with a set of *individuals* $N = \{1, 2, \dots, n\}$ and a set of *options* or *alternatives* X that the individuals can choose from. So far so good.

The spirit of social choice theory is that outcomes result purely from individual preferences, so we need a way to encode these. We do that with a so-called rational preference.

Definition 2.1. A *rational preference* on a set X is a binary relation $\preccurlyeq$ on X that satisfies

1. **Reflexivity.** $x \preccurlyeq x$ for every $x \in X$,
2. **Completeness.** For every $x, y \in X$, either $x \preccurlyeq y$ or $y \preccurlyeq x$, and
3. **Transitivity.** If $x \preccurlyeq y$ and $y \preccurlyeq z$, then $x \preccurlyeq z$.[1]

We write $x \prec y$ if $x \preccurlyeq y$ and $y \not\preccurlyeq x$ and write $x \sim y$ if $x \preccurlyeq y$ and $y \preccurlyeq x$. The set of all rational preferences on X is denoted $\mathcal{R}(X)$, or simply $\mathcal{R}$ when X is understood.[2]

We encode the preference of each individual $i \in N$ as a rational preference $\preccurlyeq_i$, where $x \preccurlyeq_i y$ is interpreted as "Individual i considers option y to be at least as good as option x." Now, we can push back on these axioms (Is it true that every person can compare every pair of alternatives? Is it reasonable to assume that preferences are transitive?)—but for now, let's accept this as a working definition of preferences.

So, the setup for social choice theory is a set N of individuals, a set X of options, and a rational preference $\preccurlyeq_i \in \mathcal{R}(X)$ for each individual $i \in N$. The collection $(\preccurlyeq_i)_{i \in N}$ is called a *preference profile*; the set of all preference profiles is $\mathcal{R}^N$. The main question is: How do we use the preference profile to make a group decision?

Definition 2.2. A *social welfare function* is a function $F \colon \mathcal{R}^N \to \mathcal{R}$. A *social choice function* is a function $C \colon \mathcal{R}^N \to \mathcal{P}(X)$. (Here, $\mathcal{P}(X)$ denotes the power set of X.) We'll often denote the resulting ranking of the social welfare function by $\preccurlyeq$ without a subscript.[3]

In essence, a social welfare function is a rule for aggregating individual preferences into a *social preference ranking*, while a social choice function is a rule for determining winner(s). In large part, the study of social choice theory is the study of these aggregation rules.

[1] Briefly, then, a rational choice is *almost* a linear order on X, except that we allow ties.
[2] The notation here differs from that used in the social choice literature; there, $\succcurlyeq$, $\succ$, and $\sim$ are denoted by R, P, and I, respectively. I've opted for something that looks a bit more traditional mathematically.
[3] In *the literature*, the binary relation for the social preference ranking is denoted R.

2.1 Arrow's theorem: There's no perfect social welfare function

Despite the despair-inducing title, let's press forward anyway. What are some properties we would want in a "good" or "democratic" social welfare function?

Definition 2.3. A social welfare function F is *dictatorial* if there is a member $i \in N$ such that $x \prec y$ whenever $x \prec_i y$ (regardless of the other preferences in the profile).

This is probably something we'd like to avoid. On the other hand, here are some properties that we'd like F to have.[4]

Definition 2.4. A social welfare function satisfies the *weak Pareto property* if $x \prec y$ whenever $x \prec_i y$ for every $i \in N$.

In short: If everyone prefers y to x, then the social preference ranking should reflect that.

Definition 2.5. We say that two social preference $R, R' \in \mathcal{R}^N$ profiles are *consistent* on two alternatives $x, y \in X$ if $x \preccurlyeq_i y$ if and only if $x \preccurlyeq'_i y$ for every $i \in N$. A social welfare function F is *independent of irrelevant alternatives* if for any $R, R' \in \mathcal{R}^N$ that are consistent on x and y, then $x \preccurlyeq y$ if and only if $x \preccurlyeq' y$.

That is, if the relative rankings between two alternatives remain the same across each individual (consistency), then the relative ranking of those two alternatives in the social preference ranking is the same. The cool kids abbreviate independence of irrelevant alternatives by IIA.

So what does our boy Kenneth Arrow have to say about social welfare functions?[5]

Theorem 2.6 (Arrow's impossibility theorem)**.** *If $|N| \geq 2$ and finite and $|X| \geq 3$, there is no social welfare function that has the weak Pareto property, is IIA, and has no dictator.*

Arrow's theorem is *devastating* for all the idealists in the audience. There is simply no rule to aggregate group preferences into a social preference ranking in a way that satisfies a handful of obviously desirable conditions. Oh, well—we mourn and move on. Let's see why it's true.

Proof of Arrow's impossibility theorem. Part 0 of the proof is defining two notions: A set of actors $V \subseteq N$ is called *semi-decisive for x against y* if $x \prec y$ whenever $x \prec_i y$ for every $i \in V$ and $y \prec_j x$ for each $j \in N \setminus V$. We call V *decisive for x against y* if $x \prec y$ whenever $x \prec_i y$ for every $i \in V$. So any decisive set is semi-decisive for the same pair of alternatives.

To prove the theorem, we will assume that F has the Pareto property and is IIA and prove from these assumptions that there is a dictator.

We first prove that if $i \in N$ is semi-decisive for some x against y, then i is actually a dictator in disguise. Pick some $z \in X \setminus \{x, y\}$ and consider a preference profile R where $x \prec_i y \prec_i z$ and $y \prec_j x, z$ for all $j \neq i$. Because i is semi-decisive, the social ranking satisfies $x \prec y$, and the Pareto property guarantees $y \prec z$. In this case $x \prec z$ and $x \prec_i z$, but the relative ranking of x and z for every person $j \neq i$ is unknown. By IIA, we can freely adjust the position of y in every preference without affecting the ranking of x and z, so i determines the ranking of x and z *no matter the rankings by other individuals*. That is, i is decisive for x against z for any $z \in X \setminus \{x, y\}$. A mirror argument shows that i is decisive for y against z for every $z \in X \setminus \{x, y\}$. If $z, w \in X \setminus \{x, y\}$, then the preference $z \prec_i x \prec_i w$ implies the social preference $z \prec x \prec w$ by decisiveness of i; then IIA again shows that i is decisive for z against w. Finally, $x \prec_i z \prec_i y$ for some $z \in X \setminus \{x, y\}$ gives the social preference $x \prec z \prec y$, and IIA shows i is decisive for x against y. Since i is decisive for all pairs of alternatives, i is a dictator.

[4] Social scientists also include the properties of *unrestricted domain*, which means that the social welfare function accepts any preference profile as input, and *collective rationality*, which means the output is a rational preference. Mathematically, these are baked into the definition of F as a function $\mathcal{R}^N \to \mathcal{R}$.

[5] All the juicy details are available in his original 1950 paper [1].

To finish the proof, we show that there exists an actor who is semi-decisive (and therefore a dictator by the previous paragraph). To do this, consider the set $\mathcal{S}$ of subsets of N that are semi-decisive for some pair of alternatives; by the Pareto property, $N \in \mathcal{S}$, so $\mathcal{S}$ is not empty. Let $V \in \mathcal{S}$ be a set of minimal size. If $|V| = 1$, then we're done. Otherwise, suppose V is semi-decisive for x and y, and write $V = V_1 \sqcup V_2$ (where neither set is empty) and $V_3 = N \setminus V$. Consider a profile where

1. $x \prec_i y \prec_i z$ for every $i \in V_1$
2. $z \prec_j x \prec_j y$ for every $j \in V_2$
3. $y \prec_k z \prec_k x$ for every $k \in V_3$.

Socially, $x \prec y$ because V is semi-decisive for x against y. If $z \prec y$, then (by IIA) V_2 is semi-decisive (for z against y), which contradicts the minimality of V. If $y \preccurlyeq z$, then $x \prec z$ by transitivity. But then V_1 is semi-decisive (for x against z), another contradiction! And that's all the possibilities—there simply isn't a consistent preference between y and z. We can only conclude that V cannot have more than one element: There must be a semi-decisive actor. □

2.2 Responses to Arrow's theorem

Arrow's theorem was the first general result on voting frameworks; prior work had looked at similar behavior for specific voting mechanism. This theorem irked social scientists enough that it spawned a whole area of research to respond. Some immediate questions were:

1. What if we relax the outcome to a set of winners (that is, we look for a *social choice function* instead)? Is there still a contradiction?
2. What if the conditions are changed?
3. What if X is a continuous set, for example $X = [0, 1]$?

Another set of policy dudes basically said, "So what, bro?" The prosaic translation of this is "How does this relate to the real world?" Approaches have included:

1. What is the probability of a troublesome preference profile?
2. Fieldwork/empirical work.
3. Interpretations and applications.

2.3 Sen joins Arrow's discontent

A few years after Arrow got everyone down, Amartya Sen managed to depress everyone further [13]. To describe why, we need some definitions. First, let's describe two weakenings of transitivity.

Definition 2.7. A rational preference $\preccurlyeq \in \mathcal{R}$ is *acyclic* if $x_1 \preccurlyeq x_k$ whenever there is a set $\{x_1, \dots, x_k\} \in X$ such that $x_1 \prec x_2 \prec \cdots \prec x_k$. The preference is *quasi-transitive* if $x \prec z$ whenever there is a $y \in X$ such that $x \prec y \prec z$. The set of all complete, reflexive, acyclic binary relations on X is denoted $\mathcal{R}_A$. A function $F \colon \mathcal{R}^N \to \mathcal{R}_A$ is called a *social decision function.*

Transitivity implies quasi-transitivity, which in turns implies acyclicity. Acyclicity rules out strict cyclic behavior: The preference doesn't have a cycle $x_1 \prec x_2 \prec \cdots \prec x_k \prec x_1$.

And now let's look at a different notion of democracy than simple non-dictatorship.

Definition 2.8. An aggregation rule is *liberal* if for every agent $i \in N$, there are two alternatives in X for which i is decisive. It is *minimally liberal* if there are two actors $i, j \in N$, each of which is decisive over two alternatives. (Both conditions imply non-dictatorship.)

The idea of liberalism is that there are some private decisions over which an individual has supreme authority. Here's an example: You're putting together a dinner party and invite a cadre of friends. Despite your strong preference for a cohesive, harmonious meal, you decide to host the party as a potluck since, honestly, those heathen friends of yours wouldn't really appreciate an

elegantly coordinated meal anyway. In the potluck, each person is decisive over the alternatives about what they, individually, bring. The potluck that results is the aggregation of these individual preferences in a liberal manner.[6]

So what was Sen's news?

Theorem 2.9 (Sen, 1970)**.** *If $|X| \geq 3$ and $|N| \geq 2$, there is no social decision function that satisfies the weak Pareto condition and is minimally liberal.*

Although minimal liberalism is stronger than non-dictatorship, there are serious weakenings of Arrow's hypotheses here: Sen doesn't assume IIA and only demands that the social outcome be acyclic (not necessarily transitive). It is a quite general theorem, and it shows that you can't hope that Arrow's theorem is an aberration. Preference aggregation is looking to be very hard.

2.4 Further bad news

Among other changes, Sen relaxed the condition of collective rationality (that the social ranking is transitive) to acyclicity. After his theorem, other researchers focused more directly on what happens when this condition is weakened in Arrow's theorem. As usual, we'll need some definitions, this time about concentrations of power.

Definition 2.10. A set of actors $L \subseteq N$ is an *oligarchy* if

1. L is decisive for all $x, y \in X$ (that is, if $x \prec_i y$ for every $i \in L$, then $x \prec y$) and
2. $x \prec_i y$ for some $i \in L$ implies $x \preccurlyeq y$.

The set L is a *collegium* if $x \prec y$ if and only if $x \prec_i y$ for every $i \in L$ and one $i \in N \setminus L$.

A dictator is an oligarchy of one. A collegium can only be overruled if everyone outside the collegium votes against it (or is indifferent). In this case, the social outcome must be indifferent (why?).

Here's the bad news dressed up in academic jargon.

Theorem 2.11 (Gibbard, 1973 [5])**.** *Suppose $|N| \geq 2$ and $|X| \geq 3$. If a preference aggregation function satisfies IIA and the weak Pareto property and it outputs quasi-transitive social preferences, then there is an oligarchy.*

Theorem 2.12 (Brown, 1975 [3])**.** *Suppose $|X| \geq |N| \geq 3$. If a preference aggregation function satisfies the weak Pareto property and outputs acyclic social preferences, then there is a collegium.*

What's the takeaway? Together with Arrow's theorem, these results illustrate a sort of tradeoff between the rationality of outcomes (transitivity and its pals) and a dispersion of power. If you want to distribute power amongst a larger group of people, you have to allow for a looser ordering in the social preference.

2.5 Arrow gets scooped

One of the main objections to Arrow's theorem is the assumption of universal domain—that the preference aggregation algorithm should be able to take *every possible* set of rational preferences as input. Oftentimes, it may be the case that there are some preference profiles that will never show up (for example, if the group is sufficiently homogeneous).

One example of this is the domain of so-called single-peaked preferences.

[6] To be more specific, in this case the alternatives look like, for example, "Esmeralda brings fruit salad, Joachim brings crostini, and Miklós brings pálinka"; an example of a pair of alternatives for which Miklós is decisive would be the previous one and "Esmeralda brings fruit salad, Joachim brings crostini, and Miklós brings Túró Rudi."

Definition 2.13. A preference profile in $R \in \mathcal{R}^N$ is *single-peaked* if there is an ordering $(x_1, \ldots, x_k)$ of the elements of X so that every individual ranking is monotone on this ordering. That is, if $\preccurlyeq_i \in R$, then there is some $s \in \{1, \ldots, k\}$ such that $x_1 \preccurlyeq_i x_2 \preccurlyeq_i \cdots \preccurlyeq_i x_s \succcurlyeq_i \cdots \succcurlyeq_i x_k$.

Essentially, this means that we can arrange the alternatives in X on a one-dimensional spectrum, each voter has a preferred option, and the further an alternative is from it, the less that voter likes it. Single-peaked preferences might be reasonable in certain left-right policy decisions (though certainly not all) or in deciding how much of the budget to allocate to a particular program.

It turns out that if everyone is guaranteed to have single-peaked preferences, we can make a good preference aggregation function.

Theorem 2.14 (Black, 1948 [2])**.** *On the domain of single-peaked preference profiles with $|X| \geq 3$, pairwise majority voting is a social welfare function.*[7]

The proof of this theorem just consists of showing that each possible ordering of three alternatives a, b, and c with $a \prec b$ and $b \prec c$ is either impossible or implies that $a \prec c$.

Pairwise majority voting satisfies all of Arrow's conditions except unrestricted domain (weak Pareto, non-dictatorship, IIA, collective rationality), so we get a good deal by relaxing the universal domain condition a bit.

You might notice something funny here: Black's result was published two years before Arrow's. So Black's theorem was not in fact a response to Arrow's theorem, though it's convenient to think about it that way. In reality, it seems that Arrow was just trying to rain on Black's parade.[8]

In any case, Black's theorem makes the point that the issues raised in Arrow's theorem may not show up in practice. So here's a followup question: How likely is a cycle to occur in pairwise majority voting? It seems that the appearance of a cycle is somewhat unlikely when $|X|$ is small and $|N|$ is large, but actual results depend on the assumptions made to compute the probability. (See this page and the references therein.)

2.6 The spatial model

Now we want to consider something different: A continuous decision space $X \subseteq \mathbb{R}^n$, called the *spatial model*. This might be more realistic, for example, if the Legislature is trying to decide how to distribute its budget amongst its various needs (payoffs, kickbacks, salaries, etc.).

Brace yourself, because this section is going to be a lot mathier. In a spatial model, we assume that the preferences are *distance-based*:

1. for each $i \in N$ there is a point $x_i \in X$ such that $x_i \succ_i y$ for every $y \in X \setminus \{x_i\}$ and
2. $x \succ_i y$ if and only if $|x_i - x| < |x_i - y|$.

In a distance-based model, $x \sim_i y$ if and only if $|x_i - x| = |x_i - y|$. These are ludicrously strict assumptions—much stronger, even, than single-peakedness. Other models exist for continuous distributions, depending on a general differentiable utility function, for example, but we'll focus on this baby model.

Given an individual $i \in N$ and a policy $x \in X$, we define $P_i(x) = \{y \in X : y \succ_i x\}$ (the P is for the verb "pining" as in "Meredith has been pining for her lover's passionate embrace since the day he left").[9] So $P_i(x)$ is a fancy way to write the open ball of radius $|x_i - x|$ with center x_i. The social equivalent is the *win set* of x, denoted $W(x) = \{y \in X : y \succ x\}$ (which of course depends on the aggregation rule).

Definition 2.15. An alternative $x^* \in X$ is a *voting equilibrium* if $W(x^*) = \emptyset$ under majority rule.

[7] A stronger version of this theorem is sometimes called the Median Voter Theorem.
[8] I have absolutely no evidence to support this accusation.
[9] Actually the P stands for "preferred," but that's boring.

A voting equilibrium is *not* a policy position that everybody is happy with. It just means that there isn't a majority that would be happier with any other position. The main questions we'll start with are the existence and characterization of equilibria.

Example 2.16. Suppose $N = \{1, 2, 3\}$ and $X = [0, 1]$, with $x_1 \leq x_2 \leq x_3$. In this case, x_2 is the unique voting equilibrium. In general, suppose $|N| = n$ and $x_1 \leq x_2 \leq \cdots \leq x_n$. If n is odd, then the unique voting equilibrium is $x_{(n+1)/2}$. If n is even, then every point in $[x_{n/2}, x_{n/2+1}]$ is an equilibrium. (This is a version of the Median Voter Theorem.)

As usual, the 1-dimensional case was pretty straightforward, but the higher-dimensional ones won't be. Here are three things we know immediately for any dimension.

Proposition 2.17. *In a spatial model with distance-based preferences:*

1. *If $|N|$ is odd, then there is at most one voting equilibrium; if it exists, it is either on the boundary of X or is the ideal point of one of the actors.*
2. *If both x and y are voting equilibria, then every point on the line segment xy is an equilibrium.*
3. *If $d \geq 2$, there is a spatial model with $X \subseteq \mathbb{R}^d$ with no voting equilibrium.*

Proof. The perpendicular bisecting hyperplane of two distinct points $x, y \in \mathbb{R}^d$ is the set of points equidistant from x and y; equivalently, it is the hyperplane that is orthogonal to the segment xy and contains its midpoint. We denote this hyperplane by $H(x, y)$. If $x, y \in X$, then all the actors in the open half-space of $H(x, y)$ that contains x will vote for x over y, and the actors in the open half-space that contains y will vote for y over x. Reformulated in the language of hyperplanes, a point $x^* \in X$ is a voting equilibrium if and only if every open half-space that contains X whose bounding hyperplane contains no ideal points also contains at least $|N|/2$ ideal points. Throughout the proofs, we assume that hyperplanes and bounding hyperplanes of half-spaces contain no ideal points.

1. If x and y are distinct voting equilibria, let H_1 and H_2 be two nonintersecting half-spaces perpendicular to the line xy that contain x and y, respectively. Both contain at least $(|N|+1)/2$ ideal points, which is impossible. So there cannot be two distinct equilibria.
   Now take any $x \in X$ that is not an ideal point and not on the boundary of X, and let H be a hyperplane through x that contains no ideal points. If $y \in X$ is sufficiently close to x, then the half-spaces formed by $H(x, y)$, $H(x, 2x - y)$, and H divide the ideal points in the same way. One of y and $2x - y$ (the reflection of y through x) must beat x, since one of the half-spaces of contains a strict majority of the ideal points.
2. Let z be any point on the line segment xy and w be any point in $X \setminus \{z\}$. The open half-space of $H(z, w)$ that contains z must contain either x or y, in which case z gets at least $|N|/2$ votes over w. So z is an equilibrium.
3. Set $X = \mathbb{R}^d$ and the ideal points of N as a triangle (or, for full-dimensionality, a simplex). □

Exercise 2.18. Show that there is no equilibrium if $|N| = 3$ and the ideal points are the vertices of a non-degenerate triangle.

Part 3 of this exercise is a bit depressing, like being consigned to parliamentary purgatory. There is, though, a (much) weaker positive result for equilibria-esque points. It takes a brief foray into combinatorial geometry to prove.

Definition 2.19. A *centerpoint* of a finite point set $Y \subseteq \mathbb{R}^d$ is a point $x \in \mathbb{R}^d$ such that any closed half-space of $\mathbb{R}^d$ that contains x contains at least $|Y|/(d + 1)$ points of Y.

A centerpoint is a generalization of the median in $\mathbb{R}$. A true median, of course, would split the point set in half with every half-space. But it turns out that this isn't possible in higher dimensions; $1/(d+1)$ is the largest fraction that works for all point sets. (You can't do better, for example, when Y is the vertices of a simplex.)

Unlike voting equilibria, centerpoints always exist.

Theorem 2.20. *Every finite point set has a centerpoint.*

Using this, we can show that every spatial model has a weak status quo. Or, well, you can.

Exercise 2.21. Suppose that $X \subseteq \mathbb{R}^d$ is convex and $|N|$ is finite. Show that there exists a point $x^* \in \mathbb{R}^d$ so that, for every point $y \in X \setminus \{x^*\}$, no supermajority will vote for y over x^* with ratio greater than $d/(d+1)$.

On the other hand, a true voting equilibrium does exist in very special circumstances.

Definition 2.22. A set of points $\{x_i\} \subseteq X$ is *radially symmetric* about $x^* \in X$ if $\{x_i\} \setminus \{x^*\}$ can be partitioned into pairs whose line segments contain x^*.

Theorem 2.23 (Charles Plott, ostensibly[10])**.** *If the set of ideal points in a distance-based spatial model is radially symmetric about x^*, then x^* is a voting equilibrium.*

Proof. Label the ideal points $x_1, \ldots, x_{2n}$; if $|N|$ is odd, then the last ideal point must be x^*—ignore it. Also, suppose that x_i and x_{i+n} are paired so that x^* is on the segment $x_i x_{i+n}$ for every $1 \leq i \leq n$. Then $P_i(x^*) \cap P_{n+i}(0) = \emptyset$, so any alternative can win at most $|N|/2$ votes over 0. □

While this is a positive result, it is both fragile and weak (when $d \geq 2$, which is the only really interesting case; every set in $\mathbb{R}$ is radially symmetric). Perturbations of any size destroy radial symmetry, and the Lebesgue measure of the set of points $(x_1, \ldots, x_n) \subseteq \mathbb{R}^{nd}$ that satisfy radial symmetry is 0—in fact, the set is contained in a subspace of dimension $(d+1)n/2$. Though not only radially symmetric distributions have a voting equilibrium.

Exercise 2.24. Construct a spatial model in $\mathbb{R}^2$ that is not radially symmetric but has a voting equilibrium.

But social choice theory has a small quota of optimism, and we're close to hitting it. So here's some more bad news:

Theorem 2.25 (McKelvey, 1976 [9])**.** *If $X = \mathbb{R}^d$ with $d \geq 2$, $|N| = n \geq 3$, and there is no equilibrium, then for any $x, y \in X$, there exists a sequence of alternatives $x = x_0, x_1, \ldots, x_t = y$ such that x_i beats x_{i-1} in majority rule for every $1 \leq i \leq t$.*

More pithily: It's possible to play individuals' preferences against each other to get a bad result for everyone. Or: No equilibrium implies that cycles engulf the entire policy space.

To get a feel for how bad this theorem should make you feel, consider a hypothetical scenario where $X = \mathbb{R}^2$ and every ideal point is in the unit square $[0,1]^2$ with no equilibrium. Then some fiendish devil can create an agenda of pairwise votes that leads the committee to decide on the policy $(10^{1000}, 10^{1000})$ which everyone absolutely abhors.

Because of this, McKelvey's theorem has been taken to mean that the agenda setter has extreme power to determine outcomes. But there are several sensible objections to this, time limits and strategic voting being the main ones. We'll address objections like these in the next section.

2.7 Social choice meets the world; also, some reactions to the soul-crushing results of social choice theory

Many people looked at all of these instability results in social choice theory, then at all the stability in the real world, then back at the social choice theory and rejected it. Others did the same double take and thought that the stability derives from institutions—the rules and structure that delimit the ability to legislate. Then so-called "structure-induced equilibria" are possible. This might be formalized with a definition like this:

[10] Can *you* find a reference?

Definition 2.26. The *motion set* of a point $x \in X$, denoted $M(x)$, is the set of all counter-proposals to x allowed by the rules.

A preference-induced equilibrium is a point $x^* \in X$ such that $W(x^*) = \emptyset$. A *structure-induced equilibrium* is a point $x^* \in X$ such that $W(x^*) \cap M(x^*) = \emptyset$.

Some methods to force structure-induced equilibria include:

- dimension-by-dimension voting
- time limit or limit on the number of proposals
- more generally, agenda-setting rules
- raise the threshold for making a decision (supermajority)

A criticism of the idea that structure stabilizes the wild, wild west of preferences is that some rules are in fact *endogenous*—they were chosen by the legislature itself. So the rules themselves could be changed; why aren't they subject to the instability social choice seems to imply?

A different explanation of real-world stability is that people simply aren't clever enough to manipulate the agenda to produce chaos. This argument is a fairly weak, since it requires only a bit of cleverness to produce instability. Moreover, the agenda-setting theorem of McKelvey accords strategic skill to only the agenda-setter and not the voters, which is unrealistic. So how did political theorists respond to all this?

Political scientist Gerry Mackie takes a contrarian view of social choice, arguing that the theorems of social choice don't reflect reality, which is much more stable than the theory predicts—so we should throw the theory out. If it doesn't match empirical fact, then the theory should go. *Adiós*, social choice.[11]

This argument takes a narrow view of the lessons of social choice. Mackie is correct that social choice is not an explanatory theory, and this is an important point. But this itself points toward an important lesson of social choice theory: Politics as it is *cannot* be a simple aggregation of social preference. There must be something else going on.

William Riker, one of the original researchers and certainly the main popularizer of social choice theory, argues this perspective, positing that the body of results in social choice theory eviscerates the philosophy of populism[12]. He specifically points to the facts that election outcomes depend so strongly on the method of aggregating votes (that is, the voting system) and some scenarios have no clear-cut winner (the appearance of cycles or the lack of a voting equilibrium). It can hardly be that the outcome of an election by itself indicates a collective will of the electorate. A more refined version of Riker's argument is a sort of argument by contradiction. *If* outcomes of elections reflect only the preferences of the people, then the outcome cannot be stable. But since we observe so much stability in the real world, election outcomes must reflect more than a simple "people's will." Thus dies populism.[13]

On the other hand, Riker argues in another book [11] that the results of social choice theory suggest a new type of political manipulation: *heresthetic*, the deliberate shaping of institutions and situations to bring about desired outcomes. Techniques include agenda-setting or the introduction of new dimensions into a debate. But Riker's arguments here rely on the same assumptions of social choice theory and ascribe differential skill in, institutional ability to, and awareness of manipulation to different actors, assumptions that don't hold up in the world at large. Moreover, Mackie points out that many of the cases Riker considers are misrepresented. In any case, if a group of legislators is being hoodwinked, they won't obstinately stick with their preference ranking—they'll start manipulating, too.

[11] He wrote a whole book [8] on this. You could read it if you want.

[12] The conceit that elections determine popular will, that the winner of an election has the moral authority of the will of the people. Very popular amongst recent winners of elections.

[13] See *Liberalism against Populism* [10], especially chapters 1 and 10, for a more detailed (though quite dry) account of Riker's argument.

Mackie argues that statistical analyses of Congressional votes regularly show that legislators' positions are one-dimensional, which would obviate all the nastiness of the spatial model. This is true in fact but not spirit. Dimensional analyses (using a program called NOMINATE) do in fact return low-dimensional output. HOWEVER, legislators have a handful of special interests on (what seems to other people to be) small issues. Because these are necessarily unique to a (possibly small handful) of legislators, they don't impact the statistics too much. But these issues can be pivotal in actual debates and persuasion; this is one explanation for why bills contain provisions on so many seemingly unrelated issues—to gain votes based on special interests.

Mackie does score a point against Riker in that social choice theories necessarily ignore crucial elements of the process. It does not account for deliberation, for example, which might homogenize preferences or limit strategic action by agenda-setters. It also doesn't account for outside institutions, the judiciary, for example. And social choice theory doesn't account for longer-term changes like shifts in public opinion. Mackie does leave open, though, to what extent deliberation affects the process. What is needed to address that question is a model that incorporates both preferences and beliefs—game theory.

3 Game theory

All games we look at will be *non-cooperative*, meaning that the players don't enter into binding contracts (like vote trading, for example)—each player makes their own choice.

3.1 Normal form games

In a normal form game, the players make simultaneous and independent choices from the options available to them. We'll also assume that all information in the game is *common knowledge*, meaning that

$$\text{each player } \underbrace{\text{knows that every player}}_{\text{repeated } n \text{ times}} \text{ knows the information} \qquad \text{for each } n \in \mathbb{N}_0.$$

Here's a formal description: There is a set of actors N with $|N| \geq 2$ (if N is a singleton, then that's *decision theory*). Each person $i \in N$ has a set A_i of feasible actions and a utility function $h_i \colon \prod_{i \in N} A_i \to \mathbb{R}$ that converts the collective actions into personal utility. (Higher utility is better) All aspects of this model are common knowledge.

For brevity, we write $\mathcal{A} = \prod_{i \in N} A_i$ and $\mathbf{a} = (a_i)_{i \in N}$ for a point $\mathbf{a} \in \mathcal{A}$.

Definition 3.1. A *Nash equilibrium* is a point $\mathbf{a} \in \mathcal{A}$ if all $i \in N$ would choose a_i given knowledge that all $j \neq i$ choose a_j. In other words, $\mathbf{a} \in \mathcal{A}$ is a Nash equilibrium if $h_i(\mathbf{a}) \geq h_i(\mathbf{a}')$ for every $i \in N$ and $\mathbf{a}' \in \mathcal{A}$ with $a_j = a'_j$ for every $j \neq i$.

Intuitively, a combination of actions is a Nash equilibrium if no individual's defection benefits them. That is, each person is doing as well as possible given the actions of the other actors.

It's time to take the formal model for a walk. Let's find some simple examples.

Example 3.2 (Coordination game)**.** There are two players; one picks the row and the other picks the column in the following chart. Both receive the utility payout in the corresponding box.

	A	B
A	1	0
B	0	1

There are two Nash equilibria: both actors choose A or both choose B.

Definition 3.3. A set of actions $\mathbf{a} \in \mathcal{A}$ is *efficient* if there is no other set of actions $\mathbf{b} \in \mathcal{A}$ such that $h_i(\mathbf{b}) \geq h_i(\mathbf{a})$ for every $i \in N$ and $h_j(\mathbf{b}) > h_j(\mathbf{a})$ for at least one $j \in N$.

The previous game has two efficient actions: both A and both B. Modifying the output to

	A	B
A	3	0
B	0	1

has only one efficient outcome—AA— but the same Nash equilibria.

Example 3.4 (Bargaining game)**.** Again two player, one choosing the row and the other column. The payoff for the row player is the first number in each box; the payoff for the column player is the second.

	A	B
A	1,3	0,0
B	0,0	3,1

There are two Nash equilibria: AA and BB. These are also both efficient outcomes, since any move from them harms at least one player.

Here's another.

Example 3.5 (Prisoner's dilemma)**.** The story here is that you and your nefarious partner in crime have been apprehended by the indefatigable city police. Both of you are presented with the potential for a plea deal. If you both reject the plea deal (cooperating with each other), you get some jail time, but not much since there's not much evidence ('cause you're good at what you do[14]). If one accepts the deal and the other doesn't (they defect), the squealer gets a slap on the wrist and the loyalist gets a terribly long sentence[15]. If you both give in, you get a sentence that's somewhere between mutual cooperation and the one foisted upon you by the turncoat.

Here's the payoff matrix, with $h > c > d > \ell$.

	Coop.	Defect
Cooperate	c, c	ℓ, h
Defect	h, ℓ	d, d

There is only one equilibrium: Both defect. And yet every outcome is an equilibrium *except* mutual defection. So the best individual outcome is the worst social outcome.

In this game, both players have a *dominant strategy*: It's always better, for any given action by the other player, to defect. And yet this is a bad overall outcome.

The Prisoner's Dilemma can model any sort of situation where a desired outcome requires some input cost. Defecting gives the advantage of enjoying the benefits of the outcome without putting up any cost—unless the other party defects, too, in which case nothing happens and everyone is sad. This more general interpretation is why the game is sometimes called the Collective Action Problem.

Example 3.6 (Something else)**.** No story here; just the coldhearted and uncaring matrix.

	A	B
A	$3, 4$	$6, 1$
B	$5, 0$	$2, 4$

[14] Just not good enough to avoid getting caught in the first place
[15] like one of James Joyce's

This game has the dubious distinction of having no Nash equilibrium. A game like this—where in every outcome one person wants to change actions—is called *adversarial*. Here's an even simpler payoff matrix with the same property.

	A	B
A	$0, 1$	$1, 0$
B	$1, 0$	$0, 1$

These examples give some indications of the inadequacy of Nash equilibria: Sometimes there are too many; sometimes they aren't efficient; sometimes there aren't any at all.

In the types of adversarial games like in Example 3.6, one player wants to coordinate actions (column) and one wants to oppose actions (row). In repeated rounds of this game, the row player would want to remain unpredictable so that the wily column player can't catch on and coordinate actions.

This motivates (at least, it motivates it enough for the author of these notes) the idea of *mixed strategies*: The assignment of a probability distribution p_i to each individual's action set A_i, instead of definitely choosing only one action (which is called a *pure strategy*).

Let M_i denote the set of mixed strategies for actor i and $\mathcal{M} = \prod_{i \in N} M_i$ denote the set of collective mixed strategy. We let $\mathbf{m} = (m_i)_{i \in N} \in \mathcal{M}$ and $m_i(a)$ be the probability of $a \in A_i$ under the mixed strategy m_i.

Definition 3.7. A *mixed Nash equilibrium* is a point $\mathbf{m} \in \mathcal{M}$ if all $i \in N$ would choose m_i given knowledge that all $j \neq i$ choose m_j. In other words, $\mathbf{m} \in \mathcal{M}$ is a mixed Nash equilibrium if $\mathbb{E}(\mathbf{m}) \geq \mathbb{E}(\mathbf{m}')$ for every $i \in N$ and $\mathbf{m}' \in \mathcal{A}$ with $m_j = m'_j$ for every $j \neq i$.

Lemma 3.8. *If $\mathbf{m} \in \mathcal{M}$ is a mixed Nash equilibrium, then the value of*

$$\mathbb{E}(h_i(\mathbf{m}') \mid m'_i(a) = 1 \text{ and } m'_j = m_j \text{ for each } j \neq i) \tag{3.1}$$

is the same for every $a \in \mathcal{A}_i$ for which $m_i(a) > 0$

Proof. If the values of (3.1) are not equal for each $a \in A_i$, then i would do better with a pure strategy, choosing the value of a which maximizes (3.1). So $\mathbf{m}$ is not a Nash equilibrium. □

We can use this result to find a mixed Nash equilibria for the second game in Example 3.6. Let's say the row and column players have probability p and q of choosing option A, respectively. For the row player, the expected utility of choosing option A and B are, respectively,

$$\mathbb{E}_R(A) = q \quad \text{and} \quad \mathbb{E}_R(B) = 1 - q.$$

Since we've ruled out the existence of a pure equilibrium, any mixed Nash equilibrium $\mathbf{m} \in \mathcal{M}$ must have both $m_R(A)$ and $m_R(B)$ positive; from the lemma, $\mathbb{E}_R(A) = \mathbb{E}_R(A)$. So any Nash equilibrium must have $q = 1/2$. The same calculation for the column player reveals that p must also be $1/2$, and a short calculation show that this is indeed a Nash equilibrium.

Exercise 3.9. Use the same technique to find a mixed Nash equilibrium for the first game in Example 3.6.

Even though Nash equilibria aren't a panacea, they have the adequate property that they do, at least, exist.

Theorem 3.10 (Nash)**.** *Every game with both $|N|$ and $|\mathcal{A}|$ finite has a mixed Nash equilibrium.*

And since at this point existence is far more than we can possibly hope for in any rational choice theory theory construct [4], we'll leave it at that.

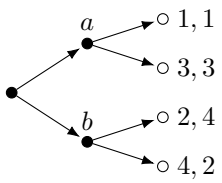

Figure 1: A decision tree. The first player makes a move at the initial node; the second player chooses at the second node. The payoffs are in the order first player, second player.

3.2 Extensive form games

An *extensive form game* models sequential interactions with common knowledge. This is modelled with a rooted tree; each leaf (or so-called *terminal node*) has payoffs for each player, something like Figure 1.

A *total strategy* for player i is a decision of which arrow to follow at every node i makes this decision. Extensive form games can be reduced to normal games by defining A_i to be the set of total strategies for player i.

If players can't communicate intentions, then the gameplay is simple, and optimal play can be discovered by working backward from the leaves. There's no need for the matrix forms and Nash equilibria (see *pterosaurs* below). However, if players *can* communicate, then their entire strategies become relevant. For example, if the second player committed to "Choose the upper arrow no matter what" (an irrational strategy if there weren't no communication), then the first player should choose the downward arrow. An analysis of all the possible strategies reveals two possible (pure) Nash equilibria:

- first player moves $\nearrow$; second player moves $\searrow$ at node a and $\nearrow$ at node b
- first player moves $\searrow$; second player moves $\nearrow$ no matter what.

The first set of strategies is the optimal one for noncommunicative gameplay. The second one illustrates the effect an ultimatum can have on the outcome.

Definition 3.11. A *subgame* of an extensive-form game is a subgraph of the decision tree that contains a single node and all of its subsequent nodes.

A subgame is like tuning into the game at some point that's not necessarily the beginning. The game in Figure 1 has three subgames (the endnodes aren't considered subgames, since those aren't decision nodes).

Definition 3.12. A *pterosaur* is a strategy profile that is a Nash equilibrium for every subgame.[16]

Pterosaurs are found using *backwards induction*, working from the terminal nodes, just as backwards induction finds winning strategies in finite combinatorial games. In some sense, pterosaurs are the most credible strategies, since each actor is acting for their own self-interest at each step. If a player broadcasts that they're going to deviate from their pterosaur strategy, this is generally less credible, though it is possible to issue a credible ultimatum, which would likely cause the game to proceed differently than the pterosaur would have it.

3.3 Repeated games

Recall the Prisoner's Dilemma / Collective Action Problem of Example 3.5, in which the single Nash equilibrium—both defect—is not efficient. This motivates a question: Under what conditions

[16] So called because this paper has dishonorably few dinosaur references. Usually a pterosaur is called a *subgame perfect Nash equilibrium* or SPNE. You could vocalize this initialism as "spine" or even "supine." People don't usually do this, but don't let that stop you.

is cooperation possible?

Here's one solution: They both sign a contract to cooperate, with a penalty of more than $h - c$ assessed to any party that defects. This makes mutual cooperation a Nash equilibrium and efficient. Alternatively, an outside actor might offer a reward of more than $h - c$ for cooperation. (To eliminate mutual defection as a Nash equilibrium, the penalty or reward must be more than $d - \ell$, as well.)

There are three restrictions on this outside actor: It must

- be able to monitor behavior,
- have the power and/or resources to fine or reward, and
- have an incentive to act.

The word for this resolution to the problem is called an *exogenous solution*. It's the only resolution for a one-shot Prisoner's Dilemma.

A practical resolution is that perhaps the model is wrong. For example, perhaps every actor is not ruthlessly self-interested; altruism would change the structure of the game and potentially make cooperation more attractive. This seems naive for a universal resolution.

Alternatively, the interaction might be repeated. In this case, actors might take into account "long-run" behavior. But in any finite number of rounds (fixed beforehand), the only pterosaur is mutual defection every time. Paints a nice picture of the potential for cooperation, huh?

On the other hand, a large enough number of interactions is psychologically indistinguishable from an infinite number of interactions. So what happens if we model that?

The wrinkle here is that the total payoff at the "end" of the game series is always ∞, which is pretty useless for comparison. But here's a question. Would you rather have \$100 now or \$100 in 10 years? Almost certainly you'd like the money now.[17] To account for this, we introduce a *discount parameter* $\delta \in [0, 1)$, where a payoff of n in k time units is worth $n\delta^k$ to you right now. One sensible value of δ is the reciprocal of the inflation rate, but it could be anything. The smaller the value of δ, the less you value future money compared to current money. Introducing this parameter makes the "end" payoff for any infinitely repeated game finite.

Let's look for some equilibria in this model. The easy one is that both players defect every single time. This is a Nash equilibrium, and both players collect $d/(1 - \delta)$ utility. In fact, it's not too hard to see that the stage game equilibrium is always an equilibrium for a repeated game. On the other hand, unconditional mutual cooperation is not an equilibrium.

Here's a different strategy, the Hardcore Punishment plan[18]: Cooperate until the other player defects; at that point, defect in every future round (thereby lowering the utility for the opposing player). If both players follow this plan, then both get $c/(1 - \delta)$ utility. If it's profitable to defect in round k, then it's at least as profitable to defect in round 1. So: If one player defects in round 1, then their utility is at most $h + d\delta/(1 - \delta)$. Therefore mutual Hardcore Punishment is an equilibrium if and only if $c \geq h(1 - \delta) + d\delta$. In particular, if δ is high, indicating some patience or forbearance, then this is an equilibrium. If δ is low, indicating impatience, then it is not, since the defecting player values a large present much more than future payments. (The exact cutoff is $\delta = (h - c)/(h - d)$.)

Here's another possible strategy. (This is beginning to feel like a Catalogue of the Prisoner's Infinite Dilemma Strategies for the Criminally Inclined.) Let's call this Lenient Tit for Tat:

- Cooperate in round 1;
- Defect in round t if the other player defected against your cooperation in round $t - 1$;
- Otherwise cooperate in round t.

Again, both taking this strategy results in mutual cooperation. If you defect at any point, then gameplay changes for the next round but resets after that; the Lenient Tit for Tat only promises one round of punishment. If you defect in one round, you want to defect in the next (to get d

[17] If not, I guess you're willing to lend it to me for 10 years with 0% interest.

[18] Game theorists frighteningly call this the "grim trigger" plan

utility rather than ℓ). So if $h + \delta d > c(1 + \delta)$, in other words if $\delta < (h - c)/(c - d)$, you'll want to defect every time. If $\delta \geq (h - c)/(c - d)$, then Lenient Tit for Tat is an equilibrium.

You could alter this strategy by changing to a k-round punishment, which is an equilibrium if and only if

$$\frac{\delta - \delta^{k+1}}{1 - \delta} = \delta + \delta^2 + \cdots + \delta^k \geq \frac{h - c}{c - d}.$$

Moving from philately to theory, there's a general result. Let he *normalized payoff* of a strategy be $1 - \delta$ times the normal discounted payout.

Theorem 3.13 (Folk Theorem)**.** *In an infinitely or indefinitely repeated game, if there is a strategy profile that results in a normalized payoff greater than the minimum that a player can guarantee themself, then there is a pterosaur that results in that payout if δ is close enough to 1.*

For example, the minimum normalized payoff that a player can guarantee themself is a d by defecting every time. In particular, mutual cooperation is a strategy profile that results in a normalized payout of c. The Folk Theorem guarantees that, so long as δ is large enough, there is a pterosaur that gives a normalized payoff of c. Of course, we already found this,[19] so perhaps it's not so exciting. Here's a different example: Supposing that $h + \ell$ is big enough, it's even better to alternate cooperation and defection, resulting in a normalized payout of

$$(1 - \delta)\left(\frac{h}{1 - \delta^2} + \frac{\ell\delta}{1 - \delta^2}\right) = \frac{h + \ell\delta}{1 + \delta}$$

for one player and

$$(1 - \delta)\left(\frac{\ell}{1 - \delta^2} + \frac{h\delta}{1 - \delta^2}\right) = \frac{\ell + h\delta}{1 + \delta}$$

for the other. If $\delta \geq 1/2$, then both payouts are greater than $\ell/2 + h/4$. So if h and ℓ are together big enough, the Folk Theorem guarantees that there's a pterosaur that results in the normalized payouts described above when δ is big enough.

3.4 Common knowledge

Much of the theory of games relies on the assumption of common knowledge. As an infinite list of conditions, this is actually quite a strong condition. Imagine that you want to meet up with a friend to watch *Movie: The Movie 2, Return 4 More Cash*, which has arrived at a movie theatre[20] near you. So you text your friend movie 2nite?. To which they respond 4 sure, c u there, because you're both insufferable. That's it—you both have common knowledge of the situation, right?

Wrong. There's a small chance that the second text won't be delivered due to the caprices of technology. So your friend doesn't know whether you know that they're going to the movie. And if you don't know that, then there's no reason for you to go to the movie; you'd think you're going alone. So you need to send a confirmation of the confirmation: righteous 👍. But now you need to know that your friend receives this, so they have to send back a message (totes), to which you must respond, which they must acknowledge, which . . . [21]

You get the problem. Somehow people manage to get along without all of this. And yet a good portion of game theory changes or simply breaks without this assumption. (See, for example, [12] and [14].) Attempts to empirically analyze this problem (for example, through psychology

[19] Almost; we proved that there's a Nash equilibrium, not necessarily a pterosaur (though Hardcore Punishment is in fact a pterosaur).

[20] Not one of those lowbrow -*er* theaters

[21] This is my modified (and quite butchered) adaptation of the description of the "coordinated attack problem," which was described in [12] as a quote from [7], which itself was a quote from [6], which is a chain that I can only assume goes on forever, since I didn't bother skimming this last source.

or behavioral economics) or create a new theoretical framework are interesting new directions for game theory and its adjacent subjects.

References

[1] Kenneth Arrow, *A difficulty in the concept of social welfare*, Journal of Political Economy **58** (1950), 328–346.

[2] Duncan Black, *On the rationale of group decision-making*, Journal of Political Economy **56** (1948), 23–34.

[3] Donald Brown, *Aggregation of preferences*, Quarterly Journal of Economics **89** (1975), 456–469.

[4] Experience, personal communication.

[5] Allan Gibbard, *Social choice and the Arrow conditions*, Economics and Philosophy **30** (2014), 269–284.

[6] Jim Gray, *Notes on data base operating systems*, Operating Systems (R. Bayer, M. Graham, R. and G. Seegmüller, eds.), Lecture Notes in Computer Science, vol. 60, Springer, 1978, pp. 393–481.

[7] Joseph Halpern, *Reasoning about knowledge: An overview*, Theoretical Aspects of Reasoning about Knowledge (Joseph Halpern, ed.), Morgan Kaufmann, 1986, pp. 1–17.

[8] Gerry Mackie, *Democracy Defended*, Cambridge University Press, 2003.

[9] Richard McKelvey, *Intransitivities in multidimensional voting models and some implications for agenda control*, Journal of Economic Theory **12** (1976), 472–482.

[10] William Riker, *Liberalism against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice*, Waveland Press, 1982.

[11] William Riker, *The Art of Political Manipulation*, Yale University Press, 1986.

[12] Ariel Rubinstein, *The electronic mail game: Strategic behavior under "almost common knowledge"*, The American Economic Review **79** (1989), 385–391.

[13] Amartya Sen, *The impossibility of a Paretian liberal*, Journal of Political Economy **78** (2020), 152–157.

[14] Hyun Song Shin and Timothy Williamson, *How much common belief is necessary for a convention?*, Games and Economic Behavior **13** (1996), 252–268.